15 Diagrammatic Monte Carlo

Nikolay Prokof'ev
University of Massachusetts
666 North Pleasant Street, Amherst, MA 01003, USA

Contents

E. Pavarini, E. Koch, and S. Zhang (eds.)
Many-Body Methods for Real Materials
Modeling and Simulation Vol. 9
Forschungszentrum Jülich, 2019, ISBN 978-3-95806-400-3
http://www.cond-mat.de/events/correl19

1 Introduction

This contribution reviews the principles and key ideas behind the Diagrammatic Monte Carlo method (DiagMC), as well as some technical details important for its efficient practical implementation. In short, DiagMC is a set of generic rules for unbiased sampling of the configuration space that involves a varying number of continuous variables. Applications of the method include series of connected Feynman diagrams and non-linear integral equations, lattice- and continuous-space path-integrals, continuous-time impurity-solvers, and any problem where the answer can be formally represented by the sum of multi-dimensional integrals. Once the configuration space to be simulated is defined, the DiagMC method will ensure that stochastic sampling is performed without systematic bias, leaving statistical error bars as the only source of uncertainty on the final answer.

Properties of large systems cannot be obtained by direct enumeration of the exponentially growing configuration/Hilbert space, or ν-space, for brevity. A variety of numerical schemes rely on mathematical formulations instead, which, if solved, would reproduce the same statistical predictions as the original model. Path integrals, high-temperature expansions, and Feynman diagrams belong to this category of methods. I will focus on the Monte Carlo (MC) sampling technique [1], which is, arguably, among the most powerful universal tools designed to deal with large and complex ν-spaces, and explain in detail how it works in the space of connected Feynman diagrams. While each implementation is model and representation specific, most rules and considerations are generic.

2 Diagrammatic Monte Carlo

In the most abstract form one is interested in knowing some quantity $Q(\mathbf{y})$ as a function of variable $\mathbf{y}$ (in general, the multi-dimensional variable $\mathbf{y}$ may include both continuous and discrete components) when the answer is expressed as a series of multi-dimensional integrals/sums

$$Q(\mathbf{y}) = \sum_{n=0}^{\infty} \sum_{\mathbb{T}} \int \cdots \int d\mathbf{x}_1 \cdots d\mathbf{x}_n \, D(n, \mathbb{T}; \mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{y}) \,, \tag{1}$$

with D being some known function of its arguments. The "diagram order" n controls the number of "internal" integration/summation variables, $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and the "topology" index $\mathbb{T}$ labels different terms of the same order in the series. The most familiar physics example, as the name of the technique suggests, would be Feynman diagrams for the many-body system illustrated in Fig. 1. Strict diagrammatic rules relate the graphical representation to the corresponding mathematical expression for the function D: up to a phase factor, it is given by the product of functions associated with the graph lines (often called propagators), $D = \prod_{\text{lines}} F_{\text{line}}$. For example, in momentum-imaginary time representation for the system of electrons interacting via the pairwise potential $V(\mathbf{r})$, the dotted lines are associated with the Fourier transform, $V(\mathbf{q})$, of the interaction potential and the solid lines with the single-particle propagators $G_0(\mathbf{p}_i, \tau)$.

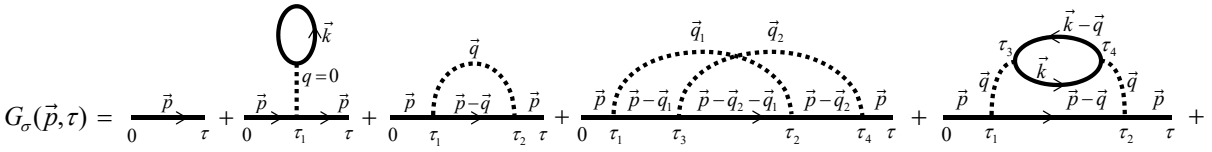

Fig. 1: *Graphical representation of the diagrammatic expansion for the Green function of an interacting many-body system.*

There are well-established diagrammatic series for other quantities of interest such as self-energies, polarization operators, pair-propagators, current-current and other correlation functions, etc. Numerous alternative representations of quantum and classical models, such as path integrals and impurity solvers, are mathematically identical to Eq. (1). Thus, regardless of the origin of Eq. (1), it can be viewed as a mathematical expression for the answer in terms of a series of multi-dimensional integrals. The real challenge is to evaluate it with high accuracy.
Let us denote the collection of all external and internal parameters that lead to a complete characterization of the diagram as $\nu = (n, \mathbb{T}; \mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{y})$, and call it the "configuration space;" a particular set of parameters has to be viewed as a point in $\{\nu\}$. Accordingly, the modulus of D_ν will be called the configuration "weight." Since, in general, the D-function is not sign-positive, we will need to introduce also the configuration "phase," $\varphi_\nu = \arg D_\nu$ (the diagram phase is not necessarily equal to 0 or π).

2.1 Updates: general principles

The MC process of generating diagrams with probabilities proportional to their weight is based on the conventional Markov-chain updating scheme [2–4] implemented directly in the space of continuous variables. All updates are broadly classified as type-I and type-II. The number of continuous variables is not changed in type-I updates that perform sampling of diagrams of the same order n. Typical examples are shown in Fig. 2. They are based on the simplest possible local modifications of the topology and line parameters allowed by the rules and conservation laws. Their implementation is straightforward; e.g., for the update illustrated in Fig. 2(a) select at random any pair of consecutive interaction vertices and exchange their places. An acceptance ratio for the corresponding update, $R_{\nu\to\nu'}$, is given by the ratio of the diagram weights,

$$R_{\nu\to\nu'} = |D_{\nu'}/D_\nu|\,, \tag{2}$$

which is easily calculated, since D_ν is the product of $F_{\rm line}$-functions and only three of them change their values in this update. Changing internal or external variables, see Figs. 2(b) and 2(c), is also standard. For example, one may select at random some interaction vertex and propose a new value for its time variable, $\tau_i \to \tau_i'$, from the (arbitrary) normalized probability density $P(\tau_i')$. The acceptance ratio for this update is given by the ratio of probabilities for suggesting the $\nu \to \nu'$ and $\nu' \to \nu$ moves times the ratio of the diagram weights

$$R_{\nu\to\nu'} = \left|\frac{D_{\nu'}}{D_\nu}\right| \frac{P(\tau_i)d\tau}{P(\tau_i')d\tau} = \left|\frac{D_{\nu'}}{D_\nu}\right| \frac{P(\tau_i)}{P(\tau_i')}\,. \tag{3}$$

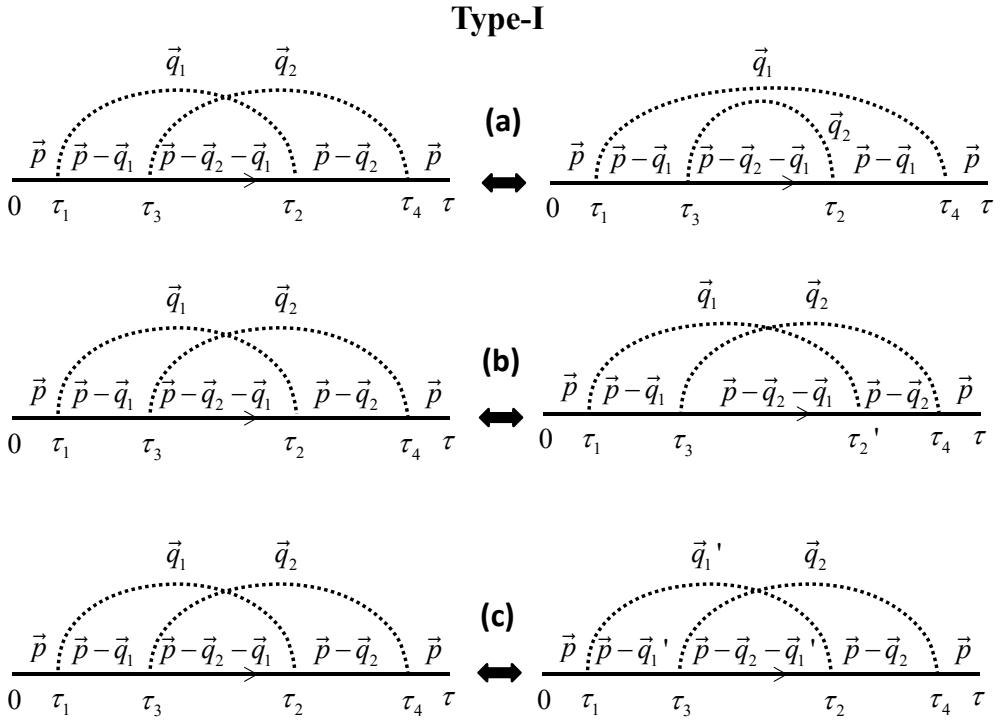

Fig. 2: *Typical type-I updates in the configuration space of Feynman diagrams for polarons: (a) changing the diagram topology by permuting the end-points of two dashed lines; (b) changing the value of the internal variable τ_2 to τ_2'; (c) changing the momentum transfer along the dashed line from $\mathbf{q}_1$ to $\mathbf{q}_1'$.*

The simplest implementation of this update would be to have non-zero $P(\tau_i)$ only on the time interval determined by the times of the previous and following interaction vertexes (times τ_3 and τ_4 in Fig. 2(b)). The probability distribution $P(\tau)$ should be optimized for the best acceptance ratio without compromising one's ability to use it for fast generation of random variables (more details will be provided when we discuss the practical implementation of the technique for a Fröhlich polaron).

Clearly, there are numerous other possibilities for type-I updates which are standard for MC simulations of multidimensional integrals. For pedagogical reasons I will keep mentioning differential measures when I first state the acceptance ratio in order to see explicitly how they cancel out in the final answer.

