A Monte-Carlo method for fermionic many-body problems

Stefan Rombouts

A Monte-Carlo method for fermionic many-body problems

Stefan Rombouts

Proefschrift ingediend tot
het behalen van de graad van
Doctor in de Wetenschappen: Natuurkunde

Promotor : Prof. Dr. K. Heyde

Universiteit Gent
Faculteit Wetenschappen
Vakgroep Subatomaire en Stralingsfysica
1996-1997

Woord Vooraf

Gedurende een kleine vier jaar heb ik gewerkt en gestudeerd met het oog op het behalen van een doctoraat in de natuurkunde. Het was een boeiende periode, in het bijzonder omdat mijn onderzoek zich situeerde op een kruispunt van vele richtingen in de fysica. Al leidde deze multidisciplinariteit onvermijdelijk tot een zekere oppervlakkigheid in elk van de disciplines, toch was het voor mij erg verrijkend om geconfronteerd te worden met facetten uit de kern-, astro-, vaste-stof- en statistische fysica, numerieke analyse, statistiek etc. Ik hoop dat dit proefschrift er in slaagt een vleugje van deze rijkdom door te geven.

Voor dit werk ben ik dank verschuldigd aan vele mensen.

In de eerste plaats aan mijn ouders, die mij de kans hebben geboden om te studeren en die ook steeds mijn nieuwsgierigheid hebben aangemoedigd voor alles wat met wetenschap te maken heeft.

Mijn promotor, professor Kris Heyde, wil ik bedanken voor de mogelijkheid die hij mij geboden heeft om te doctoreren bij de vakgroep Subatomaire en Stralingsfysica van de Universiteit Gent. Ik ben hem zeer erkentelijk voor de stimulerende wijze waarop hij mij toeliet een eigen weg te vinden in het wetenschappelijk onderzoek.

Mijn dank gaat ook uit naar alle collega's van de vakgroep voor hun hartelijke collegialiteit en behulpzaamheid. In het bijzonder ben ik dank verschuldigd aan Veerle Vandersluys, mijn sympathieke bureaugenote aan wie ik niets kon misvragen. De layout van dit proefschrift is gebaseerd op LaTex-style-files van haar hand.

Prof. Karlheinz Langanke en dr. David Dean wens ik te bedanken voor de leerrijke gesprekken. Via hen heb ik kennis gemaakt met quantum-Monte-Carlo-methodes. Ook dr. Hervé Molique bedank ik voor de interessante discussies.

Het Fonds voor Wetenschappelijk Onderzoek - Vlaanderen ben ik dankbaar voor de financiële steun in de vorm van een aspiranten-mandaat en reiskredieten.

Tenslotte wil ik de mensen bedanken van de 'MCMC Preprint Service' (te vinden op het Internet-adres http://www.stats.bris.ac.uk/MCMC/) en van het 'xxx.lanl.gov e-Print archive' (te vinden op http://xxx.lanl.gov/). Deze preprint-databases lieten mij toe de lacunes in het bibliotheekaanbod van de Universiteit Gent te overbruggen. Zonder hen

zou een groot deel van de informatie die ik voor dit proefschrift geraadpleegd heb voor mij onbereikbaar gebleven zijn. Ik heb ondervonden dat in het bijzonder voor multidisciplinair onderzoek het Internet een erg nuttig hulpmiddel is.

Gent, 29 mei 1997

Stefan Rombouts

CONTENTS

Introduction

When I started the research for my Ph.D., in October 1993, I planned to study neutrino scattering reactions on atomic nuclei. Of late years, neutrino-nucleus scattering has caught a lot of attention in connection to astrophysical topics like the supernova explosion mechanism and supernova nucleosynthesis. Especially for this last topic, neutrino-nucleus cross-sections are necessary ingredients for understanding the mechanisms at work. Though most of the nuclei are synthesised via the long known s- and r-processes (neutron capture and beta decay), the origin of some nuclei cannot be explained in this way. Neutrino induced reactions might play a central role in the synthesis mechanism of nuclei like ^{11}Be, ^{19}F, ^{180}Ta. Also in the supernova explosion mechanism neutrinos play an important role. The most abundant elements in the outer shells of a supernova are ^{4}He, ^{12}C, ^{16}O, ^{20}Ne and ^{28}Si. Here it is important to know how much energy the neutrinos can transfer from the core to the envelope of the supernova. The plan was to calculate neutrino-nucleus scattering cross-sections on ^{12}C, ^{16}O, ^{20}Ne and perhaps ^{56}Fe, using a continuum random-phase approximation (CRPA). A CRPA code developed by Jan Ryckebusch for the study of electron scattering on atomic nuclei, was modified for the study of weak-interaction processes, particularly neutrino scattering. For ^{16}O this worked fine. Because ^{16}O is a double-magic, spherical nucleus, the CRPA is a rather good approximation to the real many-body system. However, for ^{12}C, I observed that the results are quite sensitive to the specific structure that is assumed for the ground state. In the energy regime of interest (excitation energies about 10 to 20 MeV), an accurate description of the deformation and the correlations in the ground state of these nuclei is needed in order to obtain a realistic description of the neutrino-nucleus scattering reactions. The CRPA does not describe the ground state of nuclei such as ^{12}C and ^{20}Ne well enough to obtain accurate values for the neutrino scattering cross-sections. A second problem is the fact that CRPA assumes the nucleus to be in its ground state before interacting with a neutrino. In supernovae the temperature can be extremely high, of the order of 10^9K ($= 0.5$ to $1MeV$). At such high temperatures, some nuclei will be in an excited state prior to the interaction with a neutrino. This can have a considerable effect on the neutrino-scattering cross-section (see chapter 7).

In the spring of 1995, during a workshop at the ECT in Trento, I learned about the 'shell-model Monte-Carlo method'. It is a quantum many-body technique that allows the calculation of exact results, up to controllable statistical and systematical errors, in much larger model spaces than the shell-model methods based on diagonalization. Furthermore, it is a finite temperature method. The basic idea of the method is to expand the Boltzmann operator $e^{-\beta\hat{H}}$ as a sum of exponentials of one-body operators. Exponentials of one-body operators can be handeled numerically using a matrix representation for Slater determinants. The number of terms in the sum is too large to compute them all. A limited sample of terms is used instead, to obtain a statistical estimate of the true quantities. The fact that the method can take into account complicated correlations and finite-temperature effects in the initial state, makes it interesting for the study of neutrino scattering on nuclei like ^{12}C and ^{20}Ne. Disadvantages of the method are that it requires quite a lot of computer power and that, for most systems, the calculations at low temperature are spoiled by the so called 'sign problem' (see section 4.5). I decided to develop a shell-model Monte-Carlo code for the study of neutrino-nucleus scattering reactions.

In the following year I experienced that this was not at all a simple task. I spent quite

some time on the development of an accurate algorithm for the evaluation of canonical many-body traces. This lead to a new algorithm for the calculation of the coefficients of the characteristic polynomial of a general square matrix, presented in section 4.3. A second problem was the stabilization of the calculations at low temperature. A solution was found in literature (see section 4.6.1). Together with the algorithm for the canonical traces, it allowed a very accurate evaluation of the canonical trace of the exponentials of a one-body operator. At this stage, calculations were performed for the Hubbard model. (see chapter 7). Because this model has been studied extensively using quantum Monte-Carlo methods, it served as an ideal test case for the method. By the spring of 1996, I started calculations for atomic nuclei. A model was set up for $^{56}_{26}\mathrm{Fe}_{30}$, using a harmonic-oscillator plus pairing plus quadrupole Hamiltonian, and a discrete model space with 40 single-particle states. However, the Markov chain for the Monte-Carlo sampling failed to converge. Too many Markov steps would be needed to obtain accurate results and each Markov step required too much computer time. for the quantum Monte-Carlo calculation to be feasible.

It forced me to take a closer look on Markov-chain Monte-Carlo methods. This resulted in chapter 3 of this work. A study of the convergence properties of Markov-chain Monte-Carlo methods lead to a much better understanding of the convergence and to some rules of thumb for the construction of transition kernels. I implemented a practical way to determine error limits and optimized the number of Markov steps between successive evaluations of observables. The possibility of variance reduction in Markov-chain Monte-Carlo methods was studied. In order to speed up the computation of each Markov step, an improved sampling scheme was implemented (see section 4.6.2). In order to improve the performance of the method, I elaborated on decompositions based on rank-one and rank-two operators (see section 2.2.2). As a result, an alternative to the Hubbard-Stratonovich transform was found, which allowed faster calculations for the Hubbard model and the nuclear pairing Hamiltonian, among others. By the end of 1996, all these building blocks were put together to form a powerful quantum Monte-Carlo method for the fermionic many-body problem. The Hubbard model was studied as a test case. The results of these calculations are presented in chapter 5. As a first real application, the nuclear pairing Hamiltonian was studied (see chapter 6). Excellent agreement with exact results was obtained for the exactly solvable nuclear pairing model with degenerate single-particle levels. Thermal properties of a mean-field plus pairing model for nuclei in the Fe region were obtained. Up to controlable statistical (number of Markov steps) and systematical (number of inverse temperature intervals) errors, these results amount to an exact solution of the model at finite temperature. As such the method is more powerful than approximate techniques such as BCS. Furthermore, it can handle much larger model spaces than diagonalization techniques. A major drawback of the method is the 'sign problem' (see section 4.5). For most systems, it spoils the calculations at low temperature. However, I observed that for a lot of systems the method can still be used at temperatures at which the system is almost completely cooled to its ground state, such that there is no need to go to even lower temperatures.

It was not straightforward to find a good name for the quantum Monte-Carlo method.

- *"shell-model quantum Monte-Carlo"* is to restricted, because the method can be applied equally well to other fermionic many-body models than the nuclear shell

model, e.g. to the Hubbard model.

- *"auxiliary-field quantum Monte-Carlo"* is used in literature to indicate the quantum Monte-Carlo methods that are based on the Monte-Carlo integration over the auxiliary fields σ that arise in the Hubbard-Stratonovich transformation of the operator $e^{-\beta\hat{H}_2}$. (see section 2.2.1). Because alternative decompositions were developed for the operator $e^{-\beta\hat{H}_2}$, that are not based on auxiliary fields, this name is not appropriate any more.

- *"projector quantum Monte-Carlo"* is used in literature to indicate the method with ground-state projection, as discussed in section 4.4. This name is not appropriate for the application of the method in the canonical nor grand canonical ensemble.

- *"grand-canonical quantum Monte-Carlo"* is used in literature to indicate the method applied in the grand canonical ensemble. This name is not appropriate for the application of the method in the canonical ensemble nor for the ground-state-projection method.

- *"determinant quantum Monte-Carlo"* is used in literature to indicate the method applied in the grand canonical ensemble and the method with ground-state projection. The evaluation of the weights in these methods is based on the evaluation of determinants. However, in the canonical ensemble, no determinants are needed. Therefore this name is not appropriate either.

Because the method, in any form, is based on the expansion of the Boltzmann operator in a sum of terms that each can be handled easily using a matrix representation for Slater determinants, I decided to use the name *"Slater-determinant quantum Monte-Carlo"* method (SDQMC) in this work to indicate the method, in any of its forms. Thus, in this work, SDQMC is used as a general term for the shell-model quantum Monte-Carlo, auxiliary-field quantum Monte-Carlo, projector quantum Monte-Carlo, grand-canonical quantum Monte-Carlo and determinant quantum Monte-Carlo methods.
In the near future, the improved SDQMC will be applied to the study of neutrino-nucleus scattering reactions (see chapter 7). Another topic for futher research is the implementation of an algorithm for the inverse Laplace transform, in order to calculate strength functions and perhaps level densities (see section 4.1). Furthermore, for the study of atomic nuclei, attention will be paid to the separation of the spurious center-of-mass motion from the intrinsic excitations.
This work consists of two parts. In the first part, the SDQMC method is presented. Chapter 1 introduces the basic notations and a matrix representation for Slater determinants. Using this matrix representation, the exponential of a one-body operator can be handled easily in a numerical way. In chapter 2, several ways are presented to decompose the Boltzmann operator $e^{-\beta\hat{H}}$, which is generally the exponential of a two-body operator, into a sum of exponentials of one-body operators. A self-contained discussion of Markov-chain Monte-Carlo methods is given in chapter 3. The building blocks presented in chapters 1 to 3 are brought together in chapter 4 to constitute the Slater-determinant quantum Monte-Carlo method. Special attention is given to the 'sign problem'. In the second part, the application of SDQMC to several specific fermionic many-body systems is discussed.

Results for the 4×4 Hubbard model are presented in chapter 5. SDQMC calculations for the nuclear pairing Hamiltonian are discussed in chapter 6. Finally, an outlook for SDQMC calculations of neutrino-nucleus scattering cross-sections is given in chapter 7. The computer calculations for this work were performed on Digital workstations (Alpha 3000-600 and Alphastation 255/300MHz systems with a Digital-Unix operating system) and a PC (Pentium-Pro 200-MHz processor, with a Linux operating system).

The method

Overview

Chapter 1 introduces the basic notations and a matrix representation for Slater determinants. Using this matrix representation, the exponential of a one-body operator can be handled easily in a numerical way. The evaluation of the canonical or grand canoninal trace then amounts to the evaluation of the characteristic polynomial or the determinant of a matrix of moderate dimensions (the dimension equals the number of single-particle states taken into account in the model). In chapter 2, several ways are presented to decompose the Boltzmann operator $e^{-\beta\hat{H}}$, which is generally the exponential of a two-body operator, into a sum of exponentials of one-body operators. In this way, the expressions for the traces of exponentials of one-body operators can also be applied to the Boltzmann operator. Because the number of terms in the decomposition is overwhelmingly huge, a complete summation is impossible. Instead, Markov-chain Monte-Carlo methods are used to draw a sample from them and to evaluate the sum statistically. A self-contained discussion of these Markov-chain Monte-Carlo methods is given in chapter 3. The building blocks presented in these chapters are brought together in chapter 4 to constitute the Slater-determinant quantum Monte-Carlo method. This method allows the study of ground-state properties and thermodynamical properties in the canonical and grand canonical ensemble of discrete fermionic many-body systems. Special attention is given to the 'sign problem'.

1

Fermionic many-body theory with Slater determinants

1.1 Notations

Slater-determinant quantum Monte-Carlo methods (SDQMC) are based on the expansion of the thermodynamic partition function $Z_\beta = \mathrm{Tr}\left(e^{-\beta\hat{H}}\right)$ of a fermionic quantum many-body system as a sum of traces of operators that have numerically manageable form. This form is based on a matrix representation of Slater determinants. We would like to emphasize the difference between the Hilbert space of many-body states and the space of matrix representations of Slater determinants. The connection between the two will be made through the space of single-particle states. In order to avoid confusion and to allow a sound description of the SDQMC the following notations are used:

- $\Psi, \Phi, \ldots$: uppercase Greek letters for many-body states. The corresponding many-body wave functions are denoted as $\Psi(X), \Phi(X)$, with $X = (x_1, \ldots, x_A)$ a generalized coordinate. Note that many-body wave functions that differ by a constant factor represent the same many-body state.

- $\psi, \phi, \ldots$: lowercase Greek letters for single-particle states. The corresponding single-particle wave function is denoted with $\psi(x)$. Note that single-particle wave functions that differ by a constant factor represent the same single-particle state.

- $\varphi_1, \varphi_2, \ldots, \varphi_{N_\mathcal{S}}$: the basis states of the one-particle space. $N_\mathcal{S}$ is the number of basis states, $\mathcal{S} = \{\varphi_1, \varphi_2, \ldots, \varphi_{N_\mathcal{S}}\}$ is the set of basis states of the single-particle space.

- $M, \ldots$: matrices, in particular the representation matrices of Slater determinants, will be denoted with uppercase Roman letters.

- $\hat{a}_k, \hat{a}_k^\dagger$: annihilation and creation operators for a particle in state φ_k.
- $\hat{n}_k = \hat{a}_k^\dagger \hat{a}_k$: number operator for state φ_k.
- $\hat{P} = \sum_{k,l} [\hat{P}]_{kl} \hat{a}_k^\dagger \hat{a}_l$: uppercase Roman letters with a hat for one-body operators; square brackets around a one-body operator denote the matrix defined by $[\hat{P}]_{kl} = \langle \varphi_k | \hat{P} | \varphi_l \rangle$.
- $\hat{H}, \hat{V}, \ldots$: uppercase Roman letters with a hat for many-body operators.

The following symbols will occur often in this work:

- N is the number of particles.
- β is the *inverse temperature*, in literature sometimes referred to as the *imaginary time*.
- $N_{\mathcal{S}}$ is the number of single-particle basis states, in other words the dimension of the single-particle space $\mathcal{S}$.
- N_t is the number of inverse temperature intervals, in literature sometimes referred to as the number of *time slices*.

1.2 Many-body states and Slater determinants

1.2.1 The model space

Since computers can only work with finite discrete numbers, any numerical many-body technique needs a discretization at some level. Even so called '*continuum* RPA techniques' require a discretization of the coordinate or momentum space. In the nuclear shell model, one mainly neglects the degrees of freedom of the deeply bound nucleons. Only a few valence particles distributed over a number of valence-orbitals, in the outher shells of the nucleus, are taken into account as degrees of freedom of the system. One then constructs a modelspace of configurations of these valence nucleons with a definite rotational symmetry. The Hamiltonian is diagonalized in this modelspace in order to determine energy-levels and other observables. Though this is a serious truncation of the complete many-body space, the shell model, especially around half-filled shells, still leads to model spaces that can hardly be handled with present day computers.
In SDQMC, the basic discretization is done on the level of the single-particle states: one considers only a limited set $\mathcal{S}$ of discrete single-particle states. These states can be energy eigenstates in a mean-field potential, as in the nuclear shell model; they can be sites on a cristal lattice as in the Hubbard model; they could be momentum eigenstates in other applications. The many-body states are constructed by distributing N particles over these $N_{\mathcal{S}}$ single-particle states. Their wave functions are antisymmetric functions $\Psi(X)$ on $\mathcal{S}^N$, with $X = (x_1, \ldots, x_N)$ a generalized coordinate, $x_i \in \mathcal{S}$ for $i = 1, \ldots, N_{\mathcal{S}}$. This leads to a finite discrete Hilbert space $\mathcal{H}$ with dimension $N_{\mathcal{H}} = \binom{N}{N_{\mathcal{S}}} = N_{\mathcal{S}}! / [N!(N_{\mathcal{S}} - N)!]$.

This is the model space for SDQMC. On high-performant computers SDQMC can handle systems with 100 particles distributed over 200 single-particle states, leading to many-body spaces of dimension 10^{60}. Diagonalization techniques are limited to systems with a number of many-body states of the order of 10^6. This indicates the power of SDQMC methods.

1.2.2 Slater determinants

A special set of many-body states is formed by the states whose wave function can be written as an antisymmetrized product of N different single-particle wave functions. These states are called *Slater determinants*. A neccesary and sufficient condition [1] for a fermionic many-body state Ψ to be a Slater determinant is

$$\begin{aligned}\Psi(y_1, y_2, x_3, \ldots, x_N)\Psi(y_3, y_4, x_3, \ldots, x_N) + \\ \Psi(y_2, y_3, x_3, \ldots, x_N)\Psi(y_1, y_4, x_3, \ldots, x_N) + \\ \Psi(y_3, y_1, x_3, \ldots, x_N)\Psi(y_2, y_4, x_3, \ldots, x_N) = 0\end{aligned} \tag{1.1}$$

for all values of the coordinates $x_3, \ldots, x_N, y_1, \ldots, y_4$. If $\Psi(X)$ is the antisymmetrized product of N single-particle wave functions $\psi_1(x), \ldots, \psi_N(x)$ then $\Psi(X)$ can be written down as

$$\Psi(x_1, \ldots, x_N) = \det \begin{pmatrix} \psi_1(x_1) & \cdots & \psi_1(x_N) \\ \vdots & & \vdots \\ \psi_N(x_1) & \cdots & \psi_N(x_N) \end{pmatrix}, \tag{1.2}$$

a notation first used by Slater. Hence the name 'Slater determinants'. We will use this term to refer to the many-body state Ψ, not just to the determinant used in 1.2. For a Slater determinant Ψ, a set of unnormalized single-particle wave functions is given by

$$\psi_i(x) \;=\; \Psi(Y_{[y_i \to x]}) \tag{1.3}$$

with $Y = (y_1, \ldots, y_N)$ a fixed point in $\mathcal{S}^N$, $\Psi(Y) \neq 0$, and $Y_{[y_i \to x]}$ the point obtained by replacing y_i in Y with x. For every Slater determinant Ψ there exist many sets of single-particle wave functions, to every set $\{\psi_1(x), \ldots, \psi_N(x)\}$ of linearly independent single-particle wave functions corresponds one Slater determinant.

1.2.3 A matrix representation for Slater determinants

Because the single-particle space is finite and discrete, single-particle wave functions can be represented by $N_{\mathcal{S}}$-dimensional column vectors:

$$\psi \;\longleftrightarrow\; \begin{bmatrix} \langle\varphi_1|\psi\rangle \\ \langle\varphi_2|\psi\rangle \\ \vdots \\ \langle\varphi_{N_S}|\psi\rangle \end{bmatrix}. \tag{1.4}$$

This leads to an interesting matrix representation for Slater determinants: If $\Psi(X)$ is the antisymmetrized product of $\psi_1(x), \ldots, \psi_N(x)$ then Ψ can be represented by the $N_{\mathcal{S}} \times N$

matrix M given by

$$\Psi \longleftrightarrow \begin{bmatrix} \langle\varphi_1|\psi_1\rangle & \cdots & \langle\varphi_1|\psi_N\rangle \\ \vdots & & \vdots \\ \langle\varphi_{N_S}|\psi_1\rangle & \cdots & \langle\varphi_{N_S}|\psi_N\rangle \end{bmatrix}. \tag{1.5}$$

Every Slater determinant Ψ can be represented by many different $N_\mathcal{S} \times N$ matrices, to every non-singular $N_\mathcal{S} \times N$ matrix M corresponds one Slater determinant Ψ_M. Singularity of M would mean that two or more particles occupy the same single-particle state, which is forbidden by the Pauli prinicple. This matrix representation for Slater determinants is useful because of two properties:

- the overlap between two Slater determinants can be written as the determinant of the product of the representation matrices:

$$\langle\Psi_{M_1}|\Psi_{M_2}\rangle = \det\left(M_1^{\mathrm{T}} M_2\right); \tag{1.6}$$

- the result of the exponential of a one-body operator $\hat{P}$ working on a Slater determinant Ψ_m can be represented by the operation of the exponential of an $N_\mathcal{S} \times N_\mathcal{S}$ matrix P on the matrix representation M of of Ψ_M:

$$e^{\hat{P}}\Psi_M = \Psi_{M'}, \tag{1.7}$$

with the matrix M' given by

$$M' = e^{[\hat{P}]}M. \tag{1.8}$$

Here $[\hat{P}]$ is the $N_\mathcal{S} \times N_\mathcal{S}$ matrix defined by

$$[\hat{P}]_{ij} = \langle\phi_i|\hat{P}|\phi_j\rangle. \tag{1.9}$$

This last property is a corollary of the 'Thouless theorem' which states that the exponential of a one-body operator transforms Slater determinants into Slater determinants [2]. It constitutes the cornerstone of SDQMC: the representation of exponentials of operators on the many-body space $\mathcal{H}$ by operations with matrices of dimension $N_\mathcal{S} \times N_\mathcal{S}$ or $N_\mathcal{S} \times N$. Note that in general

$$\Psi_{M_1+M_2} \neq \Psi_{M_1} + \Psi_{M_2}. \tag{1.10}$$

A special set of Slater Determinants is formed by the Slater Determinants that can be represented by a matrix with one element set to 1 in every column and the other elements set to 0. This set, which we will denote with $\mathcal{D}_0$, constitutes a basis for the entire Hilbert space.

1.2.4 Many-body traces of exponentials of one-body operators

The matrix representation for Slater determinants allows a handy way to calculate the many-body trace of the exponential of a one-body operator.

Let $\hat{U}$ be an operator that transforms a Slater determinant Ψ_M represented by the $N_{\mathcal{S}} \times N$ matrix M into the Slater determinant Ψ'_M represented by the matrix $M' = UM$, where U is an $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrix. An example of such an operator $\hat{U}$ is the exponential of a one-body operator, or a product of exponentials of one-body operators. The N-particle trace of the operator $\hat{U}$ is given by

$$\hat{\mathrm{Tr}}_N\left(\hat{U}\right) = \sum_{i=1}^{N_{\mathcal{H}}} \langle \Phi_i | \hat{U} | \Phi_i \rangle, \tag{1.11}$$

where $\{\Phi_1, \ldots, \Phi_{N_{\mathcal{H}}}\}$ is a complete basis for the N-particle Hilbert space $\mathcal{H}$. One such a basis is the set $\mathcal{D}_0$. These Slater determinants are represented by matrices $\{B_1, \ldots, B_{N_{\mathcal{H}}}\}$ which have in every column one element equal to one and all the other elements equal to zero. The trace now becomes

$$\begin{aligned}
\hat{\mathrm{Tr}}_N\left(\hat{U}\right) &= \sum_{i=1}^{N_{\mathcal{H}}} \langle \Phi_i | \hat{U} | \Phi_i \rangle & (1.12)\\
&= \sum_{i=1}^{N_{\mathcal{H}}} \langle \Psi_{B_i} | \hat{U} | \Psi_{B_i} \rangle & (1.13)\\
&= \sum_{i=1}^{N_{\mathcal{H}}} \langle \Psi_{B_i} | \Psi_{UB_i} \rangle & (1.14)\\
&= \sum_{i=1}^{N_{\mathcal{H}}} \det\left(B_i^T U B_i\right). & (1.15)
\end{aligned}$$

Because of the special form of the matrices $B_1, \ldots, B_{N_{\mathcal{H}}}$, this trace is just the sum of all diagonal minors of rank N of the matrix U, which is nothing else than the coefficient of χ^N in the polynomial $\det(1 + \chi U)$. So we obtain that

$$\hat{\mathrm{Tr}}_N\left(\hat{U}\right) = \left.\frac{\left(\frac{d}{d\chi}\right)^N \det(1 + \chi U)}{N!}\right|_{\chi=0}. \tag{1.16}$$

In a thermodynamical language this N-particle trace is called the *canonical* trace, since it concerns a system with a fixed number of particles. If we extend the trace to states with any number of particles, N ranging from 0 to $N_{\mathcal{S}}$, we get the *grand canonical* trace. The grand canonical trace for a given chemical potential μ and inverse temperature β is given by

$$\begin{aligned}
\hat{\mathrm{Tr}}_{GC,\mu}\left(\hat{U}\right) &= \sum_{N=0}^{N_S} e^{\beta\mu N} \hat{\mathrm{Tr}}_N\left(\hat{U}\right) & (1.17)\\
&= \sum_{N=0}^{N_S} \left(e^{\beta\mu}\right)^N \left.\frac{\left(\frac{d}{d\chi}\right)^N \det(1 + \chi U)}{N!}\right|_{\chi=0} & (1.18)\\
&= \det\left(1 + e^{\beta\mu} U\right). & (1.19)
\end{aligned}$$

The properties of the matrix representations for Slater determinants allow us to calculate the canonical and grand canonical trace of the exponential of a one-body operator using

linear-algebra techniques for matrices of dimension $N_{\mathcal{S}}$. The amount of computing time needed for these techniques scales as $N_{\mathcal{S}}^3$. Therefore the computing time remains reasonable even for large model spaces.

1.2.5 Bosonic many-body states

Just like Slater determinants have wave functions that can be written as an antisymmetrized product of single-particle wave functions, some bosonic many-body states have wave functions that can be written as a symmetrized product of single-particle wave functions. These bosonic states can be represented by $N_{\mathcal{S}} \times N$ matrices too. Non-singularity of these matrices is not required because the Pauli principle does not apply to bosons. The property 1.7 also holds for these states, but there is no bosonic analogy to property 1.6 and no simple algebraic expression for the canonical or grand canonical traces analogous to expressions 1.16 or 1.17 exist. The evaluation of the overlap between bosonic wavefunctions would require the calculation of the 'permanent' of the matrix $M_1^T M_2$. (The permanent of a $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrix M is the symmetric analogon of the determinant. It is given by $\sum_\pi M_{1\pi_1} M_{2\pi_2} \cdots M_{N_{\mathcal{S}}\pi_{N_{\mathcal{S}}}}$, where the sum runs over all permutations π of the set $\{1, 2, \ldots, N_{\mathcal{S}}\}$.) The amount of computation time needed for the evalutation of the permanent grows exponentially with $N_{\mathcal{S}}$ [3], so that calculations for large model spaces become impracticable. Therefore the SDQMC method has no analogon for bosons.
There are however other Monte-Carlo techniques that can be applied to bosonic many-body problems. A lot of these techniques are based on a Monte-Carlo sampling of many-body states instead of a sampling of the many-body interactions [13]. Compared to Monte-Carlo methods for many-fermion systems, Monte-Carlo methods for many-boson systems have the advantage that they do not suffer from 'sign-problems' (see section 4.5).

2

Decomposition of the Boltzmann operator

2.1 Exponential of a sum of non-commuting operators

If the Hamiltonian $\hat{H}$ would be a one-body operator, we could use the expressions from the previous chapter to calculate the properties of the Boltzmann operator $e^{-\beta\hat{H}}$. This operator is interesting because it contains all the thermodynamic information of the many-body system. The thermodynamic partition function of the many-body system is given quantummechanically by the trace of this operator (in the case of zero chemical potential):

$$Z_\beta = \hat{\mathrm{Tr}}\left(e^{-\beta\hat{H}}\right). \tag{2.1}$$

Here, β has to be understood as the inverse temperature. Thermodynamic quantities as the internal energy U, the free energy A, the entropy S and more, can be derived from Z_β (see section 4.1):

$$U = -\frac{\partial \ln(Z_\beta)}{\partial\beta}, \tag{2.2}$$

$$A = -\frac{\ln(Z_\beta)}{\beta}, \tag{2.3}$$

$$S = \beta(U - A). \tag{2.4}$$

Units were chosen such that the Boltzmann constant $k = 1$. Another way to use the operator $e^{-\beta\hat{H}}$ is to see it as an operator that projects onto the many-body ground state:

$$\Psi_{E_0} \sim e^{-\beta\hat{H}}\Psi, \tag{2.5}$$

for large β and for any many-body state Ψ that has a non-vanishing overlap with Ψ_{E_0}. The components of energy eigenstates Ψ_E with a higher energy E are suppressed by a factor

$e^{-\beta(E-E_0)}$. So if we could evaluate the Boltzmann operator accurately, we would have a way to obtain both thermodynamical and groundstate information about the quantum many-body system. Our aim is to do this by decomposing $e^{-\beta\hat{H}}$ as a sum of exponentials of one-body operators. These operators can be handled easily in the Slater-determinant representation introduced in the previous section.
In many-body theory we are mostly dealing with two-body Hamiltonians, or even three-body Hamiltonians. In this work we concentrate on the former. A first step will be to separate the easy one-body part $\hat{H}_1$ of the Hamiltonian $\hat{H} = \hat{H}_1 + \hat{H}_2$ from the more difficult two-body part $\hat{H}_2$ in the expression for the operator $e^{-\beta\hat{H}}$. If $\hat{H}_1$ and $\hat{H}_2$ would commute, we could write

$$e^{-\beta(\hat{H}_1+\hat{H}_2)} = e^{-\beta\hat{H}_1}e^{-\beta\hat{H}_2}. \tag{2.6}$$

In general however $\hat{H}_1$ and $\hat{H}_2$ will not commute. For small values of β, and to second order in β we can write:

$$e^{-\beta(\hat{H}_1+\hat{H}_2)} = e^{-\beta\hat{H}_1}e^{-\beta\hat{H}_2} + \frac{\beta^2}{2}\left[\hat{H}_2, \hat{H}_1\right] + \mathcal{O}(\beta^3). \tag{2.7}$$

This means that expression 2.6 has an error of the order of β^2. We can reduce this error to third order in β by using the Suzuki-formula [4]:

$$e^{-\beta(\hat{H}_1+\hat{H}_2)} = e^{-\frac{\beta}{2}\hat{H}_1}e^{-\beta\hat{H}_2}e^{-\frac{\beta}{2}\hat{H}_1} + \beta^3\hat{R}, \tag{2.8}$$

where the error term $\hat{R}$ can be estimated by

$$\|\hat{R}\| \leq \frac{\|\left[\hat{H}_1, \left[\hat{H}_1, \hat{H}_2\right]\right]\| + 2\|\left[\hat{H}_2, \left[\hat{H}_1, \hat{H}_2\right]\right]\|}{24}. \tag{2.9}$$

Expressions that are correct to even higher order in β can be derived [5], but they require more factors in the expansion of $e^{-\beta\hat{H}}$ (already 9 factors are needed to reduce the error to order β^4), which makes these expressions practically uninteresting.
Expression 2.8 is only useful for small values of β. At higher values of β, we need to split up the inverse temperature β in a number of *inverse temperature intervals*. In literature, also the term *imaginary-time intervals* is used. Let N_t denote the number of intervals. This leads to the *Suzuki-Trotter* formula[7]:

$$\begin{aligned} e^{-\beta(\hat{H}_1+\hat{H}_2)} &= \left(e^{-\frac{\beta}{N_t}(\hat{H}_1+\hat{H}_2)}\right)^{N_t} && (2.10)\\ &= \left(e^{-\frac{\beta}{2N_t}\hat{H}_1}e^{-\frac{\beta}{N_t}\hat{H}_2}e^{-\frac{\beta}{2N_t}\hat{H}_1} + \frac{\beta^3}{N_t^3}\hat{R}\right)^{N_t} && (2.11)\\ &= e^{-\frac{\beta}{2N_t}\hat{H}_1}e^{-\frac{\beta}{N_t}\hat{H}_2}e^{-\frac{\beta}{N_t}\hat{H}_1}e^{-\frac{\beta}{N_t}\hat{H}_2}\cdots e^{-\frac{\beta}{N_t}\hat{H}_1}e^{-\frac{\beta}{N_t}\hat{H}_2}e^{-\frac{\beta}{2N_t}\hat{H}_1} \\ &\quad + \frac{\beta^3}{N_t^2}\hat{R}'. && (2.12) \end{aligned}$$

In SDQMC a balance has to be found between computational effort and accuracy. Because matrix multiplications can be quite computationally demanding, the number of factors in the expansion of $e^{-\beta\hat{H}}$ will determine the computational effort needed. It can be seen

from 2.12 that the expression 2.8 does only lead to one more factor than expression 2.7, while its error is an order β/N_t smaller. We therefore recommend to keep the error in every interval of order $(\beta/N_t)^3$, also in the decomposition of $e^{-\beta\hat{H}_2}$ that will be described in the next section.
We could use an expression with an error of the order $(\beta/N_t)^4$ per interval. If we insist on keeping the total number of factors equal to the number of factors we would have with expression 2.12, we could use only $N_t/4$ intervals (we would have 9 factors per interval; taking together the factors at the borders of two intervals would lead to 8 factors per interval, compared with 2 factors per interval for expression 2.12). The total error would be of the order of $\beta^4/(4N_t)^3$. This error would be comparable to the error in the *Suzuki-Trotter* formula if $N_t \approx 64\beta$. For most applications, a smaller value for N_t is sufficient, so that the *Suzuki-Trotter* formula is more efficient than higher order approximations.

2.2 Decomposition of $\exp(-\beta\hat{H}_2)$

In this section we will show how the exponential of the two-body Hamiltonian can be decomposed as a sum of exponentials of one-body operators:

$$\exp(-\frac{\beta}{N_t}\hat{H}_2) = \sum_{\sigma} e^{\hat{A}_\sigma}. \tag{2.13}$$

If we apply this decomposition to every factor $\exp(-\frac{\beta}{N_t}\hat{H}_2)$ in 2.12, we obtain

$$\begin{aligned}\exp(-\beta\hat{H}) &\simeq \sum_{\sigma_1,\sigma_2,\ldots,\sigma_{N_t}} e^{-\frac{\beta}{2N_t}\hat{H}_1} e^{\hat{A}_{\sigma_1}} e^{-\frac{\beta}{N_t}\hat{H}_1} e^{\hat{A}_{\sigma_2}} \cdots e^{-\frac{\beta}{N_t}\hat{H}_1} e^{\hat{A}_{\sigma_{N_t}}} e^{-\frac{\beta}{2N_t}\hat{H}_1} \\ &= \sum_{\sigma} e^{-\hat{S}_\sigma(\beta)} = \sum_{\sigma} \hat{U}_\sigma. \end{aligned} \tag{2.14}$$

The operator $\hat{U}_\sigma$ is a product of exponentials of one-body operators and as such an exponential of a one-body operator $\hat{S}_\sigma(\beta)$ itself. Therefore it can be represented by a $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrix U_σ. This allows us to use the expressions for $\hat{\mathrm{Tr}}\left(\hat{U}_\sigma\right)$ derived in chapter 1.
The decomposition of the exponential of $-\beta\hat{H}_2$ can be achieved in many different ways. Since we will have to evaluate the sum with Monte-Carlo techniques, there are a few guidelines for choosing a decomposition:

- A decomposition that is exact is preferable to a composition that is only approximate. Since the Suzuki-Trotter formula 2.12 leads to an error of the order of β^3, it is recommended to keep the error of an approximate decomposition at least of the same order.

