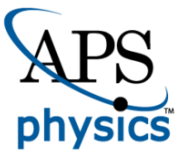

Inclusion-exclusion principle for many-body diagrammatics

Aviel Boag, Emanuel Gull, and Guy Cohen

Phys. Rev. B **98**, 115152 — Published 26 September 2018

DOI: 10.1103/PhysRevB.98.115152

Inclusion–Exclusion Principle for Many-Body Diagrammatics

Aviel Boag,[1] Emanuel Gull,[2] and Guy Cohen[3,1]

[1] *School of Chemistry, Tel Aviv University, Tel Aviv 6997801, Israel*
[2] *Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*
[3] *The Raymond and Beverley Sackler Center for Computational Molecular and Materials Science, Tel Aviv University, Tel Aviv 6997801, Israel*

Recent successes in Monte Carlo methods for simulating fermionic quantum impurity models have been based on diagrammatic resummation techniques, but are restricted by the need to sum over factorially large classes of diagrams individually. We present a fast algorithm for summing over the diagrams appearing in Inchworm hybridization expansions. The method relies on the inclusion–exclusion principle to reduce the scaling from factorial to exponential. We analyze the growth rate and compare with related algorithms for expansions in the many-body interaction. An implementation demonstrates that for a simulation of a concrete physical model at reasonable parameters and accuracy within the Inchworm hybridization expansion, our algorithm not only scales better asymptotically, but also provides performance gains of approximately two orders of magnitude in practice over the previous state-of-the-art.

I. INTRODUCTION

The accurate description of systems of many strongly interacting fermions is one of the big open problems in modern theoretical physics.[1] Apart from a few very special situations, all known exact and general solutions scale exponentially in the number of degrees of freedom. In order to make progress, approximate numerical methods that are both precise and efficient enough to describe the salient aspects of the problem need to be designed.

The solution of quantum impurity models, which describe a small interacting region (an "impurity" or "dot") coupled to a large or infinite noninteracting region ("leads" or "baths"), is much simpler than the general problem but remains a formidable challenge.[2] Quantum impurity models appear in a wide range of contexts, including in the description of magnetic atoms embedded in a host material[3] or adsorbed on a surface,[4] in the description of quantum transport through mesoscopic systems[5–8] and molecules,[9–12] and in quantum embedding algorithms.[13,14] Even greater challenges are faced where access to real-time dynamics or the description of high-lying excitations is needed. Numerical methods that are able to describe these phenomena reliably and efficiently are therefore highly desired.

The stochastic sampling of terms in a many-body perturbation theory, known as "diagrammatic"[15] or "continuous-time"[16–20] quantum Monte Carlo, has been highly successful at describing the equilibrium physics of impurity models. However, as systems are enlarged, frustration is introduced, or equations are generalized to real-time propagation,[21–25] the straightforward formulation of these algorithms scales exponentially due to either the fermionic or the dynamical sign problem. This motivates the need for formulations that either eliminate this exponential scaling entirely or delay its onset for long enough that useful results can be obtained with available resources.

Several such attempts have been made for lattice models.[26–28] They are based on using the underlying structure of many-body diagrammatics to reduce the number of diagrams that need to be considered, *e.g.* by considering connected diagrams only in a Green's function series, by considering irreducible diagrams only in a self-energy expansion, or by employing the "skeleton" technique to self-consistently resum (or "boldify") certain classes of diagrams. These techniques typically trade an alleviated sign problem (caused by the reduction of the number of diagrams) against increased algorithmic complexity and, potentially, convergence issues.[29]

In the context of impurity models, these techniques have mostly found application in the Keldysh diagrammatics for real-time propagation.[30–33] In a first implementation, partial summations (boldification) based on semi-analytic impurity model techniques[32–34] could substantially alleviate the sign problem, and in some cases allow evaluation of slow dynamics.[31,35,36] Later, the realization that the causal structure of real-time dynamics could be integrated directly into the algorithm led to the so-called Inchworm method,[37] which for several systems and expansions seems to overcome the dynamical sign problem entirely or in a wide range of physical regimes.[38–42]

However, all of these methods rely on an explicit enumeration of all allowed diagrams at a given set of n perturbation times for diagrams of order n. This enumeration is expensive, since it scales as $n!$. Access to large diagram order is therefore prohibitively expensive, and the applicability of the various methods is restricted to domains where convergence is obtained at relatively small orders.

In this paper we present a method that replaces the explicit enumeration of $n!$ diagrams in the Inchworm hybridization expansion with a fast summation algorithm based on the inclusion–exclusion principle. We develop theoretical bounds for the scaling of the algorithm and describe results from a practical implementation. We also compare our method to a reformulation of the diagram summation in the interaction expansion,[43,44]

showing that while our method is superior in the context of hybridization expansions, the method of Ref. 44 remains superior in the context of interaction expansions.

The remainder of this paper proceeds as follows: in Sec. II, we define the necessary concepts and then present our inclusion–exclusion algorithm for the hybridization expansion, as well as two optimizations. The algorithm of Ref. 44 for the interaction expansion is reviewed, and a hybridization-expansion algorithm along similar lines is presented and compared to the inclusion–exclusion algorithm. An inclusion–exclusion algorithm for the interaction expansion is then presented and compared to that of Ref. 44. Sec. III includes first a direct comparison of the direct and inclusion–exclusion summation methods, then a comparison of their performance within the Inchworm algorithm for population dynamics in an Anderson impurity model. In Sec. IV we conclude. Two appendices are also provided: appendix A presents a derivation of our main formula from the inclusion–exclusion principle, and appendix B presents the methodology behind the theoretical expressions for the computational scaling of the algorithms and optimizations we discuss.

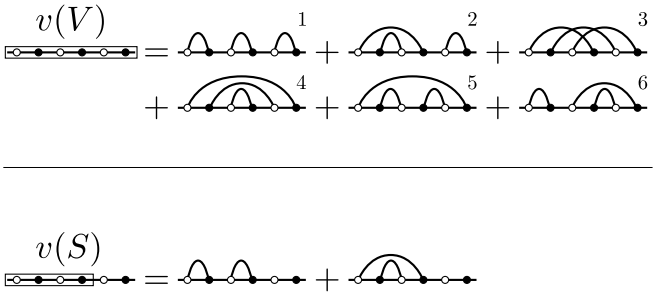

Figure 1. Elements of the bare hybridization expansion. The top panel shows the diagrammatic representation of the sum $v(V)$, marked by a box, over all diagrams generated by the complete set of vertices at all times V. The vertices are denoted by filled and empty circles, which indicate whether they consist of creation or annihilation operators, respectively. In each diagram, every creation operator is connected by a curved hybridization line to an annihilation operator. The bottom panel shows the partial sum $v(S)$ over all diagrams generated by a subset of four vertices, $S \subset V$.

II. METHOD

The standard continuous-time hybridization expansion ("bare" CTHYB) in imaginary time[18] and real time[21,22,24] has been described in the literature, and we refer readers interested in the details of the expansion to previous work. For the purposes of the present work, it is sufficient to introduce a simplified description of the diagrammatic structure and the process of evaluating diagrams. As the main idea we wish to present is general, we will do this in a form that is largely agnostic to the details of the model. Furthermore, in order to provide a self-contained description of the algorithm introduced in this paper, we will also introduce a few concepts from the Inchworm CTHYB expansion; once again, for a full discussion readers are referred to the existing literature.[37,38]

A. Definitions

Consider a generic impurity model Hamiltonian split into two parts:

$$\hat{H} = \hat{H}_0 + \hat{V}. \tag{1}$$

Here, $\hat{H}_0 = \hat{H}_D + \hat{H}_B$ is separated into "dot" and "bath" subspaces, the second of which is noninteracting (*i.e.* described by a quadratic Hamiltonian); and $\hat{V}$ is a hybridization Hamiltonian connecting the two subspaces. We assume that every element in the Hamiltonian can be written in terms of second quantization operators $\hat{a}_k$ and $\hat{a}_k^\dagger$ obeying fermionic commutation relations, with k enumerating the degrees of freedom. The time dependence of the expectation value of some observable $\hat{A}$ is then given by

$$\left\langle \hat{A}(t) \right\rangle = \left\langle \hat{U}^\dagger(t)\, \hat{A}_I(t)\, \hat{U}(t) \right\rangle, \tag{2}$$

where for any operator $\hat{O}$, $\hat{O}_I(t) \equiv e^{i\hat{H}_0 t}\hat{O}e^{-i\hat{H}_0 t}$, and $\langle \ldots \rangle$ signifies a trace on all degrees of freedom with respect to some initial density matrix. The interaction picture propagator $\hat{U}(t) \equiv e^{i\hat{H}_0 t}e^{-i\hat{H}t}$ can be written in the form

$$\begin{aligned}\hat{U}(t) = \sum_{n=0}^{\infty} (-i)^n \int_0^t \mathrm{d}t_1 \cdots \int_0^{t_{n-1}} \mathrm{d}t_n \\ \times \hat{V}_I(t_1) \cdots \hat{V}_I(t_n).\end{aligned} \tag{3}$$

In diagrammatic Monte Carlo techniques, the high-dimensional time integrals appearing when Eq. (3) is replaced into Eq. (2) are carried out stochastically by sampling the times at which the $V_I(t)$, called *vertices*, appear. This requires that we be able to efficiently evaluate the integrands

$$\left\langle \hat{V}_I(t_1) \cdots \hat{V}_I(t_n)\, \hat{A}_I(t)\, \hat{V}_I(t'_1) \cdots \hat{V}_I(t'_n) \right\rangle, \tag{4}$$

where the times $0 < t_i, t'_i < t$ come from terms in Eq. (2) for $\hat{U}^\dagger(t)$ and $\hat{U}(t)$.