Type-II updates change the diagram order $n \leftrightarrow n+m$ (they form complementary pairs of updates) and thus require that new variables be proposed from some (arbitrary) normalized probability density distribution $W(\nu; \mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m})$ when going from n to $n+m$, or erased from the diagram when going from $n+m$ to n. For example, to implement the transformation illustrated in Fig. 3 we need to propose time positions τ_3 and τ_4 and the momentum transfer $\mathbf{q}_2$ for the new dashed line. In the reverse update, these variables need to be erased. The detailed balance equation for a pair of updates which increase/decrease the diagram order by m reads

$$r_{n\to n+m}\, u_{n\to n+m}\, |D_\nu|\, W(\nu; \mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m})(d\mathbf{x})^{n+m} = r_{n+m\to n}\, u_{n+m\to n}\, |D_{\nu'}|(d\mathbf{x})^{n+m}\,, \quad (4)$$

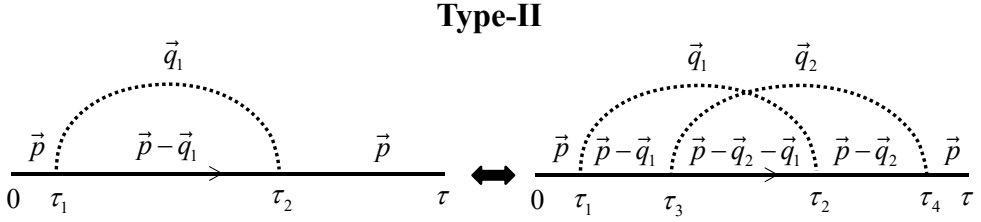

Fig. 3: *Type-II updates that increase/decrease the diagram order by one.*

where $u_{n\to n+m}$ and $u_{n+m\to n}$ are algorithm-specific probabilities of deciding which new diagram elements will be added or removed (for specific details, see the Fröhlich polaron Section), respectively, while $r_{n\to n+m}$ and $r_{n+m\to n}$ are the probabilities of accepting the update. An acceptance ratio, $R_{\nu\to\nu'} = r_{n\to n+m}/r_{n+m\to n}$, to go from configuration ν of order n to configuration ν' of order $n+m$ is then

$$R_{\nu\to\nu'} = \frac{u_{n\to n+m}}{u_{n+m\to n}} \left| \frac{D_{\nu'}}{D_\nu\, W(\nu; \mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m})} \right| . \tag{5}$$

As expected, all differential measures cancel in the acceptance ratio meaning that sampling of the configuration space with arbitrary and ever changing number of continuous variables can be done without encountering systematic errors. Note that the ratio of the diagram weights, $|D_{\nu'}/D_\nu|$, is some model-specific function of ν and the new variables $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}$. The optimal choice of W is then a compromise between the efficiency (and programming convenience) of using it for seeding new variables and the largest (on average) acceptance ratio.

It is relatively straightforward to design a set of type-I and type-II updates that satisfies the ergodicity requirement: given two arbitrary configurations ν and ν' contributing to the answer, it should take the algorithm a finite number of updates with non-zero acceptance ratios to transform one configuration into another. At this point I would like to stress that DiagMC is fundamentally different from enumerating/listing all diagrams with orders $n \le n_{\max}$ and then computing the corresponding multidimensional integrals for each diagram separately using classical MC methods. In DiagMC the diagram order, its topology, and all internal and external variables are treated on equal footing and are sampled stochastically from the probability distribution D_ν. From the DiagMC perspective, each diagram represents a point, not an integral, in the configuration space ν, see Fig. 4, and each sampled point, no matter the diagram order n, contributes equally to the statistics of the final result. For example, every diagram shown in Fig. 1 (with all internal variables specified) contributes $e^{i\varphi_\nu}$ to the statistics of $G(\mathbf{p}, \tau)$. One may wonder where did all the integrals go and why do the configuration space points with different differential measures contribute equally? Formally, this is what the detailed balance equation (5) is telling us. The other way to answer the question is as follows. One may pretend that *all* diagrams *are* of the same order(!) by interpreting unity factors in terms of the normalization integrals for W-functions in Eq. (5)

$$\int\cdots\int d\mathbf{x}_1 \cdots d\mathbf{x}_n\, D_\nu \times 1 = \int\cdots\int d\mathbf{x}_1 \cdots d\mathbf{x}_{n+m}\, D_\nu \times W(\nu; \mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}). \tag{6}$$

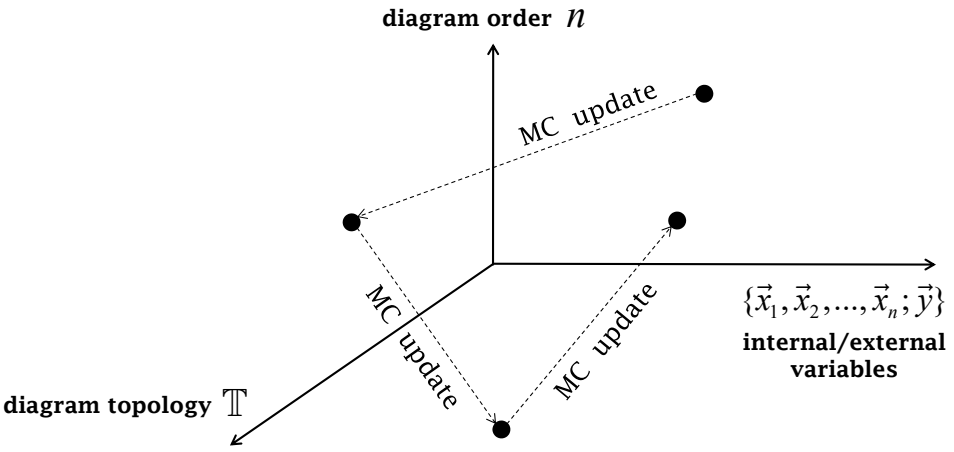

Fig. 4: *An illustration of the DiagMC process: all configuration parameters are treated in the simulation protocol on an equal footing and are subject to local Markov-chain updates.*

This point of view literally reduces type-II updates to type-I updates with the only caveat that one is free to consider any normalized W function for performing this "match of dimensions." One last note. In DiagMC the autocorrelation time is almost never a problem. When the algorithm takes the configuration to the lowest-order diagram (at this point nearly all variables are erased), it can be de-correlated in $O(1)$ updates. Given that most many body simulations are done for expansion orders $\lesssim 10$, the autocorrelation times is measured in fractions of a millisecond for local updating schemes.

2.2 Normalization

Statistics collected for $Q(\mathbf{y})$, let us denote it as $Q_{MC}(\mathbf{y})$, grows linearly with the simulation time, and needs to be properly normalized to produce a physically meaningful result. Note that Eq. (1) is not based on a ratio of two quantities and, thus, normalization is done differently from the textbook example of the Ising model. Suppose that the lowest-order term in Eq. (1) for some value $\mathbf{y}_0$ is known because it does not involve any integrals and reduces to the analytic expression for $|Q^{(0)}(\mathbf{y}_0)| = |D(0;\mathbf{y}_0)|$. Stochastic sampling eventually brings the configuration to the lowest order, and this is when we update the normalization counter, $Z_N = Z_N + \delta_{n,0}\,\delta(\mathbf{y}-\mathbf{y}_0)$. We also realize that full statistics is the sum of contributions collected from different orders, $Q_{MC}(\mathbf{y}) = Q_{MC}(n=0,\mathbf{y}) + Q_{MC}(n>0,\mathbf{y})$, and if we were to determine the lowest-order contribution to the modulus of $Q(\mathbf{y}_0)$ we would get Z_N. This immediately tells us that the properly normalized answer for the final result is

$$Q(\mathbf{y}) = \left|Q^{(0)}(\mathbf{y}_0)\right| \frac{Q_{MC}(\mathbf{y})}{Z_N}\,. \tag{7}$$

The entire protocol can be called "normalization to known result."
If there is some doubt that statistics for a given point $\mathbf{y}_0$ is representative, one can generalize the above idea by considering an integral, $I_N = \int |Q^{(0)}(\mathbf{y})|\, d\mathbf{y}$, which is assumed to be known either analytically or numerically (to any degree of accuracy). Each time the sampled diagram

order is zero, we add unity to the normalization counter, $Z_N = Z_N + \delta_{n,0}$, which is subsequently used to obtain the final result as

$$Q(\mathbf{y}) = I_N \, \frac{Q_{MC}(\mathbf{y})}{Z_N} \, . \tag{8}$$

In most cases, the lowest-order contributions are indeed trivial and there is no problem with implementing the normalization protocol. If none of the contributions to $Q(\mathbf{y})$ is known, we can still employ this protocol by adding a "fake diagram" with positive-definite weight, $D_F(\mathbf{y})\,\delta_{n,0}$, and known normalization integral, $I_N = \int D_F(\mathbf{y})\, d\mathbf{y}$, to the configuration space, and updating the Z_N counter each time this fake diagram is sampled. Equation (8) works for this setup without any modifications with the understanding that the fake diagram is used for normalization purposes only and is not contributing to $Q_{MC}(\mathbf{y})$. It is also worth noting that normalization can be always done to some positive-definite quantity; i.e., you will not face the sign-problem in the denominator.

3 Fröhlich polaron

Let us now focus on the Fröhlich polaron model and see in detail how the DiagMC technique can be used for obtaining the polaron Green function at zero temperature, see also [4]. The model Hamiltonian $H = H_{\rm e} + H_{\rm ph} + H_{\rm e-ph}$ contains three terms where

$$H_{\rm e} = \sum_{\mathbf{p}} \left(\epsilon(\mathbf{p}) - \mu\right) a^\dagger_{\mathbf{p}} a_{\mathbf{p}} \, , \quad H_{\rm ph} = \sum_{\mathbf{q}} \omega(\mathbf{q})\, b^\dagger_{\mathbf{q}} b_{\mathbf{q}} \, , \quad H_{\text{e-ph}} = \sum_{\mathbf{p},\mathbf{q}} V(q) \left(b^\dagger_{\mathbf{q}} - b_{-\mathbf{q}}\right) a^\dagger_{\mathbf{p}-\mathbf{q}} a_{\mathbf{p}} \, , \tag{9}$$

$$\epsilon(\mathbf{p}) = \frac{p^2}{2M} \, , \qquad \omega(\mathbf{q}) = \Omega \, , \qquad V(\mathbf{q}) = i\sqrt{2^{3/2}\pi\Omega^{3/2}\alpha/M^{1/2}q^2} \, , \tag{10}$$

with standard notations for creation and annihilation operators. The chemical potential, $\mu < 0$, is introduced here solely for the purpose of controlling the statistics at large times because otherwise it would diverge. The coupling between electrons and optical phonons with energy Ω can no longer be assumed weak when the dimensionless constant $\alpha > 1$. The Fröhlich polaron is the canonical model used to describe non-degenerate charge carriers in ionic semiconductors; Figure 5 explains why lattice effects can be neglected and one may proceed with the continuum description based on the parabolic dispersion relation, $\epsilon(p) = p^2/2M$, for the electron and a dispersionless optical phonon, $\omega(p) = \Omega = \text{const}$.

Connected diagrams for the Green function in imaginary time representation, $G(\mathbf{p}, \tau) = \langle c_{\mathbf{p}}(\tau) c^\dagger_{\mathbf{p}} \rangle$, are shown in Fig. 6. To convert the graphics into mathematical expressions (1) one has to use the following "conversion" rules:

- A straight line with momentum $\mathbf{p}_c$ between the time points τ_b and τ_a is associated with the bare polaron propagator, $G_0 = e^{-(\epsilon(p_c)-\mu)(\tau_b - \tau_a)}$.

- An ark connecting two dots is associated with the product of the phonon propagator, $D = e^{-\omega(p_c)(\tau_b - \tau_a)}$, the modulus of the coupling vertex squared, $|V(p_c)|^2$, and the momentum space integration factor $(2\pi)^{-3}$.