- A decomposition where all terms have a similar structure is preferable to a decomposition which has terms of very different nature, since the former can be expected to lead to smaller variances in the Monte-Carlo evaluation.

- A decomposition with less terms is preferable to one with more terms, since it can be expected to lead to smaller autocorrelation times in the Monte-Carlo evaluation. This might conflict with the previous guideline: sometimes it is preferable to have more but smoother terms in the decomposition.
- Decompositions should be devised to reduce *sign problems* in the Monte-Carlo evaluation (cfr. section 4.5). Sometimes certain symmetries guarantee the positiveness of the traces of the terms in the decompositions. In those cases it is recommended to use a decomposition that conserves this symmetry in every term.
- Since matrix multiplications take most of the time in the actual calculations, decompositions that lead to sparse matrices are preferable to decompositions that lead to dense matrices. Sometimes certain symmetries allow to calculate the trace of a term by using only part of the matrix that represents the term. This can reduce the computation time considerably. In those cases it is recommended to use a decomposition that conserves this symmetry in every term.
- For schematic interactions, the structure of the Hamiltonian often suggests a specific decomposition.

In the next sections we will highlight certain ways of constructing a decomposition for general Hamiltonians.

2.2.1 The Hubbard-Stratonovich transform

This decomposition is due to R. L. Stratonovich [8]. It was applied by J. Hubbard to the partition function of quantum many-body systems [9].

The Hubbard-Stratonovich tranform for a single quadratic Hamiltonian

If $\hat{H}_2$ is a quadratic operator, i.e.

$$\hat{H}_2 = -\hat{A}^2, \tag{2.15}$$

with $\hat{A}$ a one-body operator, then the following idendity holds:

$$e^{-\beta\hat{H}_2} = e^{\beta\hat{A}^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\sigma^2}{2}} e^{\sigma\sqrt{2\beta}\hat{A}} d\sigma. \tag{2.16}$$

This expresses the exponential of a two-body Hamiltonian as a continuous sum of exponentials of one-body operators. It is an exact decomposition, which can be seen by expanding in orders of $\hat{A}$: for the i^{th} term in the expansion one has

$$\begin{array}{rcll} 0\hat{A}^i & = & 0\hat{A}^i & \text{if } i \text{ is odd,} \\ \frac{1}{(i/2)!}\,\beta^{\frac{i}{2}}\hat{A}^i & = & \frac{2^{\frac{i}{2}}}{i!\sqrt{2\,\pi}} \int_{-\infty}^{+\infty} \sigma^i e^{-\frac{\sigma^2}{2}} d\sigma\; \beta^{\frac{i}{2}}\hat{A}^i & \text{if } i \text{ is even.} \end{array} \tag{2.17}$$

Instead of the continuous sum over the *auxiliary field* σ, one can use a discrete decomposition. This is achieved by replacing $\int_{-\infty}^{\infty} e^{-\frac{\sigma^2}{2}} d\sigma$ with a Gaussian quadrature formula.

It has been found that discrete decompositions lead to shorter autocorrelation lengths and faster convergence in the Monte-Carlo evaluation than continuous decompositions [7, 10]. If we want the error to be of order β^3, we need a Gaussian quadrature formula that is exact at least up to 5^{th} order in σ. A 3-points quadrature formula suffices. This leads to a discrete Hubbard-Stratonovich decomposition:

$$e^{-\beta\hat{H}_2} = e^{\beta\hat{A}^2} = \frac{4 + e^{+\sqrt{6\beta}\hat{A}} + e^{-\sqrt{6\beta}\hat{A}}}{6} + \hat{R}(\beta), \tag{2.18}$$

with the error given by

$$\begin{aligned} \hat{R}(\beta) &= \frac{\beta^3}{15}\hat{A}^6 + \text{higher order terms in } \beta & (2.19)\\ &= -\frac{\beta^3}{15}\hat{H}_2^3 + \text{higher order terms in } \beta. & (2.20) \end{aligned}$$

If we compare this error to the error originating from the Suzuki-Trotter formula 2.12, we see that the former will be the dominant one. It might therefore be useful to use a four-point quadrature formula in order to reduce the error and to obtain a good convergence:

$$\begin{aligned} &e^{-\beta\hat{H}_2} = e^{\beta\hat{A}^2} = \\ &\quad \frac{e^{+\sqrt{(\sqrt{6}+2)\sqrt{6}\beta}\hat{A}} + e^{-\sqrt{(\sqrt{6}+2)\sqrt{6}\beta}\hat{A}}}{2(\sqrt{6}+2)\sqrt{6}} + \frac{e^{+\sqrt{(\sqrt{6}-2)\sqrt{6}\beta}\hat{A}} + e^{-\sqrt{(\sqrt{6}-2)\sqrt{6}\beta}\hat{A}}}{2(\sqrt{6}-2)\sqrt{6}} \\ &\quad +\hat{R}(\beta), \end{aligned} \tag{2.21}$$

with the error given by

$$\begin{aligned} \hat{R}(\beta) &= \frac{\beta^4}{105}\hat{A}^8 + \text{higher order terms in } \beta & (2.22)\\ &= \frac{\beta^4}{105}\hat{H}_2^4 + \text{higher order terms in } \beta. & (2.23) \end{aligned}$$

Extension of the Hubbard-Stratonovich decomposition for a sum of squares of commuting operators

If the two-body Hamiltonian is minus a sum of squares of commuting one-body operators,

$$\hat{H}_2 = -\hat{A}_1^2 - \hat{A}_2^2 - \ldots - \hat{A}_m^2, \tag{2.24}$$

with

$$\left[\hat{A}_i, \hat{A}_j\right] = 0 \text{ for all } i \text{ and } j, \tag{2.25}$$

we can apply the Hubbard-Stratonovich decomposition 2.16 to every quadratic term seperately:

$$\begin{aligned} &e^{-\beta\hat{H}_2} \\ &\quad = e^{\beta\left(\hat{A}_1^2+\hat{A}_2^2+\ldots+\hat{A}_m^2\right)} \end{aligned}$$

$$
\begin{aligned}
&= e^{\beta\hat{A}_1^2} e^{\beta\hat{A}_2^2} \cdots e^{\beta\hat{A}_m^2} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^m} \int\int \cdots \int e^{-\frac{\left(\sigma_1^2+\ldots+\sigma_m^2\right)}{2}} e^{\sigma_1\sqrt{2\beta}\hat{A}_1} \cdots e^{\sigma_m\sqrt{2\beta}\hat{A_m}} d\sigma_1 d\sigma_2 \cdots d\sigma_m \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^m} \int\int \cdots \int e^{-\frac{\left(\sigma_1^2+\ldots+\sigma_m^2\right)}{2}} e^{\sqrt{2\beta}\left(\sigma_1\hat{A}_1+\ldots+\sigma_m\hat{A_m}\right)} d\sigma_1 d\sigma_2 \cdots d\sigma_m. \quad (2.26)
\end{aligned}
$$

Note that we made explicit use of the commutativity of the operators $\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_m$ in step 2 and 4 of 2.26. We obtain a decomposition of the form

$$
\begin{aligned}
e^{-\beta\hat{H}_2} &= \int G(\sigma) e^{\hat{A}_\sigma} d\sigma, & (2.27)\\
&\text{where} \\
\sigma &= (\sigma_1, \sigma_2, \ldots, \sigma_m), & (2.28)\\
G(\sigma) &= \frac{e^{-\frac{\left(\sigma_1^2+\ldots+\sigma_m^2\right)}{2}}}{\left(\sqrt{2\pi}\right)^m} & (2.29)\\
\hat{A}_\sigma &= \sqrt{2\beta}\left(\sigma_1\hat{A}_1 + \ldots + \sigma_m\hat{A_m}\right). & (2.30)
\end{aligned}
$$

The integral over every variable σ_i can be replaced with a three- or four-points Gaussian quadrature formula in order to obtain a sum over discrete auxiliary-field configurations. The Hubbard-interaction is a typical example of a Hamiltonian of this type (see chapter 5).

Extension of the Hubbard-Stratonovich decomposition for a sum of squares of non-commuting operators

If the two-body Hamiltonian is minus a sum of squares of non-commuting one-body operators, it is tempting to use the formula 2.26 in this case too. Now however it has an error of order β^2. We illustrate this for the case of a sum of squares of two one-body operators. Let

$$\hat{H}_2 = -\hat{A}_1^2 - \hat{A}_2^2 \quad (2.31)$$

with

$$\hat{C} = \left[\hat{A}_1, \hat{A}_2\right] \neq 0. \quad (2.32)$$

Application of formula 2.26 results in

$$
\begin{aligned}
&\frac{1}{2\pi}\int\int e^{-\frac{\left(\sigma_1^2+\sigma_2^2\right)}{2}} e^{\sqrt{2\beta}\left(\sigma_1\hat{A}_1+\sigma_2\hat{A}_2\right)} d\sigma_1 d\sigma_2 = \\
&\quad e^{\beta\left(\hat{A}_1^2+\hat{A}_2^2\right)} + \frac{\beta^2}{6}\left(\left\{\hat{A}_1, \left[\hat{A}_2, \hat{C}\right]\right\}_+ - \left\{\hat{A}_2\left[\hat{A}_1, \hat{C}\right]\right\}_+\right) - \frac{\beta^2}{2}\hat{C}^2 \\
&\quad + \text{higher order terms in } \beta. \quad (2.33)
\end{aligned}
$$

This shows clearly that 2.26 leads to an error of order β^2 if the one-body operators do not commute, even if the integration over the auxiliary fields is not replaced by a discrete sum.

The error can be reduced to order β^3 in several ways. First of all, the Suzuki formula 2.8 can be used to split the exponent:

$$\begin{aligned} e^{\beta(\hat{A}_1^2+\hat{A}_2^2)} &= e^{\frac{\beta}{2}\hat{A}_1^2}e^{\beta\hat{A}_2^2}e^{\frac{\beta}{2}\hat{A}_1^2} + \mathcal{O}(\beta^3) \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^3}\int\int\int e^{-\frac{\left(\sigma_1^2+\sigma_2^2+\sigma_3^2\right)}{2}} e^{\sigma_1\sqrt{\beta}\hat{A}_1}e^{\sigma_2\sqrt{2\beta}\hat{A}_2}e^{\sigma_3\sqrt{\beta}\hat{A}_1}d\sigma_1 d\sigma_2 d\sigma_3 \\ &\quad +\mathcal{O}(\beta^3). \end{aligned} \tag{2.34}$$

If $\hat{H}_2$ is minus a sum of squares of m non-commuting one-body operators, then we need $2m-1$ auxiliary fields and exponential factors instead of m auxiliary factors and 1 exponential factor if we want the error to be of order β^3 instead of β^2.
A similar way to reduce the error to order β^3 is obtained with

$$\begin{aligned} e^{\beta(\hat{A}_1^2+\hat{A}_2^2)} &= \frac{1}{2}e^{\beta\hat{A}_1^2}e^{\beta\hat{A}_2^2} + \frac{1}{2}e^{\beta\hat{A}_2^2}e^{\beta\hat{A}_1^2} + \mathcal{O}(\beta^3) \\ &= \frac{1}{4\pi}\int\int e^{-\frac{\left(\sigma_1^2+\sigma_2^2\right)}{2}}\left(e^{\sigma_1\sqrt{2\beta}\hat{A}_1}e^{\sigma_2\sqrt{2\beta}\hat{A}_2} + e^{\sigma_2\sqrt{2\beta}\hat{A}_2}e^{\sigma_1\sqrt{2\beta}\hat{A}_1}\right)d\sigma_1 d\sigma_2 \\ &\quad +\mathcal{O}(\beta^3). \end{aligned} \tag{2.35}$$

With this decomposition we need only m auxiliary fields and m exponential factors if we want the error to be of order β^3. But now we have to sum over different orderings for the operators $\hat{A}_1,\ldots,\hat{A}_m$. In the Monte-Carlo evaluation we will now not only have to sample the auxiliary-field configurations but also the ordering configurations for the operators $\hat{A}_1,\ldots,\hat{A}_m$ (ascending or descending order).
If the operators $\hat{A}_i$ are dense, the construction of the exponential operators will require a big computational effort. The fact that 2.34 and 2.35 lead to $2m+1$ or m exponentials, is in this case a serious disadvantage. A decomposition with an error of order β^3 which needs only one exponential can be constructed by adding terms to the Hamiltonian that lead to a cancelation of the lowest order error terms in 2.33:

$$\begin{aligned} &e^{\beta\left(\hat{A}_1^2+\hat{A}_2^2\right)} = \\ &\quad\frac{1}{\left(\sqrt{2\pi}\right)^3}\int\int\int e^{-\frac{\left(\sigma_1^2+\sigma_2^2+\sigma_3^2\right)}{2}} e^{\sqrt{2\beta}\left[\sigma_1\left(\hat{A}_1-\frac{\beta}{6}\left[\hat{A}_2,\hat{C}\right]\right)+\sigma_2\left(\hat{A}_2+\frac{\beta}{6}\left[\hat{A}_1,\hat{C}\right]\right)\right]+\sigma_3\beta\hat{C}}d\sigma_1 d\sigma_2 d\sigma_3 \\ &\quad+\mathcal{O}(\beta^3). \end{aligned} \tag{2.36}$$

The operators $\hat{A}_1$ and $\hat{A}_2$ are modified and an extra operator $\hat{C}$ and auxiliary field σ_3 are introduced. For a Hamiltonian with m squared operators $\hat{A}_1,\ldots,\hat{A}_m$, the m operators have to be modified, which poses not much problems since this can be done once for all operators before the Monte-Carlo sampling is done. But $m(m-1)/2$ additional operators $[\hat{A}_i,\hat{A}_j]$ and as much auxiliary fields have to be introduced. If the operators $\hat{A}_1,\ldots,\hat{A}_m$ form a closed algebra, i.e.

$$[\hat{A}_i,\hat{A}_j] = \sum_k f_{ijk}\hat{A}_k, \tag{2.37}$$

then these additional operators can be absorbed in the modified $\hat{A}_1,\ldots,\hat{A}_m$. In that case no additional auxiliary fields and operators are needed to reduce the error of the extended Hubbard-Stratonovich transformation to order β^3. The Hamiltonian of the pairing force in nuclear physics is a typical example of this situation (see chapter 6).

Extension of the Hubbard-Stratonovich decomposition for a general two-body Hamiltonian

In this subsection we prove that any Hermitian two-body Hamiltonian $\hat{H}_2$ can be written as a one-body operator minus a sum of squares of one-body operators. This means that the Hubbard-Stratonovich transformation can be applied to any Hermitian fermionic two-body Hamiltonian.
Let $\hat{H}_2$ be given by

$$\hat{H}_2 = \sum_{j,k,l,m} V_{jkml} \hat{a}_j^\dagger \hat{a}_k^\dagger \hat{a}_l \hat{a}_m. \tag{2.38}$$

Because $\hat{H}_2$ is Hermitian, we can impose without loss of generality

$$V_{mljk} = V_{jkml}^*. \tag{2.39}$$

Now we introduce two-dimensional indices $\lambda = (j,l)$, $\mu = (m,k)$ and the $N_\mathcal{S}^2 \times N_\mathcal{S}^2$ matrix $\bar{V}$:

$$\bar{V}_{\lambda\mu} = \bar{V}_{(j,l)(m,k)} = V_{jkml}. \tag{2.40}$$

If we interchange λ en μ we get from 2.39

$$\bar{V}_{\mu\lambda} = \bar{V}_{(m,k)(j,l)} = V_{mljk} = V_{jkml}^* = \bar{V}_{\lambda\mu}^*. \tag{2.41}$$

This means that $\bar{V}$ is a Hermitian matrix in the indices λ and μ. From linear algebra we know that a Hermitian matrix can always be brought in a form

$$\bar{V}_{\lambda\mu} = \sum_\nu \epsilon_\nu v_{\nu\lambda} v_{\nu\mu}^*, \tag{2.42}$$

with ϵ_ν a real number, e.g. by diagonalisation or by a LDL^{T} factorization [11]. Using this form for $\bar{V}$ we get

$$\begin{aligned}
\hat{H}_2 &= \sum_{j,k,l,m} \sum_\nu \epsilon_\nu v_{\nu(j,l)} v_{\nu(m,k)}^* \hat{a}_j^\dagger \hat{a}_k^\dagger \hat{a}_l \hat{a}_m \\
&= -\sum_\nu \epsilon_\nu \left(\sum_{j,l} v_{\nu(j,l)} \hat{a}_j^\dagger \hat{a}_l \right) \left(\sum_{k,m} v_{\nu(m,k)}^* \hat{a}_k^\dagger \hat{a}_m \right) + \sum_\nu \epsilon_\nu \sum_{j,k,l,m} v_{\nu(j,l)} v_{\nu(m,k)}^* \delta_{k,l} \hat{a}_j^\dagger \hat{a}_m \\
&= -\sum_\nu \frac{\epsilon_\nu}{2} \left\{ \hat{A}_\nu, \hat{A}_\nu^\dagger \right\}_+ + \hat{B}_1
\end{aligned} \tag{2.43}$$

with

$$\hat{A}_\nu = \sum_{j,l} v_{\nu(j,l)} \hat{a}_j^\dagger \hat{a}_l \tag{2.44}$$

and the one-body part

$$\hat{B}_1 = \sum_{j,l} \frac{\sum_k (V_{jklk} + V_{kjkl})}{2} \hat{a}_j^\dagger \hat{a}_l. \tag{2.45}$$

Each term $\frac{\epsilon_\nu}{2} \left\{ \hat{A}_\nu, \hat{A}_\nu^\dagger \right\}_+$ can be written as a sum of squares of operators. If $\epsilon_\nu > 0$ as a sum of squares of two Hermitian operators:

$$\frac{\epsilon_\nu}{2} \left\{ \hat{A}_\nu, \hat{A}_\nu^\dagger \right\}_+ = \left[\sqrt{\frac{|\epsilon_\nu|}{2}} \left(\hat{A}_\nu + \hat{A}_\nu^\dagger \right) \right]^2 + \left[i \sqrt{\frac{|\epsilon_\nu|}{2}} \left(\hat{A}_\nu - \hat{A}_\nu^\dagger \right) \right]^2, \tag{2.46}$$

and if $\epsilon_\nu < 0$ as a sum of squares of two anti-Hermitian operators:

$$\frac{\epsilon_\nu}{2}\left\{\hat{A}_\nu, \hat{A}_\nu^\dagger\right\}_+ = \left[i\sqrt{\frac{|\epsilon_\nu|}{2}}\left(\hat{A}_\nu + \hat{A}_\nu^\dagger\right)\right]^2 + \left[\sqrt{\frac{|\epsilon_\nu|}{2}}\left(\hat{A}_\nu - \hat{A}_\nu^\dagger\right)\right]^2. \tag{2.47}$$

We can make all ϵ_ν positive (negative) by adding a positive (negative) constant c times the unity matrix to the matrix $\bar{V}$. This corresponds to adding an operator

$$-c\sum_{\lambda=(j,l),\mu=(m,k)}\delta_{\lambda,\mu}\hat{a}_j^\dagger\hat{a}_k^\dagger\hat{a}_l\hat{a}_m = -c\sum_{j,k}\hat{a}_j^\dagger\hat{a}_k^\dagger\hat{a}_k\hat{a}_j = -c\left(\hat{N}^2 - \hat{N}\right) \tag{2.48}$$

to the Hamiltonian, with $\hat{N} \equiv \sum_j \hat{a}_j^\dagger\hat{a}_j$ the number operator. It commutes with $\hat{H}_2$ and $\hat{H}_1$. In the canonical picture it is a scalar factor, so it can be handeled easily in SDQMC. There is still more freedom to decompose $\hat{H}_2$. Because $\hat{a}_j^\dagger\hat{a}_k^\dagger\hat{a}_l\hat{a}_m = \hat{a}_k^\dagger\hat{a}_j^\dagger\hat{a}_m\hat{a}_l$, we can alter V according to

$$\begin{aligned} V_{jkml} &\rightarrow V_{jkml} + c \\ V_{kjlm} &\rightarrow V_{kjlm} - c, \end{aligned} \tag{2.49}$$

with c a complex number, without changing its physical content. The matrixelements V_{jjml}, V_{jkll} or V_{jjll} have no physical meaning for fermions, so they can be given any value. This will lead to a different decomposition of $\hat{H}_2$.

2.2.2 Decompositions based on rank one and rank two operators

Rank one operators

We use the term 'rank one operator' for a one-body operator of the form $\hat{b}_1^\dagger\hat{b}_2$ with

$$\hat{b}_1 = \sum_j b_{1j}\hat{a}_j,\ \hat{b}_2 = \sum_j b_{2j}\hat{a}_j. \tag{2.50}$$

An operator of this form can be handled easily in the matrix representation for Slater determinants introduced in section 1.2.3 because it can be represented by an operation with a matrix of rank one:

$$1 + x\hat{b}_1^\dagger\hat{b}_2 \longleftrightarrow 1 + x[\hat{b}_1^\dagger\hat{b}_2] = 1 + xb_1^\dagger b_2, \tag{2.51}$$

where b_1 and b_2 are considered as row vectors. Thus matrices of rank one are related to one-body operators of rank one. Since our aim is to construct representations for exponentials of operators, it is interesting to notice that

$$e^{-\beta\hat{b}_1^\dagger\hat{b}_2} = 1 + \left(e^{-\beta\{\hat{b}_1^\dagger,\hat{b}_2\}_+} - 1\right)\hat{b}_1^\dagger\hat{b}_2 = 1 + \left(e^{-\beta b_2 b_1^\dagger} - 1\right)\hat{b}_1^\dagger\hat{b}_2. \tag{2.52}$$

The exponential of a rank one operator can be represented by a rank one matrix operation.

Rank two operators

One can extend the expressions of the previous section to higher order operators. We use the term 'rank two operator' for a two-body operator of the form $\hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4$. The reason for this is that such an operator can be represented by the identity matrix plus a matrix of rank two, apart from a contribution of rank one operators.

Lemma 2.1 *The operation represented by the unity matrix plus a matrix of rank two can be expressed as a combination of rank one and rank two operators in the following way:*

$$1 + x b_1^\dagger b_4 + y b_2^\dagger b_3 \longleftrightarrow 1 + x \hat{b}_1^\dagger \hat{b}_4 + y \hat{b}_2^\dagger \hat{b}_3 + xy \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4. \tag{2.53}$$

Proof:

Consider the N-particle Slater determinant Ψ_M represented by the matrix M. Consider also a Slater determinant Ψ_B from the basis set $\mathcal{D}_0$, with particles in the single-particle states $\varphi_{i_1}, \varphi_{i_2}, \ldots, \varphi_{i_N}$. The overlap of Ψ_M and Ψ_B is given by

$$\langle \Psi_B | \Psi_M \rangle = \begin{vmatrix} M_{i_1 1} & M_{i_1 2} & \cdots & M_{i_1 N} \\ M_{i_2 1} & M_{i_2 2} & \cdots & M_{i_2 N} \\ \vdots & \vdots & & \vdots \\ M_{i_N 1} & M_{i_N 2} & \cdots & M_{i_N N} \end{vmatrix}. \tag{2.54}$$

The operator 2.53 transforms Ψ_M into $\Psi_{M'}$ with $M' = (1 + x b_1^\dagger b_4 + y b_2^\dagger b_3) M$. To calculate the overlap of $\Psi_{M'}$ with Ψ_B, we have to replace every column c_j in 2.54:

$$\begin{aligned} c_j = \begin{pmatrix} M_{i_1 1} \\ M_{i_2 1} \\ \vdots \\ M_{i_N 1} \end{pmatrix} \rightarrow c'_j &= \begin{pmatrix} M'_{i_1 j} \\ M'_{i_2 j} \\ \vdots \\ M'_{i_N j} \end{pmatrix} \\ &= \begin{pmatrix} M_{i_1 j} + x b^*_{i_1 1}(b_4 M_{\cdot j}) + y b^*_{i_1 2}(b_3 M_{\cdot j}) \\ M_{i_2 j} + x b^*_{i_2 1}(b_4 M_{\cdot j}) + y b^*_{i_2 2}(b_3 M_{\cdot j}) \\ \vdots \\ M_{i_N j} + x b^*_{i_N 1}(b_4 M_{\cdot j}) + y b^*_{i_N 2}(b_3 M_{\cdot j}) \end{pmatrix} \\ &= c_j + x_j \tilde{b}_1^\dagger + y_j \tilde{b}_2^\dagger, \end{aligned} \tag{2.55}$$

with

$$x_j = x \sum_{k=1}^{N_S} b_{k4} M_{kj} \tag{2.56}$$

$$y_j = y \sum_{k=1}^{N_S} b_{k3} M_{kj}. \tag{2.57}$$

The notation $M_{.j}$ denotes the vector that is given by the j^{th} column of M, $\tilde{b}$ denotes the $N \times 1$ row matrix $(b_{i_1} b_{i_2} \cdots b_{i_N})$. The overlap is then given by

$$\langle \Psi_B | \Psi_{M'} \rangle = \begin{vmatrix} c_1 + x_1 \tilde{b}_1^\dagger + y_1 \tilde{b}_2^\dagger & c_2 + x_2 \tilde{b}_1^\dagger + y_2 \tilde{b}_2^\dagger & \cdots & c_N + x_N \tilde{b}_1^\dagger + y_N \tilde{b}_2^\dagger \end{vmatrix}. \tag{2.58}$$

This determinant can be expanded as the sum of all determinants that are obtained by selecting in every column of 2.58 one of the terms c_j, $x_j \tilde{b}_1^\dagger$ or $y_j \tilde{b}_2^\dagger$. If in more than one column the term $x_j \tilde{b}_1^\dagger$ is selected, then the determinant has two linearly dependent columns, so it will vanish. The same for the term $y_j \tilde{b}_2^\dagger$. Only four types of determinants remain:

- c is selected in every column. This determinant is just $\langle \Psi_B | \Psi_M \rangle$, 2.54.
- $x_j \tilde{b}_1^\dagger$ is selected in column j, c in all others. These determinants sum up to $\langle \Psi_B | x \hat{b}_1^\dagger \hat{b}_4 | \Psi_M \rangle$ (one particle is moved from state b_4 to state b_1).
- $y_j \tilde{b}_2^\dagger$ is selected in column j, c in all others. These determinants sum up to $\langle \Psi_B | y \hat{b}_2^\dagger \hat{b}_3 | \Psi_M \rangle$ (one particle is moved from state b_3 to state b_2).
- $x_j \tilde{b}_1^\dagger$ is selected in column j, $y_k \tilde{b}_2^\dagger$ is selected in column k, c in all others. These determinants sum up to $\langle \Psi_B | xy \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4 | \Psi_M \rangle$ (particles are moved from states b_4 and b_3 to states b_1 and b_2).

Taking all these terms together, we find that

$$\langle \Psi_B | \Psi_{M'} \rangle = \langle \Psi_B | 1 + x \hat{b}_1^\dagger \hat{b}_4 + y \hat{b}_2^\dagger \hat{b}_3 + xy \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4 | \Psi_M \rangle. \tag{2.59}$$

This holds for any basis state Ψ_B, so that

$$\Psi_{M'} = \left(1 + x \hat{b}_1^\dagger \hat{b}_4 + y \hat{b}_2^\dagger \hat{b}_3 + xy \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4 \right) \Psi_M. \tag{2.60}$$

This proves 2.53.

End of proof.

The relation 2.53 can now be used in several ways to obtain a useful decomposition for the exponential of a general two-body Hamiltonian. A first approach is based on exponentials of rank two operators.

Decomposition for a single rank two operator

In order to make a connection with the exponential of a rank two operator, we look at the square of such an operator. Let the operator $\hat{P}$ be given by

$$\hat{P} = \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4. \tag{2.61}$$

Then the square of this operator is given by

$$\begin{aligned} \hat{P}^2 &= \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4 \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4 \\ &= \left(\left\{ \hat{b}_1^\dagger, \hat{b}_4 \right\}_+ \left\{ \hat{b}_2^\dagger, \hat{b}_3 \right\}_+ - \left\{ \hat{b}_2^\dagger, \hat{b}_4 \right\}_+ \left\{ \hat{b}_1^\dagger, \hat{b}_3 \right\}_+ \right) \hat{b}_1^\dagger \hat{b}_2^\dagger \hat{b}_3 \hat{b}_4 \\ &= \gamma \hat{P}, \end{aligned} \tag{2.62}$$

with γ given by

$$\gamma = (b_4 b_1^\dagger)(b_3 b_2^\dagger) - (b_4 b_2^\dagger)(b_3 b_1^\dagger). \tag{2.63}$$

By repeated application of this expression we obtain that

$$\hat{P}^n = \gamma^{n-1}\hat{P}. \tag{2.64}$$

If we apply this to every term in the series expansion of $e^{-\beta\hat{P}}$, we find

$$e^{-\beta\hat{P}} = 1 + \left[e^{-\beta\gamma} - 1\right]\hat{P}. \tag{2.65}$$

Now we can use the relation 2.53 to obtain a matrix-representation decomposition for $e^{-\beta\hat{P}}$:

$$e^{-\beta\hat{P}} = \frac{\left[1 + x b_1^\dagger b_4 + y b_2^\dagger b_3\right] + \left[1 - x b_1^\dagger b_4 - y b_2^\dagger b_3\right]}{2}, \tag{2.66}$$

if x and y are chosen such that

$$xy = e^{-\beta\gamma} - 1. \tag{2.67}$$

Decomposition based on rank two operators for a general two-body Hamiltonian: first possibility

Any two-body Hamiltonian $\hat{H}_2 = \sum_{jklm} V_{jkml}\hat{a}_j^\dagger\hat{a}_k^\dagger\hat{a}_l\hat{a}_m$ can be written as a sum of rank two operators $\hat{P}_1, \hat{P}_2, \ldots \hat{P}_M$. A trivial example, with $N_{\mathcal{S}}^4$ terms, is given by

$$\hat{P}_{j,k,l,m} = \hat{a}_j^\dagger\hat{a}_k^\dagger\hat{a}_l\hat{a}_m. \tag{2.68}$$

Expansions with less terms will often be more useful. For any Hamiltonian, decompositions with at most $N_{\mathcal{S}}^3/2$ terms can be constructed. We can factorize $e^{-\beta\hat{H}_2}$ as a product of exponentials of rank two operators, just like we did in section 2.2.1 for a Hamiltonian that is a sum of squares of one-body operators, e.g. for non-commuting $\hat{P}_{1\ldots M}$:

$$e^{-\beta\hat{H}_2} \simeq e^{-\frac{\beta}{2}\hat{P}_1} e^{-\frac{\beta}{2}\hat{P}_2} \cdots e^{-\frac{\beta}{2}\hat{P}_{M-1}} e^{-\beta\hat{P}_M} e^{-\frac{\beta}{2}\hat{P}_{M-1}} \cdots e^{-\frac{\beta}{2}\hat{P}_2} e^{-\frac{\beta}{2}\hat{P}_1}. \tag{2.69}$$

For each of these factors we can use the two-term decomposition 2.66. This leads to a decomposition for $e^{-\beta\hat{H}_2}$ with 2^{2M-1} terms of $2M-1$ factors.

The discrete Hubbard Stratonovich transformation of Hirsch

A specific example of this type of decompositions is the 'discrete Hubbard Stratonovich transformation' introduced by Hirsch for the two-body Hamiltonian of the Hubbard model [14]. The single-particle basis states φ_{is} have a spatial part that is a point on a lattice (index i) and a spin part (index s, $\uparrow$ or $\downarrow$). The two-body Hamiltonian has the form

$$\hat{H}_2 = U\sum_i \hat{n}_{i\uparrow}\hat{n}_{i\downarrow}, \tag{2.70}$$

with $\hat{n}_{is}$ the occupation number operator for the state φ_{is}:

$$\hat{n}_{is} = \hat{a}_{is}^\dagger\hat{a}_{is}. \tag{2.71}$$

The Hirsch decomposition is given by [13, 14]

$$e^{-\beta U \hat{n}_\uparrow \hat{n}_\downarrow} = \frac{1}{2} \sum_{\sigma=\pm 1} e^{2a\sigma(\hat{n}_\uparrow - \hat{n}_\downarrow) - \frac{U\beta}{2}(\hat{n}_\uparrow + \hat{n}_\downarrow)}, \tag{2.72}$$

$$\tanh^2(a) = \tanh \frac{\beta U}{4}, \tag{2.73}$$

for $U > 0$ and

$$e^{-\beta U \hat{n}_\uparrow \hat{n}_\downarrow} = \frac{1}{2} \sum_{\sigma=\pm 1} e^{2a\sigma(\hat{n}_\uparrow + \hat{n}_\downarrow - 1) - \frac{U\beta}{2}(\hat{n}_\uparrow + \hat{n}_\downarrow - 1)}, \tag{2.74}$$

$$\tanh^2(a) = -\tanh \frac{\beta U}{4}, \tag{2.75}$$

for $U < 0$. Though the name 'discrete Hubbard Stratonovich transformation' might suggest that this decomposition is a discretized approximation to the Hubbard Stratonovich transformation 2.16, it is actually an exact decomposition of the type 2.66. The values for x and y that correspond to this decomposition can be calculated by applying expression 2.52 to the right hand sides of 2.72 or 2.74. The advantage of this parametrization compared to the ones with other values for x and y, is that the resulting matrix representation U_σ (not to be confused with the interaction strenght parameter U) in 2.14 splits up in a spin-up and a spin-down part that are related to one another:

$$U_\sigma = \begin{pmatrix} U_{\uparrow\sigma} & 0 \\ 0 & U_{\downarrow\sigma} \end{pmatrix}. \tag{2.76}$$

In the case of negative U we have that $U_{\downarrow\sigma} = U_{\uparrow\sigma}$ and in the case of positive U we have that $U_{\downarrow\sigma} = \left(U_{\uparrow\sigma}^{-1}\right)^{-1}$. This makes that only half of the matrix has to be computed in order to calculate $\hat{\mathrm{Tr}}\left(\hat{U}_\sigma\right)$, which saves a lot of computing time. Furthermore, it guarantees a positive sign for systems with negative U and an equal number of particles in spin-up and spin-down states and for half-filled systems with positive U.

Decomposition based on rank two operators for a general two-body Hamiltonian: second possibility

The interesting point in the decomposition of the previous section is that the exponential of a rank two-operator is again a rank two operator, plus the identity operator. This property not only holds for rank two operators, also for any two-body operator $\hat{P}$ of the form

$$\hat{P} = \sum_{jk} V_{jk} \hat{a}_j^\dagger \hat{a}_k^\dagger \hat{b}_3 \hat{b}_4. \tag{2.77}$$

The square of such an operator is given by

$$\begin{aligned} \hat{P}^2 &= \sum_{jklm} V_{jk} \hat{a}_j^\dagger \hat{a}_k^\dagger \hat{b}_3 \hat{b}_4 V_{lm} \hat{a}_l^\dagger \hat{a}_m^\dagger \hat{b}_3 \hat{b}_4 \\ &= \sum_{jklm} V_{jk} V_{lm} \left(\left\{ \hat{a}_l^\dagger \hat{b}_4 \right\}_+ \left\{ \hat{a}_m^\dagger \hat{b}_3 \right\}_+ - \left\{ \hat{a}_m^\dagger \hat{b}_4 \right\}_+ \left\{ \hat{a}_l^\dagger \hat{b}_3 \right\}_+ \right) \hat{a}_j^\dagger \hat{a}_k^\dagger \hat{b}_3 \hat{b}_4 \end{aligned}$$

$$= \sum_{jklm} V_{jk} V_{lm} \left(b_{4l} b_{3m} - b_{4m} b_{3l} \right) \hat{a}_j^\dagger \hat{a}_k^\dagger \hat{b}_3 \hat{b}_4$$
$$= \gamma \hat{P}, \tag{2.78}$$

with γ given by

$$\gamma = \sum_{lm} V_{lm} \left(b_{4l} b_{3m} - b_{4m} b_{3l} \right) = b_4^T V b_3 - b_3^T V b_4. \tag{2.79}$$

Again we find that

$$e^{-\beta \hat{P}} = 1 + \left[e^{-\beta\gamma} - 1 \right] \hat{P}. \tag{2.80}$$

The right hand side is a sum of rank-two operators. We can decompose it as a sum of operators of type 2.53:

$$e^{-\beta \hat{P}} = \frac{\sum_{jk} G(\sigma) \left[1 + x_\sigma b_{1\sigma}^\dagger b_4 + y_\sigma b_{2\sigma}^\dagger b_3 \right]}{\sum_{\sigma'} G(\sigma')}, \tag{2.81}$$

provided that $G(\sigma)$, x_σ ,y_σ and $b_{1\sigma}, b_{2\sigma}$ fulfil the following requirements:

$$\frac{\sum_\sigma G(\sigma) x_\sigma \hat{b}_{1\sigma}^\dagger \hat{b}_4}{\sum_{\sigma'} G(\sigma')} = 0 \tag{2.82}$$
$$\frac{\sum_\sigma G(\sigma) y_\sigma \hat{b}_{2\sigma}^\dagger \hat{b}_3}{\sum_{\sigma'} G(\sigma')} = 0 \tag{2.83}$$
$$\frac{\sum_\sigma G(\sigma) x_\sigma y_\sigma \hat{b}_{1\sigma}^\dagger \hat{b}_{2\sigma}^\dagger}{\sum_{\sigma'} G(\sigma')} = \left[e^{-\beta\gamma} - 1 \right] \sum_{jk} V_{jk} \hat{a}_j^\dagger \hat{a}_k^\dagger. \tag{2.84}$$

An obvious decomposition of this type is obtained by taking all $G(\sigma) = 1$ and

$$e^{-\beta \hat{P}} = \sum_{jk} \frac{\left[1 + x_{jk} a_j^\dagger b_4 + y_{jk} a_k^\dagger b_3 \right] + \left[1 - x_{jk} a_j^\dagger b_4 - y_{jk} a_k^\dagger b_3 \right]}{2 N_{\mathcal{S}}^2}, \tag{2.85}$$

with

$$x_{jk} = N_{\mathcal{S}} \sqrt{\left| \left(e^{-\beta\gamma} - 1 \right) V_{jk} \right|} \tag{2.86}$$
$$y_{jk} = s N_{\mathcal{S}} \sqrt{\left| \left(e^{-\beta\gamma} - 1 \right) V_{jk} \right|}, \tag{2.87}$$

with $s = +1$ or -1 depending on the sign of $\left(e^{-\beta\gamma} - 1 \right) V_{jk}$. Depending on the specific form of the interaction, other choices of $G(\sigma)$, x_σ ,y_σ and $b_{1\sigma}, b_{2\sigma}$ might be better suited. E.g. it is recommendable to include as much as possible the variation of the strength V_{jk} in the factor w_σ. This will lead to a more effective and smoother Monte-Carlo sampling, with more weight attributed to the important parts of the interaction.
A general two-body Hamiltonian can always be written as a sum of operators of the form 2.77. At most $N_{\mathcal{S}}^2/2$ terms are needed. For schematic interactions less terms will be needed. E.g. the pairing Hamiltonian requires only $N_{\mathcal{S}}/2$ terms. Again, we can factorize $e^{-\beta \hat{H}_2}$ as a product of exponentials of operators ot the form 2.77, just like we did in section 2.2.1 or 2.2.2.