In bare CTHYB, Eq. (2) is finally written as

$$\begin{aligned}\left\langle \hat{A}(t) \right\rangle = \sum_{n=0}^{\infty} \sum_{\{s_1,\ldots,s_{2n}\}} v(\{s_1,\ldots,s_{2n}\}) \\ \times p(\{s_1,\ldots,s_{2n}\}).\end{aligned} \tag{5}$$

Here, p is a local propagator part that can be obtained from exact diagonalization of the isolated dot Hamiltonian; and v, which is called the lead influence functional, takes the form

$$v\left(\{s_1,\ldots,s_{2n}\}\right) \equiv \sum_{k_1,\ldots,k_{2n}\in B} \gamma_{k_1}\gamma^*_{k_2}\cdots\gamma_{k_{2n-1}}\gamma^*_{k_{2n}} \times \left\langle \hat{a}^\dagger_{I,k_1}(s_1)\,\hat{a}_{I,k_2}(s_2) \cdots \hat{a}^\dagger_{I,k_{2n-1}}(s_{2n-1})\,\hat{a}_{I,k_{2n}}(s_{2n})\right\rangle_B . \tag{6}$$

The k indices are taken from the bath subspace only, and the γ_k are parameters depending on the model. Averaging is performed only over the isolated bath subspace. Additional model-dependent local indices which may appear in the expansion have been suppressed for brevity.

The $s_i \in \left\{t_j, t'_j, t\right\}$ in Eq. (6) are a set of contour times, and together they are called a *configuration*. Since the part of $\hat{H}_0$ which includes bath operators is quadratic, Eq. (6) can be evaluated using Wick's theorem.[45] This results in a sum over $n!$ different diagrams, each of which corresponds to a permutation of n indices matching each creation operator with an annihilation operator. Nevertheless, for fermions, this sum can be evaluated at a computational cost which is cubic in n because it takes the form of a determinant:[18,21]

$$v\left(\{s_1,\ldots,s_{2n}\}\right) = \mathrm{Det}M\left(s_1,\ldots,s_{2n}\right). \tag{7}$$

The elements of the matrix M are given by a set of interaction picture correlation functions which can be easily evaluated, since they describe time evolution within a noninteracting reference system:

$$M_{ij} = \sum_{k_{2i+1},k_{2j}} \gamma_{k_{2i+1}}\gamma^*_{k_{2j}} \times \left\langle \hat{a}^\dagger_{I,k_{2i+1}}(s_{2i+1})\,\hat{a}_{I,k_{2j}}(s_{2j})\right\rangle_B . \tag{8}$$

The top panel of Fig. 1 illustrates the connection between determinants and diagrams.[18] The determinant of Eq. (7) is represented by a box, with filled (empty) circles representing the times at which creation (annihilation) operators appear in a particular configuration. We have chosen a certain 6^{th} order (*i.e.* the perturbation order $2n = 6$ or $n = 3$) configuration. The terms comprising the determinant, each of which corresponds to a particular permutation pairing the n creation operators to the n annihilation operators, delineate $n! = 6$ individual diagrams. The so-called hybridization lines in the diagrams signify pairings, and a line between operators at times s_{2i} and s_{2j+1} corresponds to a multiplicative factor of M_{ij} from Eq. 8. We denote the sum over all diagrams generated by the complete set of vertices V, with $|V| = 2n$, as $v(V)$. In the lower panel of of Fig. 1, we show a sum over all diagrams generated by some $S \subset V$, which can also be evaluated as a determinant.

The value of each diagram is a product of the hybridization functions of Eq. (8) multiplied by an additional fermion sign determined by the choice of permutation, or equivalently a term in Eq. (7); and by the local propagator $p\left(\{s_1,\ldots s_{2n}\}\right)$ which does not depend on the permutation and is therefore not of interest in the present context. The fermion sign is suppressed in our diagrammatic notation for clarity, but is crucial in order for the sum to form a determinant.

The bare CTHYB expansion of Ref. 18 benefits greatly from this determinant structure and the resulting polynomial cost of evaluating the sum of all diagrams associated with a configuration. Essentially, it means that time configurations rather than individual diagrams form the sampling space. However, the real time bare expansions,[21–25] as well as their bold counterparts,[30,32–34] suffer from a dynamical sign problem: as the propagation time t increases, the stochastic error increases exponentially.

B. Fully connected, k-connected, proper and improper diagrams

The Inchworm algorithm overcomes the dynamical sign problem (in at least some cases) by taking advantage of the causal diagrammatic properties of the expansion and the fact that evaluating propagation over short time intervals is numerically inexpensive.[37] However, this comes at a cost: within the Inchworm expansion, contributions are written in terms of dressed propagators, and the sum over diagrams for a particular configuration can no longer be written in the determinant form of Eq. (7).

It is therefore necessary to explicitly iterate over a factorial number of permutations for each configuration and filter a subset of dressed diagrams, which is typically still factorial. One then sums over the factorial number of contributions corresponding to this subset individually, resulting in an overall $O(n!)$ computational scaling in the expansion order $2n$, which should be compared to $O\left(n^3\right)$ scaling in bare expansions. Nevertheless, while the order needed to converge bare expansions always increases with time, Inchworm expansions can often be terminated at low orders. In such cases, the loss of the determinant structure may be worthwhile, as the exponential scaling in time due to the dynamical sign problem is removed.

In order to explain precisely which diagrams must be summed within the Inchworm method, we first introduce the concept of connected and k-connected diagrams. A diagram is considered (fully) connected if all hybridization lines within it are connected by crossing. Note that this differs from the more standard definition of connectedness encountered in interaction expansions, where connectivity is a property of the graph of vertices, which are connected by Green's function lines. Here, it can be thought of as a property of the graph comprising hybridization lines as nodes, with edges drawn between

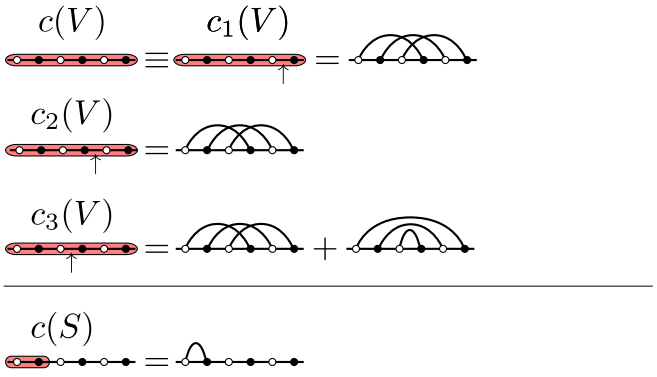

Figure 2. Elements of the Inchworm hybridization expansion. The top panel shows the sum over all connected or k-connected diagrams generated by the complete set of vertices V. This is denoted by a rounded and shaded box, with an arrow delineating the value of k, *i.e.* the boundary between proper and improper vertices. The connected and 1-connected sums are identical by definition and include a single diagram; additionally, for this particular case, no new 2-connected diagrams exist and one more 3-connected diagram exists. The lower panel shows a sum over connected diagrams generated by S, a subset of V containing two vertices.

any two hybridization lines which cross each other. Connectivity is illustrated in the top panel of Fig. 2, where the sum $c\,(V)$ over connected diagrams generated by vertices V is denoted by a rounded and shaded box. Of the six diagrams in the top panel of Fig. 1, only diagram 3 is connected.

A diagram is k-connected if each of its connected components contains at least one of k special vertices, which we refer to as *improper*, whereas the other $2n-k$ vertices are termed *proper*. In the diagrams discussed here, improper vertices are always the rightmost k vertices. 1-connectedness is identical to full connectedness, since there can be only one connected component containing the single improper vertex. In the upper panel of Fig. 2, the sum $c_k\,(V)$ over k-connected diagrams generated by all vertices V is denoted by a rounded and shaded box with an arrow to the left of the k improper vertices. For the particular configuration we have chosen to discuss, the sum $c_1\,(V)$ over 1-connected and $c_2\,(V)$ over 2-connected diagrams is the same, but the sum $c_3\,(V)$ over 3-connected diagrams contains one additional term corresponding to diagram 4 in the top panel of Fig. 1.

In analogy to Fig. 1, the bottom panel shows that it is also possible to define a sum over all connected diagrams generated by a subset S of the vertices V. In this case, we chose a two-vertex subset which contains only a single diagram.

In the most basic Inchworm expansion,[37] one extends a known propagator over some time interval $(t_i, t_\uparrow)$ into a longer propagator over the interval (t_i, t_f) with $t_f > t_\uparrow$. We set vertices in the interval $(t_\uparrow, t_f)$ to be improper, and all other vertices to be proper. Given that there are k improper vertices, the mathematical problem that needs to be addressed within the algorithm may then be reduced to the summation of all k-connected diagrams.