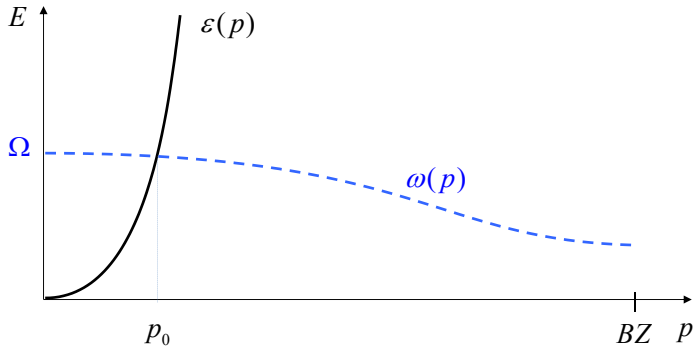

Fig. 5: *Dispersion relations for electrons and phonons are such that their intersection and important physics effects take place at momenta* $p \sim p_0 = \sqrt{2M\Omega}$*, much smaller than the Brillouin zone boundary.*

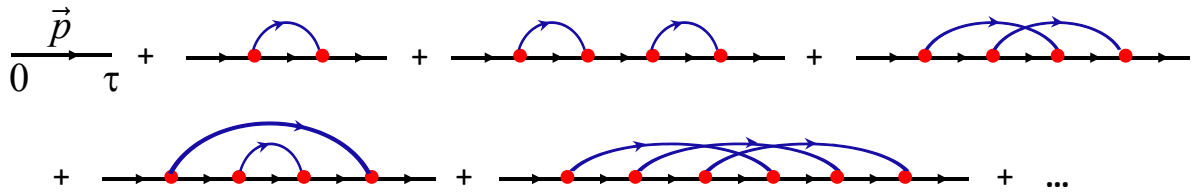

Fig. 6: *Green function diagrams obtained by expanding in the number of phonon lines.*

Thus, for this problem all diagrams are sign-positive and the series are convergent for any value of τ. Sign-positive series cannot be resummed; they are either meaningless or convergent. In our case, there are $(2n-1)!!$ diagrams of order n. On the other hand, the integration measure of $2n$ time-ordered points is $\tau^{2n}/(2n)!$, and this is sufficient to guarantee series convergence for Fröhlich polarons. The rest is a straightforward implementation of the generic DiagMC protocol. My set of updates is just one possible realization among many available.

3.1 Type-I updates

• *Global* τ *update.* The simplest update changing the global time variable τ is shown in Fig. 7(a). The probability density for the new value is a simple exponential

$$P(\tau') = Ee^{-E(\tau'-\tau_{\rm last})}\,, \qquad E = p^2/2M - \mu\,, \tag{11}$$

and the acceptance ratio is unity because $D_{\nu'}/D_\nu = P(\tau')/P(\tau)$, see Eq. (3). [If r is a random number uniformly distributed on the interval $(0,1)$, then $\tau' = \tau_{\rm last} - E^{-1}\ln(r)$ by the transformation method.] Strictly speaking, this is the only type-I update required for ergodicity! Below I present several other type-I updates that can be added in order to (i) improve efficiency and reduce autocorrelations, (ii) have an over-complete set of updates for meaningful tests of the detailed balance, (iii) have fun.

• *Internal* τ *update.* Changing the time variable of the interaction vertex is equally easy, see Fig. 7(b). The probability density for the new value of τ_b' is a simple exponential

$$P(\tau_b') = \frac{Ee^{-E(\tau_b'-\tau_a)}}{1-e^{-E(\tau_c-\tau_a)}}\,, \qquad E = (p_a^2 - p_b^2)/2M \pm \Omega\,, \tag{12}$$

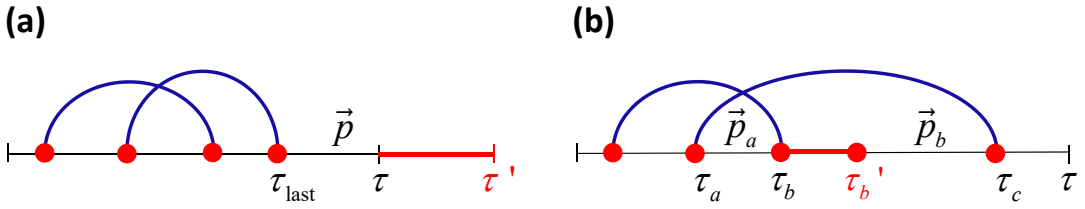

Fig. 7: *Type-I updates changing the external (a) and internal (b) time variables:*

and the acceptance ratio is unity for exactly the same reason as for the *global τ update*. [$\tau_b' = \tau_a - E^{-1}\ln[1 - r(1 - e^{-E(\tau_c - \tau_a)})]$ by the transformation method.]

• *Rescaling all τ variables update.* By introducing dimensionless time variables $s_i = \tau_i/\tau$, and paying attention that all propagators are exponential functions of time, we realize that the diagram dependence on the global time τ is given by the Poisson distribution. If we were to use

$$P(\tau') = \frac{E(E\tau')^{2n}}{(2n)!} e^{-E\tau'}, \qquad E = \sum_{i=1}^{2n+1} \Delta s_i E_i - \mu\,, \tag{13}$$

where the sum is over all time intervals in the graph, and E_i is the energy associated with each interval (counting both the polaron energy and the energy of all phonon lines covering it), to propose new values of τ', we would always accept the update. Unfortunately, for this distribution the transformation method cannot be used. However, for large $E\tau$ and n, the Poisson distribution is approaching the Gaussian. The idea then is to use

$$P(\tau') = \frac{E}{\sqrt{4\pi n}}\, e^{-(E\tau' - 2n)^2/4n}\,, \tag{14}$$

instead. [From now on I will stop mentioning the transformation method and how it is used (sometimes with tricks) for well-know distributions.] Obviously, the update is rejected, when the proposed τ' is negative—one should not be afraid of proposing unphysical values if this leads to radical simplifications of the numerical procedure with only a minor loss of efficiency. The acceptance ratio is given by

$$R = \exp\left(2n\ln\frac{\tau'}{\tau} - E\,(\tau' - \tau) + \frac{(E\tau' - 2n)^2 - (E\tau - 2n)^2}{4n}\right), \tag{15}$$

and is close to unity (on average) for large diagram orders.

• *Internal $|q|$ update.* To change the modulus of the phonon momentum (the direction is preserved) we select at random any of the phonon lines (probability of selecting a particular one is $1/n$) and propose the new value for q from the Gaussian probability density

$$P(q') = \frac{1}{\sqrt{2\pi s^2}}\, e^{-(q' - q_0)^2/2s^2}\,, \tag{16}$$

where $q_0 = \langle\mathbf{p}\rangle\cdot\mathbf{q}/q$, $s^2 = M/(\tau_b - \tau_a)$, and $\langle\mathbf{p}\rangle = (\tau_b - \tau_a)^{-1}\int_{\tau_a}^{\tau_b} d\tau[\mathbf{p}(\tau) + \mathbf{q}]$ is the "average" electron momentum on the time interval (τ_a, τ_b) *in the absence* of the updated phonon propagator, see Fig. 8(a). This update is always accepted (provided q' is non-negative) because $P(q')$ is

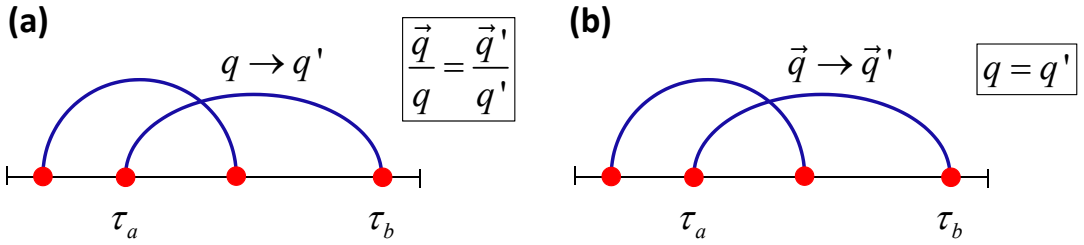

Fig. 8: *Type-I updates changing the momentum variable.*

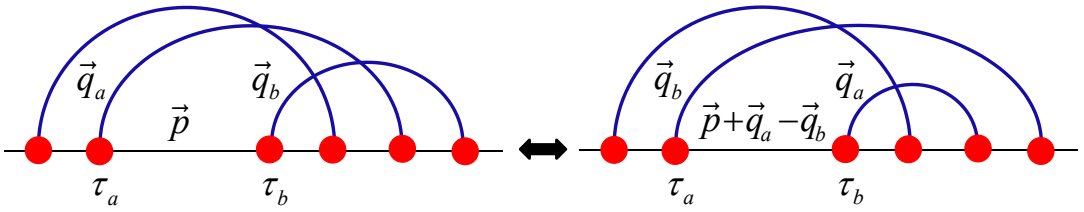

Fig. 9: *Type-I update changing the local diagram topology.*

reproducing precisely the diagram weight for the Fröhlich polaron. Indeed,

$$\frac{1}{2M}\int_{\tau_a}^{\tau_b} d\tau\left[(\mathbf{p}(\tau)+\mathbf{q}-\mathbf{q}')^2-p^2\right] = \text{const.} + \frac{\tau_b-\tau_a}{2M}\,(q'-q_0)^2\,. \tag{17}$$

• *Internal* $\mathbf{q}/q$ *update.* To change the phonon momentum direction while keeping its modulus fixed, we select at random any of the phonon lines and propose the new value for $\mathbf{q}/q$ from the uniform distribution for the azimuthal angle φ and exponential distribution for the cosine of the polar angle θ

$$P(\varphi,\theta) = \frac{A\sin(\theta)}{4\pi\sinh(A)}\,e^{A\cos(\theta)}\,, \tag{18}$$

where $A=(\tau_b-\tau_a)|\langle\mathbf{p}\rangle|q/M$. Both angles are defined relative to the axis set by the vector $\langle\mathbf{p}\rangle$ defined in the previous update, see also Fig. 8(b). This update is always accepted because (18) reproduces the functional dependence of (17) on updated angles.

• *Topology change.* Here the idea is to select at random any nearest-neighbor (n.n.) pair of vertices and swap their places, see Fig. 9. The momenta of phonon propagators remain fixed except when the selected pair is connected by the phonon line, in which case it changes sign. This proposal will change the momentum of the polaron line to $\mathbf{p}'=\mathbf{p}+\mathbf{q}_a-\mathbf{q}_b$. The acceptance ratio is directly related to the ratio of diagram weights

$$R = \exp\{-(\tau_b-\tau_a)[\epsilon(p')-\epsilon(p)\pm\omega(q_a)\pm\omega(q_b)]\}\,, \tag{19}$$

where the proper $\pm$ option has to be chosen depending on whether the corresponding phonon propagator is getting longer or shorter in time.