Decomposition based on rank two operators for a general two-body Hamiltonian: third possibility

The expression 2.53 can be used to construct a representation for the operator $1 + \epsilon \hat{H}_2$, with H_2 a general two-body Hamiltonian. Suppose that we have a decomposition for $\hat{H}_2$ as a sum of rank two operators:

$$\hat{H}_2 = \sum_{\lambda} V_{\lambda} \hat{b}^{\dagger}_{1\lambda} \hat{b}^{\dagger}_{2\lambda} \hat{b}_{3\lambda} \hat{b}_{4\lambda}. \tag{2.88}$$

Then $1 + \epsilon \hat{H}_2$ can be represented as

$$1 + \epsilon \hat{H}_2 = \frac{\sum_{\sigma} G(\sigma) \left[1 + x_{\sigma} b^{\dagger}_{1\sigma} b_{4\sigma} + y\, b^{\dagger}_{2\sigma} b_{3\sigma} \right]}{\sum_{\sigma'} G(\sigma')}, \tag{2.89}$$

provided that $G(\sigma)$, x_{σ} ,y_{σ} and the $b_{i\sigma}$ fulfil the following requirements:

$$\frac{\sum_{\sigma} G(\sigma) x_{\sigma} \hat{b}^{\dagger}_{1\sigma} \hat{b}_{4\sigma}}{\sum_{\sigma'} G(\sigma')} = 0 \tag{2.90}$$

$$\frac{\sum_{\sigma} G(\sigma) y_{\sigma} \hat{b}^{\dagger}_{2\sigma} \hat{b}_{3\sigma}}{\sum_{\sigma'} G(\sigma')} = 0 \tag{2.91}$$

$$\frac{\sum_{\sigma} G(\sigma) x_{\sigma} y_{\sigma} \hat{b}^{\dagger}_{1\sigma} \hat{b}^{\dagger}_{2\sigma} \hat{b}_{3\sigma} \hat{b}_{4\sigma}}{\sum_{\sigma'} G(\sigma')} = \hat{H}_2. \tag{2.92}$$

Furthermore all coefficients $G(\sigma)$ have to be positive real numbers in order for Monte-Carlo methods to be applicable. If $\hat{H}_2$ is given by

$$\hat{H}_2 = \sum_{jklm} V_{jkml} \hat{a}^{\dagger}_j \hat{a}^{\dagger}_k \hat{a}_l \hat{a}_m \tag{2.93}$$

then an obvious decomposition of this type is obtained by taking all $G(\sigma) = 1$ and

$$1 + \epsilon \hat{H}_2 = \sum_{jklm} \frac{\left[1 + x_{jkml} a^{\dagger}_j a_m + y_{jkml} a^{\dagger}_k a_l \right] + \left[1 - x_{jkml} a^{\dagger}_j a_m - y_{jkml} a^{\dagger}_k a_l \right]}{2 N^4_{\mathcal{S}}}, \tag{2.94}$$

with

$$x_{jkml} = N^2_{\mathcal{S}} \sqrt{|\epsilon V_{jkml}|} \tag{2.95}$$

$$y_{jkml} = s N^2_{\mathcal{S}} \sqrt{|\epsilon V_{jkml}|}, \tag{2.96}$$

with $s = +1$ or -1 depending on the sign of ϵV_{jkml}. However, this leads to large values for x or y because of the factor $N^2_{\mathcal{S}}$ in the right hand side. It is better to use decompositions with less terms. Furthermore, it is here also recommendable to include as much as possible the variation of the strength in the factor $G(\sigma)$.

Now that we dispose of an exact representation for operators of the form $1 + \epsilon\hat{H}_2$, we can approximate $e^{-\beta\hat{H}_2}$ with combinations of operators of the form $1+\epsilon\hat{H}_2$. An approximation that is correct up to second order is given by

$$e^{-\beta\hat{H}_2} \simeq \frac{1 + \left(1 - \beta\hat{H}_2\right)^2}{2}. \tag{2.97}$$

The leading error term is given by

$$\frac{\beta^3}{6}\hat{H}_2^3. \tag{2.98}$$

Though this error is of the order β^3, it is in most cases quite large compared to the error originating from the Suzuki-decomposition 2.9, since the latter is a combination of commutators of operators, while the former is a mere product of operators. We experienced a very slow convergence when increasing the number of inverse-temperature intervals, N_t, when we applied this decomposition to the repulsive Hubbard model.
Because such decompositions have errors proportional to powers of $\beta\hat{H}_2$, they can be expected to be more effective than the decompositions of the second type only if the two-body part of the Hamiltonian is small compared to the one-body part.
Better results were obtained with a fourth order decomposition:

$$e^{-\beta\hat{H}_2} \simeq \frac{19 + 50\left(1 - \frac{3}{5}\beta\hat{H}_2\right)^2 + 3\left(1 - \beta\hat{H}_2\right)^4}{72}. \tag{2.99}$$

Now the leading error term is of the form

$$\frac{\beta^5}{120}\hat{H}_2^5. \tag{2.100}$$

2.2.3 Comparing the decompositions

We discussed a variety of ways in which the operator $e^{-\beta\hat{H}_2}$ can be decomposed for SDQMC calculations. These decompositions still have many degrees of freedom. Furthermore, one can combine them and obtain hybrid decompositions. E.g. the pairing plus quadrupole Hamiltonian can be split in two parts, pairing and quadrupole. The quadrupole part lends itself naturally for a Hubbard-Stratonovich decomposition, while the pairing part is more easily handled with a decomposition based on rank two operators of the form 2.81. A comparison between all these types of decompositions would be interesting. Because of the limited possibilities of our computer systems, we could not make a comparison, at present, for general interactions. We had to restrict ourselves to schematic interactions: the Hubbard model, the Pairing Hamiltonian, ... Details about these calculations are given in the second part of this work. Concerning the different decompositions, we can make a few remarks:
For the Hubbard model it is known that Hirch's discrete Hubbard-Stratonovich transform of section 2.2.2 performs better than the continuous Hubbard-Stratonovich transform of section 2.2.1 [7]. Since it is an exact decomposition for the two-body interaction, it is clear that it should outperform the discretized Hubbard-Stratonovich transform of section 2.2.1

too. We did calculations for the Hubbard model with decompositions of the form 2.97 and 2.99. Because of their rank-one structure these decompositions lead to much faster matrix multiplications than the Hirsch decomposition. They also had less configurations per inverse-temperature slice to sample over. However, the second-order decomposition 2.97 required 1000 or more inverse-temperature slices to reduce the systematic errors. Therefore this decomposition performed much worse than the Hirsch-decomposition. The fourth-order expansion 2.99 needs less slices to converge (N_t of the order of 100), but it leads to low acceptance rates for the Metropolis sampling algorithm (see section 3.4.1). This can be explained by the fact that configurations that differ only in a few inverse-temperature slices can yield completely different values for the traces. The distribution of terms in the decomposition was not smooth enough to be sampled efficiently by Markov-chain Monte-Carlo methods. Away from half filling, the repulsive Hubbard model leads to sign problems. These were found to be more severe in the case of decomposition 2.99 than for the Hirsch decomposition. This shows that the sign-problem is closely related to the form of the decomposition. So maybe the degrees of freedom, that we still have when constructing decompositions, allow for a decomposition with less sign problems.
For the pairing decomposition we tried several decompositions of the type 2.99. They all suffered from a very slow thermalization in the Markov-chain Monte-Carlo sampling. Decompositions of type 2.81 performed very well. No comparison with Hubbard-Stratonovich calculations was made. We expect them to be less performant, because they require exponentiation and multiplication of dense matrices, and because they are based on coarser approximations to $e^{-\beta \hat{H}_2}$.
From this we are tempted to believe that decompositions of type 2.81 are the most interesting ones for general interactions too: they are based on matrix operations of rank one and two, that can be executed much faster than full-rank matrix operations. In addition they are the most exact in decomposing the operator $e^{-\beta \hat{H}_2}$ for a general two-body interaction. This is a conjecture. Evidence hopefully follows in the near future.

3

Markov-chain Monte-Carlo methods

We will limit the discussion of Markov-chain Monte-Carlo methods (MCMC) to finite systems (dimension is N). Because any numerical simulation is finite and discrete, this is general enough to cover the practical applications of MCMC, especially for SDQMC. This limitation will allow us to use results from linear algebra in order to demonstrate the convergence of MCMC.

3.1 The Monte-Carlo trick

Our aim is to compute a ratio of two sums

$$\mathrm{E}(f) = \frac{\sum_x f(x)w(x)}{\sum_{x'} w(x')}, \tag{3.1}$$

where the number of states x is very large, too large to make a complete summation feasible. Nothing is said about the sign of the $w(x)$. We will assume that all $w(x)$ are positive. The case where a fraction of the $w(x)$ is negative, will be discussed in the next chapter in the framework of SDQMC methods, when the (in)famous 'sign problem' is discussed. By normalizing the $w(x)$ to a probability distribution

$$\pi(x) = \frac{w(x)}{\sum_{x'} w(x')}, \tag{3.2}$$

$\mathrm{E}(f)$ can be interpreted as a weighted average:

$$\mathrm{E}(f) = \sum_x f(x)\pi(x). \tag{3.3}$$

Then π is a probabilty distribution so that

$$\pi(x) > 0 \ \forall x \tag{3.4}$$

$$\sum_x \pi(x) = 1. \tag{3.5}$$

We require a strict inequality in 3.4 in order to avoid problems with division by zero, but the results of this chapter can easily be extended to probability distributions that vanish for certain states x. The 'trick' of Monte-Carlo methods exists in approximating the complete sum in 3.3 with a sum over a limited sample $x^{[1]}, x^{[2]}, \ldots x^{[M]}$ drawn according to the probabilty distribution π. The central limit theorem [13] assures that for large enough M the sample average

$$\mathrm{E}_S(f) = \frac{1}{M} \sum_{j=1}^{M} f(x^{[j]}), \tag{3.6}$$

tends to the total average $\mathrm{E}(f)$. In the limit of large M, $\mathrm{E}_S(f)$ is normally distributed around $\mathrm{E}(f)$ with a standard deviation

$$\sigma = \sqrt{\frac{\mathrm{E}(f^2) - \mathrm{E}(f)^2}{M}}. \tag{3.7}$$

The main problem of course, is to draw a sample $x^{[1]}, x^{[2]}, \ldots x^{[M]}$ distributed according to π.

3.1.1 Independent sampling techniques

In order for the central limit theorem to apply, all $x^{[i]}$ of the sample have to be independent. If we have a method to draw randomly one x according to the probability distribution $\pi(x)$, then we can generate a sample by repeating this method M times independently. For a number of probability distributions one can easily generate samples.

The uniform distribution

The simplest one is the uniform distribution: all x are equally probable. If we number the states x from 1 to N, then all we have to do is to generate M random numbers between 1 and N. A lot of 'random-number-generator' routines have been developed. They generate 'pseudo-random' sequences of numbers: each number is obtained by applying a definite transformation on the previous numbers, but the numbers look as if they are completely uncorrelated. Most programming-language compilers have a built-in random-number generator. Care has to be taken with these, because often they work well for short sequences but are too simple to generate large uncorrelated samples that are needed for Monte-Carlo simulations [12, 13]. For our calculations we used the random-number generating fortran routine 'ran1' described in [12].

The transformation method

Some probability distributions with a simple analytical form can be generated by transforming a uniform distribution: Suppose we want to sample a distribution $\pi(y)$, with the variable y a real number. Let the function Π be defined by

$$\Pi(y) = \int_{-\infty}^{y} \pi(t)dt. \tag{3.8}$$

If the inverse function of Π can be calculated, then we can transform a variable x uniformily distributed between 0 and 1 into a variable $y = \Pi^{-1}(x)$ distributed according to $\pi(y)$. This method is strongly limited by the fact that one needs to know Π^{-1}. Often one knows only the unnormalized weight $w(x)$ instead of the normalized probability $\pi(x)$. In order to sample according to $w(x)$ one has to know the normalization constant $\sum_x w(x)$, which involves the computation of all terms, not just a limited sample. Often the state space is too large to calculate $w(x)$ for all states x, because this is one of the main motivations for the use of Monte-Carlo methods.

Von Neumann rejection

If one can draw a random variable from a distribution $\pi_0(x)$, then one can sample some other distribution $\pi(x)$ on the same set of states x with a method known as 'Von Neumann rejection': Let c be a constant such that

$$\pi(x) \leq c\pi_0(x) \quad \forall x. \tag{3.9}$$

- *step 1:* Draw x according to $\pi_0(x)$.
- *step 2:* Draw y according to a uniform distribution between 0 and 1.
- *step 3:* If $y > \frac{\pi(x)}{c\pi_0(x)}$ then go back to step 1. Otherwise, x is the result.

The final x will be distributed according to $\pi(x)$. Note that the probability that x is accepted in step 3 is equal to $\frac{\sum_x \pi(x)}{c}$ (averaged over all x). Therefore Von Neumann rejection will only be efficient if c is not too big. This is only possible if there exists a 'comparison function' π_0 that can easily be sampled and that differs not too much from π. The method can also be used when only the unnormalized weight $w(x)$ is known: just replace $\pi(x)$ with $w(x)$ in 3.9 and in step 3. In these cases it is often difficult to determine the constant c such that condion 3.9 is met.
The method is visualized in figure 3.1.
For Monte-Carlo simulations the rejection method is often unusable. In order to ensure that 3.9 holds one might have to calculate all $\pi(x)$. Or one has to take c large enough so that one can be sure of 3.9 without checking it for all x. Most often this leads to a very high rejection rate, which makes the method useless.

3.2 Markov-Chain Monte-Carlo sampling

What to do with complicated, unnormalized probability distributions ? Statistical methods that permit to sample these distributions do exists : 'Markov-chain Monte-Carlo methods' (MCMC). The sacrifice one has to make is that the x's in a sample are no longer independent. It has been shown that the results obtained with these samples do converge to the exact results if one takes the samples large enough.
Markov-Chain Monte-Carlo methods like the 'Gibbs sampler' or the 'Metropolis random walk' are widely used nowadays in all kinds of fields: statistical physics, statistics, econometrics, biostatistics, ... Though their use is very widespread and their basic convergence

Figure 3.1: *Illustration of Von Neumann rejection: Points are generated uniformely in the region below $c\pi_0(x)$. Points that fall in the shaded region, below $\pi(x)$, are accepted. The other ones are rejected.*

properties are known for many years [18], the issue of their convergence is adressed only superficially in physics literature. E.g. in [13] it is shown that

$$\|\pi^{[j+1]} - \pi\| < \|\pi^{[j]} - \pi\|, \tag{3.10}$$

which means that $\pi^{[j]}$ comes closer to the target distribution π with every Markov step. But this does not prove the convergence of $\pi^{[j]}$ to π:

$$\|\pi^{[j]} - \pi\| \longrightarrow 0 \qquad \text{for large j}. \tag{3.11}$$

A lot arbitrariness also still exists in the way error limits are determined. Sampling the Markov chain every m steps, where m is determined so that the sampled values are nearly independent, as is suggested in [6, 13], leads to suboptimal sampling and underestimated error limits. Though the deviations are not dramatic, little computational effort is needed to improve on them. In this chapter we introduce MCMC on a sound basis and discuss convergence and error limits thoroughly.

3.2.1 Markov chains

A Markov chain is a sequence of states $x^{[0]}, x^{[1]}, x^{[2]}, x^{[3]}, \ldots$ where every $x^{[i]}$ is drawn statistically from a probability distribution $P(x^{[i-1]}, x^{[i]})$ that depends only on $x^{[i-1]}$ and is independent of i. This does not mean that $x^{[i]}$ will be independent from $x^{[i-2]}$, $x^{[i-3]}$, $\ldots$: $x^{[i]}$ depends on $x^{[i-1]}$ and $x^{[i-1]}$ depends on $x^{[i-2]}$, therefore $x^{[i]}$ will also depend on $x^{[i-2]}$, but not as strongly as $x^{[i-1]}$.

A Markov chain is characterized by its 'transition kernel' P. $P(y_1, y_2)$ gives the probability that $x^{[i]} = y_2$ given that $x^{[i-1]} = y_1$. Because $P(y_1, y_2)$ is a probability, the following conditions hold:

$$\sum_y P(x, y) = 1 \quad \forall x, \tag{3.12}$$

and

$$0 \leq P(y_1, y_2) \leq 1 \quad \forall y_1, y_2. \tag{3.13}$$

In order to say something about the statistical properties of the $x^{[i]}$, one needs to know the probability distribution $\pi^{[0]}(x)$ of the initial value $x^{[0]}$. Let us denote the probability distribution of $x^{[i]}$ with $\pi^{[i]}(x)$. From the way the Markov chain is constructed it follows that

$$\begin{aligned} \pi^{[1]}(x) &= \sum_y \pi^{[0]}(y) P(y, x), \\ \pi^{[2]}(x) &= \sum_y \pi^{[1]}(y) P(y, x), \\ &\vdots \\ \pi^{[i]}(x) &= \sum_y \pi^{[i-1]}(y) P(y, x) \quad \forall i. \end{aligned} \tag{3.14}$$

Stationary distribution

A probability distribution $\pi(x)$ for which

$$\pi(x) = \sum_y \pi(y) P(y, x) \quad \forall x, \tag{3.15}$$

is called a 'stationary distribution' for the Markov chain with transition kernel P. From 3.15 and 3.14 it follows that if $x^{[i]}$ is distributed according to π, then $x^{[i+1]}, x^{[i+2]}, \ldots$ are distributed according to π too. Under certain conditions on the transition kernel P that will be discussed furtheron, a Markov chain will have a unique stationary distribution π and the probabilty distribution $\pi^{[i]}$ will converge to π for large enough i. 'Convergence' of the MCMC means that

$$\|\pi^{[j]} - \pi\| \longrightarrow 0 \tag{3.16}$$

for large j and for some measure $\|.\|$. Not only whether a Markov chain converges but also the rate at which $\|\pi^{[j]} - \pi\|$ approaches 0 is an important issue. In practice, one assumes that after a certain number of steps, say i_0, the Markov chain has converged and that $\pi^{[i_0]} = \pi^{[i_0+1]} = \ldots = \pi$. One then says that the Markov chain has 'thermalized'. The first i_0 steps are called the 'burn in' or 'thermalization' steps.
The idea behind MCMC is to construct a transition kernel P such that it has a given 'target distribution' π as its stationary distribution. A Markov chain is set up with this kernel. After a number of thermalization steps, the Markov chain is assumed to have converged. From then on the Markov steps are used to generate a sample $x^{[1]}, x^{[2]}, \ldots, x^{[M]}$. The Markov chain has thermalized so all $x^{[i]}$ are distributed according to π. With these $x^{[i]}$ a

sample average is calculated for an observable f

$$\mathrm{E}_S(f) = \frac{1}{M}\sum_{j=1}^{M} f(x^{[j]}). \tag{3.17}$$

We will demonstrate that $\mathrm{E}_S(f)$ is statistically distributed around an average value $\mathrm{E}(f)$ (the accuracy of MCMC) and that the standard deviation of $\mathrm{E}_S(f)$ from $\mathrm{E}(f)$ tends to 0 for large enough M (the precision of MCMC). Because the $x^{[j]}$ are not independent expression 3.7 cannot be used here to estimate the precision and will have to be modified.

3.2.2 Matrix notation for P and π

In order to discuss the convergence and the precision of MCMC, it is useful to introduce a matrix notation for P and π. This will allow the use of results from linear algebra. The space of states x is finite and discrete, so we can number the states from 1 to N: $x_1, x_2, \ldots x_N$. (the subscripts denote the position in the ordering of all the states x, not to be confused with the index of the position in the Markov chain, denoted with superscripts). We define the $N \times N$ matrix P by

$$P_{ij} = P(x_i, x_j). \tag{3.18}$$

Thus P_{ij} is the probability to go from the i^{th} to the j^{th} state in one Markov step. Probability distributions will be denoted as N-dimensional column vectors π, with elements

$$\pi_i = \pi(x_i), \tag{3.19}$$

or also as diagional $N \times N$ matrices Π with diagonal elements given by

$$\Pi_{ii} = \pi(x_i). \tag{3.20}$$

Furthermore we introduce the column vector $E = (111\ldots1)^T$.
We can immediately write down some properties:

$$\Pi E = \pi, \tag{3.21}$$

$$E^T \Pi E = E^T \pi = 1, \tag{3.22}$$

$$PE = E. \tag{3.23}$$

Property 3.14 becomes in matrix notation

$$\left(\pi^{[i]}\right)^T = \left(\pi^{[i-1]}\right)^T P. \tag{3.24}$$

Repeating this i times, we find

$$\left(\pi^{[i]}\right)^T = \left(\pi^{[0]}\right)^T P^i. \tag{3.25}$$

The condition for π to be a stationary distribution becomes

$$\pi^T = \pi^T P. \tag{3.26}$$

Note that 3.23 and 3.26 both have a specific meaning in the language of linear algebra: E is a right eigenvector of P with eigenvalue 1 and π is a left eigenvector of P with eigenvalue 1.

3.2.3 Reversible Markov chains

A Markov chain is said to be 'reversible' if a probability distribution π exists such that the following condition is fulfilled (reversibility condition):

$$\pi(x)P(x,y) = \pi(y)P(y,x) \;\; \forall x, y, \tag{3.27}$$

or in matrix notation

$$\Pi P = (\Pi P)^T . \tag{3.28}$$

In the language of linear algebra this means that ΠP is a symmetric matrix. Condition 3.27 is sometimes referred to as 'detailed balance'. It is a sufficient condition for P to have π as a stationary distribution:

Lemma 3.1 *If a probability distribution π fulfills the reversibility condition 3.27 with P, then π is a stationary distribution for the Markov chain with transition kernel P.*

Proof:

Summation over y and application of 3.12 to the left-hand side transforms the reversibility condition 3.27 into condition 3.15.

End of proof.

Nearly all Markov-Chain Monte-Carlo methods that are used in practice are based on reversible Markov chains.
For reversible Markov chains the convergence can be discussed in terms of the eigenvalues of a real symmetric matrix. To see this we rewrite expression 3.25 as

$$\begin{aligned}
\left(\pi^{[i]}\right)^T &= \left(\pi^{[0]}\right)^T P^i \\
&= \left(\pi^{[0]}\right)^T \left(\frac{1}{\sqrt{\Pi}}\sqrt{\Pi}\right) P^i \left(\frac{1}{\sqrt{\Pi}}\sqrt{\Pi}\right) \\
&= \left(\pi^{[0]}\right)^T \frac{1}{\sqrt{\Pi}} \left(\sqrt{\Pi} P \frac{1}{\sqrt{\Pi}}\right)^i \sqrt{\Pi} \\
&= \left(\pi^{[0]}\right)^T \frac{1}{\sqrt{\Pi}} \tilde{P}^i \sqrt{\Pi},
\end{aligned} \tag{3.29}$$

with

$$\tilde{P} = \sqrt{\Pi} P \frac{1}{\sqrt{\Pi}} = \frac{1}{\sqrt{\Pi}} (\Pi P) \frac{1}{\sqrt{\Pi}}. \tag{3.30}$$

$\sqrt{\Pi}$ is the diagonal matrix defined by

$$[\sqrt{\Pi}]_{ii} = +\sqrt{\Pi_{ii}} = +\sqrt{\pi_i}. \tag{3.31}$$

Because ΠP (reversibility condition 3.28) and $\sqrt{\Pi}$ are real symmetric matrices, $\tilde{P}$ is a real symmetric matrix too. From linear algebra we now that a real symmetric matrix always

can be diagonalized through an orthogonal transformation and that all its eigenvalues are real. Let

$$\tilde{P} = O^T \Lambda O, \tag{3.32}$$

with O an orthogonal matrix and Λ a real diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_N$ of $\tilde{P}$ as its diagonal elements. Then $\pi^{[j]}$ is given by

$$\left(\pi^{[j]}\right)^T = \left(O\frac{1}{\sqrt{\Pi}}\pi^{[0]}\right)^T \Lambda^j \left(O\sqrt{\Pi}\right). \tag{3.33}$$

This shows that the evolution of $\pi^{[j]}$ with j, and hence the convergence of $\pi^{[j]}$ to π, will depend on the eigenvalues of $\tilde{P}$.

3.2.4 Eigenvalues of $\tilde{P}$

In the previous subsection we introduced the symmetrized transition matrix $\tilde{P}$. We have shown that its eigenvalues are real and that they determine the convergence of the MCMC. Here we will show that $\tilde{P}$ has one eigenvalue equal to 1, with a corresponding eigenvector that is related to the target probability distribution π. We will also show that all other eigenvalues are smaller than 1 in absolute value.
First of all we note that eigenvalues of $\tilde{P}$ are also eigenvalues of P and that their eigenvectors are related:

Lemma 3.2 *If v is an eigenvector of $\tilde{P}$ with eigenvalue λ then $v_R = \frac{1}{\sqrt{\Pi_0}}v$ is a right eigenvector of P with eigenvalue λ and $v_L = \sqrt{\Pi_0}v$ is a left eigenvector of P with eigenvalue λ.*

Proof:

v is an eigenvector of $\tilde{P}$ with eigenvalue λ

$$\begin{array}{rcl} & \Updownarrow & \\ \tilde{P}v & = & \lambda v \\ & \Updownarrow & \\ \sqrt{\Pi}P\frac{1}{\sqrt{\Pi}}v & = & \lambda v \\ & \Updownarrow & \\ P(\frac{1}{\sqrt{\Pi}}v) & = & (\frac{1}{\sqrt{\Pi}}\lambda)v. \end{array}$$

and analogously:
v is an eigenvector of $\tilde{P}$ with eigenvalue λ

$$\begin{array}{rcl} & \Updownarrow & \\ v^T\tilde{P} & = & \lambda v^T \\ & \Updownarrow & \end{array}$$

$$\begin{aligned} v^T\sqrt{\Pi}P\frac{1}{\sqrt{\Pi}} &= \lambda v^T \\ &\Updownarrow \\ (\sqrt{\Pi}v)^T P &= \lambda(\sqrt{\Pi}v)^T. \end{aligned}$$

End of proof.

Lemma 3.3 *$\tilde{P}$ has an eigenvector $v_1 = \sqrt{\Pi}E$ with eigenvalue $\lambda_1 = 1$.*

Proof:

Property 3.23 tells us that E is a right eigenvector of P with eigenvalue 1. From lemma 3.2 it follows that $\sqrt{\Pi}E$ is an eigenvector of $\tilde{P}$ with eigenvalue 1.

End of proof.

An upper bound for the eigenvalues of P and $\tilde{P}$ can be found by taking the ∞-norm of P [11]:

$$\|P\|_\infty = \max_{i=1,\ldots,N}\left(\sum_{j=1}^{N}|P_{ij}|\right). \tag{3.34}$$

Lemma 3.4 $\|P\|_\infty = 1$.

Proof:

Because all P_{ij} are positive we have that

$$\|P\|_\infty = \max_{i=1,\ldots,N}\left(\sum_{j=1}^{N}|P_{ij}|\right) = \max_{i=1,\ldots,N}\left(\sum_{j=1}^{N}P_{ij}\right). \tag{3.35}$$

From 3.23 it follows that

$$\|P\|_\infty = \max_{i=1,\ldots,N}(1) = 1. \tag{3.36}$$

End of proof.

Lemma 3.5 *All eigenvalues of P have an absolute value smaller than or equal to 1.*

Proof:

If v is a right eigenvector of P with eigenvalue λ, then

$$\begin{aligned} \lambda v &= Pv \\ &\Downarrow \\ |\lambda||v_i| &= |(Pv)_i| \quad \forall i \\ &\Downarrow \end{aligned}$$

$$
\begin{aligned}
|\lambda| \max_{i=1,\ldots,N} |v_i| &= \max_{i=1,\ldots,N} |(Pv)_i| \\
&\leq \max_{i=1,\ldots,N} \sum_{j=1}^{N} |P_{ij}||v_j| \\
&\leq \max_{i=1,\ldots,N} \sum_{j=1}^{N} |P_{ij}||v_j| \\
&\leq \max_{i=1,\ldots,N} \left(\sum_{j=1}^{N} |P_{ij}| \right) \max_{k=1,\ldots,N} |v_k| \\
&= \|P\|_\infty \max_{i=1,\ldots,N} |v_i| \\
&\Downarrow \\
|\lambda| &\leq \|P\|_\infty = 1. \qquad (3.37)
\end{aligned}
$$

End of proof.

From lemma 3.2 it follows that all eigenvalues of $\tilde{P}$ have an absolute value smaller than or equal to 1 too.
In order to prove the convergence of MCMC we have to demonstrate that $\tilde{P}$ has only one eigenvalue equal to 1. This will require an additional condition on P. Therefore we introduce the notion 'irreducibility': A Markov chain with transition kernel P is said to be irreducible if any state x can be reached from any initial state $x^{[0]}$ in a finite number of Markov steps. In mathematical notation:

$$\forall i, j : \exists k : [P^k]_{ij} > 0. \qquad (3.38)$$

Lemma 3.6 *If P^2 is irreducible then $\tilde{P}$ has only one eigenvalue equal to 1.*

Proof:

From lemma 3.3 we know that $\tilde{P}$ has an eigenvector $v_1 = \sqrt{\Pi}E$ with eigenvalue $\lambda_1 = 1$. Suppose $\tilde{P}$ has a second, independent eigenvector v_2 with eigenvalue $\lambda_2 = 1$. Because $\tilde{P}$ is real symmetric, v_2 has to be orthogonal to v_1: $v_2^T v_1 = 0$. Furthermore we can take v_2 to be a real vector without loss of generality. Now consider the 2-norm of v_2 and $\tilde{P}v_2$. With 3.30, 3.23 and 3.26 we find that

$$
\begin{aligned}
\|\tilde{P}v_2\|_2^2 &= \lambda_2^2 \|v_2\|_2^2 \qquad (3.39) \\
&\Downarrow \\
\sum_{i=1}^{N} \left(\tilde{P}v_2\right)_i^2 &= 1 \sum_{i=1}^{N} (v_{2i})^2 \qquad (3.40) \\
&\Downarrow \\
\sum_{i,j,k=1}^{N} \tilde{P}_{ij} v_{2j} \tilde{P}_{ik} v_{2k} &= \sum_{i=1}^{N} \pi_i \tilde{v}_{2i}^2 \qquad (3.41) \\
&\Downarrow
\end{aligned}
$$

$$\sum_{i,j,k=1}^{N} \pi_i P_{ij} \tilde{v}_{2j} P_{ik} \tilde{v}_{2k} = \sum_{i,j,k=1}^{N} \pi_i P_{ij} P_{ik} \tilde{v}_{2i}^2 \tag{3.42}$$

$$\Downarrow$$

$$\sum_{i,j,k=1}^{N} \pi_i P_{ij} P_{ik} \tilde{v}_{2j} \tilde{v}_{2k} = \sum_{i,j,k=1}^{N} \pi_i P_{ij} P_{ik} \frac{\tilde{v}_{2i}^2 + \tilde{v}_{2k}^2}{2}. \tag{3.43}$$

Here $\tilde{v}_2$ stands for $\frac{1}{\sqrt{\Pi}} v_2$. We now that

$$ab \leq \frac{a^2 + b^2}{2} \tag{3.44}$$

and that equality is only possible if $a = b$. Therefore, the equality in 3.43 will only hold if $\tilde{v}_{2_j} = \tilde{v}_{2_k}$ for all j, k for which there exists an i such that $\pi_i P_{ij} P_{ik} \neq 0$.

From this we can draw the following corollary: if one can go from state j_1 to state j_2 in two Markov steps, then $\tilde{v}_{2j_1}$ and $\tilde{v}_{2j_2}$ must be equal. Because if one can go from state j_1 to state j_2 in two Markov steps then there must be a state i such that $P_{j_1 i} \neq 0$ and $P_{ij_2} \neq 0$. ΠP is symmetric and $\pi_i \neq 0$ and $\pi_{j_1} \neq 0$. Therefore $P_{ij_1} \neq 0$. So 3.43 can hold only if $\tilde{v}_{2j_1} = \tilde{v}_{2j_2}$ From this it follows that if any state j_1 can be reached from any other state j_2 in an *even* number of steps, then all $\tilde{v}_{2i}$ must be equal. But then $\tilde{v}_2 \sim E$ and thus $v_2 \sim \sqrt{\Pi} E = v_1$! Therefore $\tilde{P}$ has a unique eigenvalue equal to one. A completely analogous reasoning can be followed to show that $\tilde{P}$ has no eigenvalue equal to -1.

So we can conclude this by saying: if any state j_1 can be reached from any other state j_2 in an *even* number of Markov steps, then $\tilde{P}$ has a unique eigenvalue $\lambda_1 = 1$ and all its other eigenvalues have an absolute value $|\lambda_i| < 1$. We can reformulate the condition as: if P^2 is irreducible then $\tilde{P}$ has a unique eigenvalue $\lambda_1 = 1$ and all its other eigenvalues have an absolute value $|\lambda_i| < 1$.

End of proof.

We remark that if $P(x, x) \neq 0$ for some x, then irreducibility of P implies irreducibility of P^2: Suppose that one can go from y_1 to y_2 in an odd number of steps, then one can go from y_1 to x, stay 1 step in x, go back the same way to y_1 and then go to y_2 in an odd number of steps. In this way one has gone from y_1 to y_2 in an even number of steps. If any state y_2 can be reached from any other state y_1 in a finite number of steps and $P(x, x) \neq 0$ for some x, then y_2 can allways be reached from y_1 in an *even* number of steps.

A fool's MCMC sampler

Irreducibility of P is not a sufficient condition for all but one $|\lambda_i|$ to be smaller than 1, as can be seen from the following counterexample: Consider a system with n binary degrees of freedom, e.g. a nearest-neighbour Ising system with n spins that can point eather up or down [16]. A state of the system is specified by a vector $\sigma = \{\sigma_1, \ldots, \sigma_n\}$

with $\sigma_i = +1$ or -1 depending on whether the i^{th} spin is pointing up or down. Let the weight $w(\sigma)$ be given by a Boltzman factor $e^{-E_\sigma/(kT)}$. A thermal ensemble of such systems can be sampled using MCMC. The probability distribution $\pi(\sigma)$ is proportional to $w(\sigma)$. Suppose we have a transition kernel P that satisfies the reversibility condition 3.27 and that flips m spins in each Markov step. Define Σ_1 as the set of states that have an odd number of spins pointing up and Σ_2 as the set of states that have an even number of spins pointing up. If m is even then P transforms Σ_1 into Σ_1 and Σ_2 into Σ_2. This means that a state in Σ_2 can never be reached starting from a state in Σ_1 and vice versa. Thus P is not irreducible. If m is odd then P can be irreducible. P will transform Σ_1 into Σ_2 and Σ_2 into Σ_1. Therefore P will have the structure

$$P = \begin{pmatrix} 0 & P_1 \\ P_2 & 0 \end{pmatrix}. \tag{3.45}$$

Because of this structure P, for every right eigenvector

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{3.46}$$

with eigenvalue λ there will be a right eigenvector

$$\begin{pmatrix} v_1 \\ -v_2 \end{pmatrix} \tag{3.47}$$

with eigenvalue -λ. Because 1 is an eigenvalue of P, -1 is an eigenvalue to. Thus there exists a $|\lambda_i| = 1$, $i \geq 2$, even though P is irreducible. Therefore the Markov chain will not converge to the target distribution. Note that P^2 is not irreducible since it again transforms Σ_1 into Σ_1 and Σ_2 into Σ_2.