C. Application of the inclusion–exclusion principle

The inclusion–exclusion principle can be used to avoid the explicit summation over a factorial number of k-connected diagrams. To see how this works, we will first consider fully connected diagrams for the same example configuration considered above (see the top panel of Fig. 3). Every disconnected diagram contains at least one disconnected piece composed of lines fully spanning an adjacent subset of the proper vertices. We will refer to an adjacent subset as a *segment*. Therefore, to obtain the set of connected diagrams, one might start from the sum over all diagrams $v\,(V)$, calculated as a determinant in polynomial time, and subtract all terms with connected subsegments of V. To do this, one could try to sum over all possible segments, taking connected diagrams within the segment and all diagrams outside it. Only segments with the same number of creation and annihilation operators need be considered.

However, a diagram containing two disconnected pieces would be subtracted twice in this manner: once for the term in which the segment includes one piece, and once for the term where it includes the other. To cancel out this double-counting, one should now add all such diagrams, by introducing terms corresponding to all possible *pairs* of segments. This argument could be repeated indefinitely, leading to a mathematical structure analogous to the one that results from attempting to express the size of a union of N sets by summing the sets and their intersections. The formal mathematical connection with this concept, known as the inclusion–exclusion principle, is presented in appendix A. Our expression for the sum over k-connected diagrams is as follows:

$$c_k\,(V) = \sum_{j=0}^{n-k} (-1)^j \sum_{\{S_i\}} v\left(V \backslash \bigcup_{i=1}^{j} S_i\right) \prod_{i=1}^{j} c_1\,(S_i)\,. \quad (9)$$

Here, j is the number of segments, and the summation is over all possible segments comprising the $2n-k$ proper vertices. We note that the expression is given in terms of the $c_1\,(S_i)$, which can be recursively evaluated from it. While we will show several substantial optimizations, Eq. (9) describes the central result of this publication. The complete process is illustrated in Fig. 3 for our 6^{th} order configuration, with the bottom panel illustrating the evaluation of one of the elements appearing in the sum (which is in this case zero, a fact that we will take advantage of soon).

At first glance, it is not clear that this approach holds any advantage: in fact, in Fig. 3 we sum over 10 elements rather than the 6 in Fig. 9, even before taking into account the fact that we must also perform more summations to obtain the various elements appearing

in the expansion. However, consider the scaling: the sum in eq. (9) is over all sets of segments. Naively, to count them, one notes that there are 2^{2n-k} ways to decide which vertices will be included in segments (ignoring for a moment the differences between creation and annihilation operators, which decrease this number). In the worst case, if all $2n-k$ are chosen, the number of ways to construct segments from this set is the number of compositions of $2n-k$, of which there are 2^{2n-k-1}; so, at worst, the summation should scale as $2^{4n-2k-1}$. We must also compute each of the $c_1(S_i)$, each of which should be no more expensive, but there is only a quadratic number $(2n-k)(2n-k-1)$ of these. Given that each step entails the calculation of a single determinant at $O(n^3)$, even a rough estimate of the asymptotic computational complexity is $O(n^5 4^n)$, high but substantially less than factorial. In fact, as we show in appendix B, a more careful calculation shows that the correct scaling C_n in this case can be bounded from above by

$$L_n^U = O\left(n^3 \alpha^{2n}\right), \tag{10}$$

with $\alpha \approx 1.8019$. To simplify the calculation, this estimate assumes that all operators can be paired with all other operators, which results in an overestimate of the complexity. An alternative assumption is that half the vertices are creation (annihilation) operators, but they are arranged in arbitrary order. In this case it is possible to calculate a cost averaged over the orderings. We term this estimate

$$L_n^L = O\left(n^3 \beta^{2n}\right), \tag{11}$$

where $\beta \leq \alpha$. This is neither a strict upper bound nor a lower one. However, it may be expected to function as a heuristic lower bound, since one could suppose the computational complexity in most models to be more strongly influenced by the worst case ordering than by the average. We find that $\beta \approx 1.5072$ (see appendix B).

It is possible to generalize the algorithm to expansions where any two vertices might be paired, such that there is no distinction between creation and annihilation operators. In this case, the determinant is replaced by a Pfaffian. Pfaffians, like determinants, can be computed in polynomial time, and everything else in the algorithm remains essentially unmodified. Furthermore, the worst-case scaling criterion of Eq. (10) becomes exact. This generalization is of some interest from the mathematical viewpoint, and might be considered the solution of a simpler, cleaner problem. However, it is not immediately clear to us that it has utility in the physical context. We will therefore not explore it further here.

D. Optimizations

We can improve the algorithm further. In particular, for the example used in the illustration, it is possible to drop all but two (slightly modified) terms, such that the sum over connected diagrams can be obtained from the evaluation of a single order 3 determinant and two order 2 determinants which turn out to be zero. In general, however, the computational cost will remain exponential in the number of vertices. While the algorithm we have presented already produces an improvement of the scaling to exponential, the actual exponent can be reduced further with the aid of a few simple observations. This is of course worthwhile, because it provides an additional exponential improvement in performance.

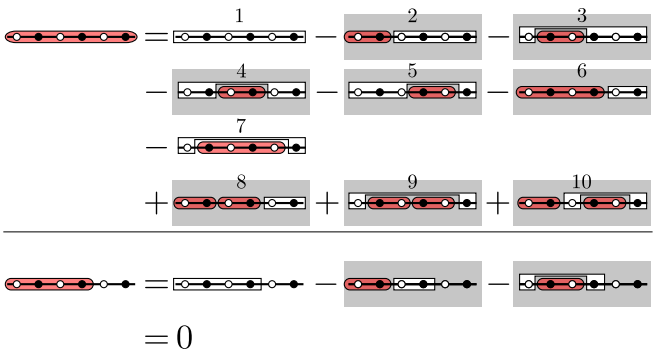

Figure 3. Illustration of the inclusion–exclusion algorithm. In the top panel, the expansion for the sum over fully connected ($k = 1$) diagrams for a 6^{th} order configuration is written in diagrammatic form. Below, the expression for a particular element within this expansion is shown. Terms inside shaded rectangles are removed by the optimizations.

1. *First optimization: adjacent segments*

First, note that we compute v on the same subset $V \backslash \bigcup_i S_i$ for many different sets of segments $\{S_i\}$, since adjacent segments occupy the same vertices as their union. For example, in Fig. 3 diagrams 6 and 8 share the same determinant, as do diagrams 7 and 9. Therefore, we can regroup the sum by first summing over non-adjacent segments, and then summing over all possible divisions of a segment into adjacent subsegments (which, once again, can be enumerated as compositions). In order to do this, we first rewrite Eq. (9) in the following form:

$$\begin{aligned} c_k(V) &= \sum_{j=0}^{n-k} \sum_{\{S_i\}} v\left(V \backslash \bigcup_{i=1}^{j} S_i\right) \prod_{i=1}^{j} \left[-c_1(S_i)\right] \\ &= \sum_{\{S_i\}} v\left(V \backslash \bigcup_i S_i\right) \prod_i \left(-c(S_i)\right). \end{aligned} \tag{12}$$

The summation now runs simultaneously over all sets of proper segments, with no regard as to how many segments are in a set. Now, if we let $a(S)$ be the sum of values of all partitions of a given segment S,

$$a(S) \equiv \sum_{\{D_i\}} \prod_i \left(-c(D_i)\right), \tag{13}$$

where the $\{D_i\}$ are all possible partitions of S into adjacent subsegments, we can write

$$c_k(V) = \sum_{\{A_i\}} v\left(V \backslash \bigcup_i A_i\right) \prod_i a(A_i), \tag{14}$$

where the $\{A_i\}$ are all disjoint non-adjacent segments comprising proper vertices. This allows us to drop diagrams 8 and 9 in Fig. 3.

$a(S)$ must now be evaluated for every possible segment S. If we choose S to be any segment $\{v_1, \ldots, v_j\}$, it is easy to see that for the case where the last segment is of length ℓ, the contribution to a is $a(\{v_1, \ldots, v_{j-\ell-1}\})\, c(\{v_{j-\ell}, \ldots, v_j\})$. Repeating this argument for all possible lengths ℓ then gives all contributions:

$$\begin{aligned} a(\{v_1, \ldots, v_j\}) = -\sum_{\ell=1}^{j} & a(\{v_1, \ldots, v_{j-\ell-1}\}) \\ & \times c(\{v_{j-\ell}, \ldots, v_j\}). \end{aligned} \tag{15}$$

As we show in appendix B, the effect of this reformulation is a reduction in the computational complexity to $\alpha \approx 1.618$ and $\beta \approx 1.4142$.

2. Second optimization: removing two-vertex segments

For a second optimization, one need only note that a hybridization line between two adjacent proper vertices never crosses any other hybridization line, and therefore can't be a part of a k-connected diagram. Given this, it is possible to eliminate the values of all such lines by setting the corresponding elements of M_{ij} to zero. After doing so, there is no longer any need to consider segments of length two, and the complexity improves to $\alpha \approx 1.4432$ and $\beta \approx 1.2676$.