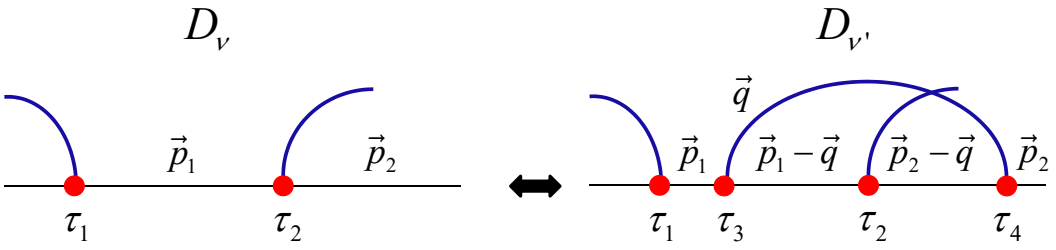

Fig. 10: *Type-II updates changing the diagram order.*

3.2 Type-II updates. Data structure

The design of type-II updates, especially in view of Eq. (6), is as flexible. I will only describe one of them following a particular strategy of selecting new variables. Formally, this type-II update and the first update described in the previous subsection, constitute an ergodic set of MC procedures capable of simulating the Green function dependence on time.

• *Increase/decrease the diagram order by one.* Type-II updates typically come in complementary pairs that satisfy the detailed balance condition within each pair. Let $p_{n\to n+1}$ and $p_{n+1\to n}$ be the probabilities to make a decision to apply one of these two updates. In the *increase* update propose the following steps:

– Select one of the polaron propagators at random (corresponding probability: $1/(2n+1)$); let the parameters of the selected interval be $\mathbf{p}_1$, τ_1, τ_2, see Fig. 10.

– propose the fist new time variable τ_3 from the uniform probability density $1/\Delta\tau$, where $\Delta\tau = \tau_2 - \tau_1$.

– propose the momentum for the new phonon propagator from the probability distribution

$$\frac{\sin(\theta)}{4\pi}\,\frac{1}{p_0(1+q/p_0)^2}\,, \tag{20}$$

where $p_0 = \sqrt{2M\Omega}$, see Fig. 5. It is uniform for the solid angle of $\mathbf{q}$ and has an easy to handle power law for the modulus of $\mathbf{q}$.

– propose the new time variable $\tau_4 > \tau_3$ from the distribution

$$\Omega(1+q/p_0)^2\, e^{-\Omega(1+q/p_0)^2(\tau_4-\tau_3)}\,. \tag{21}$$

As usual, the update is rejected if $\tau_4 > \tau$. This distribution is not a perfect match to the ratio of the diagram weights in the proposed setup

$$\left|\frac{D_{\nu'}}{D_\nu}\right| = |V(q)|^2\, e^{-\Delta E(\tau_4-\tau_3)}\,\frac{q^2\sin(\theta)}{(2\pi)^3} \propto e^{-\Delta E(\tau_4-\tau_3)}\sin(\theta)\,, \tag{22}$$

where the energy change on the updated interval is given by $\Delta E = \Omega + [q^2 - 2\mathbf{q}\cdot\langle\mathbf{p}\rangle]/2M$, and $\langle\mathbf{p}\rangle = (\tau_4-\tau_3)^{-1}\int_{\tau_3}^{\tau_4} d\tau\mathbf{p}(\tau)$. However, it is "good enough" in terms of the average acceptance ratio.

The *decrease* update is conceptually very simple: Select any of the existing phonon propagators (probability is $1/n$, where n is the *current* diagram order), and propose to remove it. We are all set to formulate the detailed balance equation

$$r_{n\to n+1}|D_\nu|\frac{p_{n\to n+1}}{2n+1}\,\frac{\Omega\sin(\theta)}{4\pi\Delta\tau p_0}\,e^{-\Omega(1+q/p_0)^2(\tau_4-\tau_3)} = r_{n+1\to n}|D_{\nu'}|\frac{p_{n+1\to n}}{n+1} \tag{23}$$

and its solution for the *increase* update

$$R_{n\to n+1} = \frac{p_{n+1\to n}}{p_{n\to n+1}}\,\frac{2n+1}{n+1}\,\frac{2\alpha\Omega\Delta\tau}{\pi}\,e^{[qp_0+\mathbf{q}\cdot\langle\mathbf{p}\rangle](\tau_4-\tau_3)/M}\,. \tag{24}$$

Notice that the second ratio contains $(n+1)$, the number of phonon propagators in the proposed configuration, in the denominator. The solution of the same detailed balance equation for the *decrease* update acceptance ratio reads:

$$R_{n\to n-1} = \frac{p_{n-1\to n}}{p_{n\to n-1}}\,\frac{2n-1}{n}\,\frac{\pi}{2\alpha\Omega\Delta\tau}\,e^{-[qp_0+\mathbf{q}\cdot\langle\mathbf{p}\rangle](\tau_4-\tau_3)/M}\,. \tag{25}$$

Again, one has to be careful in formulating it in terms of the current, order n, and proposed, order $n-1$, configuration parameters. In particular, $\Delta\tau$ is the duration of the polaron interval where the removed polaron propagator starts *after* the corresponding phonon propagator is removed (it may be the case that both τ_3 and τ_4 are smaller than τ_2 in Fig. 10). Also, the average polaron momentum $\langle\mathbf{p}\rangle$ needs to be computed for the proposed configuration.

• As far as normalization is conserved, the easiest way would be to normalize to the known integral of the bare Green function

$$I_N = \int_0^\infty d\tau e^{-(p^2/2M-\mu)\tau} = \frac{1}{p^2/2M-\mu}\,. \tag{26}$$

• At this point it is worth saying a couple of words about the data structure because recovering the necessary information for performing updates is often crucial for the efficiency of the algorithm. Most updates are designed to modify the diagram structure and its parameters locally (in terms of graph connections); i.e., only a few parameters and propagators are involved in each update. [This requirement does not apply to DiagMC algorithms based on exact summation of all diagram topologies that are discussed in the next Section.] Correspondingly, the data structure should be implemented in such a way that updates can be completed after performing $O(1)$ operations in the limit of $(n,\tau)\to\infty$. Here is one possible structure:

1. Every polaron propagator (or "interval") in the diagram has a unique label $\ell > 0$.

2. This label is used to retrieve information about the propagator momentum, as well as its initial, and final times, using $p(\ell, 1{:}3)$, $\tau_i(\ell)$, and $\tau_f(\ell)$ arrays, respectively. One can introduce additional arrays, if necessary, and update the corresponding information; e.g., $N_{\rm ph}(\ell)$ returns the number of phonon propagators covering the ℓ interval (number of phonons in the virtual state at time $\in \big(\tau_i(\ell),\tau_f(\ell)\big)$.

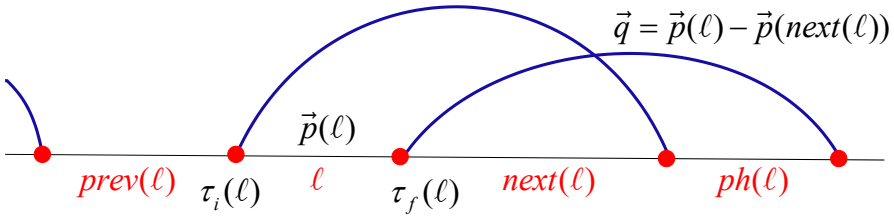

Fig. 11: *Data structure for efficient implementation of local updates.*

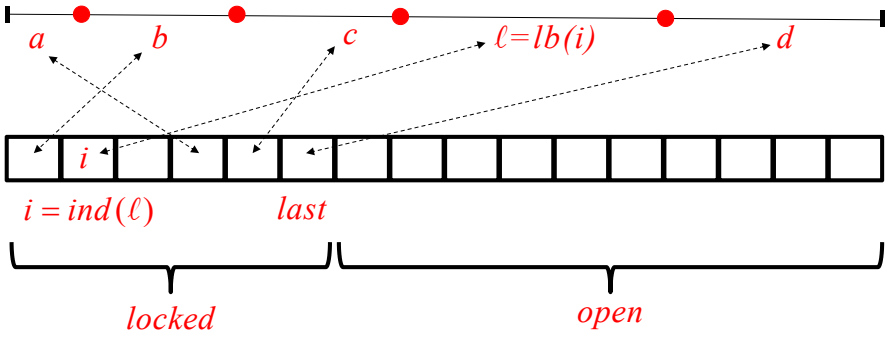

Fig. 12: *Efficient management of a finite set of labels.*

3. All labels are linked pairwise in two ways. Arrays $prev(\ell)$ and $next(\ell)$, see Fig. 11, allow one to get labels of intervals immediately preceding and following ℓ. Obviously, $next(prev(\ell)) = \ell$, and $prev(next(\ell)) = \ell$. An array $ph(\ell)$ establishes a link between the left ends of the intervals connected by the phonon propagator. Again, $ph(ph(\ell)) = \ell$. It is easy to see that any local (around ℓ) information about the graph properties can be quickly recovered without knowing the global structure. An ordered array 1, 2, 3, ..., $2n+1$ of labels cannot be used because if some label ℓ is removed from the middle of the list (see also text below), all labels with values $> \ell$ must be updated, in violation of the "locality" principle.

4. In a long simulation run new/existing intervals will be created/erased trillions of times. It is thus not practical to never use the same label twice. There is, however, a simple programming trick that allows one to use the same set of labels forever, and manage it with $O(1)$ operations. At any moment it is known which $n+1$ labels are "locked" because they are already utilized for linking the graph intervals and which ones are "open" for labeling new elements. This is achieved with the help of the $ind(\ell)$ and $lb(i)$ arrays. By definition, labels $lb(1)$, $lb(2)$, ... $lb(last = 2n + 1)$ are "locked" and labels $lb(> 2n + 1)$ are "open," while $ind(lb(i)) = i$ establishes a connection between the graph labels and an ordered array of indexes, see an illustration in Fig. 12. When a new interval is added to the diagram it is given a label $lb(last + 1)$ and the variable $last$ is increased by one. When some interval and its label ℓ are removed, one first determines its index $i = ind(\ell)$. If $i = last$, the value of $last$ is decreased by one; otherwise, it is necessary to swap an association between indexes i and $last$ and labels ℓ and $lb(last)$, and only then decrease the value of $last$ by one.