3.2.5 Non-divergence and convergence of MCMC

By using a spectral decomposition

$$\tilde{P} = \sum_{i=1}^{N} v_i \lambda_i v_i^T, \tag{3.48}$$

we can rewrite 3.33 as

$$\begin{aligned} \left(\pi^{[j]}\right)^T &= \left(\left(\pi^{[0]}\right)^T \frac{1}{\sqrt{\Pi}} v_1\right) \lambda_1^j \left(v_1^T \sqrt{\Pi}\right) + \sum_{i=2}^{N} \left(\left(\pi^{[0]}\right)^T \frac{1}{\sqrt{\Pi}} v_i\right) \lambda_i^j \left(v_i^T \sqrt{\Pi}\right) \\ &= \left(\left(\pi^{[0]}\right)^T E\right) 1^j \left(E^T \Pi\right) + \sum_{i=2}^{N} \left(\left(\pi^{[0]}\right)^T \frac{1}{\sqrt{\Pi}} v_i\right) \lambda_i^j \left(v_i^T \sqrt{\Pi}\right) \\ &= \pi^T + \sum_{i=2}^{N} c_i \lambda_i^j \bar{\pi}_i^T, \end{aligned} \tag{3.49}$$

with

$$c_i = \left(\pi^{[0]}\right)^T \frac{1}{\sqrt{\Pi}} v_i, \tag{3.50}$$

and

$$\bar{v}_i = \sqrt{\Pi} v_i. \tag{3.51}$$

Because the v_i are orthonormal vectors, the coefficients c_i are bounded by

$$\sum_{i=2}^{N} c_i^2 = \sum_{j=1}^{N} \frac{\left(\pi_j^{[0]}\right)^2}{\pi_j} - 1. \tag{3.52}$$

The $\bar{v}_i$ are no probability distributions because they can have negative elements. They are bounded by

$$\|\bar{v}_i\|_2 \leq 1, \qquad \forall i. \tag{3.53}$$

If $\pi^{[0]} = \pi$ then the coefficients $c_i = v_1^T v_i$ are equal to zero for all $i \geq 2$ because of the orthogonality of the eigenvectors of $\tilde{P}$. In that case $\pi^{[j]}$ will be equal to π for all j. So 3.27 is a sufficient condition for 'non divergence' of the Markov chain.
In practice we have $\pi^{[0]} \neq \pi$, so some of the c_i will be > 0. The Markov chain will converge only if the corresponding $|\lambda_i| < 1$. As demonstrated in the previous section, this is guaranteed if P^2 is irreducible. We see from 3.49 that $\pi^{[j]}$ will converge to π and furthermore that this convergence is *geometric*.
We can conclude:

- **condition 1:** If P is reversible then the Markov chain is non-divergent.
- **condition 2:** If, in addition, P^2 is irreducible then the Markov chain converges geometricly.

These are sufficient conditions. Also if reversibility is not fulfilled, the Markov chain can still converge. A Markov chain for which π is a stationary distribution, will converge to π if P is irreducible and not periodic [15]. P is said to be periodic if some states can be reached only in a number of Markov steps that is an multiple of some integer value, the periodicity. In the example of the fool's MCMC sampler for the Ising model given in the previous section, P was periodic with periodicity 2 for odd m.
This discussion was limited to finite, discrete Markov chains. MCMC techniques can also be applied to countable infinite or continuous state spaces. The previous results remain valid: if P is reversible and P^2 is irreducible, then P has a unique eiqenvalue equal to 1. But this does not guarantee geometric convergence of the Markov chain: P will have inifinitely many eigenvalues. It is possible that there are eigenvalues infinitesimally close to 1 that prevent the Markov chain from converging. It can also happen that the initial distribution is such that the sum of the coefficients c_i in 3.52 diverges. This can also prevent the Markov chain from converging. Establishing conditions that guarantee convergence (or even geometric convergence) for MCMC in infite state spaces is a goal of ongoing research in statistics [15].

3.2.6 Understanding the convergence of MCMC

In order to construct efficient transition kernels for MCMC, it is important to have a clear picture in mind how the convergence evolves.

Lets start with an initial distribution $\pi^{[0]}$. A good measure for the deviation of $\pi^{[0]}$ from the target distribution π is given by the sum $s^{[0]}$ of the squares of all the c_i in 3.49. It was already mentioned that

$$s^{[0]} = \sum_{k=2}^{N} c_k^2 = \sum_{i=1}^{N} \frac{\left(\pi_i^{[0]}\right)^2}{\pi_i} - 1. \tag{3.54}$$

We can extend this to every j by defining $s^{[j]}$ as

$$s^{[j]} = \sum_{i=1}^{N} \frac{\left(\pi_i^{[j]}\right)^2}{\pi_i} - 1. \tag{3.55}$$

These $s^{[j]}$ can be interpreted as the variance of

$$f_i^{[j]} = \frac{\pi_i^{[j]}}{\pi_i} \tag{3.56}$$

around its average, weighted according to π:

$$\begin{aligned}
\mathrm{E}(f^{[j]}) &= \sum_{i=1}^{N} \frac{\pi_i^{[j]}}{\pi_i}\pi_i = \sum_{i=1}^{N} \pi_i^{[j]} = 1, && (3.57)\\
\mathrm{E}\left((f^{[j]}-1)^2\right) &= \sum_{i=1}^{N} \left(\frac{\pi_i^{[j]}}{\pi_i} - 1\right)^2 \pi_i && (3.58)\\
&= \sum_{i=1}^{N} \left(\frac{(\pi_i^{[j]})^2}{\pi_i} - 2\pi_i^{[j]} + \pi_i\right) && (3.59)\\
&= \sum_{i=1}^{N} \frac{\left(\pi_k^{[j]}\right)^2}{\pi_k} - 2 + 1 = s^{[j]}. && (3.60)
\end{aligned}$$

If $s^{[j]} = 0$ then $f_i^{[j]} = 1$ for all i and thus $\pi_i^{[j]} = \pi_i$ for all i. Hence $s^{[j]}$ is a good measure for the convergence of the Markov chain.
How does f evolve under a MCMC step? From the reversibility condition 3.27 it follows that

$$\begin{aligned}
f_i^{[j+1]} &= \frac{\pi_i^{[j+1]}}{\pi_i} && (3.61)\\
&= \sum_{k=1}^{N} \frac{\pi_k^{[j]} P_{ki}}{\pi_i} && (3.62)\\
&= \sum_{k=1}^{N} \frac{\pi_k^{[j]} P_{ik}}{\pi_k} && (3.63)\\
&= \sum_{k=1}^{N} P_{ik} f_k^j. && (3.64)
\end{aligned}$$

This means that $f_i^{[j]}$ is replaced by its average over all states k that can be reached from state i, weighted according to the probability distribution P_{ik}. This local averaging results in a

smoothing of $f^{[j]}$ with every MCMC step. If P connects the whole state space well enough (irreducibility of P^2), then $f^{[j]}$ will finally have a flat distribution, i.e. $f_i^{[j]} \simeq \mathrm{E}(f^{[j]}) = 1$ for all i. At that point $\pi^{[j]}$ has converged to π.
From this it can be expected that a transition kernel that connects more states (with a non-negligible probability) will lead to a more efficient smoothing of $f^{[j]}$ and hence to a faster convergence.
We can quantify this convergence by looking at the evolution of $s^{[j]}$:

$$
\begin{aligned}
s^{[j+1]} &= \sum_{i=1}^{N} \left(f_i^{[j+1]} - 1\right)^2 \pi_i && (3.65)\\
&= \sum_{i=1}^{N} \left(f_i^{[j+1]}\right)^2 \pi_i - 1 && (3.66)\\
&= \sum_{i,k=1}^{N} f_i^{[j+1]} P_{ik} f_k^{[j]} \pi_i - 1 && (3.67)\\
&= \sum_{i,k=1}^{N} P_{ik} \frac{1}{2} \left[\left(f_k^{[j]}\right)^2 + \left(f_i^{[j+1]}\right)^2 - 2\left(f_k^{[j]} - f_i^{[j+1]}\right)^2\right] \pi_i - 1 && (3.68)\\
&= \frac{1}{2} \sum_{i,k=1}^{N} \pi_i P_{ik} \left(f_k^{[j]}\right)^2 + \frac{1}{2} \sum_{i,k=1}^{N} \pi_i P_{ik} \left(f_i^{[j+1]}\right)^2 \\
&\qquad - \sum_{i,k=1}^{N} \pi_i P_{ik} \left(f_k^{[j]} - f_i^{[j+1]}\right)^2 - 1 && (3.69)\\
&= \frac{1}{2} \left[\sum_{i,k=1}^{N} \pi_k P_{ki} \left(f_k^{[j]}\right)^2 - 1\right] + \frac{1}{2} \left[\sum_{i=1}^{N} \pi_i \left(f_i^{[j+1]}\right)^2 - 1\right] \\
&\qquad - \sum_{i,k=1}^{N} \pi_i P_{ik} \left(f_k^{[j]} - f_i^{[j+1]}\right)^2 && (3.70)\\
&= \frac{1}{2} s^{[j]} + \frac{1}{2} s^{[j+1]} - \sum_{i,k=1}^{N} \pi_i P_{ik} \left(f_k^{[j]} - f_i^{[j+1]}\right)^2 . && (3.71)
\end{aligned}
$$

If we define $\Delta_i^{[j]}$ as a kind of local variance of $f_k^{[j]}$ around its average weighted according to P_{ik},

$$\Delta_i^{[j]} = \sum_{k=1}^{N} P_{ik} \left(f_k^{[j]} - f_i^{[j+1]}\right)^2 , \tag{3.72}$$

then it follows from 3.71 that

$$s^{[j+1]} = s^{[j]} - 2\mathrm{E}(\Delta^{[j]}). \tag{3.73}$$

In other words, the amount by which $s^{[j]}$ decreases in one MCMC step is given by 2 times the average local variance of $f^{[j]}$ around its local average (weighted according to P_{ik}) $f^{[j+1]}$.

Figure 3.2: *$H(x^{[j]})$ as a function of the number Markov steps j for three different initial distributions, averaged over 5000 independent Markov chains.*

3.2.7 Monitoring the convergence of MCMC

Except for some special cases, we do not have any prior information on the eigenvalues of the transition kernel P. A practical problem in MCMC is to determine a minimal thermalization length n_0 such that one can safely assume that the Markov chain has converged.
A safe way to determine such a thermalization length would be to run a large number of independent Markov chains and to monitor how the value of an observable $f(x^{[j]})$, averaged over all the Markov chains, varies with j. Figure 3.2 illustrates this for a SDQMC calculation in the 4×4 Hubbard model.Three different initial distributions were used. With each initial distribution 5000 independent Markov chains were run. The monitored observable was $H(x) = -\ln(w(x))$. This observable can be interpreted as a kind of 'free energy' of the configurations. We used this observable because it can be evaluated with no cost from the weight $w(x)$ and because we observed that it thermalized slower than other observables like e.g. the energy. For this observable, one can also determine a lower bound for the first autocorrelation coefficient (see section 3.4.4). After 150 steps the Markov chains seem to have thermalized. In order to rely on these Markov chains for the calculation of other observables too, it would be safe to consider at least 300 thermalization steps. It is suggested by Lang et al. [37] to start the Markov chain in a state $x^{[0]}$ for which $w(x^{[0]})$ is high. Such an initial state should lead to a faster thermalization than an initial state that is drawn randomly from the whole configuration space. If we

look at the total strength 3.52 of the coefficients c_i,

$$s^{[0]} = \sum_{i=2}^{N} c_i^2 = \sum_x \frac{\left[\pi^{[0]}(x)\right]^2}{\pi(x)} - 1, \tag{3.74}$$

we see that the c_i can be very large if states for which $\pi(x)$ is small have a considerable initial probability $\pi^{[0]}(x)$. Choosing $x^{[0]}$ from a region where $w(x^{[0]})$ is high, eliminates the contribution to $s^{[0]}$ from states with a small $\pi(x)$. However, our results shown in figure 3.2, do not support this idea: The three initial distributions required an approximately equal number of thermalization steps.

Running 5000 independent Markov chains is useful for illustrating the convergence behaviour of MCMC. In practice, we cannot afford to run 5000 independent Markov chains in order to assure convergence. What one wants ideally is a way to tell, after a number of steps in a single Markov chain, whether the Markov chain has converged or not, based upon the results obtained so far with that Markov chain. Because only after the thermalization steps, computational effort and memory have to be devoted to the evaluation of observables and the accumulation of statistics. Several methods have been suggested to diagnose convergence during a single Markov chain. A review is given in [28]. Most of these methods are complicated to implement and apply only to specific MCMC methods. For our calculations we monitored convergence in a pragmatic way: Typically some 50 independent Markov chains were used, in order to obtain accurate error limits on the results (see section 3.3.3). A thermalization length n_0 was choosen on the safe side. Shorter lengths might have been sufficient, but it was easier to take the thermalization a little bit too long and to check wether it was long enough, than to take a shorter thermalization length and to redo the calculations if the thermalization length was found to be too short. We monitored the convergence by looking at the values for the observable $H(x) = -\ln(w(x))$. The value of $H(x^{[j]})$ was stored for $j = \frac{n_0}{10}, \frac{2n_0}{10}, \ldots, \frac{20n_0}{10}$. These $H(x^{[j]})$ were averaged over the 50 independent Markov chains. With a t-test these averages were compared to the average value of $H(x)$ for the total of all Markov chains. In this way the convergence of the Markov chains after n_0 steps was tested for the observable $H(x)$. If occasionally n_0 was too short, the calculation was repeated with a larger n_0. But most often we could deduce a safe n_0 before starting the calculation by looking at the thermalization of previous calculations with slightly different inputs.

3.3 Sample averages and their precision

3.3.1 Averages, variances and autocorrelations

After the Markov chain has converged, we can use the next M steps of the chain to generate a sample $x^{[1]}, x^{[2]}, \ldots, x^{[M]}$. With this sample we can estimate the expectation value $\mathrm{E}(f)$ of an observable f:

$$\mathrm{E}_S(f) = \frac{1}{M} \sum_{j=1}^{M} f(x^{[j]}). \tag{3.75}$$

$\mathrm{E}_S(f)$ is an unbiased estimate for $\mathrm{E}(f)$. This can be seen by averaging $\mathrm{E}_S(f)$ over all converged Markov chains $x^{[1]}, x^{[2]}, \ldots, x^{[M]}$. Because the chain has converged, $x^{[1]}$ will be distributed according to π. We have that

$$\begin{aligned}
&\langle \mathrm{E}_S(f) \rangle_{\text{(all MCMC samples)}} \\
&= \sum_{x^{[1]},\ldots,x^{[M]}} \pi\left(x^{[1]}\right) P(x^{[1]}, x^{[2]}) \cdots P(x^{[M-1]}, x^{[M]}) \frac{1}{M} \sum_{j=1}^{M} f(x^{[j]}) \\
&= \frac{1}{M} \sum_{j=1}^{M} \sum_{x^{[1]},\ldots,x^{[M]}} \pi\left(x^{[1]}\right) P(x^{[1]}, x^{[2]}) \cdots P(x^{[M-1]}, x^{[M]}) f(x^{[j]}) \\
&= \frac{1}{M} \sum_{j=1}^{M} \sum_{x^{[j]}} \pi\left(x^{[j]}\right) f(x^{[j]}) \\
&= \frac{1}{M} \sum_{j=1}^{M} \mathrm{E}(f) \\
&= \mathrm{E}(f). \qquad (3.76)
\end{aligned}$$

In going from 3.76 to 3.76 we used 3.12 and 3.14. The precision of this estimate can be quantified by evaluating the variance of $\mathrm{E}_S(f)$ around its mean value $\mathrm{E}(f)$:

$$\begin{aligned}
&\mathrm{Var}\left[\mathrm{E}_S(f)\right]_{\text{(all MCMC samples)}} \\
&= \langle (\mathrm{E}_S(f) - \mathrm{E}(f))^2 \rangle_{\text{(all MCMC samples)}} \\
&= \langle \left(\mathrm{E}_S(\bar{f})\right)^2 \rangle_{\text{(all MCMC samples)}} \\
&= \sum_{x^{[1]},\ldots,x^{[M]}} \pi\left(x^{[1]}\right) P(x^{[1]}, x^{[2]}) \cdots P(x^{[M-1]}, x^{[M]}) \frac{1}{M^2} \sum_{i,j=1}^{M} \bar{f}(x^{[i]}) \bar{f}(x^{[j]}) \\
&= \sum_{x^{[1]},\ldots,x^{[M]}} \pi\left(x^{[1]}\right) P(x^{[1]}, x^{[2]}) \cdots P(x^{[M-1]}, x^{[M]}) \\
&\qquad \left[\frac{1}{M^2} \sum_{j=1}^{M} \bar{f}(x^{[j]})^2 + \frac{2}{M^2} \sum_{i<j=1}^{M} \bar{f}(x^{[i]}) \bar{f}(x^{[j]}) \right] \\
&= \frac{1}{M^2} \sum_{j=1}^{M} \sum_{x^{[j]}} \pi\left(x^{[j]}\right) \bar{f}(x^{[j]})^2 \\
&\qquad + \frac{2}{M^2} \sum_{i<j=1}^{M} \sum_{x^{[i]},x^{[j]}} \pi\left(x^{[i]}\right) \bar{f}(x^{[i]}) P^{j-i}(x^{[i]}, x^{[j]}) \bar{f}(x^{[j]}) \\
&= \frac{1}{M^2} \sum_{j=1}^{M} \mathrm{E}(\bar{f}^2) + \frac{2}{M^2} \sum_{k=1}^{M-1} (M-k) \sum_{x,y} \pi(x) \bar{f}(x) P^k(x,y) \bar{f}(y) \\
&= \frac{\sigma^2(f)}{M} \left[1 + 2 \sum_{k=1}^{M-1} (1 - k/M) \rho_k(f) \right]. \qquad (3.77)
\end{aligned}$$

We introduced the notations

$$\bar{f} = f - \mathrm{E}(f) \qquad (3.78)$$

$$\sigma^2(f) = \mathrm{E}\left[(f-\mathrm{E}(f))^2\right] = \mathrm{E}\left(\bar{f}^2\right) \tag{3.79}$$

$$\rho_k(f) = \sum_{x,y} \pi(x)\bar{f}(x)P^k(x,y)\bar{f}(y)/\sigma^2(f). \tag{3.80}$$

$\rho_k(f)$ is called the k^{th} autocorrelation coefficient of f. If we introduce the vector

$$\tilde{f} = \sqrt{\Pi}\bar{f}, \tag{3.81}$$

we can rewrite $\sigma^2(f)$ and $\rho_k(f)$ as

$$\sigma^2(f) = \tilde{f}^T\tilde{f} \tag{3.82}$$

$$\rho_k(f) = \frac{\tilde{f}^T\tilde{P}^k\tilde{f}}{\tilde{f}^T\tilde{f}}. \tag{3.83}$$

From the fact that $\mathrm{E}(\bar{f}) = 0$ it follows that $\tilde{f}$ is orthogonal to $v_1 = \sqrt{\Pi}E$ and that $\rho_k(f)$ is bounded by the second largest eigenvalue of $\tilde{P}$:

$$|\rho_k(f)| \leq |\lambda_2|^k. \tag{3.84}$$

Equality will occur for $\tilde{f} = v_2$. Because $|\lambda_2| < 1$, $\rho_k(f)$ goes to 0 as k goes to infinity. $\rho_k(f)$ is a measure for the correlation between $f(x^{[i]})$ and $f(x^{[i+k]})$.

The central limit theorem

Expression 3.84 shows that this correlation decreases exponentially with increasing k. After some number of Markov steps, the $f(x)$ can be considered as almost independent samples. Hence, if M gets very big, one can expect the sample average $\mathrm{E}_S(f)$ to behave almost as an average of independently sampled values. For large M, the central limit theorem tells us that the average of independently sampled values tends to be normally distributed with a mean $\mathrm{E}(f)$ and a variance $\sigma^2(f)/M$. For a MCMC sample the central limit theorem still holds under quite generall conditions [15] (that are fulfilled for finite reversible Markov chains) but the distribution now has a variance given by

$$\sigma^2_{MC}(f) = \frac{\sigma^2(f)}{M}\left[1 + 2\sum_{k=1}^{\infty}\rho_k(f)\right]. \tag{3.85}$$

From 3.84 it follows that $\sigma^2_{MC}(f)$ is bounded by

$$\sigma^2_{MC}(f) \leq \frac{\sigma^2(f)}{M}\left(1 + 2\sum_{k=1}^{\infty}\lambda_2^k\right) = \frac{\sigma^2(f)}{M}\frac{1+\lambda_2}{1-\lambda_2}. \tag{3.86}$$

This shows that λ_2 is not only determinative for the convergence rate of MCMC, but also for the precision of the sample averages. A matrix notation for $\sigma^2_{MC}(f)$ is given by

$$\sigma^2_{MC}(f) = \frac{\sigma^2(f)}{M}\tilde{f}^T\frac{1+\tilde{P}}{1-\left(\tilde{P}-v_1v_1^T\right)}\tilde{f}/\left(\tilde{f}^T\tilde{f}\right) \tag{3.87}$$

$$= \frac{\sigma^2(f)}{M}r(f). \tag{3.88}$$

It is interesting to compare the independent and the MCMC variances. In order to obtain the same precision with MCMC as with independent sampling, one needs $r(f)$ times the number of samples needed with independent sampling, where $r(f)$ is given by

$$r(f) = \tilde{f}^T \frac{1+\tilde{P}}{1-\left(\tilde{P}-v_1 v_1^T\right)} \tilde{f} / \left(\tilde{f}^T \tilde{f}\right). \tag{3.89}$$

If P has eigenvalues close to 1, $r(f)$ can be quite large. On the other hand, if P has negative eigenvalues, then it can happen that $r(f)$ is smaller than 1. In that case, MCMC sampling will be more precise than independent sampling [19] ! This has the paradoxial consequence that a kernel with eigenvalues close to -1 will converge very slowly but can yield very precise results after thermalization. So the fool's sampler from section 3.2.4 was maybe not that foolish after all! If an other transition kernel is used first in order to make $\pi^{[j]}$ converge to π, then the fool's sampler might be superior in evaluating sample averages.

Estimating autocorrelations

The autocorrelation coefficients $\rho_k(f)$ are intrinsic properties of the Markov chain, closely related to the eigenvalues of $\tilde{P}$. Since they determine the error limits on the sample averages and also the optimal sampling interval lentgh (see below), it is important to have a reasonable estimate for them. This is obtained by evaluating the correlation between the sampled values:

$$C_k(f) = \frac{1}{M-k} \sum_{i=1}^{M-k} f_i f_{i+k} - \mathrm{E}_S^2(f). \tag{3.90}$$

The expectation value of $C_k(f)$ is given by

$$\begin{aligned} \langle C_k(f) \rangle_{(\text{all MCMC samples})} &= \rho_k(f)\sigma^2(f) - \mathrm{Var}\left[\mathrm{E}_S(f)\right]_{(\text{all MCMC samples})} & (3.91) \\ &= \left[\rho_k(f) - \frac{r(f)}{M}\right] \sigma^2(f). & (3.92) \end{aligned}$$

If M is large enough the second term can be neglected. Then we get

$$\rho_k(f) \simeq \frac{C_k(f)}{C_0(f)}, \tag{3.93}$$

where $C_0(f)$ is the variance among the sampled values. Note that this requires $\rho_k(f) \gg \frac{r(f)}{M}$ so that the estimate is bad for small ρ_k (large k) or small M (small samples). Even if M is very large, the estimates for the small ρ_k are unreliable. We observed that the variance of the C_k is almost independent of k. Therefore the relative error increases for larger k or smaller ρ_k. It often requires much more samples to obtain reliable estimates for the C_k than to obtain reliable estimates for $\mathrm{E}_S(f)$. Therefore methods to obtain error limits on $\mathrm{E}_S(f)$ that are based on autocorrelation coefficients are often either unreliable (too big errors in the C_k) or inefficient (too many samples are needed in order to get reliable values for the C_k).

First-order-autocorrelated series

We say that the series $f(x^{[1]}), f(x^{[2]}), \ldots$ is a 'first-order-autocorrelated' series if

$$\rho_k(f) = [\rho_1(f)]^k \,, \quad \forall k. \tag{3.94}$$

This is the case if $\rho_k(f)$ depends only on one non-zero eigenvalue of $\tilde{P}$. In other cases, the assumption of a first-order-autocorrelated series can be a good approximation. For such a series, $r(f)$ given by

$$r(f) = \frac{1+\rho_1(f)}{1-\rho_1(f)}. \tag{3.95}$$

3.3.2 Sampling with intervals

Until now we assumed that $f(x)$ was evaluated for every $x^{[j]}$. If the evaluation of $f(x)$ requires a considerable amount of computing time, it might be more efficient to evaluate $f(x)$ only after every n^{th} Markov step. This amounts to replacing P by P^n. It reduces the sample size M by a factor n. At the same time it can reduce the factor $r(f)$ considerably. In this way $\sigma^2_{MC}(f)$ increases only slightly, while the computation time is strongly reduced. This can enhance the efficiency of the method.
It is often suggested to take the interval such that the first autocorrelation coefficient between sampled values is about 0.1 [13, 6]. This value is not motivated by efficiency. It should ensure that the sampled values are almost independent, such that the rules of statistics for independent samples can be applied to fix the error limits for the sample average. However, this interval length leads to suboptimal sampling: most of the generated configurations are not used for the evaluation of the sample average. Furthermore, the samples that are used are considered as independent, while they still have an autocorrelation of 10%. Therefore the obtained error limits might be misleading.
What interval leads to the most efficient sampling? To set the idea, we assume that the $f(x^{[j]})$ form a first-order-autocorrelated series. Let γ be the ratio between the computing time need for the evaluation of $f(x)$ in one point x and the computing time needed for one Markov step. Let M, the number of Markov steps, be large. The thermalization steps are neglected. Let ρ be the autocorrelation coefficient of f for two successive Markov-chain states. Then the first autocorrelation coefficient of f in the case of sampling with n-step intervals is given by

$$\rho_1^{[n]}(f) = \rho^n. \tag{3.96}$$

To determine the optimal sampling interval, we look at the computation time that is needed to get error limits smaller than a given value σ_g. Suppose a Markov chain of M steps is used to generate sample values every n^{th} step. Then $\mathrm{E}_{S,n}(f)$ is the average of M/n sample values. The error on $\mathrm{E}_{S,n}(f)$ is related to its variance:

$$\sigma^2_{MC,n}(f) \simeq \frac{\sigma^2(f)}{M/n} r_n(f). \tag{3.97}$$

Expression 3.89 is adapted to the fact that the Markov chain is sampled every n^{th} step:

$$r_n(f) = \tilde{f}^T \frac{1 + \tilde{P}^n}{1 - \left(\tilde{P}^n - v_1 v_1^T\right)} \tilde{f} / \left(\tilde{f}^T \tilde{f}\right). \tag{3.98}$$

To reach the desired level of precision, $\sigma_{MC,n}(f) = \sigma_g$, we need a number of sample values given by

$$\frac{M}{n} = \frac{\sigma^2(f)}{\sigma_g^2} r_n(f). \tag{3.99}$$

The needed amount of computer time is given by

$$T = M + \frac{M}{n}\gamma = \frac{M}{n}(n + \gamma) = \frac{\sigma^2(f)}{\sigma_g^2} r_n(f)(n + \gamma), \tag{3.100}$$

where the time unit is equal to the time needed for one Markov step. The efficiency ϵ is inversely proportional to T:

$$\epsilon \propto \frac{1}{(n + \gamma)\, r_n(f)}. \tag{3.101}$$

For a first-order-autocorrelated series this becomes

$$\epsilon \propto \frac{1}{(n + \gamma)\, \frac{1+\rho^n}{1-\rho^n}}. \tag{3.102}$$

The optimal interval length $n$is the one that maximizes ϵ. Here, ϵ depends furthermore on two parameters: γ and ρ. Instead of the variable n we take $x = \rho^n$ as the independent variable for the optimization of ϵ. This has the advantage that there is a scaling in ϵ:

$$\epsilon \propto \frac{1 - x}{(\ln(x) + \ln(\rho)\gamma)(1 + x)}, \tag{3.103}$$

such that the optimal x depends only on one parameter $z = \ln(\rho)\gamma$. The values of the parameters depend strongly on the system that is studied. Typical values are $\gamma \simeq 1$, $\rho \simeq 0.95$. This leads to $z \simeq -0.05$. Figure 3.3 shows the efficiency as a function of x for several values of z. The efficiency is optimal around the suggested value of $x = 0.1$ only if $z \leq -2$. This is the case if the Markov chain has a very short autocorrelation length or if the values for the observable require a lot of computing time. Such a low value for z is seldomly encountered in practice. Hence one can expect that the optimal value for x will be greater than 0.1. We made the assumption that the series of values was a first-order-autocorrelated series. In reality the autocorrelation coefficient and the efficiency are related to a weighted average over the eigenvalues of $\tilde{P}$:

$$x = \rho_1^{[n]}(f) = \frac{c_2^2 \lambda_2^n + c_3^2 \lambda_3^n + \cdots + c_N^2 \lambda_N^n}{c_2^2 + c_3^2 + \cdots + c_N^2}, \tag{3.104}$$

with the coefficients $c_1, c_2, \ldots c_N$ given by

$$c_i = v_i^T \tilde{f}. \tag{3.105}$$

Figure 3.3: *The efficiency* $\epsilon = \frac{1-x}{(\ln(x)+\ln(\rho)\gamma)(1+x)}$ *for a first-order-autocorrelated series as a function of the first autocorrelation coefficient* x*, for several values of the parameter* z *defined in the text.*

With these coefficients, $r_n(f)$ can be expressed as

$$r_n(f) = \frac{c_2^2 \frac{1+\lambda_2^n}{1-\lambda_2^n} + c_3^2 \frac{1+\lambda_3^n}{1-\lambda_3^n} + \cdots + c_N^2 \frac{1+\lambda_N^n}{1-\lambda_N^n}}{c_2^2 + c_3^2 + \cdots + c_N^2}. \tag{3.106}$$

The efficiency is now given by

$$\epsilon = \frac{c_2^2 + c_3^2 + \cdots + c_N^2}{(n+\gamma)\left[c_2^2 \frac{1+\lambda_2^n}{1-\lambda_2^n} + c_3^2 \frac{1+\lambda_3^n}{1-\lambda_3^n} + \cdots c_N^2 \frac{1+\lambda_N^n}{1-\lambda_N^n}\right]}. \tag{3.107}$$

The averaging over all λ's will shift the optimal x to lower values. An estimate for the optimal x can be obtained by fitting the autocorrelation coefficients ρ_j with an expression of the form

$$\rho_j \simeq a_1 l_1^j + a_2 l_2^j + \cdots + a_m l_m^j, \text{with } a_1 + a_2 + \cdots + a_m = 1, \tag{3.108}$$

where m, the number of terms, is quite small so that a reasonable fit can be obtained. Then we can substitute the c_i and λ_i in 3.107 with the a_i and l_i and determine the n that maximizes ϵ. This procedure is illustrated in the following example.

Example for the optimal sampling interval

The internal energy of the half-filled 4×4 Hubbard model was calculated with a SDQMC method. The interaction strength was $U = 8|t|$, the inverse temperature $\beta = 2/|t|$,

i	a_i	l_i
1	0.053888	0.99898
2	0.744906	0.98757
3	0.201206	0.95933

Table 3.1: *The coefficients for the fit of the autocorrelation coefficients with a function of the form 3.108.*

Figure 3.4: *Autocorrelation coefficients for a SDQMC calculation of the energy for the half-filled* 4×4 *Hubbard model with* $U = 8|t|$,$\beta = 2/|t|$,$N_t = 80$.

the number of inverse temperature slices $N_t = 80$. The energy was evaluated after every Markov step. The first 1000 autocorrelation coefficients for this observable were calculated. They were fitted with a function of the form 3.108 with 3 terms. The addition of a fourth term did not improve the fit, so only three terms were retained. The coefficients a_i and l_i are listed in table 3.1 The autocorrelations and the fitted function are shown in figure 3.4. With the algorithm we used, the evaluation of the energy required 1.4 times the amount of computer time needed for 1 Markov step. Using the fitted coefficients a_i and l_i, an estimate for $r_n(E)$ was obtained

$$r_n(E) = a_1 \frac{1 + l_1^n}{1 - l_1^n} + a_2 \frac{1 + l_2^n}{1 - l_2^n} + a_3 \frac{1 + l_3^n}{1 - l_3^n}. \tag{3.109}$$

This estimate was used to calculate the efficiency $\epsilon(n)$ as a function of the length n of the interval with which the energy should be sampled.

$$\epsilon(n) = \frac{1}{(n + 1.4) r_n(E)}. \tag{3.110}$$

Figure 3.5: *Efficiency $\epsilon_n(E)$ versus the first autocorrelation coefficient $\rho_1^{[n]}(E)$ for the SDQMC calculation described in the text, with a varied sampling-interval length n.*

In figure 3.5 the efficiency $\epsilon(n)$ is plotted as a function of the first autocorrelation coefficient $\rho_1^{[n]}(E)$

$$\rho_1^{[n]} = a_1 l_1^n + a_2 l_2^n + a_3 l_3^n, \tag{3.111}$$

that would be obtained if the energy was evaluated only every n Markov steps. The optimal efficiency is obtained at $\rho_1^{[n]} \simeq 0.55$. This corresponds to an interval length of some 40 Markov steps. If the interval length would be chosen such that $\rho_1^{[n]} \simeq 0.1$, which corresponds to $n \simeq 200$, the efficiency would be only 75% of its maximal value. Figure 3.6 shows the efficiency for several values of γ. It is observed that the efficiency remains within 5% of its maximal value at $\rho_1^{[n]} \simeq 0.5$ for a broad range of values for γ. Only for extreme values of $\gamma \simeq 200$, the maximum in the efficiency curve is found at $\rho_1^{[n]} \simeq 0.1$. Around $\rho_1^{[n]} \simeq 0.5$ the effiency is observed to vary very smoothly. From this we conclude that in most cases, choosing the sampling-interval length n such that $\rho_1^{[n]} \simeq 0.5$, is a good choice. Only in extreme cases, it will lead to a lower efficiency than the often suggested value $\rho_1^{[n]} \simeq 0.1$.

Guidelines for choosing the sampling interval

We want to make two more remarks on interval sampling. In most cases, the optimal interval length n can be expected to be quite large such that the optimal $n \gg \gamma$. In those cases the efficiency is not very sensitive to n. If the evaluation of f requires 2 times more computation time than 1 Markov step, then a run where f is evaluated every 20 steps will require a fraction $\frac{20+2}{20+2\times 2} = 0.92$ of the time that is required for evaluating every 10 steps.

Figure 3.6: *Efficiency $\epsilon_n(E)$ versus the first autocorrelation coefficient $\rho_1^{[n]}(E)$ for several values of the parameter γ.*

Evaluating every 10 steps will be at least as precise as evaluating every 20 steps, so the efficieny will differ not more than 9% and probably less. Therefore it is safer to take the interval length not too long, if γ is still much smaller.
A second remark is connected to the better-than-independent-sampling paradox mentioned at the end of section 3.3.1. Suppose that f is evaluated every n Markov steps. This amounts to Markov-chain sampling with a kernel P^n instead of P. If n is even, P^n will have only positive eigenvalues, while $P^{(n-1)}$ can have negative ones. Therefore it can happen that an interval of n Markov steps leads to a lower precision than an interval of $n-1$ Markov steps, although the former scheme leads to less correlated samples. This suggests that it is preferable to take an odd number of Markov steps in between sampled values. In practice, the negative eigenvalues of $\tilde{P}$ tend to be small so that the effect will be small too.
To conclude we can say that, generally speaking, taking the interval length n such that the first autocorrelation coefficient is smaller than 0.1 leads to too long intervals and a suboptimal efficiency. A rule of thumb that will do better in most cases, is to take n such that the first autocorrelation coefficient is approximately 0.5. We also recommend to take n odd (if it is of no avail, it is no drawback either).

3.3.3 Error limits on sample averages

From the MCMC sample $x^{[1]}, \ldots, x^{[M]}$ we can calculate the sample average $\mathrm{E}_S(f)$ of a function f. This can be considered as a 'measurement' of an 'observable' f (e.g. energy, densities, ...). Because MCMC is a statistical technique, it is important to determine error limits for the measured values. Otherwise the obtained results are meaningless. Several

approaches can be followed, the one more reliable or efficient than the other.