Let us now revisit Fig. 3. In the top panel, with the second optimization, every term except 1 and 7 (*i.e.* all terms outlined by shaded rectangles) can be immediately dropped. Note that in this example, both optimizations discard diagrams 8 and 9, but at higher orders the overlap decreases. The first term is a third order determinant of a modified M_{ij} in which some of the elements have been set to zero. The seventh term, similarly to the sixth term in the bottom panel, is a second order determinant of a similarly modified submatrix of M_{ij}, which turns out to be zero. We therefore see that even for this $n = 3$ example, the optimized inclusion–exclusion method requires the computation of fewer terms than the direct algorithm.

E. Inverted algorithm

Recently, an algorithm was found that allows for summing all connected diagrams in interaction expansions in exponential rather than factorial time.[44] It was shown that this leads to polynomial complexity for evaluating thermodynamic quantities in certain regimes,[43] and was later extended to the summation of irreducible diagrams.[46–48] Eq. (9) is reminiscent of the main result of Ref. 44. Rephrased in a slightly modified form for easy comparison with the expressions presented here, Ref. 44 proposed the following formula for the sum over all connected diagrams within an interaction expansion for a Hubbard model:

$$c(E, V) = v(E, V) - \sum_{S \subsetneq V} c(E, S)\, v(\emptyset, V \backslash S). \tag{16}$$

Here, V and E are sets of internal and external vertices, respectively; $v(A, B)$ is the sum over all (interaction) diagrams generated by external vertices A and internal vertices B (given by a certain determinant); and $c(A, B)$ is the sum over all connected diagrams with external vertices A and internal vertices B. This result is seemingly much simpler than Eq. (9): there is no inclusion–exclusion hierarchy and the summation is terminated at the level of single subsets rather than sets of subsets. However, since there is a sum over subsets rather than segments, the resulting computational scaling is $O(3^n)$, exponentially worse than in our case. Inspired by this work, we set out to see if our algorithm could be formulated in a similar way, and if any advantage might be gained by this for either problem.

1. Inverted algorithm for the hybridization expansion

Comparing Eqs. (9) and (16), if we let internal (external) vertices correspond to proper (improper) vertices, the expression is inverted: while in Eq. (16) the subtracted contributions are connected to the external part, in Eq. (9) they are disconnected from it. With this in mind, it is possible to derive a different way of evaluating $c_k(V)$, where the improper vertices are always enclosed in a k-connected element:

$$\begin{aligned} c_k(V) = {} & v(V) \\ & - \sum_{\{A_i\} \backslash \{\emptyset\}} c_k\left(V \backslash \bigcup_i A_i\right) \prod_i v(A_i). \end{aligned} \tag{17}$$

Here, the summation is over all sets of one or more non-adjacent segments comprising proper points. This is in much closer analogy to Eq. (16). It is even more similar to Eq. (14), other than in the signs and the reversal of roles between c and v; what appeared as the first optimization in the inclusion–exclusion algorithm is necessary here for correctness.

The computational scaling of this algorithm is less than factorial, but unfortunately remains higher than

Optimization Level	α	β
Unoptimized, Eq. (9)	1.8019	1.5072
1st optimization, Eq. (14)	1.6180	1.4142
2nd optimization, section II D 2	1.4432	1.2676
Inverted algorithm, Eq. 17	2.1935	1.7321
Inverted alg. with 2nd optimization	1.8718	1.4861

Table I. Theoretical complexity of the two proposed algorithms for the hybridization expansion at different levels of optimization. $O\left(n^3\alpha^{2n}\right)$ is an overestimating simplification assuming all operators can be connected to each other, while $O\left(n^3\beta^{2n}\right)$ provides an average cost assuming the operators are randomly ordered and is most likely an underestimate of the cost.

that of the inclusion–exclusion algorithm: as discussed appendix B, it can be bound at $\alpha \approx 2.1935$ and $\beta \approx 1.7321$ as presented; the second optimization still applies to it, at which point we obtain $\alpha \approx 1.8718$, still higher than even the unoptimized algorithm based on Eq. (3); and $\beta \approx 1.4861$, larger than the smallest α for the inclusion–exclusion case. The inverted algorithm is therefore less suitable than the inclusion–exclusion algorithm for the hybridization expansion.

We present a summary of the theoretical computational complexities characterizing the different algorithms and optimizations in Table I, and refer the reader to appendix B for details.

2. *Inclusion–exclusion algorithm for the interaction expansion*

The algorithm of Sec. II C is substantially more efficient than the one of Sec. II E 1, which is reminiscent of the one in Ref. 44. It is intriguing to consider whether the inclusion–exclusion principle might also be useful in the context of the interaction expansion. On one hand this is a conceptually simpler problem, because there is less mathematical structure to it; but on the other hand a computationally harder one, because one must consider subsets of vertices rather than segments.

It is easy to see that, as an alternative to Eq. (16), the sum over connected diagrams can be recast in the following form:

$$c\left(E,V\right) = \sum_{j=0}^{\infty} (-1)^j \sum_{\{S_i\}} v\left(E, V\backslash\bigcup_i S_i\right) \times \prod_i c\left(\emptyset, S_i\right). \tag{18}$$

Here, we perform the summation over all possible sets of disjoint subsets of internal vertices $\{S_i\} \subset V$. This is analogous to Eq. (9).

Let us now consider the computational complexity of Eq. (18). The asymptotically dominant contribution in this case is not the evaluation of the determinants, which is $O\left(n^3 2^n\right)$ for $n = |V| + |E|$, but the sum itself. The disjoint subsets of a set with n elements are known as its partitions. The sequence of numbers counting the partitions of sets of increasing size are the Bell numbers B_n, which are asymptotically bound by[49]

$$B_n < \left(\frac{0.792n}{\ln\left(n+1\right)}\right)^n. \tag{19}$$

This is better than factorial complexity, but worse than exponential (as is the corresponding lower bond). Therefore, the inclusion–exclusion algorithm is better than the brute force approach to the interaction expansion, but not nearly as efficient as that of Ref. 44. Nevertheless, the inclusion–exclusion principle may turn out to be of interest within interaction expansions with more detailed structure, such as in cases where one neglects long-ranged correlations; and may also turn out to be more amenable to fast update schemes. As this is beyond the scope of the present work, we leave it for future study.

III. RESULTS

A. Comparison with direct algorithm

To analyze the algorithm, we will begin by considering the computational cost of the summation itself, with no regard to any physical context. This allows for a cleaner exploration of the scaling and for a well-defined comparison with the theoretical exponents α and β. For this purpose, we implemented a brute-force summation over over all k-connected permutations for a given configuration ("Direct algorithm") and the inclusion–exclusion algorithm with both optimizations for performing the same task ("Fast algorithm"). We applied these implementations to all possible vertex configurations at different perturbation orders. Importantly, we verified that the results given by the two algorithms are identical within numerical accuracy in all cases. We also measured the average evaluation time per configuration. While the absolute value of this time depends on the implementation details and hardware, the scaling with the perturbation order should be largely independent of such details and can be explored systematically. The result depends to some degree on the details of the model, which may feature symmetries limiting the possible configurations; for the present purpose, we assume no such symmetries, and we have found (not shown) that enforcing symmetries has a relatively small quantitative effect on the results.

Fig. 4 presents the average evaluation time of the sum over all 1-connected diagrams as a function of the perturbation order $2n$ (1-connected diagrams are the worst case for our algorithm, and more general summations over the k-connected diagrams appearing in the Inchworm expansion perform quantitatively, if not qualitatively, better.). In comparison to the brute-force

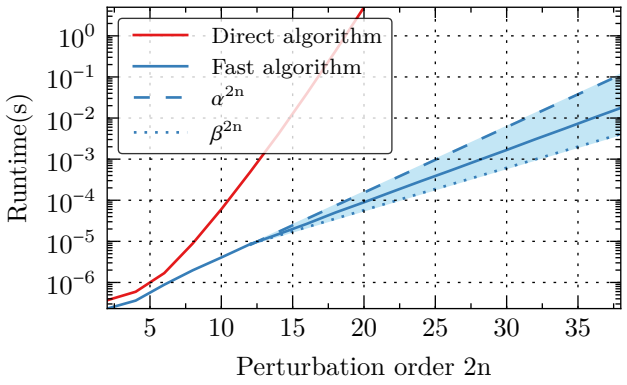

Figure 4. Comparison of runtimes for summing all 1-connected diagrams, using implementations of both the direct algorithm and optimized inclusion–exclusion algorithm. The results are averaged over all possible operator orderings for a model without any symmetries. The theoretical upper bound α^{2n} and approximate lower bound β^{2n} are also shown as dashed and dotted lines, respectively, with the region between them shaded.

method, the inclusion–exclusion algorithm exhibits superior scaling, which appears to be asymptotically exponential as expected. The effective exponent is $\gamma \approx 1.33$, which lies below α, as it must; and also lies above β. We note that since we are averaging over operator orderings, the exponent β must be exact in the asymptotic limit, and any deviation from it is due to the n^3 factor in the complexity Eq. 11. This shows that the inclusion–exclusion algorithm works in practice: it not only scales better than its direct counterpart, but also does not feature a prohibitive prefactor that keeps it from being used for small perturbation orders. In fact, the new algorithm appears to always be faster, even at order 1.