3.3 Illustrative results: polaron Green function

One can employ DiagMC for in-depth studies of polaron properties, including direct access to every coefficient (by modulus) in the Lehmann expansion of the exact wave function with momentum $\mathbf{p}$

$$|\Psi_{\mathbf{p}}\rangle = c_{\mathbf{p}} a^{\dagger}_{\mathbf{p}}|0\rangle + \sum_{\mathbf{q}} c_{\mathbf{p},\mathbf{q}} b^{\dagger}_{\mathbf{q}} a^{\dagger}_{\mathbf{p}-\mathbf{q}}|0\rangle + \sum_{\mathbf{q}_1\mathbf{q}_2} c_{\mathbf{p},\mathbf{q}_1,\mathbf{q}_2} b^{\dagger}_{\mathbf{q}_1} b^{\dagger}_{\mathbf{q}_2} a^{\dagger}_{\mathbf{p}-\mathbf{q}_1-\mathbf{q}_2}|0\rangle + \dots \tag{27}$$

With this one can determine the probability of finding an electron along with a given number of phonons in the polaron cloud [4]. Here I will only review how one computes the Green function and extracts the polaron Z-factor, $Z_{\mathbf{p}} = |c_{\mathbf{p}}|^2$, and energy from its asymptotic expression

$$G_{\mathbf{p}}(\tau \to \infty) \to Z_{\mathbf{p}}\, e^{-(E_{\mathbf{p}}-\mu)\tau}\ . \tag{28}$$

The simplest way of collecting statistics for $G_{\mathbf{p}}(\tau)$ would be to split the entire τ axis into small bins $\Delta_i = \tau_i - \tau_{i-1}$ and update bin counters for an integral of the function over the bin size (for brevity I will suppress the momentum index)

$$I^{(i)} = \int_{\tau_{i-1}}^{\tau_i} d\tau\, G(\tau)\ , \tag{29}$$

$$I^{(i)}_{MC} = I^{(i)}_{MC} + 1\,, \qquad \text{if } \tau \in (\tau_i,\, \tau_{i-1})\ . \tag{30}$$

For small bins, an estimate for the Green function at point $\bar{\tau}_i = (\tau_i + \tau_{i-1})/2$ can be made as

$$G(\bar{\tau}_i) \approx \frac{I_N}{Z_N}\,\frac{I^{(i)}_{MC}}{\Delta_i}\ . \tag{31}$$

While integrals (29) are free of systematic errors, the most straightforward final step (31) does involve a finite bin size error.

There are several ways for eliminating this systematic error. The bin hierarchy method [5, 6] takes the limit of very small bins and overcomes the problem of large statistical noise for small bins by restoring the entire function $G(\tau)$ using splines with self-adaptive nodes. This protocol is rather technical to reproduce here but it is very efficient and ensures that the systematic errors are always smaller than the statistical ones.

One may also keep the bin sizes fixed, but collect several integrals over bins to improve the accuracy of restoring $G(\tau)$. Any smooth function on a given interval can be expanded as

$$G(\tau) = \sum_{j=1}^{\infty} \alpha^{(i)}_j e^{(i)}_j(\tau)\,, \qquad \text{if } \tau \in (\tau_i,\, \tau_{i-1})\,, \tag{32}$$

where $\{e^{(i)}_j(\tau)\}$ is an ortho-normal basis (ONB) on the interval $(\tau_i,\, \tau_{i-1})$. It may be Legendre polynomials, but any other ONB with an inner product defined as

$$\langle f|g\rangle = \int_{\tau_{i-1}}^{\tau_i} d\tau\, w^{(i)}(\tau) f(\tau) g(\tau)\,, \tag{33}$$

can be used instead. In practice, the series is truncated at some finite order N_i, that is determined to provide an accurate description of $G(\tau)$ with unmeasurable (within statistical errors) systematic bias. To account for divergencies in $G(\tau)$ one may need to use singular basis functions and $w > 0$ weights. For example, if $G(\tau \to 0) \propto 1/\tau^{1/2}$, the first basis function in the set may be $e_1(\tau) = A_a/\tau^{1/2}$. In this case, the $w(\tau)$ is required to ensure that the normalization integral for e_1 is finite, and $w(\tau) = \tau^{1/2}$ will do the job. It is also possible to consider infinite size bins; if the leading asymptotic decay is of power-law type $G(\tau \to \infty) \propto 1/\tau^a$, the ONB on the time interval (τ_h, ∞) may be constructed from the set of functions $e_j \propto 1/\tau^{a+b_j}$ with $b_1 = 0$ and $b_{j>1} > 0$ to account for the dominant term and several subleading corrections.
According to the theory of Hilbert spaces, the coefficients of expansion are determined by the integrals

$$\alpha_j^{(i)} = \int_{\tau_{i-1}}^{\tau_i} d\tau\, w^{(i)}(\tau) e_j^{(i)}(\tau) G(\tau)\ , \tag{34}$$

with unbiased MC estimators

$$\alpha_{j,MC}^{(i)} = \alpha_{j,MC}^{(i)} + w^{(i)}(\tau) e_j^{(i)}(\tau)\,, \qquad \text{if } \tau \in (\tau_i,\, \tau_{i-1}). \tag{35}$$

After appropriate normalization of statistics, one obtains the Green function from

$$G(\tau) = \frac{I_N}{Z_N} \sum_{j=1}^{N_i} \alpha_{j,MC}^{(i)}\, e_j^{(i)}(\tau). \tag{36}$$

Obviously, the conventional procedure described by Eqs. (29)-(31) is nothing but the special case when there is only one constant basis function.
Finally, one can use the reweighing method for an unbiased estimate of the function at a specified set of points $\bar{\tau}_i$. For each point one decides on the interval $(\tau_i,\, \tau_{i-1})$ that will be used to collect statistics for $G(\bar{\tau}_i)$; there are no formal restrictions on the sizes and locations of these intervals or their overlaps for different points. An optimal choice would be to have $\bar{\tau}_i$ roughly in the middle of the interval, and the interval width Δ_i to be small enough to avoid multi-scale variations of G within the interval. For any simulated point than falls within the interval, a factor $D_\nu(\bar{\tau}_i)/D_\nu(\tau)$ accounts for the difference between the $G(\bar{\tau}_i)$ and $G(\tau)$ functions. Thus

$$G(\bar{\tau}_i)\Delta_i = \int_{\tau_{i-1}}^{\tau_i} d\tau\, G(\tau)\, \frac{D_\nu(\bar{\tau}_i)}{D_\nu(\tau)}\ , \tag{37}$$

implying that an unbiased estimator for $G(\bar{\tau})$ is given by

$$G_{MC}(\bar{\tau}_i) = G_{MC}(\bar{\tau}_i) + \frac{D_\nu(\bar{\tau}_i)}{D_\nu(\tau)\Delta_i}\,, \qquad \text{if } \tau \in (\tau_i,\, \tau_{i-1})\ . \tag{38}$$

The normalization of statistics using I_N/Z_N does not change.
Once the data for $G(\tau)$ are collected they are analyzed according to Eq. (28), see left panel in Fig. 13. The dependence of the quasiparticle residue on the coupling constant is shown in the right panel of Fig. 13. It is not unusual to have statistical errors for this problem at the level of 10^{-6} in relative units. Many more results and direct MC estimators for polaron properties can be found in Ref. [4].

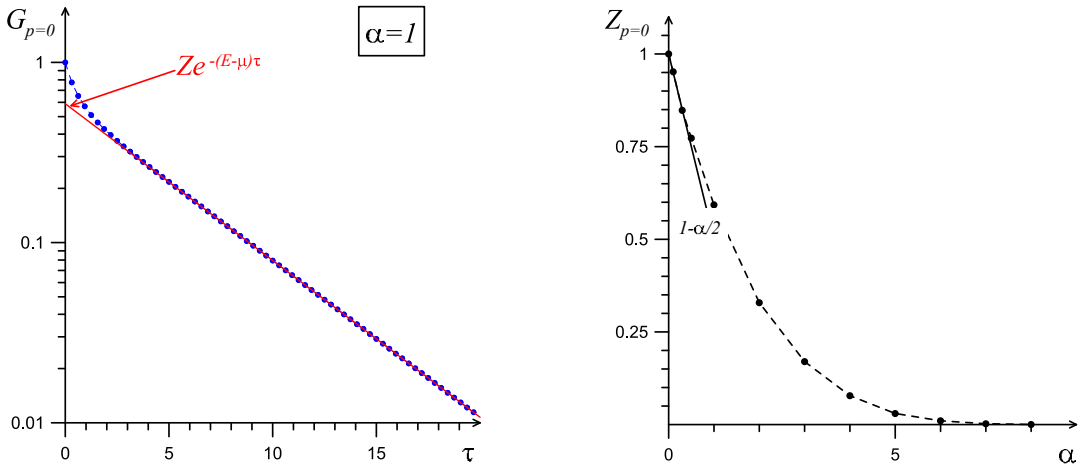

Fig. 13: *Left panel: Fröhlich polaron Green function at zero momentum for* $\alpha = 1$ *(we use units such that* $M = 1$ *and* $\Omega = 1$*) and its asymptotic single exponential behavior. Right panel: Quasiparticle residue at* $p = 0$ *as a function of the coupling constant.*

4 Fermionic sign blessing I

DiagMC for a generic interacting fermionic system follows the same rules. Clearly, the diagrams themselves are different, see Fig. 1, the differences in the interaction Hamiltonian. Also, the diagrams are no longer sign-positive. Even in the absence of gauge fields, the diagram sign may change for the following reasons: (i) for repulsive interactions and expansion in the interaction potential the diagram sign contains $(-U)^n$, (ii) fermionic propagators are subject to anti-periodic boundary conditions, $G(\tau{<}0) = -G(\tau{+}\beta)$, (iii) the diagram sign contains $(-1)^L$ where L is the number of fermionic loops in a given topology. This raises two important questions: "Are the diagrammatic series convergent and under what conditions?" and "What is the role of the fermionic sign?" In what follows, I will explain that the two questions are closely related. The bottom line is that the DiagMC technique works and can be made very efficient thanks to the fermionic sign; i.e., it is a "blessing," not a problem.

4.1 Convergence of diagrammatic series for fermions

Dyson's argument for perturbative expansions in the coupling constant states that the convergence radius of the series is zero for continuous-space systems, no matter whether for bosons or fermions. Indeed, if it were finite, system properties would be analytic functions of the coupling constant near the origin. However, when the sign of the interaction is flipped from repulsion to attraction, continuous space systems collapse to a point (infinite density state) because even for fermions the kinetic energy increase $\propto n^{5/3}$ cannot overcome the potential energy gain $-|g|n^2$. To understand how this physical intuition is reflected in the mathematical structure of the series, notice that the number of different topologies for diagrams of order $n \gg 1$ is proportional to $n!$. Since $n!$ always beats the power law dependence on the coupling constant g^n, there is also a formal reason for suspecting that the convergence radius might be zero.

It appears then that Dyson's argument makes the entire diagrammatic approach for many body systems nearly useless in the strongly correlated regime. This is where the Pauli principle and the fermionic sign come to rescue. To begin with, the collapse argument does not apply to lattice systems, such as the famous single-band tight-biding Hubbard model. At best, one can place two, not more, fermions with opposite spins on the site and the system cannot collapse to infinite density. Thus, we expect that finite temperature properties of the Hubbard model are analytic functions of the coupling constant U in the limit of $U \to 0$ and this is confirmed by exact mathematical considerations. Likewise, Dyson's argument can be refuted for continuous space fermions with high momentum cutoff that acts similarly to the Brillouin zone boundary. Extrapolating converged calculations to the infinite cutoff limit may be a well-defined procedure. Finally, the diagrammatic technique admits an infinite number of alternative formulations when certain geometric series of the original diagrams are accounted for right from the beginning are incorporated self-consistently into the new "expansion point;" typical examples include mean-field and dynamic mean-field theories, ladder summations, screening, and solutions for the low-order skeleton set. The new expansion is no longer in terms of the coupling constant, and the original Dyson argument does not apply directly. Self-consistent mean-field and skeleton set solutions, on top of which the new expansion is made, can easily a incorporate non-analytic dependence on the bare coupling constant, as, e.g., in the famous BCS solution.
Going back to the mathematical structure of the diagrammatic expansion, we realize that the only possibility for the series to converge, despite factorial scaling of the number of allowed topologies, is to have massive cancellations between the diagrams within the same order. This is "sign blessing I:" the DiagMC method relies on the fermionic sign because this is the necessary condition for having series with nice properties. In what follows I will discuss simple illustrative examples demonstrating how resummation techniques allow one to extract accurate answers from divergent sign-alternating series. [If series converge, it is time to publish the solution.]