Pseudo-independent samples

A method for establishing error limits that is often suggested in literature [13, 6] is based on sampling with intervals (see section 3.3.2): evaluate $f(x)$ every n Monte-Carlo steps and take n large enough such that the samples can be considered as independent ones. Error limits can then be obtained from standard statistical rules. If we label the sampled values as $f_1, f_2, \ldots f_m$, $m = M/n$, then the outcome of the 'measurement' is the sample average

$$\mathrm{E}_S(f) = \frac{1}{m}\sum_{i=1}^{m} f_i. \tag{3.112}$$

The variance on $\mathrm{E}_S(f)$ is given by

$$\sigma_S^2(f) = \mathrm{E}_S\left[\left(f - \mathrm{E}_S(f)\right)^2\right]/(m-1). \tag{3.113}$$

A 95%-confidence interval for the average is given by the appropriate $t_{95\%}$ coefficient ($m-1$ degrees of freedom) times the standard deviation σ_S.
The arbitrary point in this method is how to decide when samples are independent. It is suggested to take n such that $\rho_n \leq 0.1$. As we showed before, this often leads to intervals that are longer than optimal. Furthermore, we do not know the coefficients ρ_n. We have to use the estimates C_n/C_0. These estimates can become unreliable for values as small as 0.1.
The obtained samples are not completely independent: there is still a correlation of 10%, so it is more accurate to call them 'pseudo-independent'. If the series of sample-values is first-order autocorrelated, this correlation of $\rho = 10\%$ leads to an increase with a factor $\sqrt{r(f)} = \sqrt{\frac{1+\rho}{1-\rho}} \simeq 1.11$ in the standard deviation of $\mathrm{E}_S(f)$. So the suggested error limits are 10% too small. In general the series are not first-order autocorrelated, so that the underestimation will be less than 10%. The factor 1.11 is a maximum value. We suggest to multiply the pseudo-independent error limits with this factor, since it guarantees an 'at least 95%'-confidence level.

Estimation of $r(f)$

In most cases it is advantageous to sample with shorter intervals. Here the sampled values can no longer be considered as independent. The independent-sample variance has to be multiplied with the factor $r(f)$ to obtain the true sample variance (see expression 3.88). Estimating error limits amounts to estimating $r(f)$.
Because $\rho_k \simeq C_k/C_0$, one could naively be tempted to estimate $r(f)$ by

$$r(f) \simeq 1 + 2\sum_{k=1}^{m}\left(1 - \frac{k}{m}\right)\frac{C_k(f)}{C_0(f)}. \tag{3.114}$$

However, this estimate turns out to be exactly 0. The reason for this is the second term in expression 3.92. This term makes that C_k/C_0 is not a good estimate of ρ_k for larger

k. Because that second term is proportional to $r(f)$, we can account for it by modifying 3.114 into

$$r(f) \simeq 1 + 2\sum_{k=1}^{m}\left(1 - \frac{k}{m}\right)\left(\frac{C_k(f)}{C_0(f)} + \frac{r(f)}{M}\right). \tag{3.115}$$

If the summation over k is carried out, one obtains the trival result

$$r(f) \simeq 0 + r(f). \tag{3.116}$$

Again, this does not lead us to a good estimate for $r(f)$. A way to obtain an estimate for $r(f)$ is to neglect the contibutions of the ρ_k for k bigger than a certain number K. Because the ρ_k decrease exponentially with increasing k, this seems plausible. One can bring the fraction of $r(f)$ on the right hand side of equation 3.115 to the other side, and divide out the prefactor of $r(f)$. This leads to [18]

$$r(f) \simeq \frac{m^2}{(m-K)(m-K+1)}\left[1 + 2\sum_{k=1}^{K}\left(1 - \frac{k}{m}\right)\frac{C_k(f)}{C_0(f)}\right]. \tag{3.117}$$

The number K may not be too small, because in that case a substantial fraction of the ρ_k would be neglected and $r(f)$ would be underestimated. Taking K too large makes the first factor in the right hand side of 3.117 large and the second factor small, such that the errors on the second factor are multiplied drastically and spoil the estimate. For applications in SDQMC, we found this estimate to be impractical because it suffered too much from the high uncertainties on the C_k for larger k. Sample sizes that allowed accurate determination of sample averages did not allow accurate estimation of the errors on these averages with this estimate of $r(f)$.
A different approach to estimate $r(f)$ is to estimate the dominant eigenvalues of $\check{P}$ as was explained in section 3.3.2. An estimate for $r(f)$ is then given by expression 3.106. Again, for practical use this estimate is unreliable: the estimates for the eigenvalues of $\check{P}$ are spoiled by the uncertainties in the C_k. Furthermore the determination of the eigenvalues from the C_k can be quite complicated, because it amounts to an inverse Laplace transform. In the case of a first-order-autocorrelated series, $r(f)$ can be estimated as

$$r(f) = \frac{1+\rho}{1-\rho} \simeq \frac{1 + C_1/C_0}{1 - C_1/C_0}. \tag{3.118}$$

Because the C_1 are most often well determined, this estimate is useful, as far as the approximation of a first-order-autocorrelated series is applicable.

Repeated runs

A simple way to obtain reliable error limits on the sample averages, is to restart the Markov chain a number of times (say N) with independent starting values. The obtained sample averages will be approximately normally distributed, because of the central limit theorem. Furthermore, they are indepedent because they are the results of independent runs. The total average of the N sample averages is taken as the final result. The error on this value can be obtained from standard statistics, since it is an average of N independent values.

Instead of one Markov chain of M steps, one runs N Markov chains of M/N steps. The accuracy of the final value will remain the same (it remains an average over M values). The estimated error limits will be more reliable. Restarting the Markov chain a number of times also avoids the risk that the Markov chain gets stalled in a limited region of the configuration space. Furthermore it allows a reliable monitoring of the convergence of the Markov chain. For most of our calculations we determined error limits in this way, with $N = 50$.
The disadvantage of this procedure is that one has to thermalize the Markov chain N times. If thermalization is quick, this is no big problem. But if the thermalization requires a lot of Markov steps, this procedure becomes inefficient. An alternative in this case is not to restart the Markov chain every M/N steps while still calculating an average every M/N steps. One then obtains N values that can be expected to be normally distributed. But now these values are not independent. If M/N is large enough, the correlations among these values will be small, so that the series of values can be considered as a first-order-autocorrelated series. If one determines the first autocorrelation coefficient C_1 of these values, then accurate error limits can be obtained by multiplying the error limits for independent case with the factor $\sqrt{(}r(f)$, with $r(f)$ given by 3.118. Still it is recommendable to restart the Markov chain a few times, in order to avoid the risk that the Markov chain gets stalled in a limited region of the configuration space.

Error limits on ratios of observables.

In SDQMC we often have to calculate ratios of observables. Determining error limits on ratios of averages is more complicated than determining error limits on the averages themselves. Especially if the sampled values are autocorrelated, as is the case in MCMC. Suppose that we need to evaluate a ratio f/g,

$$f/g = \frac{\sum_x f(x)w(x)}{\sum_{x'} g(x)w(x')}. \tag{3.119}$$

A situation often encountered in SDQMC, is that $g(x)$ is the sign, $w(x)$ the absolute value of $\hat{\mathrm{Tr}}\left(\hat{U}_x\right)$ and $f(x)$ an observable (e.g. the energy E_x) times $g(x)$. The ratio f/g is equal to

$$f/g = \frac{\mathrm{E}(f)}{\mathrm{E}(g)}. \tag{3.120}$$

One could determine $\mathrm{E}_S(f)$ and $\mathrm{E}_S(g)$ from a sample obtained in a independent MCMC run. Note that the obtained values are not independent. Then we take as an estimate for f/g

$$\mathrm{E}_S(f/g) = \frac{\mathrm{E}_S(f)}{\mathrm{E}_S(g)}. \tag{3.121}$$

If $\mathrm{E}_S(f)$ and $\mathrm{E}_S(g)$ would be independent, the variance on $\mathrm{E}_S(f/g)$ could be estimated as

$$\sigma^2(f/g) = \frac{\mathrm{E}_S^2(f)}{\mathrm{E}_S^2(g)}\left[\frac{\sigma^2(f)}{\mathrm{E}_S^2(f)} + \frac{\sigma^2(f)}{\mathrm{E}_S^2(f)}\right]. \tag{3.122}$$

Figure 3.7: *Scatter plot of $E_S(f)$ versus $E_S(g)$ obtained from 400 SDQMC runs of 8000 Markov steps each. The system studied here was the 4×4 Hubbard model, with $U = 4|t|$, $\beta = 8$, $N_t = 160$, 6 spins up and 6 spins down.*

	f	g	f/g (energy)
value	0.0988	1.749	17.70
standard deviation	0.0040	0.074	0.20

Table 3.2: *Standard deviations on $E_S(f)$, $E_S(g)$ and $E_S(f/g)$.*

In figure 3.7 we show the scatter plot of $\mathrm{E}_S(f)$ versus $\mathrm{E}_S(g)$ obtained from 400 SDQMC runs. Here, g is the sign and f the energy times g of the terms in the expansion of $\hat{\mathrm{Tr}}\left(e^{-\beta\hat{H}}\right)$. It is clear form this figure that variances on $\mathrm{E}_S(f)$ are strongly correlated to variances on $\mathrm{E}_S(g)$, in such a manner that they are divided out to a large extent in the ratio f/g. The standard deviations on $\mathrm{E}_S(f)$, $\mathrm{E}_S(g)$ and $\mathrm{E}_S(f/g)$ are listed in table 3.2. If f and g would be independent, the resulting standard deviation on f/g would be of the order of 1.0 instead of 0.2. If $\mathrm{E}_S(f)$ and $\mathrm{E}_S(g)$ would be calculated from independently sampled states $x^{[j]}$, then the correlations between f and g could be taken into account with the estimate for the variance given by [18]

$$\sigma^2(f/g) = \frac{\mathrm{E}_S\left[(f - \mathrm{E}_S(f/g)g)^2\right]}{\mathrm{E}_S^2(g)}. \tag{3.123}$$

This expression can be applied to Monte-Carlo data if one assumes (pseudo-) independent samples, as in section 3.3.3. Or in the case of repeated runs, as in section 3.3.3, if one applies the expression only to the averaged values of every run.

We remark that if repeated runs are used, then one should estimate f/g as

$$\frac{\mathrm{E}_{S_1}(f)+\mathrm{E}_{S_2}(f)+\cdots+\mathrm{E}_{S_N}(f)}{\mathrm{E}_{S_1}(g)+\mathrm{E}_{S_2}(g)+\cdots+\mathrm{E}_{S_N}(g)}, \tag{3.124}$$

and not as

$$\frac{\mathrm{E}_{S_1}(f)}{\mathrm{E}_{S_1}(g)}+\frac{\mathrm{E}_{S_2}(f)}{\mathrm{E}_{S_2}(g)}+\cdots+\frac{\mathrm{E}_{S_N}(f)}{\mathrm{E}_{S_N}(g)}, \tag{3.125}$$

because the former estimate is less biased than the latter.

3.3.4 Variance reduction

In this section we explain a method to reduce the variance on the Monte-Carlo sample averages. For generality, we introduce the formulas for a ratio of observables f/g. By putting $g=1$ one obtains the formulas for a single observable f. This discussion is a generalization of ideas presented in [18] and [12].
Instead of evaluating $\mathrm{E}(f/g)$ by drawing a sample $S=(x^{[1]},x^{[2]},\ldots x^{[M]})$ according to $w(x)$ and evaluating

$$\mathrm{E}_S(f/g)=\frac{\sum_{j=1}^M f(x^{[j]})}{\sum_{j=1}^M g(x^{[j]})} \tag{3.126}$$

one could draw a sample $S'=(y^{[1]},y^{[2]},\ldots y^{[M]})$ according to an alternative weight distribution $v(y)$ and evaluate

$$\mathrm{E}_{S'}(g)=\frac{\sum_{j=1}^M f(y^{[j]})w(y^{[j]})/v(y^{[j]})}{\sum_{j=1}^M g(y^{[j]})w(y^{[j]})/v(y^{[j]})}. \tag{3.127}$$

This is a good estimate of $\mathrm{E}(f/g)$ too, because

$$\begin{aligned}
\mathrm{E}(f/g) &= \frac{\sum_x f(x)w(x)}{\sum_{x'} g(x')w(x')} \\
&= \frac{\sum_x f(x)[w(x)/v(x)]v(x)}{\sum_{x'} g(x')[w(x')/v(x')]v(x')} \\
&= \frac{\sum_x f(x)[w(x)/v(x)]v(x)}{\sum_{x'} v(x')} \Big/ \frac{\sum_y g(y)[w(y)/v(y)]v(y)}{\sum_{y'} v(y')} \\
&= \frac{\mathrm{E}_v(fw/v)}{\mathrm{E}_v(gw/v)}.
\end{aligned} \tag{3.128}$$

Sampling according to a different distribution will lead to a different variance on the obtained sample average. In the case of independent sampling we can calculate the variance on the ratio 3.127 using expression 3.123:

$$\begin{aligned}
\sigma_v^2(f/g) &\simeq \frac{\mathrm{E}_v\left[(fw/v-\mathrm{E}_v(f/g)gw/v)^2\right]}{\mathrm{E}_v^2(gw/v)} \\
&= \frac{\mathrm{E}_w(v/w)\mathrm{E}_w\left[(f-\mathrm{E}_w(f/g)g)^2\,w/v\right]}{\mathrm{E}_w^2(g)}.
\end{aligned} \tag{3.129}$$

What form for $v(x)$ results in the smallest variance? Minimizing expression 3.129 by varying $v(x)$ leads to

$$v(x)^2 \propto \left[f(x) - \mathrm{E}(f/g)g(x)\right]^2 w^2(x). \tag{3.130}$$

To be useful for Monte-Carlo sampling, $v(x)$ has to be positive. Apart from a normalization, we obtain

$$v(x) = \left|f(x) - \mathrm{E}(f/g)g(x)\right| w(x). \tag{3.131}$$

The corresponding variance on f/g is given by

$$\sigma_v^2(f/g) \simeq \frac{\mathrm{E}_w^2\left(\left|f(x) - \mathrm{E}(f/g)g(x)\right|\right)}{\mathrm{E}_w^2(g)}. \tag{3.132}$$

Sampling according to an alternative distribution $v(x)$ can reduce the error on the obtained sample averages. The error will be at least as large as 3.132. Practically, one does not know the ratio $\mathrm{E}(f/g)$ in advance. One can make a guess for it, $R \simeq \mathrm{E}(f/g)$, and sample according to $|f(x) - Rg(x)|w(x)$.

For Markov-chain Monte-Carlo methods, variance reduction is not always an improvement: the error of the results depends not only on $\sigma(f/g)$, but also on the autocorrelations in the sample, expressed by the factor $r(f)$ in expression 3.88. Sampling according to the weight $|f(x) - Rg(x)|w(x)$ instead of $w(x)$ can enhance these autocorrelations because of two reasons. First of all, the weight $|f(x) - Rg(x)|w(x)$ has nodes where $f(x) \simeq Rg(x)$. For these configurations x, the weight becomes small. It is not very probable that the Markov chain will pass through a region with small weight. It can be expected that, compared to sampling according to $w(x)$, the chain will stay for a longer period in the region where $f(x) < Rg(x)$, before passing on to the region where $f(x) > Rg(x)$. Hence, autocorrelations will be larger. A second reason, that plays a role in case of Metropolis-sampling (see section 3.4.1), is that the autocorrelations are related to the deviation between the target distribution $w(x)$ and the stationary distribiution of the proposition Kernel. The extra factor $|f(x) - Rg(x)|$ in front of the target distribution can enhance these deviations.

We tried to apply variance reduction to SDQMC calculations for the energy of the 4×4 repulsive Hubbard model. We observed a reduction in the errors of the sample averages, but longer Markov-chain runs where needed because of the larger autocorrelations. Furthermore, the evaluation of the energy with every Markov-chain made that a Markov chain took twice as much time as in the case without variance reduction. Therefore, variance reduction was no improvement for these calculations. We think that variance reduction could improve the efficiency of Markov-chain Monte-Carlo methods in cases where the evaluation of $v(x)$ does not require much more time than the evaluation of $w(x)$,

3.4 Construction of transition kernels

Given a target distribution π, we want to construct a transition kernel P such that P is reversible (condition 3.27) with π and such that P^2 is irreducible. Before we go into detail

on some particular kernel types, we want to mention two ways in which we can combine reversible kernels into new ones. These kernels are called *hybrid* kernels. They can be useful in order to construct irreducible kernels from reversible but reducible kernels.
The first way is to choose randomly between several kernels. This amounts to a weigthed average of kernels:

Lemma 3.7 *If P_1 and P_2 are transition kernels reversible with π, then $P = (P_1 + P_2)/2$ is a transition kernel reversible with π too.*

Proof:

P is a transition kernel since al its elements fall between 0 and 1 and

$$PE = \frac{P_1 + P_2}{2}E = \frac{P_1E + P_2E}{2} = \frac{E + E}{2} = E. \tag{3.133}$$

Furthermore P satisfies the reversibility condition 3.28 since

$$\begin{aligned} \Pi P &= \Pi\frac{P_1 + P_2}{2} = \frac{\Pi P_1 + \Pi P_2}{2} \\ &= \frac{P_1^T\Pi + P_2^T\Pi}{2} = \left(\frac{P_1 + P_2}{2}\right)^T \Pi \\ &= P^T\Pi. \end{aligned}$$

End of proof.

This lemma can be extended to any linear combination of transtion kernels,

$$P = \alpha_1P_1 + \alpha_2P_2 + \ldots \alpha_nP_n, \tag{3.134}$$

with $0 < \alpha_1, \alpha_2, \ldots, \alpha_n < 1$ and $\alpha_1 + \alpha_2 + \ldots + \alpha_n = 1$.
We can also apply two different kernels consecutively. This does not necesserally lead to a reversible kernel.

Lemma 3.8 *If P_1 and P_2 are transition kernels reversible with π, then $P = P_1P_2$ is a transition kernel too. It has π as its stationary distribution, but it will only be reversible if $P_1P_2 = P_2P_1$.*

Proof:

P is a transition kernel since al its elements fall between 0 and 1: P_{ij} can be seen as the weighted average of the elements of the j^{th} column of P_2, weighted according to the elements of the i^{th} row of P_1. Since all the elements of P_2 fall between 0 and 1, P_{ij} does too. Furthermore it is clear that

$$PE = P_1P_2E = P_1E = E. \tag{3.135}$$

P has π as it stationary distribution because of condition 3.26:

$$\pi P = \pi P_1P_2 = \pi P_2 = \pi. \tag{3.136}$$

From

$$\Pi P = \Pi P_1 P_2 = P_1^T \Pi P_2 = P_1^T P_2^T \Pi = (P_2 P_1)^T \Pi \tag{3.137}$$

we see that the reversibility condition 3.28 will only be fulfilled if

$$P_2 P_1 = P_1 P_2. \tag{3.138}$$

End of proof.

Some MCMC methods are based on transition kernels of this type. An example is the 'deterministic scan Gibbs sampler' (cfr. section 3.4.2). A simple way to make a product transition kernel reversible, is by repeating it in reversed order: instead of applying two times $P = P_1 P_2 \cdots P_n$ one can use $P' = P_1 P_2 \cdots P_n P_n \cdots P_2 P_1$. For a non-reversible kernel $P = P_1 P_2$ that is a product of reversible kernels, the convergence of the Markov chain can still be understood as a local smoothing of $f^{[j]} = \pi^{[j]}/\pi$ as discussed in section 3.2.6, with the minor modification that the smoothing proceeds according to $P_2 P_1$ instead of $P = P_1 P_2$. For a discussion of the convergence properties of hybrid kernels in terms of the convergence of the constituent kernels, see [23, 24].
In the next section we will discuss a method to construct reversible kernels which have a given target distribution as their stationary distribution. To obtain irreducible kernels, one sometimes has to combine several reversible kernels. This is a practical problem that will depend on the details of the system under study. We will discuss this for SDQMC in more detail later on.

3.4.1 The Metropolis-Hastings method

The most important method for building reversible kernels was introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller [17] for computing thermodynamical properties of molecules. It was extended to a general sampling method by Hastings [18]. Most methods for building reversible kernels used nowadays, like e.g. the Gibbs sampler, can be understood as special cases of this method.
The method works as follows: suppose that the Markov chain is in a state $x^{[j]}$. For the next Markov step a trial state x^T is proposed by making a small, random change to the configuration of $x^{[j]}$. This has to be done in such a way that the probability of generating x^T from $x^{[j]}$, denoted with $Q(x^{[j]}, x^T)$, has to be equal to the probability of generating $x^{[j]}$ from x^T, $Q(x^{[j]}, x^T) = Q(x^T, x^{[j]})$. For the Ising system of section 3.2.4, e.g., such an x^T could be generated by flipping a few randomly chosen spins on the lattice. Here $x^{[j+1]}$ is chosen in the following way:

- If $w(x^T) \geq w(x^{[j]})$, then take $x^{[j+1]} = x^T$.
- If $w(x^T) < w(x^{[j]})$, then take $x^{[j+1]} = x^T$ with probability $w(x^T)/w(x^{[j]})$, otherwise take $x^{[j+1]} = x^{[j]}$.

This defines a transition kernel $P(x^{[j]}, x^{[j+1]})$ for the Markov chain. It is easy to verify that P fulfills the reversibility condition 3.27:

$$w(x^{[j]}) P(x^{[j]}, x^{[j+1]}) = w(x^{[j+1]}) P(x^{[j+1]}, x^{[j]}). \tag{3.139}$$

It is important to notice that the method is based on the ratio $w(x^T)/w(x^{[j]})$. Hence the normalization of $w(x)$ does not have to be known. This method defines a random walk through the configuration space, where trial moves are accepted or rejected according to the ratio $w(x^T)/w(x^{[j]})$.
The method can be generalized to asymmetric proposition kernels Q [13]. Generate the trial move x^T from $Q(x^{[j]}, x^T)$. Let q be given by

$$q = \frac{w(x^T)Q(x^T, x^{[j]})}{w(x^{[j]})Q(x^{[j]}, x^T)}. \tag{3.140}$$

- If $q \geq 1$, then take $x^{[j+1]} = x^T$.
- If $q < 1$, then take $x^{[j+1]} = x^T$ with probability q, otherwise take $x^{[j+1]} = x^{[j]}$.

If Q is a reversible kernel whose stationary distribution $w_Q(x)$ is known (not necessarely normalized), then q can be calculated as

$$q = \frac{w(x^T)w_Q(x^{[j]})}{w(x^{[j]})w_Q(x^T)}. \tag{3.141}$$

This can be interesting if x^T is generated in a complicated way such that it is difficult to determine $Q(x^T, x^{[j]})$ explicitly. An example of such a situation is given in section 3.4.5. Other rules for accepting or rejecting x^T can work too. Hastings [18] mentions the following rule, a generalization of a method proposed by Barker:

- Take $x^{[j+1]} = x^T$ with probability $\frac{q}{1+q}$, otherwise take $x^{[j+1]} = x^{[j]}$.

The method can be formulated in a general way by writing the transition kernel $P(x, y)$ as the product of the proposition kernel $Q(x, y)$ with an acceptance function $A(x, y)$,

$$P(x, y) = Q(x, y)A(x, y). \tag{3.142}$$

The acceptance function must fulfill $0 < A(x, y) \leq 1$, $\sum_y Q(x, y)A(x, y) = 1, \forall x$ and the reversibility condition

$$w(x)Q(x, y)A(x, y) = w(y)Q(y, x)A(y, x). \tag{3.143}$$

This acceptance function shifts the stationary distribution of the Markov chain from $w_Q(x)$ to the target distribution $w(x)$. Possible forms for $A(x, y)$ are

- generalized Metropolis:

$$A(x, y) = \min\left(1, \frac{w(y)Q(y, x)}{w(x)Q(x, y)}\right), \quad \forall x \neq y \tag{3.144}$$

$$A(x, x) = 1 - \sum_{y \neq x} Q(x, y)(1 - A(x, y)), \quad \forall x. \tag{3.145}$$

n	a	S	σ_S	E	σ_E	E/S	$\sigma_{E/S}$
1	0.544	0.329	0.062	5.84	1.12	17.72	0.15
2	0.382	0.305	0.045	5.37	0.82	17.63	0.11
3	0.297	0.325	0.037	5.75	0.67	17.69	0.14
4	0.242	0.317	0.032	5.59	0.57	17.65	0.13
5	0.203	0.314	0.038	5.55	0.68	17.66	0.16

Table 3.3: *Comparison of the efficiency of several proposition kernels. n is the number of spins updated per Markov step, a the acceptance rate.*

- generalized Barker:

$$A(x,y) = \frac{w(y)Q(y,x)}{w(y)Q(y,x) + w(x)Q(x,y)}, \quad \forall x \neq y \tag{3.146}$$

$$A(x,x) = 1 - \sum_{y \neq x} Q(x,y)(1 - A(x,y)), \quad \forall x. \tag{3.147}$$

The second lines, 3.145 and 3.147, have to assure that $\sum_y P(x,y) = 1$. Peskun has shown that the Metroplis rules give the opimum choice for A [19].
The optimum choice for the proposition kernel Q depends strongly on the system under consideration. In most applications, x^T is constructed by making small moves from $x^{[j]}$. If x denotes a spin configuration (as in the Ising model), x^T can be generated by flipping some spins. If x denotes a real vector, x^T can be generated by adding to one of its components a small step that is uniformely sampled from an interval $[-L, L]$. If the moves are small, then the Markov chain will stay in the same region of the configuration space for a long while. This leads to large autocorrelations and hence to a low efficiency. If the moves are big, then a lot of trial moves will be rejected. A lot of time is spend on the evaluation of unused configurations, which also makes the method inefficient. A balance has to be found between low autocorrelation and high acceptance. A measure that is often used to quantify this is the acceptance rate a. It is the ratio of the number of accepted trial moves to the number of Markov steps in a run of the Markov chain. It is often suggested to make the trial moves of such a size that $a \simeq 0.5$. This value is quite arbitrary. It is a useful rule of thumb, but it does not guarantee a near to optimal efficiency for the MCMC. Gelman, Roberts and Gilks showed that for a class of systems with high dimensional state spaces the optimal acceptance rate is $a \simeq 0.23$ [21]. For our SDQMC calculations, we observed that the efficiency of the Markov chain did not vary very much for acceptance rates between 0.25 and 0.5. Table 3.3 shows results from a calculation for the repulsive Hubbard model on a 4×4 lattice, with $6 + 6$ electrons, $U = 4|t|$, $\beta = 6/|t|$. The average sign S and an energy observable E (such that E/S gives the internal energy), were calculated for several proposition-kernel parameter settings which each lead to a different acceptance rate a. 20 runs of 50000 (correlated) samples were done. From this an average and a standard error for the observables were calculated. Table 3.3 shows that the optimal acceptance rate for E and S is 24%, remarkably close to the value of Gelman et al. The standard error of E/S on the other hand, does not seem to be very sensitive to a.

3.4.2 The Gibbs sampler

A special case of the Metropolis-Hastings method is the Gibbs sampler [22]. It can be used when the configurations are vectors $x = (x_1, x_2, \ldots x_n)$, and when the conditional probabilities $\pi(x_j|x_1, x_2, \ldots x_{j-1}, x_{j+1}, \ldots, x_n)$ are known (i.e. can be sampled directly) but the total propability density $\pi(x)$ is not. Transition kernels $P_1, \ldots P_n$ are defined as follows:
For $P_j(x, y)$: generate y from x by taking $y_i = x_i$ for $i \neq j$. Draw y_j according to $\pi(x_j|x_1, x_2, \ldots x_{j-1}, x_{j+1}, \ldots, x_n)$.
Each of these P_j is reversible with $\pi(x)$. The P_j can be considered as Metropolis kernels with the proposition probability given by the conditional probability of the j^{th} component of x. Because $\pi(x|y)\pi(y) = \pi(y|x)\pi(x)$, the ratio q in 3.140 equals 1 and the trial moves are always accepted. The P_j^2 are not irreducible. In order to obtain a transition kernel P whose square is irreducble, one has to combine the P_j using lemma 3.7 or lemma 3.8. This leads to the 'random-scan Gibbs sampler' and the 'deterministic-scan Gibbs sampler' respectively. One can expect the random-scan sampler to lead to shorter autocorrelation lengths, because its transition kernel connects more states at once than the kernel of the deterministic-scan sampler. This will lead to a more efficient smoothing of the function $f^{[j]}$ defined in 3.56. As explained in section 3.2.6, this means that the Markov chain convergences faster.
If the conditional probabilities $\pi(x_j|x_1, x_2, \ldots x_{j-1}, x_{j+1}, \ldots, x_n)$ are not known explicitly, Monte-Carlo techniques can be used to evaluate them. Because the configuration spaces for one variable x_j, with the other variables fixed, is much smaller than the total configuration space, the sampling of one variable can often be done in an efficient way. This leads to hybrid algorithms like the 'Metropolis in Gibbs' sampler, where the x_j are sequentially sampled using the Metropolis algorithm.
In statistical physics the Gibbs sampler is known as the "heat bath" algorithm [13]. A canonical ensemble for n classical particles is sampled according to the Boltzmann factor $e^{-\beta H(x_1, x_2, \ldots, x_n)}$. This is done by bringing one particle into equilibrium with a "heat bath" at inverse temperature β, while keeping the other particles fixed. Bringing x_j into equilibrium with a "heat bath" at inverse temperature β amounts to sampling x_j according to $\pi(x_j|x_1, x_2, \ldots x_{j-1}, x_{j+1}, x_n) \propto e^{-\beta H(x_1, x_2, \ldots x_{j-1}, y, x_{j+1}, \ldots, x_n)}$. Solving the one-body problem for every variable x_j sequentially in a heat bath with all the other variables fixed brings the total many-body system in thermal equilibrium with the heat bath. The heat-bath algorithm is often applied to spin systems. These can be seen as systems where the components x_j of the state vector x can take on only two values: up or down. At every Markov step, one spin x_j is flipped with a probability $\frac{q}{1+q}$, where q is the ratio between the weight of the configuration with x_j flipped to the weight of the given configuration. This probability $\frac{q}{1+q}$ is also the probability with which a randomly generated trial move for x_j has to be accepted according to Barkers rule (see section 3.4.1). As discussed before, the Metropolis acceptance rule is more efficient. Therefore, the heat-bath algoritm for spin systems can easily be improved by flipping the spin x_j with probability $\min(1, q)$. Liu has shown how the heat-bath algorithm or Gibbs sampler can be improved in an analogous way for systems whit components x_j that can take on more than two values [29].

3.4.3 The independence Metropolis sampler

An interesting type of Metropolis algorithm is the independence Metropolis sampler. It is often used as a building block for hybrid kernels. An interesting point is that the eigenvalues of its transition kernel can be computed explicitely. From these, the convergence properties for the kernel can be determined. The independence sampler is a Metropolis sampler whose proposition kernel is independent of x, $Q(x,y) = Q(y)$. This means that whatever the state $x^{[j]}$ is, the trial move x^T is drawn independently from a distribution $Q(x^T)$. Note that $Q(y)$ is also the stationary distribution for such a proposition kernel. The acceptance of the trial move does depend on $x^{[j]}$: x^T is accepted with probability

$$\min\left(1, \frac{\pi(x^T)Q(x^{[j]})}{\pi(x^{[j]})Q(x^T)}\right). \tag{3.148}$$

This leads to a transition kernel P whose second largest eigenvalue is given by [25]

$$\lambda_2 = 1 - \frac{1}{\max_x\left(\pi(x)/Q(x)\right)}. \tag{3.149}$$

This shows that the convergence behaviour of the Independence Metropolis sampler is closely related to the deviation between the target distribution and the stationary distribution of the proposition kernel. Also the transition probabilities $P^n(x,y)$ can be computed exactly for this case [26].

3.4.4 A limitation on Metropolis algorithms

Except for the independence Metropolis sampler of the previous section, it is difficult to predict the convergence behaviour of Metropolis kernels. An interesting result was obtained by Caracciolo et al. [27]. They showed that a lower limit for the second largest eigenvalue λ_2 of the transition kernel can be derived in terms of an energy-like observable H defined by

$$H(x) = \ln(w_Q(x)/w(x)). \tag{3.150}$$

The first autocorrelation coefficient for this observable is bounded by

$$\rho_1(H) \geq 1 - \frac{4/e^2}{\mathrm{var}(H)}, \tag{3.151}$$

with $\mathrm{var}(H) = \mathrm{E}\left[(H - \mathrm{E}(H))^2\right]$. Because $\lambda_2 \geq \rho_1(f)$ for any f , we have a lower bound for λ_2:

$$\lambda_2 \geq 1 - \frac{4/e^2}{\mathrm{var}(H)}. \tag{3.152}$$

This again illustrates that the convergence of the Markov chain is closely related to the deviation between the target distribution and the stationary distribution of the proposition kernel.

This insight allows us to give an intuitive answer to the following question: Suppose that we have to sample configurations of the from 2.18:

$$\frac{4 + \hat{\text{Tr}}\left(e^{+\sqrt{6\beta}\hat{A}}\right) + \hat{\text{Tr}}\left(e^{-\sqrt{6\beta}\hat{A}}\right)}{6}. \tag{3.153}$$

We could generate a trial move by choosing between these three terms with equal probability. The ratios between $w(x)$ and $w_Q(x)$ are then given by $2\hat{\text{Tr}}(1)$, $\hat{\text{Tr}}\left(e^{+\sqrt{6\beta}\hat{A}}\right)/2$ and $\hat{\text{Tr}}\left(e^{-\sqrt{6\beta}\hat{A}}\right)/2$ respectively. Another possibility is to choose the first term with probability $4/6$, the second and the third term each with probability $1/6$. The ratios between $w(x)$ and $w_Q(x)$ are then given by $\hat{\text{Tr}}(1)$, $\hat{\text{Tr}}\left(e^{+\sqrt{6\beta}\hat{A}}\right)$ and $\hat{\text{Tr}}\left(e^{-\sqrt{6\beta}\hat{A}}\right)$ respectively. The question is: what way of proposing a trial move is the most efficient ? If β is small, then the ratios between $w(x)$ and $w_Q(x)$ for the latter method will all be close to 1, while for the former method they have a larger variation. Therefore it can be expected that the latter way of choosing the trial move will lead to a more efficient sampling algorithm. In general we recommend to construct the proposition kernel Q such that its stationary distribution is as similar to the target distribution as possible, while still allowing a quick generation of trial moves.

3.4.5 Guided Metropolis sampling

Suppose that we want to sample from a distribution $w(x)$ that requires a considerable amount of computation time for the evalutation of $w(x)$, and that we also have a function $w_0(x) \simeq w(x)$ that is much easier to evaluate. An efficient way to generate trial moves for the Metropolis sampling of $w(x)$, is to run a Metropolis Markov chain for $w_0(x)$ of a certain number of steps, say n, and to use the resulting state y as trial move. If $Q(x, y)$ is the Metropolis transition kernel for $w_0(x)$, then the proposition kernel for this algorithm is given by $Q^n(x, y)$. Calculating $Q^n(x, y)$ and $Q^n(y, x)$ explicitly is practically impossible. We cannot evaluate the ratio q from section 3.4.1 using expression 3.140. But $Q^n(x, y)$ is a Metropolis transition kernel for $w_0(x)$. Hence it has $w_0(x)$ as its stationary distribution. So we can evaluate q using expression 3.141:

$$q = \frac{w(x^T)w_0(x^{[j]})}{w(x^{[j]})w_0(x^T)}. \tag{3.154}$$

If $w_0(x) \simeq w(x), \forall x$, then $q \simeq 1$. This means that most of the trial moves will be accepted. So using $w_0(x)$ to 'guide' the proposal moves for the Metropolis sampling $w(x)$ results in a high acceptance rate. If $w_0(x)$ is much easier to sample than $w(x)$, this guided Metropolis sampling will be much more efficient than the Metropolis sampling based on $w(x)$ only. In the limit of large n, the Metropolis Markov chain for $w_0(y)$ will thermalize. The trial moves y can then be considered to be independent from x. They are distributed according to $w_0(y)$ The guided sampler becomes an independence Metropolis sampler.
In SDQMC, where the $w(x)$ are given by the trace of a product of exponentials of one-body operators,

$$w(x) = \hat{\text{Tr}}\left(e^{\hat{A}(x_1)}e^{\hat{A}(x_2)} \dots e^{\hat{A}(x_{N_t})}\right), \tag{3.155}$$

a function $w_0(x)$ that is sometimes useful for guided sampling is obtained by retaining only the diagonal elements of the $\hat{A}(x)$. The product of the exponentials of the diagonal matrices is again a diagonal matrix. Its trace can be evaluated with $\mathcal{O}\left((N_S^2\right)$ operations. We applied this method in SDQMC calculations for the 4×4 Hubbard model. The guided sampling improved the efficiency of the SDQMC drastically for low inverse temperatures, $\beta \simeq 1$, and high interaction strengths, $U \simeq 8$. At higher values of β the method improved the efficiency only slightly.