We reiterate that the evaluation time per configuration shown in Fig. 4 is the average over possible operator orderings, as this more closely reflects the use of the algorithm in a physical context. However, it is also possible to consider the worst case. A similar analysis (not shown) then leads to an exponent of $\gamma \approx 1.4$, still within the theoretical bounds but closer to the upper limit. We further note that while asymptotically the factor n^3 in Eqs. (10) and (11) becomes irrelevant, it is straightforward to take it into account at finite n using nonlinear function fitting. While we verified that this procedure has a small quantitative effect on the result, we did not use it in practice.

B. Effect within Inchworm Monte Carlo

Next, we consider what happens when we apply the inclusion–exclusion algorithm to a concrete simulation of a physical model within Inchworm Monte Carlo. We choose the Anderson impurity model addressed in the original Inchworm paper:[37]

$$H = \sum_{\sigma\in\{\uparrow,\downarrow\}} \varepsilon d_\sigma^\dagger d_\sigma + U n_\uparrow n_\downarrow + \sum_{\sigma k} \varepsilon_k a_{\sigma k}^\dagger a_{\sigma k} + \sum_{a\sigma k} \left(\gamma_k a_{\sigma k}^\dagger d_\sigma + \text{H.C.}\right). \tag{20}$$

Here, we set the dot's single-particle energy ε to $\varepsilon = -\frac{U}{2}$, where U is the Hubbard interaction energy, such that the system is particle–hole symmetric. The d_σ and $d_\sigma^\dagger$ are dot fermionic annihilation and creation operators, and the $a_{\sigma k}$ and $a_{\sigma k}^\dagger$ are corresponding operators on the lead. The lead single-particle energies ε_k and the dot–lead hybridization terms γ_k are determined so as to produce a flat band with overall coupling strength Γ, cutoff energy Ω_C and cutoff width $\frac{1}{\nu}$:

$$\begin{aligned}\Gamma(\omega) &\equiv 2\pi \sum_k \gamma_k^* \gamma_k \delta(\omega - \varepsilon_k) \\ &= \Gamma / \left[\left(1 + e^{\nu(\omega-\Omega_c)}\right)\left(1 + e^{-\nu(\omega+\Omega_c)}\right)\right].\end{aligned} \tag{21}$$

Our choice of physical parameters will be motivated by our interest in exploring a problem where high perturbation orders are important. We will therefore arbitrarily select parameters which are particularly difficult for the hybridization expansion. Throughout this work, we set $U = 3\Gamma$, $\Omega_C = 100\Gamma$, $\nu\Gamma = 10$. Additionally, the inverse temperature of the bath is set to $\beta\Gamma = 100$ and its chemical potential is $\mu = 0$. The dot is initially in the unoccupied state and decoupled from the bath; at time zero the coupling is suddenly activated.

In Fig. 5, we plot the time dependence of the dot's probability to be in the unoccupied state in which it began, $P_0(t)$. The dynamics are evaluated using Inchworm Monte Carlo, with the summations over k-connected diagrams performed either directly ("Direct") or by using the inclusion–exclusion algorithm ("Fast"), using the same total amount of computer time. The maximum order of diagrams sampled is limited to either 2 (where the result is not converged) or 14 (where we will soon show that it is converged). Statistical error estimates evaluated by the methods introduced in Ref. 37 are marked by the width of the different curves. Both implementations of the method produce the same result to within numerical accuracy, but the inclusion–exclusion algorithm provides greatly improved accuracy at the higher order.

Next, we consider convergence with the maximum diagram order. In Fig. 6, the results from the inclusion–exclusion-based Inchworm method are plotted at a series of maximum orders. The inset zooms in on the result at the maximum time reached here, $\Gamma t = 2$, where it can be seen that to obtain convergence within the error bars it is necessary to go to orders $2n \gtrsim 12$ or 14. In this case, convergence corresponds to relative errors of $\gtrsim 0.5\%$ in P_0.

The computer time used to obtain each line in Figs. 5 and 6 is constant, and the errors clearly increase with order. We now turn to studying how these errors,

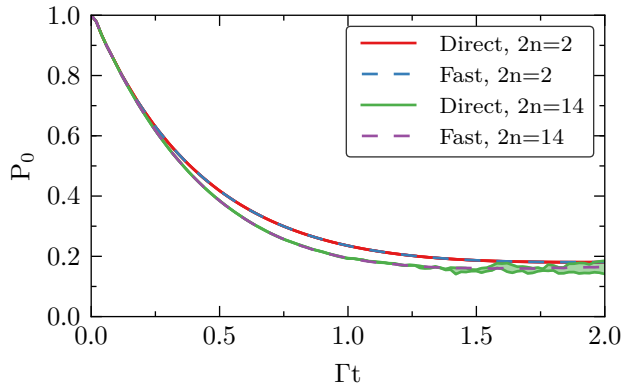

Figure 5. Time dependent population of the unoccupied state in an Anderson impurity model under a coupling quench, using the direct and optimized inclusion–exclusion algorithms. For both algorithms, we perform calculations up to perturbation orders of $2n = 2$ and $2n = 14$.

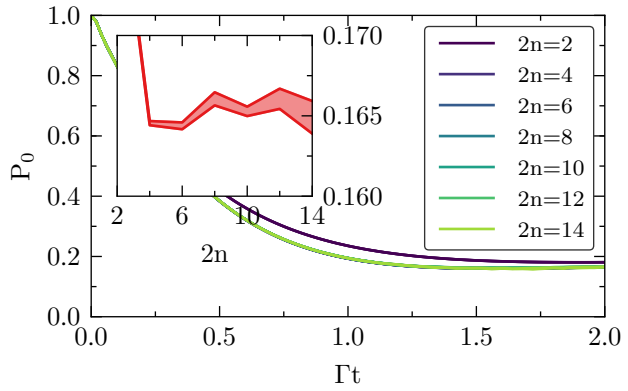

Figure 6. Time dependent population of the unoccupied state in an Anderson impurity model under a coupling quench using the optimized inclusion–exclusion algorithm at several perturbation orders $2n$. The inset shows the data at the final time as a function of the order, showing that order $2n \sim 10 - 14$ is needed to achieve converged results at relative errors of $\sim 0.5\%$.

which become approximately constant at long times, vary with the maximum perturbation order. This procedure is ultimately what will determine the usefulness of the inclusion–exclusion algorithm within the Inchworm method: in practice, a faster summation method allows us to sample more diagrams using the same computational resources, thus reducing the statistical errors.

In Fig. 7, we plot the average error at times $1.8 \leq t \leq 2$ as a function of the maximum perturbation order $2n$, using both algorithms. Outside a small region at $2n = 4$, which is most likely due to statistical fluctuations in our sampling, the new algorithm is substantially faster. At the highest perturbation order we were able to reach using the direct algorithm, $2n = 14$, the inclusion–exclusion algorithm provides errors smaller by approximately an order of magnitude (at higher orders so few diagrams are sampled that the result becomes

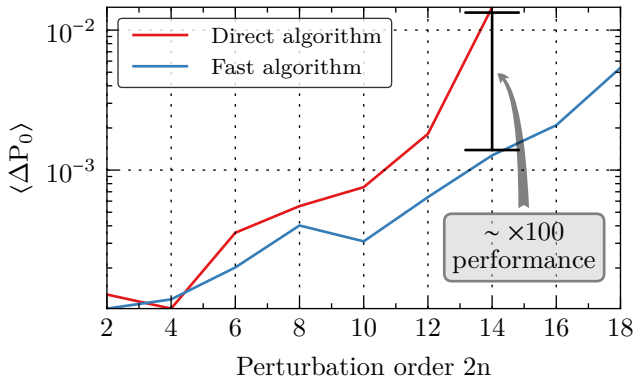

Figure 7. Average errors at long times as a function of the perturbation order $2n$, using the direct and inclusion–exclusion algorithms, for the same parameters used in Figs. 6 and 5. Since Monte Carlo errors scale with the inverse square root of the computation time, an order of magnitude reduction in the error, which is reached at $2n \gtrsim 14$, corresponds to a two order of magnitude enhancement in the computational efficiency.

unreliable without using more computer time). As errors in Monte Carlo procedures scale with the computer time T as $\frac{1}{\sqrt{T}}$, obtaining the same reduction in error with the previous algorithm would entail using approximately two orders of magnitude more computational resources. At even higher orders, we expect this factor to increase rapidly.

IV. CONCLUSIONS

We proposed, analyzed and tested an algorithm based on the inclusion–exclusion principle. The algorithm sums all connected (or k-connected) diagrams in continuous time hybridization expansions, which are needed within Inchworm Monte Carlo methods, in exponential time instead of the previous factorial time. In practice, with two additional optimizations that we proposed, the exponent we found depends to some degree on the model in question, but if no symmetries are taken advantage of the algorithm is $O\left(\gamma^{2n}\right)$ where $\gamma \approx 1.33$ and $2n$ is the perturbation order (odd orders $2n + 1$ can be ruled out for models where the number of fermions is conserved). We also derived a rigorous upper bound and an approximate lower bound for this exponent.