4.2 Resummation techniques

One way to deal with divergent sign-alternating series, $Q = \sum_{n=1}^{\infty} d_n$, outside their finite convergent radius is as follows. Introduce a smooth function, $f(n, \epsilon)$, that satisfies two conditions: for $\epsilon \to 0$ and finite n it approaches unity, $f(n, 0) = 1$; for $n \to \infty$ and finite ϵ it goes to zero faster than an exponential function, $f(n \to \infty, \epsilon)a^n \to 0$ for any $a > 1$. The resummed series

$$Q_\epsilon = \sum_{n=1}^{\infty} d_n \, f(n, \epsilon) \ , \tag{39}$$

is guaranteed to converge because $f(n, \epsilon)$ suppresses the geometrical divergence of the original series, while in the limit of $\epsilon \to 0$ the original and resummed series coincide. By extrapolating Q_ϵ to $\epsilon = 0$ one effectively performs an analytic continuation of the sign-alternating series outside of its convergence radius, see Fig. 14(a). The $f(n, \epsilon)$ function is up to you to design, because apart from the conditions specified it is rather arbitrary; different choices for f provide a good estimate for the error introduced by extrapolation from finite values of ϵ to zero.

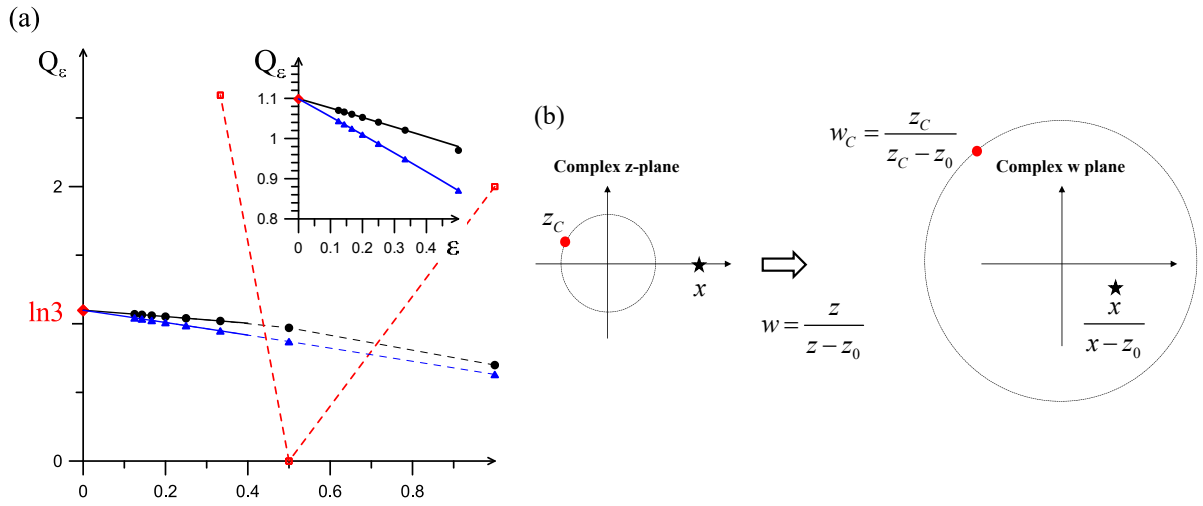

Fig. 14: *(a) Resummation of divergent series for* $\ln(1+x)$ *with* $x = 2$ *using* $f(n,\epsilon) = e^{-\epsilon n^2}$ *(circles) and* $f(n,\epsilon) = e^{-\epsilon n^{3/2}}$ *(triangles). Extrapolation to* $\epsilon = 0$ *was performed using parabolic fits. Partial sums* $\sum_1^{[\epsilon]}(-x)^{n+1}/n$ *are shown by open squares. The value* $\ln 3$ *is marked by the diamond on the vertical axis. (b) Moving a simple-pole and increasing the convergence radius using conformal mapping.*

The above protocol is blind to specific properties of the series and may require knowledge of many terms in the series for reliable extrapolation, especially for less "aggressive" f-functions. More efficient methods exist when the reason for reaching the convergence radius is known better. Suppose that the series behaves as $d_n = \gamma_n x^n$ with $\gamma_n = (-1)^n$ and $M = 10$ terms are known. The goal is get an answer for $x = 3$, well outside of the radius of convergence. From available information one can roughly estimate the convergence radius, and produce constant phase lines $y(x)$ for the complex function $Q(z = x + iy) = \sum_n^M \gamma_n z^n$ to establish that a simple pole is located close to the real axis, say at $z_0 \approx -1.05$. Next, one performs a conformal mapping $w = z/(z - z_0)$, or $z = -wz_0/(1 - w)$ and constructs the Taylor series for $Q_w(w) = \sum_n^M \sigma_n w^n$. The final answer is given by $Q_w(x/(x - z_0))$; with extraordinary accuracy it reproduces $1/(1 + x) = 0.25$. Under conformal mapping the singularities are moved away from the origin of the expansion and the point of interest ends up well within the radius of convergence, see the illustration in Fig. 14(b).

Similarly, it is possible to handle poles of higher order or several poles, but high accuracy rests on the number of known terms in the series. A slightly different version of the method is known as extrapolation by Padé approximants. One assumes that the function behind the series is given by the ratio of two polynomials, $Q(z) = P_k(z)/P_m(z)$, with $k + m \leq M$. For each (k, m) pair the polynomials are determined by matching the coefficients of the Taylor series for the ratio to γ_n. The final answer is determined by examining how $P_k(x)/P_m(x)$ depends on (k, m) when we increase the order of polynomials.

Conformal mappings can be also used to improve the convergence properties of series by moving branch cuts away from the origin. The ratio of polynomials can be replaced by the ratio of hypergeometric functions to achieve efficient extrapolation in cases when the convergence radius is limited by the branch cuts [7]. The mathematical and physical literature on the topic

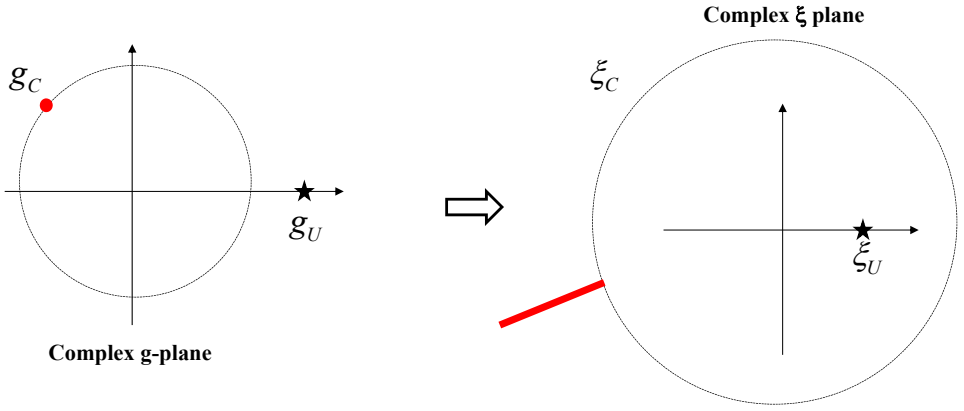

Fig. 15: *Illustration of the shifted action trick. It amounts to changing the origin of expansion and introducing a different expansion parameter.*

is vast, and many methods carry names of famous mathematicians. The bottom line is that divergent series outside their radius of convergence are almost as valuable for extracting the required information as convergent series, provided enough terms in the series are known and singularities are reasonably well understood.

To conclude this subsection, I would like to mention that series with convergence radius zero (e.g., when branch cuts originate from the center of expansion) are also subject to efficient resummation methods that guarantee that the final answer is unique in the limit. The analysis starts with establishing the asymptotic behavior of the Taylor series coefficients γ_n for large n by employing the method developed by Lipatov [8]. Since all singularities are encoded in $\gamma_{n\to\infty}$ (any finite number of terms is just a polynomial), the corresponding information should be used for designing the optimal resummation method and guaranteeing that it performs the unique analytic continuation for the physical parameters of the problem. A beautiful example of such an analysis can be found in Ref. [9].

4.3 Shifted action

Yet another way of manipulating series convergence is shifting the expansion point, as illustrated in Fig. 15. Of course, it can not be done by solving exactly an interacting problem for some value of the coupling constant, say g_1, and then expanding in $g - g_1$ (but this protocol may be tried numerically if g_1 is inside the convergence radius). However, it is relatively "cheap" to expand on top of the mean-field solution or solutions based on a limited subset of diagrams, such as ladders, random phase approximation, and self-consistent skeleton graphs, as well as any set of dressed propagators. The new expansion still provides an exact solutions to the problem. The tool goes under the name of "shifted action."

I will explain how shifted action works by considering the case of the Green function. Further generalizations are mentioned in Ref. [10]; in brief, the shifted action protocol can be applied at the level of any multi-point correlation function (with the help of Stratonovich-Hubbard transformations). Let the original interacting problem be described by the action (in terms of the

Grassmann field ψ)

$$S[\psi] = \langle\psi|G_0^{-1}|\psi\rangle + gS_{\rm int}[\psi]\,, \tag{40}$$

where G_0 is the bare fermion propagator, and g is the coupling constant. [For brevity, I will use vector-space notations to suppress space, time, spin, etc., indices and integrals/sums over them, and loosely call the corresponding kernels "functions."] Instead of expanding e^{-S} in powers of g, one can introduce an auxiliary action

$$S_\xi^{(\mathcal{N})}[\psi] = \langle\psi|\tilde{G}_{\mathcal{N}}^{-1} + \xi\Lambda_1 + \ldots + \xi^{\mathcal{N}}\Lambda_{\mathcal{N}}|\psi\rangle + \xi g S_{\rm int}[\psi] \tag{41}$$

with auxiliary expansion parameter ξ. Despite the fact that the first term is harmonic, we will treat $\tilde{G}_{\mathcal{N}}$ as a new "shifted" bare propagator and expand $e^{-S_\xi^{(\mathcal{N})}}$ in powers of ξ using all ξ-dependent terms. For the two actions to represent the same physical system at $\xi = 1$ we demand that

$$\tilde{G}_{\mathcal{N}}^{-1} + \sum_{n=1}^{\mathcal{N}} \Lambda_n = G_0^{-1}\,. \tag{42}$$

Given that the final answer for the Green function can be expressed in terms of the proper self-energy Σ as

$$G^{-1} = G_0^{-1} - \Sigma = \tilde{G}_{\mathcal{N}}^{-1} - \left[\Sigma - \sum_{n=1}^{\mathcal{N}} \Lambda_n\right]\,, \tag{43}$$

the $\{\Lambda_n\}$ functions act as counter-terms with respect to the n-th order proper self-energy diagrams generated by the interaction term $gS_{\rm int}[\psi]$. There are no restrictions on the number of counter terms or their functional dependence; this freedom can be used to optimize the convergence of the series for $\xi = 1$. Even a simple self-energy shift such as $\Lambda_1 = \mu_1$, equivalent to a change in the chemical potential, can help to solve the problem by moving the $\xi = 1$ point inside the radius of convergence [11].