4

The Slater-determinant quantum Monte-Carlo method

The concepts introduced in the previous chapters are brought together in this chapter to form a powerful quantum Monte-Carlo method for the study of fermionic many-body problems.
In chapter 1 we showed that exponentials of one-body operators can be represented by $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrices, with $N_{\mathcal{S}}$ the dimension of the one-body space. In chapter 2 we showed how the exponential of a general two-body Hamiltonian $\hat{H}$ can be decomposed as a sum of exponentials of one-body operators,

$$\begin{aligned} & e^{-\beta\hat{H}} \\ &= \sum_{\sigma} e^{-\frac{\beta}{2}\hat{H}_0} e^{-\hat{A}_{\sigma_1}(\beta)} e^{-\beta\hat{H}_0} e^{-\hat{A}_{\sigma_2}(\beta)} e^{-\beta\hat{H}_0} \cdots e^{-\beta\hat{H}_0} e^{-\hat{A}_{\sigma_{N_t}}(\beta)} e^{-\frac{\beta}{2}\hat{H}_0} \\ &= \sum_{\sigma} e^{-\hat{S}_{\sigma}(\beta)} = \sum_{\sigma} \hat{U}_{\sigma}. \end{aligned} \tag{4.1}$$

In chapter 3 we showed how a sum over a large number of terms can be approximated by statistical sampling using Markov-chain Monte-Carlo methods. In the present chapter we will show how these building blocks can be combined into a method for the study of fermionic many-body systems. The cornerstone of the method is the Boltzmann operator $e^{-\beta\hat{H}}$. It can be used in two ways: it can be seen as the many-body density-matrix operator of a statistical ensemble of quantum many-body systems at a temperature $T = 1/\beta$, or it can be seen as an operator that projects states onto the ground state of the many-body system, if β is large. Hence the Boltzmann operator can be used to study the thermodynamical (internal energy, specific heat, ...) and ground-state (ground-state energy, density and momentum distributions, correlations, ...) properties of quantum many-body systems. Spectroscopic information such as energies and level densities of excited states can be obtained from thermodynamical data using an inverse Laplace transform. The inverse Laplace transform is an ill-conditioned numerical problem. This makes the calculation

of spectroscopic quantities using SDQMC a lot more complicated than the calculation of thermodynamical or ground-state properties. Therefore, we restricted this work to calculations of thermodynamical and ground-state properties. Implementing a routine for the inverse Laplace transform in order to be able to calculate spectroscopic quantities, will be one of the first goals of our future research.

4.1 Statistical quantummechanics and thermodynamics

In statistical mechanics, one studies the properties of an 'ensemble' of states. An ensemble can be defined as a set of possible states of the system. To each state in the enemble a weight is attributed that is proportional to the probability for the system to be observed in that particular state. Several types of ensembles are used in statistical mechanics: the grand canonical, the canonical and the microcanonical ensemble.
The grand canonical ensemble represents a system that is in equilibrium with an infinitely large heat and particle reservoir at a temperature T and a chemical potential μ. The system can exchange energy with the reservoir. This leads to fluctuations in the energy of the system. The system can also exchange particles with the reservoir, leading to fluctuations in the number of particles. Statistical mechanics learns us that, at thermal equilibrium at a temperature T and a chemical potential μ, the probability for the system to be in a state with energy E and a number of particles equal to N, is proportional to $e^{\frac{-E+\mu N}{kT}}$. Here, k is the Boltzmann constant. In what follows, we choose temperature units such that $k = 1$. Instead of the temperature T, we mostly use the variable $\beta = 1/T$. With these notations, the weight attributed to a state with energy E and N particles in the grand canonical ensemble is given by $e^{-\beta E+\beta\mu N}$.
One could isolate the system from the reservoir in such a way that particles can no longer be exchanged, while energy exchange is still possible. Then the system has a fixed number of particles, N, while the energy of the system fluctuates. The ensemble that represents such a system is the canonical ensemble. The probability to find the system in a state with energy E is proportional to the Bolzmann factor $e^{-\beta E}$.
One can further isolate the system, in such way that no energy is exchanged anymore. Because the energy is a constant of motion for an isolated system, the system will have a fixed energy. The ensemble that represents such a system is called the microcanonical ensemble.
In statistical quantummechanics an ensemble of states can be represented by a statistical density matrix $\hat{\rho}$,

$$\hat{\rho} = \sum_i w_i |\phi_i\rangle\langle\phi_i|, \tag{4.2}$$

with w_i the weight attributed to the many-body state ϕ_i. This density matrix represents a mixture of states, not to be confused with a superposition of states. In order to see w_i as a probability for the system to be found in state ϕ_i, $\hat{\rho}$ should be normalized such that $\hat{\mathrm{Tr}}(\hat{\rho}) = 1$. For the canonical ensemble, the unnormalized statistical density matrix is given by the Boltzmann operator $e^{-\beta\hat{H}}$. This can be seen by expanding the operator $e^{-\beta\hat{H}}$

in the basis of N-particle energy eigenstates:

$$e^{-\beta\hat{H}} = \sum_i e^{-\beta E_i} |E_i\rangle\langle E_i|. \tag{4.3}$$

The probability to find the system in a state with energy E_i is proportional to the Boltzmann factor $e^{-\beta E_i}$. The normalized density matrix is given by

$$\hat{\rho} = \frac{1}{Z_\beta} e^{-\beta\hat{H}}, \tag{4.4}$$

with Z_β defined as

$$Z_\beta = \hat{\mathrm{Tr}}_N \left(e^{-\beta\hat{H}} \right). \tag{4.5}$$

In what follows, we will use the notation $\hat{\mathrm{Tr}}_N$ for the trace over the N-particle states and $\hat{\mathrm{Tr}}$ for the trace over the complete many-body space; thus

$$\hat{\mathrm{Tr}} = \sum_{N=1}^{N_S} \hat{\mathrm{Tr}}_N. \tag{4.6}$$

Z_β is called the 'partition function'. It contains all thermodynamic information on the system in the canonical ensemble. For the grand canonical ensemble, a term $\beta\mu\hat{N}$, with $\hat{N}$ the particle number operator, has to be included in the exponent of the Boltzmann operator The operator now acts on the whole many-body space without restrictions on the number of particles:

$$\hat{\rho}_{GC} = \frac{1}{Z_{GC,\beta}} e^{-\beta\hat{H}+\beta\mu\hat{N}}, \tag{4.7}$$

with the grand partition function $Z_{GC,\beta}$ now given by

$$Z_{GC,\beta} = \hat{\mathrm{Tr}} \left(e^{-\beta\hat{H}+\beta\mu\hat{N}} \right). \tag{4.8}$$

With the SDQMC method, we can calculate the thermodynamic expectation values of operators. The following expressions, apply to the canonical ensemble. The extension to the grand canonical ensemble is straightforward. In the canonical ensemble, the thermodynamic expectation value for an operator $\hat{A}$ is given by

$$\langle\hat{A}\rangle_C = \frac{1}{Z_\beta} \hat{\mathrm{Tr}}_N \left(\hat{A} e^{-\beta\hat{H}} \right). \tag{4.9}$$

One can interpret $\langle\hat{A}\rangle_C$ as the ensemble average of the matrix element $\langle E_i|\hat{A}|E_i\rangle$.
Particularly interesting quantities to study with SDQMC, are thermodynamical quantities such as the internal energy, the specific heat, the entropy and the free energy of the system. The internal energy U is the ensemble average of the energy, or in other words, the thermodynamical expectation value of the Hamiltonian:

$$U = \langle\hat{H}\rangle_C = \frac{1}{Z_\beta} \hat{\mathrm{Tr}}_N \left(\hat{H} e^{-\beta\hat{H}} \right) = -\frac{\partial \ln |Z_\beta|}{\partial\beta}. \tag{4.10}$$

The specific heat C gives the amount by which the internal energy increases if the temperature of the system is increased by a small amount:

$$C = \frac{\partial U}{\partial T} = -\beta^2 \frac{\partial U}{\partial \beta} = \beta^2 \frac{\partial^2 \ln |Z_\beta|}{(\partial \beta)^2} = \beta^2 \left(\langle \hat{H}^2 \rangle_C - \langle \hat{H} \rangle_C^2 \right). \tag{4.11}$$

The entropy S of the system is given by

$$\begin{aligned} S &= -\hat{\mathrm{Tr}}_N \left[\ln(\hat{\rho}) \hat{\rho} \right] \\ &= -\frac{1}{Z_\beta} \hat{\mathrm{Tr}}_N \left[\left(-\beta \hat{H} - \ln(Z_\beta) \right) e^{-\beta \hat{H}} \right] \\ &= \beta \langle \hat{H} \rangle_C + \ln(Z_\beta) \\ &= \beta U + \ln(Z_\beta). \end{aligned} \tag{4.12}$$

From the thermodynamical relation for the free energy $F = U - TS$, one finds that

$$F = -\frac{\ln(Z_\beta)}{\beta}. \tag{4.13}$$

The partition function is also related to the level density $g(E)$ of excited states of the N-particle system: Z_β is the Laplace transform of $g(E)$:

$$Z_\beta = \sum_i e^{-\beta E_i} = \sum_E e^{-\beta E} g(E). \tag{4.14}$$

In an analogous way, the thermodynamic expectation value $\langle \hat{A} \rangle_C$ of an operator $\hat{A}$ in the canonical ensemble can be seen as the Laplace transform of its expectation value $\langle \hat{A} \rangle_E$ in the microcanonical ensemble:

$$\langle \hat{A} \rangle_C = \frac{1}{Z_\beta} \sum_i \langle E_i | \hat{A} | E_i \rangle e^{-\beta E_i} = \frac{1}{Z_\beta} \sum_E \langle \hat{A} \rangle_E g(E) e^{-\beta E}. \tag{4.15}$$

This shows that, in principle, information obtained from the canonical ensemble can be transformed to information in the microcanonical ensemble. However, this transformation is equivalent to an inverse Laplace transform, which is known to be numerically very unstable. Though in recent years maximum-entropy techniques have proven to be very useful to stabilize the inverse Laplace transform, for this application the method remains inaccurate and unstable.
The thermodynamic response function $R_{\hat{A}}(\tau)$ for an operator $\hat{A}$ in the canonical ensemble is defined as [6]

$$R_{\hat{A}}(\tau) = \langle e^{\tau \hat{H}} \hat{A}^\dagger e^{-\tau \hat{H}} \hat{A} \rangle = \frac{1}{Z_\beta} \hat{\mathrm{Tr}} \left[e^{-(\beta - \tau) \hat{H}} \hat{A}^\dagger e^{-\tau \hat{H}} \hat{A} \right]. \tag{4.16}$$

Inserting a complete set of N-particle energy eigenstates ($|\Psi_i\rangle$, $|\Psi_f\rangle$ with energies E_i, E_f) shows that

$$R_{\hat{A}}(\tau) = \frac{1}{Z_\beta} \sum_{i,f} e^{-\beta E_i} \left| \langle \Psi_f | \hat{A} | \Psi_i \rangle \right|^2 e^{-\tau (E_f - E_i)}. \tag{4.17}$$

The thermodynamic response function $R_{\hat{A}}(\tau)$ can be seen as the Laplace transform of the strength function $S_{\hat{A}}(\omega)$:

$$R_{\hat{A}}(\tau) = \sum_{\omega} e^{-\tau\omega} S_{\hat{A}}(\omega) \tag{4.18}$$

$$S_{\hat{A}}(\omega) = \frac{1}{Z_\beta} \sum_{i,f} e^{-\beta E_i} \left| \langle \Psi_f | \hat{A} | \Psi_i \rangle \right|^2 \delta(\omega - E_f + E_i). \tag{4.19}$$

Here, $S_{\hat{A}}(\omega)$ gives the strength with which the action of the operator $\hat{A}$ can excite the system at a temperature $T = 1/\beta$ with an excitation energy ω. In the limit of large β, $S_{\hat{A}}(\omega)$ gives the strength function for excitation out of the ground state. The inverse Laplace transform needed to obtain $S_{\hat{A}}(\omega)$ from $R_{\hat{A}}(\tau)$ is not as troublesome as the inversion of the Laplace transform in expression 4.15. Maximum-entropy methods yield useful results here [6, 32]. It has been asserted that maximum-entropy methods form the only consistent way to take into account the statistical errors on the Monte-Carlo data in an inverse Laplace transform. [33].

4.2 SDQMC for the grand canonical ensemble

The thermodynamical expectation value for an operator $\hat{A}$ in the grand canonical ensemble can be expressed as a sum of terms that can be handled easily in a numerical way using the decomposition 4.1. If we write this decomposition as

$$\hat{\text{Tr}} \left[e^{-\beta\hat{H}} e^{\beta\mu\hat{N}} \right] = \hat{\text{Tr}} \left[\sum_{\sigma} \hat{U}_\sigma e^{\beta\mu\hat{N}} \right] = \sum_{\sigma} w(\sigma). \tag{4.20}$$

then the expectation value of the operator $\hat{A}$ can be written as

$$\begin{aligned} \langle \hat{A} \rangle &= \frac{\hat{\text{Tr}} \left[\hat{A} e^{-\beta\hat{H}} e^{\beta\mu\hat{N}} \right]}{\hat{\text{Tr}} \left[e^{-\beta\hat{H}} e^{\beta\mu\hat{N}} \right]} \\ &= \frac{\sum_\sigma \hat{\text{Tr}} \left[\hat{A} \hat{U}_\sigma e^{\beta\mu\hat{N}} \right]}{\sum_{\sigma'} \hat{\text{Tr}} \left[\hat{U}_{\sigma'} e^{\beta\mu\hat{N}} \right]} \\ &= \frac{\sum_\sigma f_A(\sigma) w(\sigma)}{\sum_{\sigma'} w_{\sigma'}}, \end{aligned} \tag{4.21}$$

with

$$w(\sigma) = \hat{\text{Tr}} \left[\hat{U}_\sigma e^{\beta\mu\hat{N}} \right] \tag{4.22}$$

$$f_A(\sigma) = \hat{\text{Tr}} \left[\hat{A} \hat{U}_\sigma e^{\beta\mu\hat{N}} \right] / w(\sigma). \tag{4.23}$$

These quantities can be evaluated exactly for any given configuration σ. However, there are too many configurations σ to sum them all up. Expression 4.21 has a form that can be

evaluated using the Markov-chain Monte-Carlo methods from chapter 3. How to proceed if part of the weights $w(\sigma)$ are negative, is explained in section 4.5.
The Monte-Carlo technique that is obtained by evaluating expression 4.21 using Markov-chain Monte-Carlo techniques, is often called the 'Grand-Canonical Monte-Carlo Method' or the 'Determinant Monte-Carlo method' [7]. The method has been used extensively for the study of condensed matter systems such as the Hubbard model. For the decomposition of the Boltzmann operator, use is made of the Hubbard-Stratonovich transform from section 2.2.1 or Hirsch's discrete Hubbard-Stratonovich transform from section 2.2.2. Because of the 'auxiliary fields' σ arising in the Hubbard-Stratonovich transform 2.26, the method is sometimes also denoted as 'auxiliary field Monte-Carlo', though this name is also used for other Monte-Carlo techniques that rely on the Hubbard-Stratonovich transform.

4.2.1 Evaluation of weights and observables in the grand canonical ensemble.

In section 1.2.4 it was shown that the grand canonical trace of an operator $\hat{U}$ that transforms Slater determinants into Slater determinants, is given by

$$\hat{\mathrm{Tr}}\left[\hat{U}e^{\beta\mu\hat{N}}\right] = \det\left(1+e^{\beta\mu}U\right), \tag{4.24}$$

where U is the $N_{\mathcal{S}}\times N_{\mathcal{S}}$ matrix representation of the operator $\hat{U}$. The operator $\hat{U}_\sigma = e^{\hat{S}_\sigma(\beta)}$ in the decomposition 4.1 for the Boltmann operator is such an operator. Its matrix representation is given by $U_\sigma = e^{\left[\hat{S}_\sigma(\beta)\right]}$. This leads to a matrix expression for $w(\sigma)$:

$$\begin{aligned} w(\sigma) &= \det\left(1+e^{\beta\mu}e^{\left[\hat{S}_\sigma(\beta)\right]}\right) \\ &= \det\left(1+\chi U_\sigma\right), \end{aligned} \tag{4.25}$$

with $\chi = e^{\beta\mu}$. This determinant can easily be evaluated using standard linear algebra techniques. The well known 'LU'-decomposition method requires about $2N_{\mathcal{S}}^3$ floating point operations (flops) [11]. For ill conditioned matrices, the singular value decomposition is more suited. It requires about $\frac{8}{3}N_{\mathcal{S}}^3$ flops for the calculation of the determinant [11].
In order to calculate the grand canonical expectation value of an operator $\hat{A}$, we not only need to calculate the trace of $\hat{U}_\sigma$ but also of $\hat{A}\hat{U}_\sigma$. In general this last operator has a different nature from the former one, because $\hat{A}\hat{U}_\sigma$ is not a product of exponentials of one-body operators. The expression 4.24 cannot be used. If $\hat{A}$ is a one-body operator, we can get around this problem. We define the operator $\hat{Q}_{\hat{A}}(\epsilon)$ as the operator which transforms a Slater determinant Ψ_M, represented by the $N_{\mathcal{S}}\times N_{\mathcal{S}}$ matrix M, into the Slater determinant $\Psi_{M'}$ represented by

$$M' = (1+\epsilon A)\,M, \tag{4.26}$$

where $A = \left[\hat{A}\right]$ is the matrix representation of $\hat{A}$ in the one-particle space. Note that

$$\left.\frac{d}{d\epsilon}\hat{Q}_{\hat{A}}(\epsilon)\right|_{\epsilon=0} = \hat{A}, \tag{4.27}$$

because the operator $\hat{Q}_{\hat{A}}(\epsilon)$ is equal to the operator $e^{\epsilon\hat{A}}$ up to the first order in ϵ. From this we obtain the expression

$$\begin{aligned}\hat{\mathrm{Tr}}\left(\hat{A}\hat{U}_\sigma e^{\beta\mu\hat{N}}\right) &= \frac{d}{d\epsilon}\hat{\mathrm{Tr}}\left[\hat{Q}_{\hat{A}}(\epsilon)\hat{U}_\sigma e^{\beta\mu\hat{N}}\right]\Big|_{\epsilon=0} \\ &= \frac{d}{d\epsilon}\det\left[1+\chi\left(1+\epsilon A\right)U_\sigma\right]\Big|_{\epsilon=0}. \end{aligned} \tag{4.28}$$

The derivative in ϵ can be evaluated by taking a small but finite value for ϵ. This leads to

$$\hat{\mathrm{Tr}}\left(\hat{A}\hat{U}_\sigma\right) \simeq \frac{\det\left[1+\chi\left(1+\bar{\epsilon}A\right)U_\sigma\right]-\det\left[1+\chi U_\sigma\right]}{\bar{\epsilon}}, \tag{4.29}$$

with $\bar{\epsilon}$ a small constant, small enough to make the systematic error on 4.29 negligible compared to the statistical error originating from the Monte-Carlo procedure. A compact notation for $f_A(\sigma)$ is

$$f_A(\sigma) = \hat{\mathrm{Tr}}\left[\hat{A}\hat{U}_\sigma\right]/w(\sigma) = \frac{d}{d\epsilon}\ln\left|\hat{\mathrm{Tr}}\left[\hat{Q}_{\hat{A}}(\epsilon)\hat{U}_\sigma\right]\right|\Big|_{\epsilon=0}. \tag{4.30}$$

Another way to calculate $\hat{\mathrm{Tr}}\left(\hat{A}\hat{U}_\sigma e^{\beta\mu\hat{N}}\right)$ is obtained by manipulating the determinant in expression 4.28 such that one gets

$$\begin{aligned}\hat{\mathrm{Tr}}\left(\hat{A}\hat{U}_\sigma e^{\beta\mu\hat{N}}\right) &= \frac{d}{d\epsilon}\det\left[1+\chi\left(1+\epsilon A\right)U_\sigma\right]\Big|_{\epsilon=0} \\ &= \det\left(1+\chi U_\sigma\right)\frac{d}{d\epsilon}\det\left(1+\epsilon A\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)\Big|_{\epsilon=0} \\ &= \det\left(1+\chi U_\sigma\right)\mathrm{Tr}\left(A\frac{\chi U_\sigma}{1+\chi U_\sigma}\right). \end{aligned} \tag{4.31}$$

The notation Tr is used for the matrix trace, in contrast to the notation $\hat{\mathrm{Tr}}$ that is used for the many-body trace. Because $\det\left(1+xU_\sigma\right) = w(\sigma)$, it follows from 4.23 and 4.31 that

$$f_A(\sigma) = \mathrm{Tr}\left(A\frac{\chi U_\sigma}{1+\chi U_\sigma}\right). \tag{4.32}$$

A particular type of one-body operator is $\hat{A} = \hat{a}_j^\dagger\hat{a}_k$. Its expectation value gives an element of the one-body density matrix ρ^1:

$$\rho^1_{kj} = \langle\hat{a}_j^\dagger\hat{a}_k\rangle_{GC}. \tag{4.33}$$

From expression 4.32 we obtain that for this operator

$$\begin{aligned} f_A(\sigma) &= \mathrm{Tr}\left(a_j^\dagger a_k\frac{\chi U_\sigma}{1+\chi U_\sigma}\right) \\ &= \left(\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)_{kj} \end{aligned} \tag{4.34}$$

Note that here a_k denotes the row vector which has a 1 on the k^{th} entry and zeros in all other entries. Using the notation defined in expression 3.1, we can write that ρ^1 is the expectation value of the matrix R^1_σ

$$\rho^1 = \mathrm{E}\left(R^1_\sigma\right), \tag{4.35}$$

with the matrix R^1_σ given by

$$R^1_\sigma = \frac{\chi U_\sigma}{1+\chi U_\sigma}. \tag{4.36}$$

Expectation values for a two-body operator $\hat{B}$ can be obtained by decomposing the two-body operator as a sum of products of one-body operators:

$$\hat{B} = \sum_i \hat{A}_{1i}\hat{A}_{2i}. \tag{4.37}$$

The grand canonical trace for a product of two one-body operators $\hat{B} = \hat{A}_1\hat{A}_2$, can be evaluated as

$$\hat{\mathrm{Tr}}\left[\hat{A}_1\hat{A}_2\hat{U}_\sigma e^{\beta\mu\hat{N}}\right] = \frac{d}{d\epsilon_1}\frac{d}{d\epsilon_2}\,\hat{\mathrm{Tr}}\left[\hat{Q}_{\hat{A}_1}(\epsilon_1)\,\hat{Q}_{\hat{A}_2}(\epsilon_2)\,\hat{U}_\sigma e^{\beta\mu\hat{N}}\right]\Big|_{\epsilon_1=0,\,\epsilon_2=0}. \tag{4.38}$$

Again, the derivatives can be evaluated by taking small but finite values for ϵ_1 and ϵ_2. Alternatively, one can elaborate the formula further by manipulating the determinants:

$$\begin{aligned}&\hat{\mathrm{Tr}}\left(\hat{A}_1\hat{A}_2\hat{U}_\sigma e^{\beta\mu\hat{N}}\right)\\ &= \frac{d}{d\epsilon_1}\frac{d}{d\epsilon_2}\det\left[1+\chi\,(1+\epsilon_1A_1)\,(1+\epsilon_2A_2)\,U_\sigma e^{\beta\mu\hat{N}}\right]\Big|_{\epsilon_1=0,\,\epsilon_2=0}.\end{aligned} \tag{4.39}$$

Analogously to the reasoning in 4.31, we can transfrom this expression in the form

$$\begin{aligned}&\hat{\mathrm{Tr}}\left(\hat{A}_1\hat{A}_2\hat{U}_\sigma\right)\\ &= \frac{d}{d\epsilon_2}\det\left[1+\chi\,(1+\epsilon_2A_2)\,U_\sigma\right]\mathrm{Tr}\left[A_1\frac{\chi\,(1+\epsilon_2A_2)\,U_\sigma}{1+\chi\,(1+\epsilon_2A_2)\,U_\sigma}\right]\Big|_{\epsilon_2=0}.\\ &= \det\left(1+\chi U_\sigma\right)\,f_B(\sigma),\end{aligned} \tag{4.40}$$

with $f_B(\sigma)$ for a two-body operator $\hat{B} = \hat{A}_1\hat{A}_2$ given by

$$\begin{aligned}f_B(\sigma) &= \mathrm{Tr}\left(A_1\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)\mathrm{Tr}\left(A_2\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)\\ &\quad+\mathrm{Tr}\left(A_1A_2\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)-\mathrm{Tr}\left(A_1\frac{\chi U_\sigma}{1+\chi U_\sigma}A_2\frac{\chi U_\sigma}{1+\chi U_\sigma}\right).\end{aligned} \tag{4.41}$$

This procedure can be extended to operators of any order. The calculations will require more and more computational effort as the rank gets higher.
A particular type of two-body operator is $\hat{B} = \hat{a}^\dagger_j\hat{a}^\dagger_k\hat{a}_l\hat{a}_m$. Its expectation value gives an element of the two-body density matrix ρ^2:

$$\rho^2_{mlkj} = \langle\hat{a}^\dagger_j\hat{a}^\dagger_k\hat{a}_l\hat{a}_m\rangle. \tag{4.42}$$

The operator B can be written as a product of one-body operators in the following way:

$$\hat{B} = \left(\hat{a}_j^\dagger \hat{a}_m\right)\left(\hat{a}_k^\dagger \hat{a}_l\right) - \delta_{mk}\hat{a}_j^\dagger \hat{a}_l. \tag{4.43}$$

Using expressions 4.31, 4.34, 4.36 and 4.40, one obtains that

$$\rho^2_{mlkj} = \mathrm{E}(R^2_{mlkj}), \tag{4.44}$$

with R^2_{mlkj} given by

$$R^2_{mlkj} = R^1_{mj}R^1_{lk} - R^1_{lj}R^1_{mk}. \tag{4.45}$$

Relation 4.45 shows a strong analogy with the relation between the two-body density matrix and the one-body density matrix for a pure Slater determinant Ψ, derived by Löwdin [34] in the framework of Hartree-Fock theory:

$$\varrho^2_{mlkj} = \varrho^1_{mj}\varrho^1_{lk} - \varrho^1_{lj}\varrho^1_{mk}, \tag{4.46}$$

with ρ^1 and ρ^2 here defined as:

$$\varrho^1_{kj} = \langle\Psi|\hat{a}_j^\dagger \hat{a}_k|\Psi\rangle \tag{4.47}$$

$$\varrho^2_{mlkj} = \langle\Psi|\hat{a}_j^\dagger \hat{a}_k^\dagger \hat{a}_l \hat{a}_m|\Psi\rangle. \tag{4.48}$$

In fact, relation 4.45 indicates that relation 4.46 also holds for the grand canonical one-body and two-body density matrix for a system with a one-body Hamiltonian (a mean field) at finite temperature (in the grand canonical ensemble). As such, it is a finite-temperature extension of relation 4.46.
An alternative way to calculate the weigth for an operator $\hat{A}$, is based on the derivative of the exponential of an operator.

$$\begin{aligned}\frac{de^{\hat{X}(\epsilon)}}{d\epsilon} = &\left\{\frac{d\hat{X}}{d\epsilon}(\epsilon) + \frac{1}{2}\left[\hat{X}(\epsilon), \frac{d\hat{X}}{d\epsilon}(\epsilon)\right] + \frac{1}{6}\left[\hat{X}(\epsilon), \left[\hat{X}(\epsilon), \frac{d\hat{X}}{d\epsilon}(\epsilon)\right]\right]\right. \\ &\left. + \frac{1}{24}\left[\hat{X}(\epsilon), \left[\hat{X}(\epsilon), \left[\hat{X}(\epsilon), \frac{d\hat{X}}{d\epsilon}(\epsilon)\right]\right]\right] + \ldots\right\} e^{\hat{X}(\epsilon)}.\end{aligned} \tag{4.49}$$

On taking the trace of both sides, the terms with commutators vanish, so one gets the result

$$\hat{\mathrm{Tr}}\left(\frac{d}{d\epsilon}e^{\hat{X}(\epsilon)}\right) = \hat{\mathrm{Tr}}\left(\frac{d\hat{X}}{d\epsilon}(\epsilon)e^{\hat{X}(\epsilon)}\right). \tag{4.50}$$

This expression can be used to derive the following expression for the grand canonical expectation value of an operator $\hat{A}$:

$$\hat{\mathrm{Tr}}\left(\hat{A}e^{-\beta\hat{H}+\beta\mu\hat{N}}\right) = \frac{d}{d\epsilon}\hat{\mathrm{Tr}}\left(e^{-\beta\hat{H}+\beta\mu\hat{N}+\epsilon\hat{A}}\right)\Big|_{\epsilon=0}. \tag{4.51}$$

To evaluate the trace on the right hand side one needs a decomposition analogous to the decomposition 4.1 for $e^{-\beta\hat{H}}$. If $\hat{A}$ is a one-body operator or if $\hat{A}$ is a part of the two-body Hamiltonian that is treated seperately in the decomposition of $e^{-\beta\hat{H}}$, then a decomposition

for $e^{-\beta\hat{H}+\epsilon\hat{A}}$ exists that is based on the same configurations σ as 4.1, with slightly modified operators $\hat{H}'_0(\epsilon)$ or $\hat{A}'_{\sigma_i}(\beta,\epsilon)$. Then the weights for the Monte-Carlo sampling are given by

$$w(\sigma) = \hat{\mathrm{Tr}}\left[\hat{U}_\sigma e^{\beta\mu\hat{N}}\right] \tag{4.52}$$

$$\begin{aligned} f'_A(\sigma) &= \frac{d}{d\epsilon}\hat{\mathrm{Tr}}\left[\hat{U'}_\sigma(\epsilon)e^{\beta\mu\hat{N}}\right]\Big|_{\epsilon=0} / w(\sigma) \\ &= \frac{d}{d\epsilon}\det\left[1+\chi U'_\sigma(\epsilon)\right]\Big|_{\epsilon=0} / w(\sigma) \end{aligned} \tag{4.53}$$

with $\hat{U'}_\sigma(\epsilon)$ given by

$$\begin{aligned} &\hat{U'}_\sigma(\epsilon) \\ &= e^{-\hat{S'}_\sigma(\beta,\epsilon)} \\ &= e^{-\frac{\beta\hat{H'}_0(\epsilon)}{2}} e^{-\hat{A'}_{\sigma_1}(\beta,\epsilon)} e^{-\beta\hat{H'}_0(\epsilon)} e^{-\hat{A'}_{\sigma_2}(\beta,\epsilon)} e^{-\beta\hat{H'}_0(\epsilon)} \\ &\quad \cdots e^{-\beta\hat{H'}_0(\epsilon)} e^{-\hat{A'}_{\sigma_{N_t}}(\beta,\epsilon)} e^{-\frac{\beta\hat{H'}_0(\epsilon)}{2}}. \end{aligned} \tag{4.54}$$

Because $\hat{U'}_\sigma(\epsilon) = \hat{U}_\sigma$ for $\epsilon = 0$, we can rewrite $f'_A(\sigma)$ as

$$f'_A(\sigma) = \frac{d}{d\epsilon}\ln\left|\det\left[1+\chi U'_\sigma(\epsilon)\right]\right|\Big|_{\epsilon=0}. \tag{4.55}$$

Again, the derivative in ϵ can be evaluated by taking a small but finite value for ϵ. At first glance, it looks like this method for the evaluation of observables is far less efficient than the method based on expression 4.32 or 4.41 because one has to calculate the matrix U'_σ. This requires $2N_t$ matrix-matrix multiplications. On the other hand, expression 4.55 exploits the permutational symmetry of the trace:

$$\hat{\mathrm{Tr}}(\hat{A}\hat{B}\hat{C}) = \hat{\mathrm{Tr}}(\hat{C}\hat{A}\hat{B}) = \hat{\mathrm{Tr}}(\hat{B}\hat{C}\hat{A}). \tag{4.56}$$

Cyclic permutation of the components $\sigma_1, \sigma_2, \ldots \sigma_{N_t}$ of the configuration vector σ, corresponds to a cyclic permutation of the inverse-temperature slices in the Suzuki-Trotter decomposition 2.10. Because of the permutational symmetry of the trace, such a permutation will not alter the value of $w(\sigma)$. The factor $\hat{Q}_{\hat{A}}$ in expression 4.30 breaks this symmetry. Therefore, $f_A(\sigma)$ is not invariant under cyclic permutations of the components of σ, while $f'_A(\sigma)$ is. $f'_A(\sigma)$ equals the value obtained by averaging $f_A(\sigma)$ over all cyclic permutations of σ. It corresponds to inserting the factor $\hat{Q}_{\hat{A}}$ in every inverse-temperature slice of $\hat{U}_\sigma$. Therefore the variance on $f'_A(\sigma)$ will be much smaller than the variance on $f_A(\sigma)$. Furthermore, the method based on $f'_A(\sigma)$ has the advantage that it is easy to code in computer-programming language. The fact that it requires more computer time is not a big disadvantage, because most computer time in SDQMC calculations goes to the evaluation of the weights $w(\sigma)$, if the sampling interval is well chosen (see section 3.3.2). The systematic error on $f'_A(\sigma)$ will be somewhat bigger than the systematic error on $f_A(\sigma)$, because of the systematic errors in the decomposition of $e^{-\beta\hat{H}+\beta\mu\hat{N}+\epsilon\hat{A}}$.

A particular operator for which the method based on expression 4.55 is useful, is the Hamiltonian $\hat{H}$. Its grand canonical expectation value is related to the internal energy of the system:

$$U = \langle \hat{H} \rangle_{GC}. \tag{4.57}$$

Taking the derivative to an auxiliary variable ϵ is equivalent to taking the derivative to β. Expression 4.55 becomes

$$f'_H(\sigma) \;=\; -\frac{d}{d\beta'} \ln\left[1 + \chi(\beta) U_\sigma(\beta')\right]\Big|_{\beta'=\beta}. \tag{4.58}$$

This is also the direct analog of the thermodynamical expression 4.10. The thermodynamic quantities cited in section 4.1 can be determined from the thermodynamical relations among these quantities. The logarithm of the grand partition function can be obtained by numerically integrating $-U + \mu N = \frac{d}{d\beta} \ln(Z_\beta)$ from 0 to β. Expression 4.13 immediately gives the free energy F, while expression 4.12 can be used to obtain the entropy. Obtaining the specific heat in the grand canonical ensemble is more complicated, because it requires the derivative of U to β under the condition that N remains constant. Changing the temperature will also change the number of particles, so a correction has to be made on the expression

$$C = -\beta^2 \frac{dU}{d\beta}. \tag{4.59}$$

These corrections require the evaluation of $\langle \hat{H}\hat{N} \rangle$ and $\langle \hat{N}^2 \rangle$. From the total derivative

$$\left.\frac{dU}{d\beta}\right|_{N \text{ constant}} = \frac{\partial U}{\partial \beta} + \frac{\partial U}{\partial \mu} \left.\frac{d\mu}{d\beta}\right|_{N \text{ constant}}, \tag{4.60}$$

and from the condition

$$\left.\frac{d\langle \hat{N} \rangle}{d\beta}\right|_{N \text{ constant}} = 0, \tag{4.61}$$

one can obtain an expression for the corrections term on 4.59. The evaluation of the specific heat is easier in the canonical ensemble, where the number of particles is fixed.

Rank one updates for MCMC sampling in the grand canonical ensemble

In SDQMC based on a Metropolis sampling algorithm, the weight $w(\sigma')$ from expression 4.25 has to be evaluated for configurations σ' that differ only in a few components from a previous configuration σ. If the decomposition of the exponential of the two-body Hamiltonian is based on rank-one or rank-two operators, as explained in section 2.2.2, then a fast updating scheme for the $w(\sigma')$ can be used [35]. It requires that $U_{\sigma'}$ can be written as

$$U_{\sigma'} = U_\sigma + x b_1^\dagger b_2, \tag{4.62}$$

where b_1 and b_2 are row vectors. We discuss it here for Hirsch's discrete Hubbard-Stratonovich transformation (see section 2.2.2), but it applies more generally to decompositions based on rank-one or rank-two operators. In the case of Hirsch's discrete

Hubbard-Stratonovich transformation, flipping the component of σ for the i^{th} lattice point in the j^{th} inverse-temperature slice transfroms $\hat{U}_\sigma$ in the following way:

$$\begin{aligned} \hat{U}_\sigma &= \hat{U}_{L\sigma} e^{2a\sigma_{ij}(\hat{n}_{\uparrow i} - \hat{n}_{\downarrow i})} \hat{U}_{R\sigma} \\ &\downarrow \\ \hat{U}_{\sigma'} &= \hat{U}_{L\sigma} e^{-2a\sigma_{ij}(\hat{n}_{\uparrow i} - \hat{n}_{\downarrow i})} \hat{U}_{R\sigma}. \end{aligned} \tag{4.63}$$

The index L (R) denotes the part operator $\hat{U}_\sigma$ that is obtained by multiplying all the operators for the inverse-temperature slices left (right) of the j^{th} slice. The corresponding transform on $U_{\uparrow\sigma}$ is given by 4.62, with

$$x = -2\sinh(2a\sigma_{ij}), \tag{4.64}$$

$$b_1^\dagger = U_{\uparrow L\sigma} e_i^\dagger, \tag{4.65}$$

$$b_2 = e_i U_{\uparrow R\sigma}, \tag{4.66}$$

where e_i is the row vector with 1 on the i^{th} entry and zero's on the other entries. The weight attributed to the new configuration is now given by

$$\begin{aligned} w_\uparrow(\sigma') &= \det(1 + \chi U_{\uparrow\sigma'}) \\ &= \det\left(1 + \chi U_{\uparrow\sigma} + \chi x b_1^\dagger b_2\right) \\ &= \det(1 + \chi U_{\uparrow\sigma}) \det\left(1 + \frac{1}{1 + \chi U_{\uparrow\sigma}} \chi x b_1^\dagger b_2\right) \\ &= w_\uparrow(\sigma)\left(1 + \chi x b_2 \frac{1}{1 + \chi U_{\uparrow\sigma}} b_1^\dagger\right) \end{aligned} \tag{4.67}$$

Thus $w_\uparrow(\sigma')$ can be calculated without computing $U_{\uparrow\sigma'}$. Instead of N_t matrix matrix multiplications, the calculation involves only N_t matrix vector multiplications. This reduces the number of needed flops with a factor $N_\mathcal{S}$. If the Metropolis trial move σ' is accepted, then $(1 + \chi U_{\uparrow\sigma})^{-1}$ has to be determined in order to apply expression 4.67 for the new trial moves. This can be done in $\mathcal{O}(N_\mathcal{S}^2)$ flops if the matrix inversion is based on a QR decomposition [12]. Note that if this decomposition is known, the inverse has not to be calculated explicitly in order to evaluate 4.67. Besides, the matrix $(1 + \chi U_{\uparrow\sigma})^{-1}$ can be used for the evaluation of expectation values of observables using expressions 4.32 or 4.41. After a number of updates, the factorization of $(1 + \chi U_{\uparrow\sigma})^{-1}$ degrades due to rounding errors and has to be recomputed from scratch.
This scheme can be improved further if one updates σ only in the last inverse-temperature slice. Then, in the case where b_1 and b_2 are given by 4.65, 4.66, one has that

$$b_1^\dagger = e_i^\dagger, \tag{4.68}$$

$$b_2 = e_i U_{\uparrow\sigma}. \tag{4.69}$$

Expression 4.67 can now be written as

$$w_\uparrow(\sigma') = w_\uparrow(\sigma)\left[1 + x\left(1 - \frac{1}{1 + \chi U_{\uparrow\sigma}}\right)_{ii}\right]. \tag{4.70}$$

In this case no matrix vector multiplications are needed and $(1+\chi U_{\uparrow\sigma})^{-1}$ needs to be updated only if the trial move is accepted. This update can be done in $\mathcal{O}\left(N_{\mathcal{S}}^2\right)$ flops. After a few steps by changing the last slice only, the last-but-one slice can be brought to the last position using the cyclic permutation symmetry of the grand canonical trace. This requires only two matrix matrix multiplications. If the decomposition of $e^{-\beta\hat{H}}$ is based on diagonal and rank-one matrices, this cyclic permutation of one inverse temperature slice requires only $\mathcal{O}\left(N_{\mathcal{S}}^2\right)$ flops too. A minor disadvantage of these cyclical updates is that they amount to a deterministic-scan Gibbs sampler, while a random-scan sampler can be expected to lead to shorter autocorrelations (see section 3.4.2).
In some cases, several steps of the form 4.62 will be needed to transform U_σ in $U_{\sigma'}$. The procedure then will have to be repeated a number of times for one update. The major limitation of this scheme is that it only allows trial moves where one component of σ is changed. In a lot of cases, trial moves where more components are changed, will lead to a more efficient sampling. Instead of repeating the rank-one updating scheme a number of times, it can be more efficient to calculate U_σ from scratch for every trial move. A scheme that reduces the number of matrix multiplications considerably in such cases, is given in section 4.6.2.