We applied the algorithm to a physical problem requiring high perturbation orders, and showed that at reasonable parameters and accuracy it provides a practical speedup of two orders of magnitude when compared to our previous implementation. We note that this speedup is implementation and optimization dependent: in particular, caching optimizations may speed up the direct algorithm at low orders. We believe overall performance can be improved even further by optimizing parts of the code which had been of negligible com-

putational importance until now. However, the scaling with problem size is universal. Furthermore, a variety of other calculations, in particular those involving Green's functions, will greatly benefit from generalizations of the algorithm introduced here. This will be the subject of future work.

Our algorithm is reminiscent of one which was introduced in Ref. 44 in order to sum connected diagrams in other Monte Carlo methods based on the interaction expansions, where the definition of connectedness is very different. We showed that an idea along similar lines, which we called the "inverted" algorithm, is correct but less efficient than our algorithm for the hybridization expansion. We further showed that our inclusion–exclusion algorithm can be applied to the interaction expansion, but—at least naively—is less efficient than the inverse algorithm in that case. As both ideas are very general in their applicability, it will be of interest to explore their relative merits within other expansions, methods and models in the future.

Looking forward, improving the computational efficiency of the Inchworm method by a practical two orders of magnitude is a major step towards making real-time Monte Carlo a viable alternative to imaginary time techniques. We believe further improvement will stem from this work, such as fast update schemes, and expect the inclusion–exclusion principle to be even more beneficial in Inchworm hybridization expansions for multiorbital impurity models. The same ideas should also be applicable to other Inchworm expansions. The method does not generalize to bosons, where Wick's theorem phrases the sum as a permanent rather than a determinant—exact computation of permanents in polynomial time is thought to be impossible.[50] On the other hand, bosons do not suffer from a fermionic sign problem, and Monte Carlo algorithms for summing boson diagrams work well.[51] It would therefore be of interest to consider the usefulness of the inclusion–exclusion principle within mixed bose–fermi systems.[52] We further believe it will find applications beyond Inchworm—for example, in the evaluation of self energies within bare hybridization expansions, or within bold-line Monte Carlo[30,32–34] and DiagMC techniques.[26–28,53]

Acknowledgements We are grateful to Olga Goulko for directing our attention to Ref. 44. G.C. acknowledges support by the Israel Science Foundation (Grant No. 1604/16). E.G. was supported by DOE ER 46932. This research was supported by Grant No. 2016087 from the United States-Israel Binational Science Foundation (BSF).

[1] J. P. F. LeBlanc, A. E. Antipov, F. Becca, I. W. Bulik, G. K.-L. Chan, C.-M. Chung, Y. Deng, M. Ferrero, T. M. Henderson, C. A. Jiménez-Hoyos, E. Kozik, X.-W. Liu, A. J. Millis, N. V. Prokof'ev, M. Qin, G. E. Scuseria, H. Shi, B. V. Svistunov, L. F. Tocchio, I. S. Tupitsyn, S. R. White, S. Zhang, B.-X. Zheng, Z. Zhu, and E. Gull, Physical Review X **5**, 041041 (2015).

[2] R. Bulla, T. A. Costi, and T. Pruschke, Reviews of Modern Physics **80**, 395 (2008).

[3] P. W. Anderson, Physical Review **124**, 41 (1961).

[4] R. Brako and D. M. Newns, Journal of Physics C: Solid State Physics **14**, 3065 (1981).

[5] L. P. Kouwenhoven, C. M. Marcus, P. L. McEuen, S. Tarucha, R. M. Westervelt, and N. S. Wingreen, in *Mesoscopic Electron Transport*, NATO ASI Series (Springer, Dordrecht, 1997) pp. 105–214.

[6] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge UniversityPress, 1997).

[7] D. Goldhaber-Gordon, H. Shtrikman, D. Mahalu, D. Abusch-Magder, U. Meirav, and M. A. Kastner, Nature **391**, 156 (1998).

[8] R. M. Potok, I. G. Rau, H. Shtrikman, Y. Oreg, and D. Goldhaber-Gordon, Nature **446**, 167 (2007).

[9] A. Aviram and M. A. Ratner, Chem. Phys. Lett. **29**, 277 (1974).

[10] A. Aviram, M. A. Ratner, and E. F. (U.S.), *Molecular electronics: science and technology* (New York Academy of Sciences, 1998).

[11] A. Nitzan, Ann. Rev. Phys. Chem. **52**, 681 (2001).

[12] A. Nitzan and M. A. Ratner, Science **300**, 1384 (2003).

[13] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, Reviews of Modern Physics **68**, 13 (1996).

[14] D. Zgid and E. Gull, New Journal of Physics **19**, 023047 (2017).

[15] N. V. Prokof'ev and B. V. Svistunov, Physical Review Letters **81**, 2514 (1998).

[16] E. Gull, A. J. Millis, A. I. Lichtenstein, A. N. Rubtsov, M. Troyer, and P. Werner, Reviews of Modern Physics **83**, 349 (2011).

[17] A. N. Rubtsov, V. V. Savkin, and A. I. Lichtenstein, Physical Review B **72**, 035122 (2005).

[18] P. Werner, A. Comanac, L. de' Medici, M. Troyer, and A. J. Millis, Physical Review Letters **97**, 076405 (2006).

[19] P. Werner and A. J. Millis, Physical Review Letters **99**, 146404 (2007).

[20] E. Gull, P. Werner, O. Parcollet, and M. Troyer, EPL (Europhysics Letters) **82**, 57003 (2008).

[21] L. Mühlbacher and E. Rabani, Physical Review Letters **100**, 176403 (2008).

[22] P. Werner, T. Oka, and A. J. Millis, Physical Review B **79**, 035320 (2009).

[23] P. Werner, T. Oka, M. Eckstein, and A. J. Millis, Physical Review B **81**, 035108 (2010).

[24] M. Schiró and M. Fabrizio, Physical Review B **79**, 153302 (2009).

[25] M. Schiró, Physical Review B **81**, 085126 (2010).

[26] N. Prokof'ev and B. Svistunov, Physical Review Letters **99**, 250201 (2007).

[27] N. V. Prokof'ev and B. V. Svistunov, Physical Review B **77**, 125101 (2008).

[28] N. Prokof'ev, in *Strongly Correlated Systems*, Springer Series in Solid-State Sciences No. 176, edited by A. Avella and F. Mancini (Springer Berlin Heidelberg, 2013) pp. 273–292.

[29] E. Kozik, M. Ferrero, and A. Georges, Physical Review Letters **114**, 156402 (2015).
[30] E. Gull, D. R. Reichman, and A. J. Millis, Physical Review B **84**, 085134 (2011).
[31] G. Cohen, E. Gull, D. R. Reichman, A. J. Millis, and E. Rabani, Physical Review B **87**, 195108 (2013).
[32] G. Cohen, D. R. Reichman, A. J. Millis, and E. Gull, Physical Review B **89**, 115139 (2014).
[33] G. Cohen, E. Gull, D. R. Reichman, and A. J. Millis, Physical Review Letters **112**, 146802 (2014).
[34] E. Gull, D. R. Reichman, and A. J. Millis, Physical Review B **82**, 075109 (2010).
[35] G. Cohen and E. Rabani, Physical Review B **84**, 075150 (2011).
[36] G. Cohen, E. Y. Wilner, and E. Rabani, New Journal of Physics **15**, 073018 (2013).
[37] G. Cohen, E. Gull, D. R. Reichman, and A. J. Millis, Physical Review Letters **115**, 266802 (2015).
[38] A. E. Antipov, Q. Dong, J. Kleinhenz, G. Cohen, and E. Gull, Physical Review B **95**, 085144 (2017).
[39] Q. Dong, I. Krivenko, J. Kleinhenz, A. E. Antipov, G. Cohen, and E. Gull, Physical Review B **96**, 155126 (2017).
[40] H.-T. Chen, G. Cohen, and D. R. Reichman, The Journal of Chemical Physics **146**, 054105 (2017).
[41] H.-T. Chen, G. Cohen, and D. R. Reichman, The Journal of Chemical Physics **146**, 054106 (2017).
[42] M. Ridley, V. N. Singh, E. Gull, and G. Cohen, Physical Review B **97**, 115109 (2018).
[43] R. Rossi, N. Prokof'ev, B. Svistunov, K. V. Houcke, and F. Werner, EPL (Europhysics Letters) **118**, 10004 (2017).
[44] R. Rossi, Physical Review Letters **119**, 045701 (2017).
[45] J. W. Negele and H. Orland, *Quantum many-particle systems* (Westview Press, 1998).
[46] R. Rossi, arXiv:1802.04743 [cond-mat] (2018), arXiv: 1802.04743.
[47] A. Moutenet, W. Wu, and M. Ferrero, Physical Review B **97**, 085117 (2018).
[48] F. Simkovic and E. Kozik, arXiv:1712.10001 [cond-mat] (2017), arXiv: 1712.10001.
[49] D. Berend and T. Tassa, Probability and Mathematical Statistics-Poland **30**, 185 (2010), wOS:000208490200001.
[50] L. G. Valiant, Theoretical Computer Science **8**, 189 (1979).
[51] L. Pollet, Reports on Progress in Physics **75**, 094501 (2012).
[52] H.-T. Chen, G. Cohen, A. J. Millis, and D. R. Reichman, Physical Review B **93**, 174309 (2016).
[53] N. V. Prokof'ev, B. V. Svistunov, and I. S. Tupitsyn, Physics Letters A **238**, 253 (1998).
[54] A. Schiller and K. Ingersent, Physical Review Letters **75**, 113 (1995).
[55] M. H. Hettler, A. N. Tahvildar-Zadeh, M. Jarrell, T. Pruschke, and H. R. Krishnamurthy, Physical Review B **58**, R7475 (1998).
[56] T. N. Lan and D. Zgid, The Journal of Physical Chemistry Letters **8**, 2200 (2017).
[57] J. Vučičević, N. Wentzell, M. Ferrero, and O. Parcollet, Physical Review B **97**, 125141 (2018).
[58] G. D. Mahan, *Many-Particle Physics* (Plenum Press, New-York, 1990).
[59] A. L. Fetter and J. D. Walecka, *Quantum theory of many-particle systems* (Dover Pubns, 2003).