Of special practical interest is the case when the counter-term Λ_n is exactly the n-th order contribution to the self-energy coming from $S_{\rm int}$. The resulting expansion—standard for effective-field theories—is then identical to the semi-skeleton series based on Dyson summation of infinite sets of irreducible diagrams associated with the first $\mathcal{N}$ orders of the perturbative expansion of the original action (40). For example, if this protocol is followed for $\mathcal{N} = 1$, then the expansion will be done on top of the self-consistent Hartree-Fock solution. If shifted action is applied to the screening channel as well, by selecting the $\mathcal{N} = 1$ skeleton set to define counter-terms one is setting the expansion on top of the self-consisted GW-approximation. Next, one can account for the leading vertex corrections, etc. In view of the exact cancellation of all contributions up to order $\mathcal{N}$, the expansion starts at order $n = \mathcal{N} + 1$ and only then the counter-terms Λ_n enter the diagrammatic expansion explicitly. To find $\tilde{G}_{\mathcal{N}}$ and all counter terms for a given G_0 one has to perform the so-called "bold" DiagMC simulation, or BDMC. This leads to the numeric protocol consisting of two independent parts:

- Part I is the BDMC simulation of the truncated order-$\mathcal{N}$ skeleton sum with the goal of solving for $\tilde{G}_{\mathcal{N}}$ and $\{\Lambda_n[\tilde{G}_{\mathcal{N}}]\}$ satisfying Eq. (42);
- Part II is the DiagMC simulation of higher-order terms using $\tilde{G}_{\mathcal{N}}$ as the bare propagator.

5 Fermionic sign blessing II

Apart from the massive cancellation of contributions from diagrams of the same order, the fermionic sign is also key for having efficient algorithms to account for all possible topologies. Indeed, consider the case of density-density interparticle interactions. If diagrams are formulated in the real-space, imaginary-time representation (to eliminate restrictions imposed by the energy-momentum conservation laws) then the sum over all possible graph topologies, both connected and disconnected, has the form of a determinant for each spin component

$$\det \left| G_\sigma(\mathbf{r}_i, \tau_i; \mathbf{r}_j, \tau_j) \right| . \tag{44}$$

Thus, $n!$ terms can be summed in $O(n^3)$ number of operations. As I will discuss below, it takes much longer to compute contributions from *connected* diagrams, but it is still possible to do it much faster than in $n!$ operations, see [12]. Ultimately, this observation allows one to say that DiagMC for convergent series or series subject to resummation, generically solves the computational complexity problem for interacting fermions.

5.1 Sum or sample?

So far we discussed the original MC approach to sampling the configuration space of connected Feynman diagrams. As illustrated in Fig. 4, one option is to sample various topologies within a given order. However, knowing that the sum of all topologies features massive cancellations of contributions, this is not necessarily the best strategy. To make the point, consider the problem of determining an answer for a large sum of sign-alternating terms

$$A = Z^{-1} \sum_{i=1}^{M} c_i \,, \qquad Z = \sum_{i=1}^{M} |c_i| \,. \tag{45}$$

In M operations one can know the answer with machine precision by performing the sum. If the sum is sampled from the probability distribution $p_i = |c_i|/Z$, the error bar on the result after M operations will be (assuming that the algorithm is perfect in eliminating the autocorrelation time problem)

$$\sigma_A = \sqrt{(1-A^2)/M} \,. \tag{46}$$

Since we assume in this discussion that $|A| \ll 1$, the answer to the dilemma of what is the best method of dealing with (45) is clear: for $|A| \ll M^{-1/2}$ one is better off by doing the sum; otherwise, the answer will be known with good accuracy faster by sampling.

Enumerating terms and performing the sum over all topologies for high-order graphs is certainly possible for $n \leq 10$. At the same time, for convergent series, it is expected that the average sign is factorially small. It is thus plausible that in many cases problem (1) has to be reformulated as

$$Q(\mathbf{y}) = \sum_{n=0}^{\infty} \int \cdots \int d\mathbf{x}_1 \cdots d\mathbf{x}_n \, \bar{D}(n; \mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{y}) \,, \tag{47}$$

$$\bar{D}(n; \mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{y}) = \sum_{\mathbb{T}} D(n, \mathbb{T}; \mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{y}) \,. \tag{48}$$

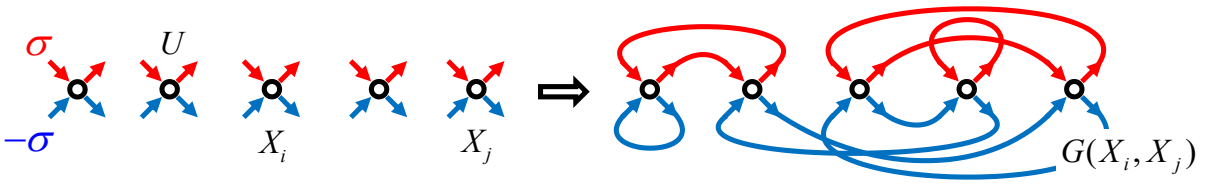

Fig. 16: *By connecting all outgoing arrows to incoming ones with the same spin index, one obtains a Feynman diagram for the partition function. Free energy density diagrams must form a connected graph.*

5.2 Determinant method for connected diagrams

An efficient method for computing $\bar{D}$ for expansions in the coupling constant was developed by Rossi in Ref. [12]. It rests on the simple observation that the sum of all connected topologies can be obtained from the sum of all topologies by subtracting disconnected ones. To be specific, consider the fermionic Hubbard model and Feynman diagrams for the free energy density. Given space-time positions of interaction vertexes $X_1, \ldots X_n$, where $X_i = (\mathbf{r}_i, \tau_i)$, all topologies are generated by establishing pairwise associations between the incoming and outgoing arrows with the same spin index, as in Fig. 16. Apart from the global factor $(-U)^n$, the diagram contribution is given by the product of all Green functions and the sign rule based on the number of fermionic loops. According to this rule, each time one swaps the destination points for two propagators the number of loops changes by ± 1 and this leads to an additional factor of -1. Thus, the sum over all possible topologies forms a determinant (44).

Let us introduce a short-hand notation for the collection of all vertex coordinates, $V = \{X_i\}$, any proper subset of coordinates, $S \subsetneq V$, the sum over all topologies (determinant) for a given set of coordinates, $\det(V)$, and the sum over all connected topologies, $C(V)$. Then, by subtracting from $\det(V)$ all disconnected cases, we obtain $C(V)$

$$C(V) = \det(V) - \sum_{S \subsetneq V} C(S) \det(V \backslash S)\,. \tag{49}$$

This is a set of recursive equations for connected contributions after similar equations are written for subsets of V. Its coefficients are based on determinants and the cost of computing all of them scales as $n^3\, 2^n$, where 2^n comes from the combinatorial number of possible proper subsets, $\sum_{m=1}^{n-1} n!/m!(n-m)!$. The number of arithmetic operations required to solve these recursive equations is $\propto 3^n$—in the large n limit this is the main computational cost.

In this scheme, the effort is exponential in the diagram order and this is certainly an enormous improvement compared to the $(n!)^2$ scaling of the total number of connected graphs. After summation over $\{X_i\}$ one should not forget to divide the n-th order contribution by $n!$ to account for the indistinguishability of the vertices. One can use this scheme (or its generalizations) to compute connected diagrams for any correlation function [12], proper self-energy [13, 14], and even semi-skeleton series; in the latter case, however, the computational cost will increase to roughly 6^n.

One final remark. When the cost of computing the diagram weight is minimal, as for Fröhlich polarons, local updates changing a couple of variables with large acceptance ratio are preferred because of their efficiency and simplicity. However, when getting the diagram weight is computationally very costly, there is nothing wrong in spending at least as much CPU time on designing efficient global updates changing all diagram variables. The gain in reduced autocorrelation time may more than overcompensate the loss in the acceptance ratio. This is where machine learning techniques hold a great promise for further improving the efficiency of DiagMC simulations. By learning typical model-dependent statistical properties of connected Feynman diagrams with n multi-dimensional coordinates, such as "gyration radius," "dipole," "quadrupole," and "multi-pole" correlations, as well as asymptotic laws for moving one or more vertices well outside of the gyration radius, global updates can achieve large enough acceptance ratios. I am not aware of systematic work done in this direction for connected Feynman diagrams.

5.3 Computational complexity problem for interacting fermions and its solution

To begin with, interacting fermions do not suffer from the sign problem, as any experimentalist measuring their properties is certainly aware of. It is a problem for some, but not all, theoretical methods used to simulate their properties. In general, sign-alternation of contributions simulated by MC methods is neither sufficient nor necessary to state that the problem is intractable. For precise quantitative discussion, one needs to define the "computational complexity problem" (CCP). The most relevant practical question to answer is "How easily can one increase the accuracy of the computed thermodynamic-limit answer?" This leads to the following definition of the CCP that can be applied to any numerical scheme. Let Q be the intensive quantity of interest in the thermodynamic limit.

A numerical scheme has a CCP if the computational time t required to obtain Q with error ϵ diverges faster than any polynomial function of $\epsilon^{-1}\to\infty$. The CCP is considered to be solved if

$$t(\epsilon) = O(\epsilon^{-a}). \tag{50}$$

I discuss only unbiased methods, for which the difference between the computed and exact values can be made arbitrary small. For methods containing an unknown systematic bias, the accuracy cannot be increased indefinitely (but it may be very small).