4.3 SDQMC in the canonical ensemble

The canonical ensemble is obtained by restricting the grand canonical ensemble to states with a fixed number of particles N. The SDQMC can be applied in the canonical ensemble in an analogous way as in the grand canonical ensemble. Analogons for the grand canonical expressions 4.21, 4.22 and 4.23 are found by replacing the grand canonical trace operator with the canonical trace operator for the N-particle states and by omitting the factor with the chemical potential μ:

$$\langle\hat{A}\rangle_C = \frac{\sum_\sigma f_A(\sigma)w(\sigma)}{\sum_{\sigma'} w(\sigma')} \tag{4.71}$$

$$w(\sigma) = \hat{\mathrm{Tr}}_N\left[\hat{U}_\sigma\right] \tag{4.72}$$

$$f_A(\sigma) = \hat{\mathrm{Tr}}_N\left[\hat{A}\hat{U}_\sigma\right]/w(\sigma). \tag{4.73}$$

Now we need a way to evaluate the N-particle trace. Formally, one can obtain expressions for the canonical trace from expressions for the grand canonical trace by taking the derivative of the latter to the variable $\chi = e^{\beta\mu}$. Let the grand canonical trace for an operator $\hat{U}$ be given by

$$\hat{\mathrm{Tr}}\left(\hat{U}e^{\beta\mu\hat{N}}\right) = \hat{\mathrm{Tr}}\left(\hat{U}\chi^{\hat{N}}\right), \tag{4.74}$$

then the canonical expectation value of $\hat{A}$ can be written as

$$\hat{\mathrm{Tr}}_N\left(\hat{U}\right) = \left(\frac{d}{d\chi}\right)^N \hat{\mathrm{Tr}}\left(\hat{U}\chi^{\hat{N}}\right)\Big|_{\chi=0}. \tag{4.75}$$

For the weight $w(\sigma)$ this leads to the formal expression

$$w(\sigma) = \hat{\mathrm{Tr}}_N\left[\hat{U}_\sigma\right]$$

$$= \left(\frac{d}{d\chi}\right)^N \hat{\mathrm{Tr}}\left(\hat{U}_\sigma \chi^{\hat{N}}\right)\Big|_{\chi=0}$$
$$= \left(\frac{d}{d\chi}\right)^N \det\left(1+\chi U_\sigma\right)\Big|_{\chi=0}. \tag{4.76}$$

The canonical weight $w(\sigma)$ is given by the coefficient of χ^N in the polynomial $\det(1+\chi U_\sigma)$. This polynomial is closely related to the characteristic polynomial of the matrix U_σ, as is explained in the next sections. Thus the calculation of canonical traces for SDQMC amounts to the evaluation of the characteristic polynomial of the matrices U_σ. For this purpose we developed an accurate and fast algorithm. Accuracy is very important here, because we also want to evaluate the first and second derivatives of the canonical traces, for the calculation of observables. Speed is important, because canonical weights have to be determined for every Markov step. The Metropolis sampling scheme can be optimized in such a way that the evaluation of the trace becomes one of the bottlenecks of SDQMC programs.

4.3.1 Numerical evaluation of canonical traces

In order to calculate the canonical trace numerically, several methods have been suggested by Lang *et al* [37]. One can start from the relation

$$\det\left(1+\chi U_\sigma\right) = e^{\mathrm{Tr}[\ln(1+\chi U_\sigma)]} \tag{4.77}$$
$$= \exp\left[\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}\chi^n \mathrm{Tr}\left(U_\sigma^n\right)\right]. \tag{4.78}$$

Picking out the coefficient of χ^N in the series expansion of both sides gives a relation between $\hat{\mathrm{Tr}}_N\left(\hat{U}_\sigma\right)$ and $\mathrm{Tr}\left(U_\sigma\right), \mathrm{Tr}\left(U_\sigma^2\right), \ldots, \mathrm{Tr}\left(U_\sigma^N\right)$. Though mathematically elegant, this formula is unpractical: it is inaccurate because it is very sensitive to roundoff errors, especially if the eigenvalues of U_σ differ by several orders of magnitude, which is common in SDQMC. It is also inefficient because it requires $N/2$ matrix multiplications. Therefore the method is only useful for small N. Lang *et al* also suggest another method: calculation of the canonical trace using coherent states. This method requires the integration over $N_{\mathcal{S}}^2$ additional fields. However, since the integration is carried out with a Monte-Carlo algorithm, this requires only slightly more computational effort than the previous method. But it aggravates the 'sign problem' (see section 4.5). A third method, suggested by the same group in another paper [38] and claimed to be better than the previous ones, uses the operator

$$\hat{P}_N = e^{-\beta\mu N}\int_0^{2\pi}\frac{d\phi}{2\pi}e^{-i\phi N}e^{(\beta\mu+i\phi)\hat{N}}, \tag{4.79}$$

that projects the ensemble with N particles out of the grand canonical ensemble. Here, $\hat{N}$ is the number operator. The parameter μ is arbitrary here and chosen to minimize numerical instabilities. The canonical trace is then given by

$$\hat{\mathrm{Tr}}_N\left(\hat{U}_\sigma\right) = e^{-\beta\mu N}\int_0^{2\pi}\frac{d\phi}{2\pi}e^{-i\phi N}\hat{\mathrm{Tr}}\left(e^{(\beta\mu+i\phi)\hat{N}}\hat{U}_\sigma\right) \tag{4.80}$$

$$= e^{-\beta\mu N} \int_0^{2\pi} \frac{d\phi}{2\pi} e^{-i\phi N} \prod_\lambda \left(1 + e^{i\phi} e^{\beta(\mu-\epsilon_\lambda)}\right), \tag{4.81}$$

where ϵ_λ is the λ^{th} eigenvalue of the matrix U_σ. This matrix U_σ is used in diagonalized form because this is numerically favorable: the grand canonical trace can now be evaluated by multiplying $N_\mathcal{S}$ scalar factors, otherwise one has to evaluate a $N_\mathcal{S} \times N_\mathcal{S}$ determinant for every value of ϕ. The integration can be carried out exactly with an $N_\mathcal{S}$-point quadrature formula. This method is stable if μ is well chosen. The value $\mu = \left(\mathrm{Re}\left(\epsilon_N\right) + \mathrm{Re}\left(\epsilon_{N+1}\right)\right)/2$ is suggested. However, if U_σ is diagonalized, the canonical trace can be evaluated much more easily by explicit construction of the polynomial

$$\det\left(1 + \chi U_\sigma\right) = \prod_{i=1}^{N_\mathcal{S}} \left(1 + \chi\epsilon_i\right). \tag{4.82}$$

If this polynomial in χ is constructed from the smallest up to the largest eigenvalue, it can be computed in an easy and stable way. $\hat{\mathrm{Tr}}_N\left(\hat{U}_\sigma\right)$ is then given by the coefficent of χ^N. The polynomial can be constructed even more efficiently without diagonalization of the matrix U_σ, as is explained in the next section.

4.3.2 Algorithm for the calculation of the characteristic polynomial of a general square matrix

The characteristic polynomial of an $N_\mathcal{S} \times N_\mathcal{S}$ matrix U is given by

$$P_U\left(\chi\right) = \det\left(U - \chi\right). \tag{4.83}$$

The coefficient of χ^N in $P_U\left(\chi\right)$ is equal to $(-1)^N$ times the coefficient of $\chi^{(N_\mathcal{S}-N)}$ in

$$P'_U\left(\chi\right) = \det\left(1 + \chi U\right). \tag{4.84}$$

The basic idea of the algorithm is to consider $1 + \chi U$ as a matrix of polynomials in χ. The determinant in equation 4.84 can be calculated using Gaussian elimination, with polynomials instead of scalars as matrix elements. Because the multiplication of two polynomials of degree N requires about $2N^2$ flops and the calculation of a determinant about $N_\mathcal{S}^3/3$ polynomial multiplications, the calculation would require a number of the order of $N_\mathcal{S}^5$ flops, which is too much for an efficient implementation. This number can be drastically reduced if U is transformed to an upper-Hessenberg form by a similarity transformation (a Householder reduction to Hessenberg form requires approximatly $\frac{10}{3}N_\mathcal{S}^3$ flops [11]). This leaves the coefficients of $P'_U\left(\chi\right)$ unchanged because

$$\det\left[1 + \chi\left(Q^{-1}UQ\right)\right] = \det\left[Q^{-1}\left(1 + \chi U\right)Q\right] \tag{4.85}$$

$$= \det\left(1 + \chi U\right). \tag{4.86}$$

In order to calculate the determinant we transform $1 + \chi U$ to upper diagonal form by Gaussian elimination, requiring now only $N_\mathcal{S}^2$ polynomial multiplications. The Gaussian elemination is performed from the right bottom corner of the matrix up to the top left corner

because in SDQMC the right bottom corner often contains the smallest elements, so that the summations involved are performed from small to large terms. This is less sensitive to roundoff errors than the summation the other way round. We start with T^{N_S}=$1+\chi U$. Now we bring column after column in upper triangular form. Suppose that T^j has columns j to $N_{\mathcal{S}}$ already in upper triangular form, i.e.

$$T^j_{i\,k} = 0, \tag{4.87}$$

for $i > k$ and $k \geq j$. Now we calculate

$$T^{j-1} = T^j G^j, \tag{4.88}$$

where

$$G^j_{i\,k} = \delta_{i,k}, \tag{4.89}$$

except for

$$\begin{array}{rcl} G^j_{j-1\,j-1} & = & T^j_{j\,j}, \\ G^j_{j\,j-1} & = & -T^j_{j\,j-1}. \end{array} \tag{4.90}$$

In the end we obtain the upper triangular matrix

$$T^1 = T^{N_S}\, G^{N_S}\, G^{N_S-1} \cdots G^2, \tag{4.91}$$

so that

$$\begin{array}{rcl} P'_U(\chi) & = & \det(1+\chi U) \\ & = & \det\left(T^{N_S}\right) \\ & = & \dfrac{\det(T^1)}{\det\left(G^{N_S}\, G^{N_S-1} \cdots G^2\right)} \\ & = & \dfrac{\prod_{i=1}^{N_S} T^1_{ii}}{\prod_{i=2}^{N_S} T^i_{ii}} \\ & = & T^1_{1\,1}. \end{array} \tag{4.92}$$

because $T^1_{ii} = T^i_{ii}$. The operations can be ordered to minimize memory use. This leads to the following algorithm ($t_{k\,i}$ corresponds with the coefficient of χ^k in $T^j_{i\,j}$) :

algorithm for calculating the coefficients of the characteristic polynomial of a $\mathbf{N_S \times N_S}$ matrix U

reduce U to upper Hessenberg form
DO $j = N_{\mathcal{S}}, 1, -1$
  DO $i = 1, j$
    DO $k = N_{\mathcal{S}} - j, 1, -1$
      $t_{k+1\,i} = U_{i\,j}\, t_{k\,j+1} - U_{j+1\,j}\, t_{k\,i}$
    ENDDO
    $t_{1\,i} = U_{i\,j}$
  ENDDO
  DO $k = 1, N_{\mathcal{S}} - j$
    $t_{k\,j} = t_{k\,j} + t_{k\,j+1}$
  ENDDO
ENDDO

(4.93)

In the end $t_{k\,1}$ is the coefficient of χ^k in $P'_U(\chi)$. This algorithm cannot break down and requires $N_{\mathcal{S}}^3/2 + N_{\mathcal{S}}^2 - N_{\mathcal{S}}/2$ flops. If one needs only the coefficient of χ^N, e.g. for the calculation of an N-particle trace in SDQMC, the number of flops can be reduced further by calculating the polynomials only up to degree N. This restricts the loops over k to values smaller than or equal to N:

$$\text{DO } \; k = \text{MAX}\left(N_{\mathcal{S}} - j, N\right), 1, -1. \tag{4.94}$$

Together with the Householder reduction to the upper Hessenberg form this makes less than $4N_{\mathcal{S}}^3$ flops. Diagonalization of the matrix U would require about $10N_{\mathcal{S}}^3$ flops with the QR algorithm [11].

Numerical tests

We have tested our algorithm numerically on its speed and accuracy. Al the tests were done in Fortran77 (DEC Fortran V3.8) on a Digital Alphastation 255/300MHz workstation running Digital Unix 3.2D. For the reduction to Hessenberg form and the diagonalization optimized Lapack routines were used [40]. For the part of the algorithm listed in the previous section only the standard optimizations of the Fortran compiler were used.
The speed was tested by calculating, for several matrix sizes, all the coefficients of the characteristic polynomial of 100 matrices with random elements. This was done with our algorithm and with complete diagonalization. The speed was measured by counting the number of cycles executed by the procedures of the algorithms (fewer cycles means faster calculation) using the 'prof -pixie' command (see reference [41]). Table 4.1 lists the results. It is clear that our algorithm is much faster than complete diagonalization: from a factor 4.5 for small matrices to a factor 1.8 for large matrices. The decrease of this factor for large matrices can be understood by the fact that the routines for the reduction to Hessenberg form and diagonalization are strongly optimized while the routine for the algoritm 4.93 is not, and that these optimizations become more and more efficient with larger matrix sizes.
In order to test the accuracy, we calculated 200000 random samples with a SDQMC program for the 4×4 Hubbard model with 8 up and 8 down electrons, with $U = 4$ and $\beta = 6$, following the method of reference [35], but taking the canonical trace instead of the grand-canonical one (see chapter 5). The calculation was done in double precision and in single precision using our algorithm and complete diagonalization. As a measure for the accuracy we used the average absolute value of the difference between the single- and double-precision result divided by the double-precision result. For our algorithm we found a value of 0.00186 ± 0.00005 and for the complete diagonalization we found 0.00607 ± 0.00010 (error limits at 95% confidence level), indicating that our algorithm is more accurate. This could be expected since it requires less operations on the data. Furthermore complete diagonalization was much more sensitive to overflow errors than our algorithm. At values of $\beta > 6$ complete diagonalization (in single precision) was not usable anymore.

Matrix dimension	*Algorithm 4.97*	*Complete diagonalization*	Ratio
4	451400	1983818	4.39
6	1009400	4413843	4.37
8	1760300	8062663	4.58
10	2870100	12511637	4.36
15	7261300	29676436	4.09
20	14224100	55696656	3.93
25	25524300	93696774	3.67
30	41735900	144177197	3.45
35	63852100	209670202	3.28
40	90395400	290658105	3.22
45	126513800	388488344	3.07
50	171095900	512056714	2.99
60	284484900	794032492	2.79
70	447113900	1163945220	2.60
80	652709400	1630550332	2.50
90	926006900	2207209655	2.38
100	1251268900	2923248380	2.34
150	4268580600	8925077120	2.09
200	10018384500	20050929483	2.00
300	32993383700	63384810388	1.92
400	77249914100	145321243773	1.88
500	149825926400	278218705522	1.86
600	257427888100	474763616745	1.84
700	407433443000	745631287828	1.83
800	607094132500	1104878051129	1.82
900	863225666500	1564619645628	1.81

Table 4.1: *Comparision of the number of cycles needed for the calculation of the coefficients of the characteristic polynomial of 100 matrices with random elements, for several matrix dimensions.*

Application in SDQMC

The algorithm presented in the previous section can be used to calculate the canonical trace of the operator $\hat{U}_\sigma$. Because of equation 1.16 it is clear that if the algorithm is applied to the matrix U_σ, the canonical trace is given by

$$\hat{\mathrm{Tr}}_N\left(\hat{U}_\sigma\right) = t_{N1}. \tag{4.95}$$

Care has to be taken to preserve stability in low-temperature SDQMC. Problems with loss of significant digits and overflow can occur for large β. In order to avoid the former problem, the product in equation 4.1 can be orthonormalized after every few factors, as explained in section 4.6.1, resulting in a decomposition

$$U_\sigma = Q\,D\,R, \tag{4.96}$$

where Q is unitary, D diagonal with real positive elements on the diagonal, and R unitary or triangular, according to the orthonormalization technique used. V and W have determinant 1 and are well conditioned matrices. The elements of D can vary over many orders of magnitude because of the exponential nature of $\hat{U}_\sigma$. The algorithm 4.93 can be modified so that it keeps the elements of D separated from the well-conditioned parts Q and R. This enhances the stability. It leeds to the following algorithm:

algorithm for calculating the canonical trace of $\hat{\mathbf{U}}_\sigma$ ($\mathbf{U}_\sigma = \mathbf{Q\,D\,R}$)
reorder the diagonal elements of D in descending order
permute the columns of Q and the rows of R accordingly
$U = R\,Q$
reduce U to upper Hessenberg form
DO $j = N_{\mathcal{S}}, 1, -1$
  DO $i = 1, j$
    DO $k = N_{\mathcal{S}} - j, 1, -1$
      $t_{k+1\,i} = U_{i\,j}\,t_{k\,j+1} - U_{j+1\,j}\,t_{k\,i}$
    ENDDO
    $t_{1\,i} = U_{i\,j}$
  ENDDO
  DO $k = 1, N_{\mathcal{S}} - j$
    $t_{k\,j} = t_{k\,j} + t_{k\,j+1}\,(D_{k+j}/D_j)$
  ENDDO
  DO $k = N_{\mathcal{S}} - j, 1, -1$
    $d_{k+1} = d_k\,D_j$
  ENDDO
  $d_1 = D_j$
ENDDO
$\hat{\mathrm{Tr}}_N\left(\hat{U}_\sigma\right) = t_{N1}\,d_N.$ (4.97)

Also in the reduction of $D\,U$ to upper Hessenberg form the elements of D and U can be kept separated. Overflow can be avoided by working with the logarithms instead of the actual values of the elements of D and d.

4.3.3 Observables in the canonical ensemble

For the evaluation of the canonical expectation value of an operator $\hat{A}$, we need to calculate the canonical trace $\hat{\mathrm{Tr}}_N\left(\hat{A}U_\sigma\right)$. The expressions derived in section 4.2.1 for the evaluation of observables in the grand canonical ensemble, can be adapted to the canonical ensemble. For a one-body operator $\hat{A}$, we can use the analogon of expression 4.30:

$$f_A(\sigma) = \hat{\mathrm{Tr}}_N\left[\hat{A}\hat{U}_\sigma\right]/w(\sigma) = \left.\frac{d}{d\epsilon}\ln\left|\hat{\mathrm{Tr}}_N\left[\hat{Q}_{\hat{A}}(\epsilon)\,\hat{U}_\sigma\right]\right|\right|_{\epsilon=0}. \tag{4.98}$$

Now the canonical trace of $\hat{A}_{\hat{O}}(\epsilon)\,\hat{U}_\sigma$ can be evaluated using the algorithm presented in the previous sections. The derivative in ϵ can be evaluated by taking a small but finite value for ϵ as in expression 4.29. Another way to calculate $\hat{\mathrm{Tr}}_N\left(\hat{A}\hat{U}_\sigma\right)$ is obtained by taking the N^{th} derivative to χ in expression 4.31

$$\begin{aligned}
\hat{\mathrm{Tr}}_N\left(\hat{A}\hat{U}_\sigma\right) &= \left.\frac{1}{N!}\left(\frac{d}{d\chi}\right)^N \hat{\mathrm{Tr}}\left(\hat{A}\hat{U}_\sigma\chi^{\hat{N}}\right)\right|_{\chi=0} \\
&= \left.\frac{1}{N!}\left(\frac{d}{d\chi}\right)^N \left[\det\left(1+\chi U_\sigma\right)\,\mathrm{Tr}\left(A\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)\right]\right|_{\chi=0} \\
&= \left.\frac{1}{(N-1)!}\left(\frac{d}{d\chi}\right)^{N-1} \det\left(1+\chi U_\sigma\right)\,\mathrm{Tr}\left(O\frac{U_\sigma}{1+\chi U_\sigma}\right)\right|_{\chi=0}.
\end{aligned} \tag{4.99}$$

If U_σ is diagonalized to $Q^\dagger E Q$, $E_{ii} = e_i$, we obtain

$$\hat{\mathrm{Tr}}_N\left(\hat{A}\hat{U}_\sigma\right) = \sum_{i=1}^{N_S} c_{N,i}\,A'_{ii}, \tag{4.100}$$

where

$$c_{N,i} = \frac{e_i}{(N-1)!}\left(\frac{d}{d\chi}\right)^{N-1}\left.\prod_{j\neq i}\left(1+\chi e_j\right)\right|_{\chi=0}, \tag{4.101}$$

and

$$A' = QAQ^\dagger. \tag{4.102}$$

The coefficients $c_{N,i}$ can be calculated efficiently by constructing the polynomials

$$p_{L,i}(\chi) = \prod_{j=1}^{i-1}\left(1+\chi e_j\right), \tag{4.103}$$

and

$$p_{R,i}(chi) = \prod_{j=i+1}^{N}\left(1+\chi e_j\right), \tag{4.104}$$

so that $c_{N,i}$ is given by e_i times the coefficient of χ^N in the polynomial $p_{L,i}(\chi)\;p_{R,i}(\chi)$. To calculate $c_{N,1}\ldots c_{N,N}$ we need about $6NN_S$ flops. The time-consuming steps are the diagionalization of U_σ and the calculation of A'.

Expectation values for a two-body operator $\hat{B}$ can be obtained by decomposing the two-body operator as a sum of products of one-body operators:

$$\hat{B} = \sum_i \hat{A}_{1i}\hat{A}_{2i}. \tag{4.105}$$

The canonical trace for a product of two one-body operators $\hat{B} = \hat{A}_1\hat{A}_2$, can be evaluated as

$$\hat{\mathrm{Tr}}_N\left[\hat{A}_1\hat{A}_2\hat{U}_\sigma\right] = \frac{d}{d\epsilon_1}\frac{d}{d\epsilon_2}\,\hat{\mathrm{Tr}}_N\left[\hat{Q}_{\hat{A}_1}(\epsilon_1)\,\hat{Q}_{\hat{A}_2}(\epsilon_2)\,\hat{U}_\sigma\right]\Big|_{\epsilon_1=0,\,\epsilon_2=0}. \tag{4.106}$$

Again, the derivatives can be evaluated by taking small but finite values for ϵ_1 and ϵ_2. Alternatively, an expression for $\hat{\mathrm{Tr}}_N\left(\hat{A}\hat{U}_\sigma\right)$ can be obtained by taking the N^{th} derivative to χ in expression 4.40:

$$\begin{aligned}
&\hat{\mathrm{Tr}}_N\left(\hat{A}_1\hat{A}_2\hat{U}_\sigma\right)\\
&= \frac{1}{N!}\left(\frac{d}{d\chi}\right)^N \hat{\mathrm{Tr}}\left(\hat{A}_1\hat{A}_2\hat{U}_\sigma\chi^{\hat{N}}\right)\Big|_{\chi=0}\\
&= \frac{1}{N!}\left(\frac{d}{d\chi}\right)^N \left\{\det(1+\chi U_\sigma)\left[\mathrm{Tr}\left(A_1\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)\mathrm{Tr}\left(A_2\frac{\chi U_\sigma}{1+\chi U_\sigma}\right)\right.\right.\\
&\quad \left.\left.+\mathrm{Tr}\left(A_1A_2\frac{\chi U_\sigma}{1+\chi U_\sigma}\right) - \mathrm{Tr}\left(A_1\frac{\chi U_\sigma}{1+\chi U_\sigma}A_2\frac{\chi U_\sigma}{1+\chi U_\sigma}\right).\right]\right\}\Big|_{\chi=0}
\end{aligned} \tag{4.107}$$

After diagonalization of U_σ this becomes

$$\hat{\mathrm{Tr}}_N\left(\hat{A}_1\hat{A}_2\hat{U}_\sigma\right) = \sum_{i=1}^{N_S} c_{N,i}\,(A'_1A'_2)_{ii} + \sum_{i=1}^{N_S}\sum_{j=1}^{N_S} d_{N,ij}\,(A'_{1\,ii}A'_{2\,jj} - A'_{1\,ji}A'_{2\,ij}), \tag{4.108}$$

where the coefficients $d_{N,ij}$ are given by

$$d_{N\,i\,j} = e_ie_j\,\frac{1}{(N-2)!}\left(\frac{d}{d\chi}\right)^{N-2}\prod_{k\neq i,j}(1+\chi e_k)\Big|_{\chi=0}. \tag{4.109}$$

The $d_{N\,i\,j}$ can be evaluated in an analogous way as the $c_{N,i}$, requiring now about $6NN_S^2$ flops. Expectation values for multi-body operators can be calculated analogously, but will require more and more flops as the rank gets higher. Note that the relation 4.45, that can be seen as a finite temperature version of Löwdins' expression 4.46, has no simple analogon in the canonical ensemble. The correct expression is obtained by taking the N^{th} derivative to χ of both sides of relation 4.45.
Also the expression 4.51 for the grand canonical expectation value of an operator $\hat{A}$ can be adapted to the canonical ensemble:

$$\hat{\mathrm{Tr}}_N\left(\hat{A}e^{-\beta\hat{H}}\right) = \frac{d}{d\epsilon}\hat{\mathrm{Tr}}_N\left(e^{-\beta\hat{H}+\epsilon\hat{A}}\right)\Big|_{\epsilon=0}. \tag{4.110}$$

Again, this method is particularly useful for the calculation of the internal energy

$$U = \langle\hat{H}\rangle. \tag{4.111}$$

Expression 4.58 can be adopted without modification:

$$f'_H(\sigma) = -\frac{d}{d\beta'} \ln\left[w(\sigma,\beta')\right]\Big|_{\beta'=\beta}. \tag{4.112}$$

The canonical partition function is obtained by integrating $-U = \frac{d}{d\beta}\ln(Z_\beta)$ from 0 to β. The entropy and the free energy can be obtained in the same way as in the grand canonical ensemble. The specific heat C can be evaluated from

$$C = \beta^2\left(\langle\hat{H}^2\rangle - \langle\hat{H}\rangle^2\right). \tag{4.113}$$

For the observable $\hat{H}^2$, the factor $f_{H^2}(\sigma)$ from 4.73 is given by

$$f'_{H^2}(\sigma) = \frac{d^2}{(d\beta')} \ln\left[w(\sigma,\beta')\right]\Big|_{\beta'=\beta} + \left[f'_H(\sigma)\right]^2. \tag{4.114}$$

4.3.4 Canonical or grand canonical ensemble?

In the thermodynamic limit, i.e. when the system size is made infinitely large while density and temperature are kept constant, the canonical and grand canonical ensemble yield the same physical results. Because of this equivalence, one could raise the question which ensemble to choose for SDQMC calculations.
From the results in the previous sections it is clear that the grand canonical ensemble has a number of advantages:

- The evaluation of the determinant 4.25 for the weight $w(\sigma)$ requires $2N_{\mathcal{S}}^3$ flops, while the evaluation of $w(\sigma)$ in the canonical ensemble requires about $4N_{\mathcal{S}}^3$ flops.
- SDQMC in the grand canonical ensemble can be speeded up even more, using the rank-one updating scheme of section 4.2.1. No such scheme for the canonical ensemble exists.
- The evaluation of the factors $f_A\sigma$ for the observables is done faster in the grand canonical ensemble.
- For applications in condensed matter physics, e.g. for the Hubbard model, the grand canonical ensemble is physically more relevant because it allows fluctuations in the particle density. In real systems, local fluctuations of the density will occur.
- Because the grand canonical trace sums over more states than the canonical trace, the average sign of the weights $w(\sigma)$ can be expected to be higher in the grand canonical ensemble than in the canonical ensemble. The 'sign problem' (see section 4.5) is less severe.

Altough the computations are more time consuming, the canonical ensemble has its advantages over the grand canonical ensemble too:

- For mesoscopic systems, e.g. atomic nuclei, the canonical and grand canonical ensemble are not at all equivalent. A clear illustration of this, are the odd-even energy differences in nuclei caused by pairing correlations [46]. Odd nuclei have relatively higher ground-state energies. At low temperatures, their weight in the grand canonical ensemble will be suppressed compared to the even nuclei. Therefore, the grand canonical ensemble gives little information on the ground-state properties of odd nuclei, at any value of the chemical potential μ. To study low temperature properties of odd nuclei, the canonical ensemble has to be used.

- For physical systems with a fixed number of particles, like atomic nuclei, the canonical ensemble is more natural.

- Even for systems that ideally should be studied in the thermodynamic limit, SDQMC are only possible at mesoscopic sizes. Shell effects can influence the results. For the 4×4 Hubbard model, e.g., states with 5 spin-up and 5 spin-down particles or 8 spin-up and 8 spin-down particles dominate the grand canonical results because of the shell structure of the single-particle part of the Hamiltonian. States with other particle numbers migth be more representative for the properties in the thermodynamic limit. This topic is discussed more extensively in chapter 5.

- Ground-state properties can be studied at lower values of β in the canonical ensemble, because low-lying excited states with different particle numbers are projected out. With decreasing temperature, expectation values for observables converge faster to their ground-state values in the canoncial than in the grand canonical ensemble.

- The specific heat is more easily evaluated in the canonical ensemble than in the grand canonical enemsble.

Because the final aim of this work is the application of SDQMC to atomic nuclei, we only did calculations in the canonical ensemble.
A way to combine the speed of grand canonical SDQMC calculations with the advantages of canonical SDQMC calculations could be given by the guided Metropolis sampler of section 3.4.5. The grand canonical trace, with an appropriately chosen chemical potential, could be used as a guiding weight for the canonical trace. This has not yet been investigated and is a topic for further research.

4.4 SDQMC with ground-state projection

4.4.1 The Boltzmann operator as a ground-state filter

In the limit of low temperature or high β, the weight $e^{-\beta E_i}$ of the excited states in the canonical or grand canonical ensemble becomes negligible compared to the weight $e^{-\beta E_0}$ of the ground state. Hence, low temperature thermodynamic expectation values of operators are equal to their ground-state expectation values. Instead of taking the low

temperature limit of the thermodynamic expectation values, one can also obtain ground-state expectation values by applying the operator $e^{-\beta\hat{H}}$ to a Slater determinant Φ that has a considerable overlap with the ground state. The Boltzmann operator strongly suppresses the amplitudes of the components of the excited states Ψ_i in Φ by a factor $e^{-\beta(E_i-E_0)}$. Therefore, the ground state is approximately given by

$$|\Psi_0\rangle \simeq e^{-\beta\hat{H}}|\Phi\rangle. \tag{4.115}$$

An obvious choice for the trial state Φ is the N-particle Hartree-Fock ground state of the system. The expectation value of an operator $\hat{A}$ is given by

$$\left\langle \Psi_0 \left| \hat{A} \right| \Psi_0 \right\rangle \simeq \frac{\left\langle \Phi \left| e^{-\frac{\beta}{2}\hat{H}} \hat{A} e^{-\frac{\beta}{2}\hat{H}} \right| \Phi \right\rangle}{\left\langle \Phi \left| e^{-\beta\hat{H}} \right| \Phi \right\rangle}, \quad \text{for large } \beta. \tag{4.116}$$

Again, this expression can be evaluated using the decomposition 4.1 and MCMC sampling. To see this, we write expression 4.116 as

$$\frac{\left\langle \Phi \left| e^{-\frac{\beta}{2}\hat{H}} \hat{A} e^{-\frac{\beta}{2}\hat{H}} \right| \Phi \right\rangle}{\left\langle \Phi \left| e^{-\beta\hat{H}} \right| \Phi \right\rangle}, = \frac{\left\langle \Psi_L \left| \hat{A} \right| \Psi_R \right\rangle}{\langle \Psi_L | \Psi_R \rangle}, \tag{4.117}$$

with

$$\begin{aligned} |\Psi_L\rangle &= e^{-\frac{\beta}{2}\hat{H}} |\Phi\rangle , \\ |\Psi_R\rangle &= e^{-\frac{\beta}{2}\hat{H}} |\Phi\rangle . \end{aligned} \tag{4.118}$$

Clearly, $|\Psi_L\rangle = |\Psi_R\rangle$. But the decomposition of the operator $e^{-\frac{\beta}{2}\hat{H}}$ will lead to different terms for the left and the right state when sampling the configurations.

$$\begin{aligned} |\Psi_{\sigma_L}\rangle &= \sum_{\sigma_L} \hat{U}_{\sigma_L} |\Phi\rangle , \\ |\Psi_{\sigma_R}\rangle &= \sum_{\sigma_R} \hat{U}_{\sigma_R} |\Phi\rangle . \end{aligned} \tag{4.119}$$

The expectation value $\left\langle \Psi_0 \left| \hat{A} \right| \Psi_0 \right\rangle$ can now be evaluated as

$$\left\langle \Psi_0 \left| \hat{A} \right| \Psi_0 \right\rangle \simeq \frac{\sum_\sigma f_A(\sigma) w(\sigma)}{\sum_{\sigma'} w_{\sigma'}}, \tag{4.120}$$

with

$$\begin{aligned} w(\sigma) &= \langle \Psi_{\sigma_L} | \Psi_{\sigma_R} \rangle \\ &= \left\langle \Phi \left| \hat{U}_{\sigma_l} \hat{U}_{\sigma_R} \right| \Phi \right\rangle \end{aligned} \tag{4.121}$$

$$\begin{aligned} f_A(\sigma) &= \left\langle \Psi_{\sigma_L} \left| \hat{A} \right| \Psi_{\sigma_R} \right\rangle / w(\sigma). \\ &= \left\langle \Phi \left| \hat{U}_{L\sigma} \hat{A} \hat{U}_{R\sigma} \right| \Phi \right\rangle / w(\sigma). \end{aligned} \tag{4.122}$$

The configuration σ denotes here a pair of configuration (σ_L, σ_R). The SDQMC obtained in this way, is often called 'Projector Quantum Monte-Carlo' method [7].

4.4.2 Evaluation of weights and observables with ground-state projection.

The weight $w(\sigma)$ can be evaluated using expression 1.6. Let M denote the $N_{\mathcal{S}} \times N$ matrix that represents the Slater determinant Φ. Then $w(\sigma)$ is given by

$$w(\sigma) = \det\left(M^T U_{L\sigma} U_{R\sigma} M\right). \tag{4.123}$$

For the evaluation of $f_A(\sigma)$ for a one-body operator $\hat{A}$, we need once more the operator $\hat{Q}_{\hat{A}}(\epsilon)$ from section 4.2.1. The factor $f_A(\sigma)$ is given by

$$\begin{aligned} f_A(\sigma) &= \frac{d}{d\epsilon}\left\langle \Phi \left| \hat{U}_{L\sigma} \hat{Q}_{\hat{A}}(\epsilon) \hat{U}_{R\sigma} \right| \Phi \right\rangle \Big|_{\epsilon=0} / w(\sigma). \qquad (4.124) \\ &= \frac{d}{d\epsilon} \det\left(M^T U_{L\sigma}(1+\epsilon A) U_{R\sigma} M\right). \Big|_{\epsilon=0} / w(\sigma). \\ &= \det\left(M^T U_{L\sigma} U_{R\sigma} M\right) \mathrm{Tr}\left(\frac{1}{M^T U_{L\sigma} U_{R\sigma} M} M^T U_{L\sigma} A U_{R\sigma} M\right) / w(\sigma) \\ &= \mathrm{Tr}\left(\frac{1}{M^T U_{L\sigma} U_{R\sigma} M} M^T U_{L\sigma} A U_{R\sigma} M\right). \qquad (4.125) \end{aligned}$$

Expressions for expectation values of products of one-body operators can be obtained by inserting more operators $\hat{Q}_{A_j}(\epsilon_j)$ in 4.124, analogous to the reasoning that was followed in section 4.2.1 for the grand canonical trace of higher order operators.
The cyclical permutation symmetry of the (grand) canonical trace is lost here, so there is no analogon for the method based on expression 4.55. Because it is a ground-state method, thermodynamical quantities like entropy or specific heat obviously cannot be obtained using ground-state projection.