Appendix A: Derivation

In this appendix, we will introduce a precise phrasing of the celebrated inclusion–exclusion principle, and show how it can be used to derive Eq. 3. This principle is most often stated in terms of counting the size of a union. For example,consider two sets A and B. The size of their union can be written

$$|A \cup B| = |A| + |B| - |A \cap B| \,. \tag{A1}$$

However, if one is given three sets A, B and C, the union is:

$$\begin{aligned} |A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| \\ - |A \cap C| - |B \cap C| + |A \cap B \cap C| \,, \end{aligned} \tag{A2}$$

and for N sets S_i, one can write a general expression in the form

$$\begin{aligned} \left|\cup_{i=1}^{N} S_i\right| = \sum_{i_1} |S_{i_1}| - \sum_{i_1<i_2} |S_i \cap S_j| \\ + \cdots \sum_{i_1<\cdots<i_N} (-1)^{N-1} \left|S_{i_1} \cap \cdots \cap S_{i_n}\right| . \end{aligned} \tag{A3}$$

The inclusion–exclusion principle has a long history and many uses in combinatorics. Interestingly, it has also found applications in the context of nonlocal extensions to dynamical mean theory,[14,54–57] though these works did not explicitly call it by this name.

Here, we use a trivial generalization from the size of the sets to a generic scalar property. The inclusion–exclusion principle as we use it deals with a set S, a collection of subsets thereof $\{A_i\}$, and a function $f\colon s \in S \to \mathbb{C}$. It states that the sum of f over elements of S that are not the elements of any A_i can be computed by first taking the sum over the values of f for all elements of S, then subtracting the sums of values of f for all subsets A_i, then adding the values of f for those elements which we have subtracted twice (the elements of all sets $A_i \cap A_j$), and so on. This leads to the following equality:

$$\begin{aligned} f \equiv \sum_{x \in S \setminus (\cup_i A_i)} f(x) = \sum_{x \in S} f(x) - \sum_i \sum_{x \in A_i} f(x) \\ + \sum_{i<j} \sum_{x \in A_i \cap A_j} f(x) - \ldots \\ = \sum_{j=0}^{j_{\max}} (-1)^j \sum_{i_0<\cdots<i_{j-1}} \sum_{x \in A_{i_0} \cap \cdots \cap A_{i_{j-1}}} f(x) \,. \end{aligned} \tag{A4}$$

To obtain our algorithm, we set D to be the set of all diagrams over vertices S and f to be the function that associates values with diagrams. We define A_S to be the set of connected diagrams over S, and obtain $\{A_s\}$ for every j by collecting all sets of j disjoint subsegments of S comprising only proper vertices. Since (a) every diagram which isn't k-connected has a connected fully

proper segment; and (b) sets of connected segments are necessarily disjoint, this leaves only proper diagrams. Applying the inclusion–exclusion principle we immediately get Eq. (9), using the fact that the sum over values of diagrams for which the segments $S_0, \ldots, S_{j-1}$ are connected is $c(S_0) \cdots c(S_{j-1})\, v\left(V \backslash \cup_{i=0}^{j-1} C(s_i)\right)$.

Appendix B: Computational efficiency

In this appendix, we show how the theoretical bounds α and β can be derived for the different approximations discussed above. A fully analytical combinatorial computation is possible, but lengthy. Since we are interested only in the asymptotic scaling, we will limit ourselves to scaling calculations based on the pole structure of the generating functions of the relevant sequences.

1. Upper bound

To simplify the derivation, we will ignore the distinction between creation and annihilation operators and analyze the complexity of the generic algorithm in which every vertex can be paired to other vertex. Of course, it is possible to implement the physical expansion with this algorithm by setting elements of of M_{ij} between operators of the same type to zero. However, the number of diagrams that needs to be summed is exponentially larger and it is clear that this will provide an overestimate of the fermionic algorithm; the result will therefore be useful as an upper bound.

Throughout this subsection, we will assume that V is the set of proper vertices, of size m, and that there exist k additional improper vertices. In the physical case, one would have $m = 2n - k$.

a. Unoptimized algorithm

Consider first how Eq. (9) is used in practice: we must compute $c(S)$ for all segments $S \subset V$ in increasing order of size, as each segment will depend on results involving smaller segments. Finally, $c(V)$ will be evaluated. Each step takes a number of evaluations of v equal to the number of ways to choose sets of disjoint segments. Let a_m denote this number for a set of size m The complexity of each step is then $O\left((m+p)^3 a_m\right)$, since the most expensive part for each set of segments is the evaluation of v. Since the number of steps of each size smaller than the final m is polynomial, while a_m turns out to be exponential in m, the final step dominates the complexity.

We now continue to the combinatorial calculation of a_m. Each set of segments either contains a segment including the last point, or does not. If it does, and this segment is of length ℓ (which is even, as in a subset of odd length not all vertices can be paired), then we are left with $a_{m-\ell}$ options to choose the rest of the subsets. If it has no segment including the last point, there are a_{m-1} options. Therefore, we get

$$a_m = a_{m-1} + a_{m-2} + a_{m-4} + \cdots + a_0 \tag{B1}$$

for any $m > 0$, and for convenience we set $a_0 = 1$.

To find the asymptotic growth rate of the sequence, it is useful to consider its generating function $f(x) \equiv \sum_{m=0}^{\infty} a_m x^m$. Using Eq. (B1),

$$\sum_{m=1}^{\infty} \left(a_m - a_{m-1} - a_{m-2} - a_{m-4} - \ldots\right) x^m = 0, \tag{B2}$$

or

$$f(x) - a_0 - \left(x + x^2 + x^4 + \ldots\right) f(x) = 0. \tag{B3}$$

Using our value for a_0 and summing the series, we obtain

$$\left(1 - x - \frac{x^2}{1-x^2}\right) f(x) = 1, \tag{B4}$$

such that

$$f(x) = \frac{1-x^2}{1-x-2x^2+x^3}. \tag{B5}$$

If asymptotically $a_m \sim \alpha^m$, $f(x)$ will have a pole with absolute value $\frac{1}{\alpha}$ and no poles with smaller absolute value. The smallest pole by absolute value is at $|x_{\min}| \approx 0.55495$, such that $\alpha = \frac{1}{|x_{\min}|} \approx 1.8019$.

b. Effect of optimizations

We can obtain an analogous formula for the number a_m of ways to choose non-adjacent disjoint subsegments of m vertices by considering three options at every stage: (a) there is no segment containing the last point, giving a_{m-1} possible choices; (b) there is a segment of length ℓ containing the last point, before which there is a vertex which is not an element of any segment, giving $a_{m-2\ell-1}$ choices; and (c) there is a single subsegment that contains all vertices, giving 1 option if m is even. Therefore,

$$a_m = \sum_{\ell=0}^{m/2} a_{m-2\ell-1} + (1 \text{ if } m \text{ is even}). \tag{B6}$$

As before, we can show that the generating function $f(x)$ for this sequence satisfies

$$f(x) = \frac{1}{1-x-x^2}, \tag{B7}$$

for which the growth rate is the golden ratio $\alpha \approx 1.6180$.

After the second optimization, we need not count segments of size 2. By analogous considerations this gives a sequence generated by

$$f(x) = \frac{1-x^2+x^4}{1-x-x^2+x^3-x^5}, \tag{B8}$$

which has the growth rate $\alpha \approx 1.4432$.

c. Inverted algorithm

To evaluate the performance of the algorithm implied by Eq. 17, we need to calculate the number of ways to choose sets of subsegments containing a total of ℓ vertices, from a set with n vertices. We will denote this number by s_m^ℓ. Given that, the runtime of the algorithm is given by $r_m = \sum_{\ell \leq m} s_m^\ell a_\ell$, where a_ℓ is the same runtime we evaluated for the inclusion–exclusion algorithm with the first optimization. Since we only care about the asymptotic growth rate, and we found that $a_\ell = O\left(\beta^\ell\right)$, then

$$r_m = O\left(\sum_{\ell \leq m} s_m^\ell \beta^\ell\right). \tag{B9}$$

Let $S(x,\beta) \equiv \sum_{m,\ell} s_m^\ell x^m \beta^\ell$ be the generating function of s_m^ℓ. The growth rate of r_m is therefore given by that of the coefficient of x^m in $S(x,\beta)$.