I will skip the discussion of how methods that suffer from the sign-problem also generically have CCP in dimensions $d \geq 2$, see Ref. [15] for details. On the contrary, for convergent (or subject to resummation) diagrammatic series, the CCP is solved by DiagMC. Indeed, for convergent series the accuracy of truncated sums $Q_N = \sum_{n=0}^{N} d_n$ is improved exponentially fast with the largest diagram order accounted for, $|Q - Q_N| \propto Q\alpha^N$, with $0 < \alpha < 1$. Thus, given some small value of ϵ, the required accuracy is reached by simulating diagrams up to order $N_\epsilon \sim \ln(\epsilon)/\ln(\alpha)$. For the determinant scheme described above the simulation time required to compute all diagrams up to order N_ϵ is an exponential function of N_ϵ, leading to an estimate

$$t(\epsilon) = t(N_\epsilon) \propto b^{N_\epsilon} = \epsilon^{\ln(b)/\ln(\alpha)} \quad \longrightarrow \quad \text{CCP solved} . \tag{51}$$

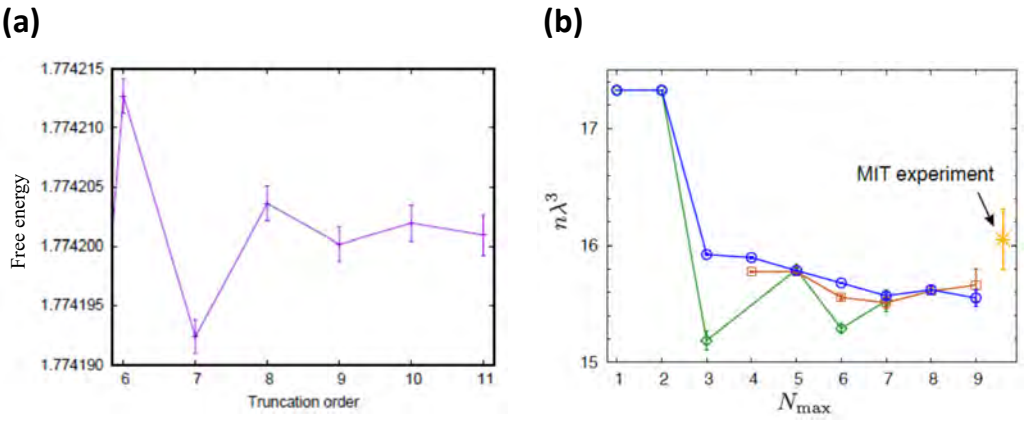

Fig. 17: *(a) Free energy density for the fermionic Hubbard model as a function of truncation order, at $T/t = 0.125$, $U/t = 2$, and $n = 0.87500(2)$ (reproduced from Ref. [12]). (b) Density of the unitary fermi gas (λ is the thermal wavelength) vs. maximal diagram order at $T/\mu = 0.5$ (or $T/T_F = 0.2$). The bold diagrammatic series is resummed by three variants of the conformal-Borel transformation (see Ref. [9]).*

Apart from the cost of the solving recursive equations for connected contributions, the value of b may also include the cost of performing an integral over the space-time variables.

5.4 Illustrative results: Hubbard model and unitary Fermi-gas

It is rare to see sign-free path integral simulations for bosons to be performed with accuracy better than 5 significant digits even for finite size systems. The remarkable plot in Fig. 17(a) proves that for convergent series one can reach an accuracy of 6 significant digits for a generic interacting fermionic system. In this example, the determinant method for connected diagrams was used to simulate the fermionic Hubbard model away from half-filling at density $n = 0.87500(2)$ and relatively low temperature $T/t = 0.125$, where t is the n.n. hopping amplitude. Better convergence was achieved through the Hartree diagram shift of the chemical potential (the convergence radius in the on-site coupling U was estimated to be about $5.1\ t$). The selected parameter set corresponds to the Fermi liquid regime; for larger values of U and on approach to half-filling the situation is much worse and further work is required to improve the performance of DiagMC. Nevertheless, a number of interesting results concerning the nature of magnetic correlations and the pseudogap regime were already obtained.

Fig. 17(b) shows another remarkable outcome of the BDMC simulations done for the unitary fermi gas. This system features a number of universal properties and is relevant for understanding properties of ultra-cold atomic gases and dilute neutron matter. Microscopically, imagine that fermions interact via a short-range attractive potential of radius R and strength U_0 that is fine tuned to the threshold of having a shallow bound state. For two-body collisions at zero energy this situation corresponds to a large s-wave scattering length, $a_s \gg R$. Next, consider a many-body system at finite density, $n = k_F^3/3\pi^2$, where k_F is the Fermi momentum, in the so-call "zero-range" limit, $k_F R \to 0$ when the interparticle distance vastly exceeds the potential radius. In this limit system properties become universal in the sense that all microscopic potentials with the same $k_F a_s$ parameter should be considered equivalent to each other. If $k_F a_s$

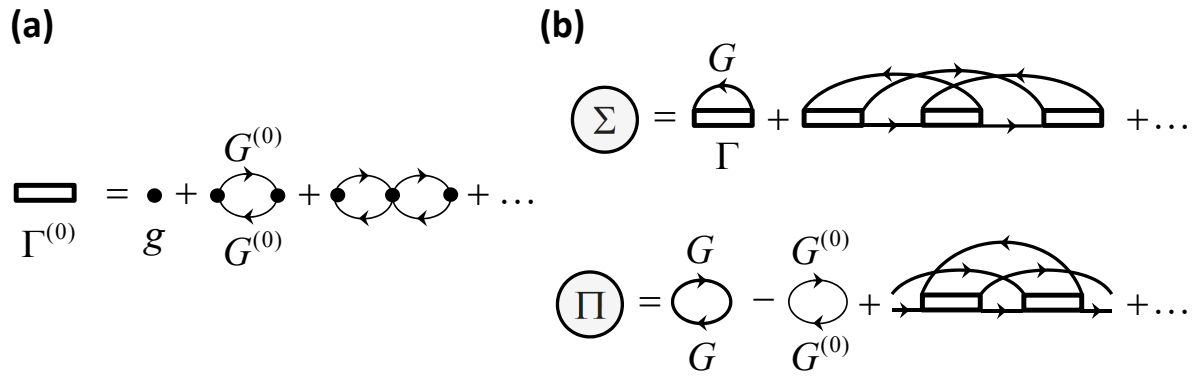

Fig. 18: *(a) Bare pair propagator $\Gamma^{(0)}$ based on the summation of ladder diagrams in terms of the zero-range potential $g\,\delta(\mathbf{r})$ (dots) and non-interacting Green functions (thin lines). (b) An order n skeleton diagram consists of n fully dressed pair propagators and Green functions (bold lines) connecting them.*

is finite, one talks about resonant fermions; this is the canonical model for discussing the BCS-BEC crossover within the superfluid fermionic state.

The unitary fermi gas corresponds to $k_F a_s = \infty$. At low temperature $T < T_F$ it is a strongly interacting system (every spin-up fermion is "contemplating" to form a bound state with every other spin-down fermion) without small parameters to justify a perturbative or mean-field treatment because the only meaningful length (energy) scale in the problem is kF (T_F). Its solution by the DiagMC method involves nearly all the tricks mentioned in this contribution:

(i) To eliminate ultra-violet divergences and to take the zero-range limit analytically, one has to the perform summation of the ladder diagrams prior to the DiagMC simulation and formulate the expansion in the number of pair propagators, see Fig. 18(a).

(ii) To reduce the number of sampled topologies, the simulation is performed for proper self energies $\Sigma[G,\Gamma]$ and $\Pi[G,\Gamma]$ within the self-consistent skeleton formulation, see Fig. 18(b). The self-consistent loop is closed by Dyson equations:

$$1/G = 1/G^{(0)} - \Sigma[G,\Gamma]\,, \qquad 1/\Gamma = 1/\Gamma^{(0)} - \Pi[G,\Gamma]\,.$$

(iii) Asymptotic properties of Γ and G in the limit of short time and distance should be taken care of using exact analytic relations, see Ref. [16] for details.

(iv) Since the resulting series have zero convergence radius, one has to study the nature the of non-analytic behavior at the origin of the expansion by employing Lipatov's technique and construct the appropriate conformal Borel resummation method, see Ref. [9].

The amount of analytic and numeric work may seem daunting, but it can hardly be avoided: the goal is to produce results with guaranteed accuracy bounds. Currently, theoretical error bars are smaller than the most precise experimental data [17] for the equation of state at low temperature right above the transition point to the superfluid state, see Fig. 17(b). At high temperature, $T > T_F$, the BDMC results are far more accurate than experimental data and provide the most stringent test for high-order virial expansion coefficients.

6 Conclusions

This contribution reviewed key principles of DiagMC, as well as some of the recent developments for interacting fermionic (or fermionized) systems that radically improve the efficiency of simulations for high orders of expansion. The number of successful applications is already very large and covers both lattice and continuous-space systems, short- and long-range interaction potentials, effective field theories, interacting topological insulators, frustrated quantum magnets, Bose and Fermi polarons, continuous-time impurity solvers, lattice- and continuous-space path integrals, point-contact dynamics, etc.

There are no known fundamental restrictions on the applicability of the method. However, its efficiency strongly depends on the convergence properties of the series and a deep theoretical/mathematical understanding of the singularities that control this convergence. Gaining such analytic understanding is, arguably, the most important and urgent direction for future work. Numerically, the field will expand in terms of applications to cover models with gauge fields, spin-orbit coupling, and more complex forms of interaction potentials. Codes have to be developed and tested for a variety of effective field theories, shifted action protocols, multi-point correlation functions, and, ultimately, for material science applications.

References

[1] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953)

[2] N.V. Prokof'ev, B.V. Svistunov, and I.S. Tupitsyn, Sov. Phys. JETP **87**, 310 (1998); *ibid.* Phys. Lett. A **238**, 253 (1998)

[3] N.V. Prokof'ev and B.V. Svistunov, Phys. Rev. Lett. **81**, 2514 (1998)

[4] A.S. Mishchenko, N.V. Prokof'ev, A. Sakamoto, and B.V. Svistunov, Phys. Rev. B **62**, 6317 (2000)

[5] O. Goulko, A. Gaenko, E. Gull, N. Prokof'ev, and B. Svistunov, Comp. Phys. Comm. 0010-4655 (2018)

[6] O. Goulko, N. Prokof'ev, and B. Svistunov, Phys. Rev. E **97**, 053305 (2018)

[7] H. Mera, T.G. Pedersen, and B.K. Nikolić, Phys. Rev. B **94**, 165429 (2016); *ibid.* arXiv:1802.06034

[8] L.N. Lipatov, Zh. Eksp. Teor. Fiz. **72**, 411 (1977) [Sov. Phys. JETP **45**, 216 (1977)]

[9] R. Rossi, T. Ohgoe, K. Van Houcke, and F. Werner, Phys. Rev. Lett. **121**, 130405 (2018)

[10] R. Rossi, F. Werner, N. Prokof'ev, and B. Svistunov, Phys. Rev. B **93**, 161102(R) (2016)

[11] W. Wu, M. Ferrero, A. Georges, and E. Kozik, Phys. Rev. B **96**, 041105 (2017)

[12] R. Rossi, Phys. Rev. Lett. **119**, 045701 (2017)

[13] F. Šimkovic IV., E. Kozik, arXiv:1712.10001

[14] R. Rossi, arXiv:1802.04743

[15] R. Rossi, N. Prokof'ev, B. Svistunov, K. Van Houcke, and F. Werner, EPL **118**, 10004 (2017)

[16] K. van Houcke, F. Werner, T. Ohgoe, N. Prokof'ev, and B. Svistunov, Phys. Rev. B **99**, 035140 (2019)

[17] M.J.H. Ku, A. Sommer, L.W. Cheuk, and M.W. Zwierlein, Science **335**, 563 (2012)