Therefore, we are only left with the task of evaluating $S(x,t)$. By similar arguments to those used before, s_n^ℓ satisfies

$$s_m^\ell = \sum_j \left(s_{m-2j-1}^{\ell-2j} + (1 \text{ if } m = \ell = 2j)\right), \tag{B10}$$

so

$$S(x,\beta) = \frac{1}{1 - x - x^2\beta^2}. \tag{B11}$$

Substituting $\beta \approx 1.8019$ from the first optimization case and repeating the procedure from before, we get that $\alpha \approx 2.1935$. Similarly, with the second optimization, $\alpha \approx 1.8718$.

2. Average over operator orders

We will now consider a case closer to the one which is of physical interest, by taking into account the fact that vertices consist of either creation or annihilation operators and that pairing can only occur between operators of different type. This is a more complex calculation, and we will only show how it is performed for the unoptimized case. The effect of the optimizations and of the growth rate of the inverted algorithm can be similarly calculated.

To address this case, we will first count the number of ways $a_{n,n}$ to partition $2n$ vertices into two subsets of n vertices corresponding to creation and annihilation operators and then choose subsegments of the full set containing the same number of operators of each type. To perform the averaging over the possible operator orders, we will then divide this result by the number of ways to partition the operators into the two types, $\binom{2n}{n}$. Since $a_{n,n}$ will turn out to be exponential in n and $\binom{2n}{n} \simeq \frac{4^n}{\sqrt{n}}$, the growth rate for the average will be the growth rate of $a_{n,n}$ divided by 4.

It turns out that it is easier to solve a slightly more general combinatorial problem: the number of ways $a_{m,n}$ to partition $m+n$ vertices into two subsets, one containing m vertices and the other containing n vertices, and then choose subsegments accordingly. This obeys the following recurrence relation:

$$a_{m,n} = a_{m-1,n} + a_{m,n-1} + \sum_{j=1}^{m} \binom{2j}{j} a_{m-j,n-j}, \tag{B12}$$

for $m, n > 0$ and $a_{0,0} = 1$. Therefore, the generating function of $a_{m,n}$, $g(x,y) \equiv \sum_{m,n} a_{m,n} x^m y^n$, satisfies the equation

$$\begin{aligned} g(x,y) &= (x+y)\, g(x,y) \\ &\quad + \left(\frac{1}{\sqrt{1-4xy}} - 1\right) g(x,y) + 1, \end{aligned} \tag{B13}$$

where we have used the fact that

$$\sum_{k=0}^{\infty} \binom{2k}{k} x^k = \frac{1}{\sqrt{1-4x}}. \tag{B14}$$

Solving this, we obtain

$$\begin{aligned} g(x,y) &= \frac{1}{2 - x - y - (1-4xy)^{-\frac{1}{2}}} \\ &= \frac{1}{2 - (1-4xy)^{-\frac{1}{2}}} \cdot \frac{1}{1 - \frac{x+y}{2-(1-4xy)^{-\frac{1}{2}}}}. \end{aligned} \tag{B15}$$

We are actually interested in the sequence $a_{n,n}$, and its generating function $f(x) = \sum_n a_{n,n} x^n$. However, $f(xy)$ contains the terms of $g(x,y)$ which have the same power of x and y. Since the only term in $g(x,y)$ that can contribute differently in Eq. (B15) is $x+y$, we can expand the second fraction in a series:

$$\begin{aligned} \frac{1}{1 - \frac{x+y}{2-(1-4xy)^{-\frac{1}{2}}}} &= \sum_{k=0}^{\infty} \left(\frac{x+y}{2 - (1-4xy)^{-\frac{1}{2}}}\right)^k \\ &= \sum_{k=0}^{\infty} \left(2 - (1-4xy)^{-\frac{1}{2}}\right)^{-k} \\ &\quad \times \sum_{i=0}^{k} \binom{i}{j} x^i y^{k-i}. \end{aligned} \tag{B16}$$

As we only want the terms with equal powers of x and y, we need only take the terms with $i = k - i$, *i.e.* $k = 2i$. With this,

$$f(xy) = \frac{1}{2-(1-4xy)^{-\frac{1}{2}}} \times \sum_{i=0}^{\infty} \left(2-(1-4xy)^{-\frac{1}{2}}\right)^{-2i} \begin{pmatrix} 2i \\ i \end{pmatrix} (xy)^i , \tag{B17}$$

from which we can obtain

$$f(x) = \left(\left(2-(1-4x)^{-\frac{1}{2}}\right)^2 - 4x\right)^{-\frac{1}{2}} . \tag{B18}$$

Finally, substituting $x = \frac{1}{4\beta^2}$, we get that the growth rate is the largest solution of $\frac{1}{\beta} = 2 - \left(1 - \frac{1}{\beta^2}\right)^{-\frac{1}{2}}$, such that $\beta \approx 1.5072$.

Appendix C: Terminology

Below, a brief summary of relevant diagrammatic terminology is provided as a convenient reference to the reader. Some redundancy with the main text exists. Our choice of wording for the various terms was motivated by an attempt to be as consistent as possible with the existing many-body physics literature, and we offer some background to explain our considerations below. This section should not be considered an introduction to the formalism: it is simply a review of definitions.

Interaction diagrams Most textbooks on many-body theory discuss Feynman diagrams for many-body correlation functions, which are commonly called Green's functions, within a ("weak-coupling") perturbation series in the Coulomb interaction.[58,59] While this is not the main focus of the present work outside of Sec. II E, we will briefly discuss these more widely familiar objects to establish the source of the terminology we used.

Interaction diagrams are essentially graphs: they comprise a set of vertices connected by edges. Each edge corresponds to a noninteracting single-particle correlation function. Other kinds of edges may be present which describe interactions, depending on the details of the model. The operators that we wish to calculate the correlation function of are denoted by vertices with one edge, which are called *external*. Other vertices describe an interaction and have four edges, and are called *internal*. Diagrams may be *connected* or *disconnected*, in precisely the graph theoretical sense.

Subgraphs of a diagram will be called *pieces* in the present context (the terms "insertions" or "inclusions" are also used in the literature). A piece formed by taking a graph with two external vertices, and removing those external vertices and the adjoining edges, is called a *self-energy piece*. A self-energy piece is then called *proper* (thus being a contribution to the *proper self-energy*) if it is one-particle irreducible, *i.e.* it cannot be separated into two disconnected pieces by removing a single edge. Such proper self-energy pieces are used within Dyson resummations.

Hybridization diagrams The hybridization (or "strong-coupling") expansion is somewhat more specialized and the corresponding diagrammatic language is less standardized. Our choice of terminology is therefore by no means unique. A more detailed explanation of the diagrammatic structure and its source was given in sec. II, and below we only review the main terms in an abstract fashion.

A diagram in the hybridization expansion is described by a graph in which some set of vertices (called the *configuration*) is drawn along a contour symbolizing time (see Fig. 1). The vertices represent an equal number of creation and annihilation operators, and accordingly half are colored black, and the other half white. A configuration with $2n$ vertices induces a factorial number of diagrams, uniquely identified by a corresponding pairings of the vertices. Each pair is connected by a *hybridization line*.

A subset of the vertices in a diagram, along with the hybridization lines connecting them, will be referred to as a *piece* of the diagram if it is itself a legal diagram (*e.g.*, diagram 1 in the top panel of Fig. 1 comprises three pieces, while diagram 3 comprises only one). A piece comprising only vertices which are adjacent (along the time axis) is a *segment*. A segment comprising only vertices contained within a second segment is its *subsegment*. Two segments are *adjacent* if they are not separated by any vertices.

Next, we say that two hybridization lines, which we denote by $E_1 = (v_1, v_2)$ and $E_2 = (v_3, v_4)$ where the v_i are vertices, are *crossing* if the time intervals defined by E_1 and E_2 overlap. We note that this is consistent with the alternative definition that the drawn hybridization lines must cross for the particular way in which we draw hybridization diagrams, *i.e.* as a planar graph with all lines above the time line (*e.g.*, diagram 1 in the top panel of Fig. 1 comprises three noncrossing lines, while all three lines in diagram 3 are crossing).

Two lines, and their corresponding vertices, are then said to be *connected* if they cross. Connectedness is transitive, so a segment, subset or entire diagram can also be said to be connected if all its lines are connected (*e.g.*, diagram 3 in Fig. 1 is connected). In the hybridization expansion, all connected diagrams contribute to the self-energy, and therefore connected diagrams or pieces are also said to be *proper*.

In the Inchworm expansion, it is necessary to generalize propriety and connectedness to k-connectedness. A diagram is *k-connected* if each of its connected components contains at least one of k special vertices, which we call *improper* vertices; all other vertices are *proper*. If there is only one improper vertex, k-connectedness corresponds exactly to normal connectedness and the choice of special vertex is arbitrary. A 1-connected diagram is therefore also connected and proper. k-connected diagrams with $k > 1$ have some degree of "impropriety", which increases with the number of improper vertices; hence the name.