4.4.3 Ground-state projection or (grand) canonical ensemble?

Compared to the SDQMC in the grand canonical or the canonical ensemble, SDQMC with ground-state projection has the following advantages:

- The evaluation of the weight $w(\sigma)$ in the ground-state projection method requires N_t multiplications of an $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrix with an $N_{\mathcal{S}} \times N$ matrix, while in the (grand) canonical method, it requires N_t multiplications of an $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrix with an $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrix. Therefore the ground-state projection method requires a factor $N/N_{\mathcal{S}}$ less flops.

- If the overlap of the trial state Φ with the true ground state Φ_0 is large, the method will converge fast. Ground-state properties can be obtained at lower values of β than in the (grand) canonical method.

- Rank-one updates analogous to the rank-one updates for the grand canonical ensemble from section 4.2.1 are possible. However, the cyclical updating procedure for the inverse-temperature slices explained in section 4.2.1 cannot be used here because of the breakdown of the cyclical permutation symmetry.

The major disadvantages are:

- The results are only physically meaningful in the limit of large β.
- The ground-state projection method suffers more from sign problems (see section 4.5) than the canonical or the grand canonical method.
- The ground-state projection method can only be used if there exists a Slater determinant Φ with a considerable overlap with the true ground state, in other words, if the Hartree-Fock method gives a good approximation of the ground state. So the ground-state projection method can only be used to study ground-state properties beyond Hartree-Fock in those cases where Hartree-Fock gives already a reasonable description. The more interesting cases where the Hartree-Fock method does not give satisfactory results, are much more difficult to study with the ground-state projection method. For the 4×4 Hubbard model, with strength $U = 4|t|$, we calculated the overlap of the Hartree-Fock ground state Φ with the true ground state Φ_0, using the canonical ensemble:

$$|\langle \Phi | \Phi_0 \rangle|^2 = \frac{\langle \Phi | e^{-\beta \hat{H}} | \Phi \rangle}{\hat{\mathrm{Tr}}_N \left(e^{-\beta \hat{H}} \right)}, \tag{4.126}$$

  in the limit of large β. For a system with 5 spin-up particles and 5 spin-down particles ($5 \uparrow 5 \downarrow$), the overlap was about 0.7, while for the $6 \uparrow 6 \downarrow$ system the overlap was too small to be determined accurately (smaller than 0.05). Note that the $5 \uparrow 5 \downarrow$ system corresponds to a closed shell configuration of the one-body Hamiltonian.

These last two remarks are closely related to a major problem of the ground-state projection method. The method has been used extensively for the study of the Hubbard model. These calculations where mostly done at particle densities for which the ground state of the one-body part of the Hamiltonian is a closed shell configuration, because at theses densities the sign problems are least and the Hartree Fock ground state has a large overlap with the true ground state. But exactly at these densities the influence of the shell structure related to the mesoscopic scale of the system, is strongest. Therefore, these densities are the least suited to extrapolate to the thermodynamic limit. The system has qualitatively different properties at these densities compared to other densities, as is illustrated in chapter 5. The fact that the ground-state projection method works well for one class of densities but not for a class of densities with qualitatively different properties, bears the risk that properties for the former densities might be extrapolated blindly to all densities, and that properties specific for the latter densities might be overseen.
The ground-state method is a fixed-N method. As disussed in section 4.3.4, this is in some cases an advantage, in other cases a disadvantage.
To our opinion, the canonical methods are to be preferred over the ground-state projection method for most applications. Only for the study of ground-state properties of a system with a small number of particles, preferably a closed shell configuration, the ground-state projection method will perform better.

4.5 The sign problem

Until now we always assumed that the weights $w(\sigma)$ where positive, so that they could be sampled using MCMC methods. However, this is generally not the case. The weights are related to the nature of the operators $\hat{U}_\sigma = e^{-\hat{S}_\sigma(\beta)}$ from expression 4.1. If $\hat{S}_\sigma(\beta)$ were a Hermitian operator, then the operator $\hat{U}_\sigma$ would be positive definite and the weight $w(\sigma)$ would allways be positive, as well for the canonical, grand canonical or the ground-state projection method. Because of the Trotter breakup 2.10 of $e^{-\beta\hat{H}}$ in inverse-temperature slices, $\hat{S}_\sigma(\beta)$ is generally not Hermitian.
Though negative weights can be handled in MCMC, as explained in the next section, they pose a fundamental problem: the variance of the results becomes infinitely large if the average sign of the weights goes to zero.
It is a problem encountered in all quantum Monte-Carlo methods for fermions. It is related to the antisymmetric nature of the fermion wave functions. For SDQMC, it is not as worse as for other quantum Monte-Carlo methods. In the canonical ensemble, for a number of systems, the SDQMC results converged to their ground-state values before the sign problem got too severe.

4.5.1 MCMC with non-negative weights

If the weight $w(\sigma)$ becomes negative for some configurations σ, MCMC methods cannot be applied directly in order to calculate expectation values of the form

$$\mathrm{E}(f) = \frac{\sum_\sigma f(\sigma)w(\sigma)}{\sum_{\sigma'} w(\sigma')}. \tag{4.127}$$

Instead, one evaluates

$$\begin{aligned} \mathrm{E}(f) &= \frac{\sum_\sigma f(\sigma)s(\sigma)|w(\sigma)|}{\sum_{\sigma'}|w(\sigma')|} \Big/ \frac{\sum_\sigma s(\sigma'')|w(\sigma'')|}{\sum_{\sigma'''}|w(\sigma''')|} && (4.128)\\ &= \frac{\mathrm{E}_{|w|}(fs)}{\mathrm{E}_{|w|}(s)}, && (4.129) \end{aligned}$$

with

$$s(\sigma) = \frac{w(\sigma)}{|w(\sigma)|}. \tag{4.130}$$

The quantity

$$\bar{s} = \mathrm{E}_{|w|}(s) = \frac{\sum_\sigma s(\sigma)|w(\sigma)|}{\sum_{\sigma'}|w(\sigma')|}, \tag{4.131}$$

is called the 'average sign'.
To evaluate $\mathrm{E}(f)$ using Monte-Carlo techniques, we have to generate a sample $S = \sigma^{[1]}, \sigma^{[2]}, \ldots, \sigma^{[M]}$ where the $\sigma^{[j]}$ are distributed according to $|w(\sigma)|$. The expectation value $\mathrm{E}(f)$ is then approximated as

$$\mathrm{E}(f) \simeq f_S = \frac{\mathrm{E}_S(fs)}{\mathrm{E}_S(s)}. \tag{4.132}$$

This estimate becomes exact when the sample size M goes to infinity. Because the sample average $\mathrm{E}_S(s)$ of the sign appears in the denominator of expression 4.132, it can be expected that the error on f_S will be large large when the average sign $\bar{s}$ is small.
If the sample S is obtained using independent sampling, an estimate for the error on f_S can be obtained from expression 3.123:

$$\begin{aligned}
&\mathrm{Var}(f_S) \\
&\quad\simeq \quad \frac{\left\langle \left(\mathrm{E}_S(fs) - \langle f_S\rangle \mathrm{E}_S(s)\right)^2 \right\rangle}{\langle \mathrm{E}_S(s)\rangle}. \\
&\quad= \quad \frac{\left\langle \frac{1}{M^2}\sum_{i,j=1}^{M} \left[f(\sigma^{[i]})s(\sigma^{[i]}) - \langle f_S\rangle s(\sigma^{[i]})\right]\left[f(\sigma^{[j]})s(\sigma^{[j]}) - \langle f_S\rangle s(\sigma^{[j]})\right]\right\rangle}{\bar{s}^2}
\end{aligned} \tag{4.133}$$

The notation $\langle . \rangle$ denotes for the weighted average over all possible samples S of size M that can be generated by independent sampling according to $w(\sigma)$. Because $\sigma^{[i]}$ and $\sigma^{[j]}$ are independent for $i \neq j$, and because

$$\begin{aligned}
\left\langle f(\sigma^{[j]})s(\sigma^{[j]})\right\rangle &= \mathrm{E}_{|w|}(fs), \\
\left\langle s(\sigma^{[j]})\right\rangle &= \mathrm{E}_{|w|}(s), \\
\frac{\mathrm{E}_{|w|}(fs)}{\mathrm{E}_{|w|}(s)} = \mathrm{E}(f) &\simeq \langle f_S\rangle,
\end{aligned}$$

expression 4.133 simplifies to

$$\begin{aligned}
\mathrm{Var}(f_S) \quad&\simeq \quad \frac{\left\langle \frac{1}{M^2}\sum_{i=1}^{M} \left[f(\sigma^{[i]})s(\sigma^{[i]}) - \langle f_S\rangle s(\sigma^{[i]})\right]^2\right\rangle}{\bar{s}^2} \\
&= \quad \frac{\left\langle \frac{1}{M^2}\sum_{i=1}^{M} \left[f(\sigma^{[i]}) - \langle f_S\rangle\right]^2 s(\sigma^{[i]})^2\right\rangle}{\bar{s}^2} \\
&= \quad \frac{\frac{1}{M}\mathrm{E}_{|w|}\left[\left(f - \langle f_S\rangle\right)^2\right]}{\bar{s}^2} \\
&\simeq \quad \frac{\mathrm{E}_{|w|}\left[\left(f - \mathrm{E}(f)\right)^2\right]}{M\bar{s}^2}
\end{aligned} \tag{4.134}$$

This shows that the variance on f_S is proportional to $\bar{s}^{-2}$. Thus the error on f_S is inversely proportional to $\bar{s}$.
In the case of dependent sampling, using Markov chains, expression 4.134 has to be multiplied with an appropriate factor r as discussed in section 3.3.1 In this case the factor is given by $r(fs - \mathrm{E}(f)s)$, with r defined by expression 3.89.
Expression 4.134 shows the fundamental limitations of MCMC methods for the sampling of non-negative weights: the error on the Monte-Carlo results is inversely proportional to the average sign of the weight $w(\sigma)$. This is the famous 'sign problem'.

4.5.2 The sign problem and the Hubbard-Stratonovich transform

The 'sign problem' arises not only in SDQMC, but also in most other fermionic quantum Monte-Carlo methods. The origin of the sign problem in methods based on the Hubbard-

Stratonovich transform (see section 2.2.1) was elucidated by Fahy and Hamann [42]. The operator $e^{-\beta\hat{H}}$ transforms an initial Slater determinant $|\Phi^{[0]}\rangle$ into a state $e^{-\beta\hat{H}}|\Phi^{[0]}\rangle$. Using the Hubbard-Stratonovich transfrom, this state can be written as a sum of Slater determinants

$$e^{-\beta\hat{H}}|\Phi^{[0]}\rangle = \int w(\sigma)d\sigma e^{-\beta\hat{A}_\sigma}|\Phi^{[0]}\rangle. \tag{4.135}$$

A single Slater determinant is 'diffused' to a distribution of Slater determinants given by 4.135. This diffusion proceeds at each inverse-temperature interval, with a rate proportional to β/N_t. Fahy and Hamann showed how this diffusion of Slater determinants, in the limit of $\beta/N_t \to 0$, can be described by a differential equation of motion for a distribution of Slater determinants $f(\Phi,\beta)$, with the inverse temperature β playing the role of a 'time' variable. This equation of motion takes the form of a diffusion equation with drift and branching terms on the manifold of Slater determinants. One can define a 'parity' transformation $\mathcal{P}$ on the Slater determinants by

$$\mathcal{P}|\Phi\rangle = -|\Phi\rangle. \tag{4.136}$$

This parity transformation commutes with the diffusion equation operator. Therefore, the eigenfunctions of the diffusion equation will have a definite parity. Fahy and Hamann point out that the eigenfunction with the highest eigenvalue will be even under $\mathcal{P}$. This eigenfunction is denoted as $f^+(\Phi)$. Because of the nature of the diffusion equation, the distribution $f(\Phi,\beta)$ will tend to $f^+(\Phi)$ for large β. But this distribution is related to a vanishing many-body state, because it contains $|\Phi\rangle$ and $-|\Phi\rangle$ with equal weight:

$$\sum_\Phi f^+(\Phi)|\Phi\rangle = \sum_\Phi f^+(\Phi)\frac{|\Phi\rangle + |-\Phi\rangle}{2} = 0. \tag{4.137}$$

Only odd-parity distributions can give rise to nonzero many-body states. Therefore, the physical many-body state $e^{-\beta\hat{H}}|\Phi^{[0]}\rangle$ that we want to describe using the Hubbard-Stratonovich transform, is related to the odd-parity eigenfunction $f^-(\Phi)$ that has the highest eigenvalue. For the ground-state projection algorithm, the average sign $\bar{s}(\Phi^{[0]})$ is given by

$$\bar{s}(\Phi^{[0]}) = \frac{\sum_\Phi f(\Phi,\beta)\langle\Phi^{[0]}|\Phi\rangle}{\sum_{\Phi'} f(\Phi',\beta)|\langle\Phi^{[0]}|\Phi'\rangle|}. \tag{4.138}$$

The denumerator couples only to the odd-parity component of $f(\Phi,\beta)$, while the denominator couples only to the even-parity component of $f(\Phi,\beta)$. If the eigenvalue E^- related to f^- is smaller than the eigenvalue E^+ related to f^+, then for large enough values of β, the average sign can be expected to decrease exponentially with increasing β, proportional to $e^{-\beta(E^+-E^-)}$. Only if f^- and f^+ are degenerate, the average sign will not go to 0. This reasoning extends directly to the canonical and the grand canonical ensemble, since there the average sign is given by the trace of $\bar{s}(\Phi^{[0]})$ over a complete set of initial Slater determinants $\Phi^{[0]}$.

This discussion shows that the sign problem is an intrinsic property of any quantum many-body method based on the Hubbard Stratonovich transfrom. One can try to deal with it, in the following ways:

- The sign problem is avoided if f^+ and f^- are degenerate. In some cases, this is guaranteed by underlying symmetries of the system (see next section).
- If the diffusion is slow, the decrease of the average sign $\bar{s}$ will be slow too. Then $\bar{s}$ might still be large enough at values of β that are physically relevant, such that SDQMC calculations are possible. Slowing down the diffusion amounts to reducing the non-Hermitian components of the operator $\hat{S}_\sigma(\beta)$. Another way to slow down the diffusion, is to use a discrete form of the Hubbard Stratonovich transform instead of a continuous one.
- The discussion of Fahy and Hamann only applies to the Hubbard-Stratonovich transfrom, other decompositions of $e^{-\beta\hat{H}}$, such as the ones discussed in section 2.2.2, might in some cases lead to a larger average sign (however, in other cases they might lead to a smaller average sign).

4.5.3 Decompositions with good sign characteristics

In a number of cases an underlying symmetry of the interacting many-fermion system guarantees a strictly positive weight $w(\sigma)$.
A first example is the attractive Hubbard model (see also chapter 5). As explained in section 2.2.2, Hirsch's discrete Hubbard-Stratonovich transform leads to a factorization of the matrices U_σ in a spin-up and a spin-down part:

$$U_\sigma = \begin{pmatrix} U_{\uparrow\sigma} & 0 \\ 0 & U_{\downarrow\sigma} \end{pmatrix}. \tag{4.139}$$

The operator $\hat{U}_\sigma$ factors correspondingly in $\hat{U}_{\uparrow\sigma}\hat{U}_{\downarrow\sigma}$, where $\hat{U}_{\uparrow\sigma}$ acts only on spin-up particles and $\hat{U}_{\downarrow\sigma}$ only on spin-down particles. The trace of $\hat{U}_\sigma$ can be written as a product of a spin-up and a spin-down trace:

$$\hat{\mathrm{Tr}}\left(\hat{U}_\sigma e^{\beta\mu\hat{N}}\right) = \hat{\mathrm{Tr}}_\uparrow\left(\hat{U}_{\uparrow\sigma}e^{\beta\mu\hat{N}_\uparrow}\right)\hat{\mathrm{Tr}}_\downarrow\left(\hat{U}_{\downarrow\sigma}e^{\beta\mu\hat{N}_\downarrow}\right), \tag{4.140}$$

for the grand canonical trace, or

$$\hat{\mathrm{Tr}}_{N_\uparrow N_\downarrow}\left(\hat{U}_\sigma\right) = \hat{\mathrm{Tr}}_{\uparrow N_\uparrow}\left(\hat{U}_{\uparrow\sigma}\right)\hat{\mathrm{Tr}}_{\downarrow N_\downarrow}\left(\hat{U}_{\downarrow\sigma}\right), \tag{4.141}$$

for the canonical trace. For the attractive Hubbard model, $U_{\uparrow\sigma} = U_{\downarrow\sigma}$. Therefore, if $N_\uparrow = N_\downarrow$, the traces 4.140 and 4.141 are squares of real numbers and thus always positive. The repulsive Hubbard model with $N_\uparrow$ spin-up and $N_\downarrow$ spin-down particles can be transformed into a attractive Hubbard model with $N_\uparrow$ spin-up and $N_\mathcal{S} - N_\downarrow$ spin-down particles by a particle-hole transformation of the spin-down particles (see chapter 5). Therefore, the repulsive Hubbard model with $N_\uparrow + N_\downarrow = N_\mathcal{S}$, a fortiori the half-filled Hubbard model, has strictly positive weights too.
Another class of systems that have no sign problems, was found by Lang *et al.* [37]. Because of spherical symmetry and time-reversal invariance, a general two-body Hamiltonian $\hat{H}_2$ for the nuclear shell model can be decomposed in the following way:

$$\hat{H}_2 = \sum_{\alpha,J,M} \lambda_{\alpha,J}(-1)^M \hat{A}_{\alpha JM}\hat{A}_{\alpha J-M}, \tag{4.142}$$

where $\lambda_{\alpha,J}$ is a real constant and the $\hat{A}_{\alpha JM}$ are one-body operators of definite multipolarity. Furthermore, $(-1)^{J+M}\hat{A}_{\alpha J-M}$ is the time-reversed operator of $\hat{A}_{\alpha JM}$. For every set of values αJM we obtain a term

$$\begin{aligned}\hat{H}_{\alpha,J,M} &= \lambda_{\alpha,J}(-1)^M \hat{A}_{\alpha JM}\hat{A}_{\alpha J-M} \\ &= \lambda_{\alpha,J}(-1)^J \left(\hat{A}_{\alpha JM} + (-1)^{J+M}\hat{A}_{\alpha J-M}\right)^2 \\ &\quad + \left(i\hat{A}_{\alpha JM} - i(-1)^{J+M}\hat{A}_{\alpha J-M}\right)^2 . \end{aligned} \tag{4.143}$$

If $\lambda_{\alpha,J}(-1)^J$ is negative, then we can make a decomposition for $\hat{H}_{\alpha,J,M}$ of the form 2.24. The Hubbard-Stratonovich transfrom 2.30 leads to a decomposition

$$\begin{aligned}\hat{H}_{\alpha,J,M} &= \int_\sigma G(\sigma)e^{\hat{A}_\sigma}, \\ \text{with} & \\ \hat{A}_\sigma &= \chi_{\alpha J}\left[(\sigma_1 + i\sigma_2)\,\hat{A}_{\alpha JM} + (\sigma_1 - i\sigma_2)\,(-1)^{J+M}\hat{A}_{\alpha J-M}\right], \end{aligned} \tag{4.144}$$

where $\chi_{\alpha J} = \sqrt{-\lambda_{\alpha,J}(-1)^J}$. Time-reversed operators couple to complex-conjugated auxialiary fields. If we arrange the single-particle states such that the states with $\hat{J}_z$ quantum number $m > 0$ come first and then their time reversed states, the matrix representation of $\hat{A}_\sigma$ will have a structure

$$A_\sigma = \begin{pmatrix} A_{1\sigma} & A_{2\sigma} \\ -A^*_{2\sigma} & A^*_{1\sigma} \end{pmatrix}. \tag{4.145}$$

Matrices with this structure have some particular properties:

- The product of two such matrices conserves this structure.
- The exponential of such a matrix conserves this structure.
- If $\begin{pmatrix} u \\ v \end{pmatrix}$ is an eigenvector with eigenvalue e of such a matrix, then $\begin{pmatrix} -v^* \\ u^* \end{pmatrix}$ is an eigenvector with eigenvalue e^* of this matrix.
- Only half of the matrix has to be computed explicitely. The symmetry of the structure 4.145 can be exploited to obtain the other half. Because matrix multiplications are the most time consuming part of SDQMC calculations, this can almost double the speed of the calculations.

From these properties, and the fact that the matrix representation of a time-reversal invariant one-body Hamiltonian also has this structure, it follows that the eigenvalues of the matrix U_σ used in the decomposition 4.1 for the Boltmann operator, come in complex-conjugated pairs. This guarantees the positiveness of the grand canonical trace of $\hat{U}_\sigma$.

$$\begin{aligned}\hat{\mathrm{Tr}}\left(\hat{U}_\sigma e^{\beta\mu\hat{N}}\right) &= \det(1+\chi U_\sigma) \\ &= \prod_{i=1}^{N_S/2} (1+\chi e_i)\,(1+\chi e_i^*) \\ &= \prod_{i=1}^{N_S/2} |\,(1+\chi e_i)\,|^2 \geq 0. \end{aligned} \tag{4.146}$$

The canonical trace is not bound to be positive. However, the close relation between the grand canonical and the canonical ensemble makes that the average sign will remain well-behaved in certain cases. For attractive interactions, the grand canonical trace will be dominated by contributions of states with $N_+ = N_-$, where N_+ (N_-) denotes the number of particles in single-particle states with $m > 0$ ($m < 0$). The grand canonical trace is positive at any value of the chemical potential, in other words, at any particle density. Therefore, the overlap $\langle \Phi | \hat{U}_\sigma | \Phi \rangle$ can be expected to be positive for states Φ for which $N_+ = N_-$. These states will also dominate the canonical trace, provided that $N = N_+ + N_-$ is even. Therefore, the average sign is well behaved for the canonical ensemble with an *even* number of particles. Also the ground-state projection method will have no sign problem, provided that the trial state Φ in expression 4.121 has $N_+ = N_-$. If the operator $\hat{N}_+ - \hat{N}_-$ commutes with the Hamiltonian, then $N_+ - N_-$ is a good quantum number. The requirement that Φ has a non vanishing overlap with the ground state implies the supposition that for the true ground state $N_+ = N_-$ holds too. Otherwise the ground-state projection method cannot be applied without sign problems.
In this picture, the basic condition for the sign to be well behaved is the condition

$$\lambda_{\alpha J} (-1)^J \leq 0, \tag{4.147}$$

for all $\{\alpha J\}$ terms in the decomposition 4.142. Examples of such systems are even-even nuclear systems with a quadrupole-quadrupole or pairing interaction [37]. Also the absence of sign problems with the attractive Hubbard model discussed above can be understood in this way.
The decompositions of $e^{-\beta \hat{H}_2}$ based on rank-one and rank-two operators of the form 2.81 or 2.89 will also lead to matrices of the form 4.145, provided that $x_\sigma = y_\sigma$ and that the operators $\hat{b}_{1\sigma}$ and $\hat{b}_{2\sigma}$ and the operators $\hat{b}_{3\sigma}$ and $\hat{b}_{4\sigma}$ are related by a structure

$$\begin{cases} b_{1\sigma} &= \begin{pmatrix} c_{1\sigma} & c_{2\sigma} \end{pmatrix}, \\ b_{2\sigma} &= \begin{pmatrix} -c^*_{2\sigma} & c^*_{1\sigma} \end{pmatrix}. \end{cases} \tag{4.148}$$

$$\begin{cases} b_{3\sigma} &= \begin{pmatrix} c_{3\sigma} & c_{4\sigma} \end{pmatrix}, \\ b_{4\sigma} &= \begin{pmatrix} -c^*_{4\sigma} & c^*_{3\sigma} \end{pmatrix}. \end{cases} \tag{4.149}$$

The pairing interaction for nuclear systems (see chapter 6) and the attractive Hubbard model are examples of systems that can be decomposed in such a way. Therefore these systems can be studied without sign problems, using SDQMC based on a decomposition of the form 2.81 or 2.89, for even numbers of particles.

4.5.4 Practical solutions to the sign problem ?

To circumvent the sign problem, Sorella *et al.* [43, 44] have proposed to ignore the sign and to use $|w(\sigma)|$ instead of $w(\sigma)$. This amounts to using $\mathrm{E}_{|w|}(f)$ as an estimate for $\mathrm{E}(f)$. From expression 4.129 it is clear that this approach is only valid if

$$\mathrm{E}_{|w|}(f)\mathrm{E}_{|w|}(s) = \mathrm{E}_{|w|}(fs), \tag{4.150}$$

Figure 4.1: *Internal neutron energy for a $^{57}_{26}Fe_{31}$ system with a mean field and pairing interaction, obtained using a SDQMC method for the 11 neutrons in the fp shell. Correct results and results obtained by neglecting the sign of the weights $w(\sigma)$. are shown.*

i.e. if f and s are uncorrelated. While this may hold approximately in some cases, as was observed by Sorella *et al.* [43, 44], it certainly does not hold generally. Therefore, this approach does not solve the sign problem. Furthermore, even if relation 4.150 holds, the method does not allow to check this relation. In the picture of Fahy and Haymann (see section 4.5.2) the method amounts to replacing f^- by f^+ [42].
We calculated the internal neutron energy for a $^{57}_{26}\mathrm{Fe}_{31}$ system, with a mean field and pairing interaction (see chapter 6) using a SDQMC method for the 11 neutrons in the fp shell. Figure 4.1 shows the internal energy obtained from expression 4.132 and the internal energy obtained by ignoring the sign of $w(\sigma)$. For inverse temperatures β higher than $1.5\mathrm{MeV}^{-1}$, the method suggested by Sorella *et al.* clearly gives incorrect results.
A different solution for the sign problem was proposed by Alhassid *et al.* [45]. They suggest to extrapolate results for a series of Hamiltonians with 'good' sign properties to results for the full Hamiltonian, that can have 'bad' sign properties. Their approach is devised for the nuclear shell model, for which the two-body Hamiltonian can be decomposed in the form 4.142. As discussed in the previous section, these systems have no sign problem for even-even particle numbers, provided that the coefficients $\lambda_{\alpha J}(-1)^J$ are negative. In general, some of the $\lambda_{\alpha J}(-1)^J$ will be positive. The $\{\alpha J\}$ terms for which $\lambda_{\alpha J}(-1)^J$ is positive, constitute the 'bad' part of the Hamiltonian.

$$\hat{H} = \hat{H}_{\mathrm{good}} + \hat{H}_{\mathrm{bad}}, \tag{4.151}$$

$$\hat{H}_{\mathrm{good}} = \hat{H}_1 + \sum_{\substack{\alpha, J, M \\ \lambda_{\alpha J}(-1)^J < 0}} \lambda_{\alpha, J}(-1)^M \hat{A}_{\alpha JM} \hat{A}_{\alpha J-M}, \tag{4.152}$$

$$\hat{H}_{\rm bad} = \sum_{\substack{\alpha, J, M \\ \lambda_{\alpha J}(-1)^J > 0}} \lambda_{\alpha,J}(-1)^M \hat{A}_{\alpha JM} \hat{A}_{\alpha J-M}. \quad (4.153)$$

A new Hamiltonian is constructed as

$$\hat{H}_g = \hat{H}_{\rm good} + g\hat{H}_{\rm bad}. \quad (4.154)$$

For $g < 0$ this Hamiltonian is free of sign problems (for even-even systems). Then, results are calculated for a system with Hamiltonian $\hat{H}_g$ at several values of g between -1 and 0. These results are then extrapolated to results for $\hat{H}_g$ with $g = 1$, the full Hamiltonian. Though for some systems [6] this approach gives good agreement with results obtained using diagonalization techniques, one should be cautious with it. In general, it can be expected that the relation between the bad part of the Hamiltonian and values for the observables will be highly non-linear. To validate the method, it would be interesting to compare results for $\hat{H}_{g=-1}$ with results for $\hat{H}_{g=1} = \hat{H}$ at values of β where the average sign is large enough to allow accurate calculations for both Hamiltonians. As far as we know, results of such a comparision have not been published yet.

The approach of Alhassid *et al.* can be applied to the Hubbard model too (see chapter 5). For the repulsive Hubbard model, the 'bad' part of the Hamiltonian is the complete two-body part of the Hamiltonian. By multiplying this part with a negative value g, the Hamiltonian is transformed into the Hamiltonian of the attractive Hubbard model, that has good sign properties, as discussed in the previous section. At half filling, both the attractive and the repulsive Hubbard model have good sign problems, so the validity of the g extrapolation can be verified. We calculated the ground-state energy for several values of two-body interaction strength U, ranging from $U = -8|t|$ to $U = 8|t|$. A second-order polynomial in U was fitted to the results for $U < 0$. An excellent fit was obtained. This polynomial was used to extrapolate the ground-state energy to positive values of U. It can be seen from figure 4.2 that the extrapolated values differ strongly from the true values. Now the energy is the observable most directly related to the Hamiltonian. Therefore, for other observables the deviations can be expected to be even bigger. Clearly, for the Hubbard model the g-extrapolation does not work. For other systems, where only a fraction of the Hamiltonian is 'bad', the method might do better. To our opinion, the method ought to be checked case by case, by looking at the validity of the g-extrapolation at low values of β, before it is used to study observables at large values of β.

It seems to us that, for a general interaction, sign problems cannot be avoided. The freedom in the decomposition of the Hamiltonian (see sections 2.2.1 and 2.2.2) might allow for a decomposition with which the average sign becomes too small only at very low temperatures. If the system is almost completely cooled to its ground state at temperatures for which the average sign $\bar{s}$ is still large enough, (practically, $\bar{s} \geq 0.1$), the relevant physics of the system can be studied with not too bad statistics. As stated before, the SDQMC would be free of sign problems if the operators $\hat{S}_\sigma(\beta)$ were Hermitian. The freedom in the decomposition of the Hamiltonian can be used to make the operators $\hat{A}_\sigma$ in expression 4.1 Hermitian for each inverse-temperature slice. This does not make $\hat{S}_\sigma(\beta)$ Hermitian, but it ensures that the non-Hermitian components in $\hat{S}_\sigma(\beta)$ are given by commutators of the operators $\hat{A}_\sigma$, so that the Hermitian part will dominate $\hat{S}_\sigma(\beta)$. In

Figure 4.2: *Extrapolation of the ground-state energy of the half-filled* 4×4 *Hubbard model from negative to positive values of the interaction strength* U.

this way, calculations might be possible at low values of β. These results could then be extrapolated to higher values of β (lower temperatures). This extrapolation seems, to us, more reliable than the g-extrapolation discussed above. A consistent extrapolation scheme, probably has to be based on an inverse-Laplace transform and maximum entropy techniques. The development of such a scheme could be a topic for future research.

4.6 Practical considerations

4.6.1 Stabilization at low temperatures.

At low temperatures, SDQMC tends to become unstable. This is caused by the fact that the matrices U_σ tend to become nearly singular for large β. The columns of U_σ become similar to one another. This means that the particles all tend to occupy the same single-particle state. If the system would be bosonic, this could be understood as a Bose condensation in one single-particle state. However, the Pauli principle blocks this. Fermionic systems are described by the linearly-independent components of the columns of U_σ. If the columns of U_σ are nearly linearly dependent, the linearly-independent compononents can become intractable due to the limited computer precision. This leads to a first practical consideration: SDQMC computer programs should always use double-precision variables. But this is not enough. To stabilize the SDQMC at low temperatures, the linearly-dependent components of the columns of U_σ have to be projected out. Several approaches have been suggested for this.

A first approach [6] is based on the singular value decomposition. The singular value decomposition (SVD) of a matrix U is a decomposition of the form

$$U = QDR^T, \tag{4.155}$$

where D is a diagonal matrix and Q and R are orthogonal matrices [11]. Because of its orthogonality, the matrix Q accurately represents the linearly-independent components of the columns of U. For the application in SDQMC, we need a way to obtain an accurate representation of the SVD of the matrix U_σ The matrix U_σ is build up as a product of matrices

$$U_\sigma = U_{\sigma_{N_S}} \dots U_{\sigma_2} U_{\sigma_1}, \tag{4.156}$$

where the matrix U_{σ_j} represents the j^{th} inverse-temperature slice in the decomposition of the Boltzmann operator 4.1. Let the matrix T_n be the product

$$T_n = U_{\sigma_n} U_{\sigma_{n-1}} \dots U_{\sigma_2} U_{\sigma_1}. \tag{4.157}$$

Suppose that we have a SVD for T_n,

$$T_n = Q_n D_n R_n^T. \tag{4.158}$$

Then the SVD for T_{n+1} can be computed accurately in the following way:

$$\begin{aligned} T_{n+1} &= U_{\sigma_{n+1}} T_n = U_{\sigma_{n+1}} Q_n D_n R_n^T \\ &= \left(U_{\sigma_{n+1}} Q_n D_n \right) R_n^T. \end{aligned} \tag{4.159}$$

The matrices $U_{\sigma_{n+1}}$ and Q_n are well conditioned so they can be multiplied without loss of significant information. The multiplication with the diagonal matrix D_n, whose diagonal elements can be huge, scales the columns but does not mix them up, so the information on the linearly-independent components of the columns of T_{n+1} remains intact. For this product $U_{\sigma_{n+1}} Q_n D_n$ a new SVD can be computed,

$$U_{\sigma_{n+1}} Q_n D_n = Q_{n+1} D_{n+1} R^T. \tag{4.160}$$

All we have to further do to obtain the SVD $T_{n+1} = Q_{n+1} D_{n+1} R_{n+1}^T$, is to multiply the two orthogonal matrices R_n and R,

$$R_{n+1} = R_n R. \tag{4.161}$$

After N_t steps, the SVD for U_σ is obtained. It can then be used to calculate accurately the grand canonical and canonical trace (see algorithm 4.97). We presented this method as if a SVD should be computed after every inverse-temperature slice. This was merely for ease of notation. In practice, a number of inverse temperature slices can be taken together when multiplying 4.159.

A second approach [35] is based on the QR decomposition [11]. The QR decomposition of a matrix U is a decomposition of the form $U = QR$, where Q is an orthogonal matrix and R an upper triangular matrix. Just as for the SVD-based method, the discussion can be

based on the matrix product T_n, expression 4.157. The linearly-independent components of the matrices T_n are now obtained from a QR-like decomposition,

$$T_n = Q_n D_n R_n, \tag{4.162}$$

where D is a diagonal matrix, Q an orthogonal matrix and R an upper triangular matrix whose diagonal elements are all equal to 1. The reasoning is completely analogous as above, except that a QR-like decomposition 4.162 instead of a SVD has to be computed for $U_{\sigma_{n+1}} Q_n D_n$ This can be done using a Gram-Schmidt orthogonalization procedure [11]. As such, this method amounts to a reorthogonalization of the single-particle states after every (or every few) inverse temperature slice(s). Advantages of this approach compared to the SVD-based approach are

- A QR-like decomposition is computed much faster than a SVD. The Gram-Schmidt procedure requires approximately $2N_{\mathcal{S}}^3$ flops, whereas a SVD requires approximately $21N_{\mathcal{S}}^3$ flops [11].
- A QR-like decomposition can easily be adapted to matrix structures of the type 4.145, such that only half of the matrix needs to be orthogonalized.
- A QR-like decomposition is easy to program.

An advantage of the SVD-based approach is

- With a SVD, the matrices R_n are always well conditioned. This is not assured with a QR-like decomposition. In practice, we never experienced such problems with the QR-like decomposition.

For our calculations, if stabilization was necessary, we always used the QR-like decomposition. The orthogonalization was done after every m^{th} inverse-temperature slice, with m such that $m\beta/N_t \simeq 1$ for the Hubbard model (see chapter 5), or $m\beta/N_t \simeq 2$ for the nuclear pairing model (see chapter 6). Using this technique, the SDQMC became very stable. By storing the logarithms instead of the actual values of the diagonal elements of the matrices D_n, the trace of $\hat{U}_\sigma$ could be calculated at any value of β.

4.6.2 Efficient MCMC sampling - Hybrid samplers.

In section 3.4.1 it was already mentioned that the proposition kernel for Metropolis sampling should be devised in such a way that about 25% of the trial moves is accepted. For SDQMC, this can be arranged in several ways. Let the configuration be given by a vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_{N_t})$, with one component for each inverse-temperature slice. Each of these components σ_j will have a number of subcomponents σ_{ji} related to the decomposition of $e^{-\beta \hat{H}_2}$ (see section 2.2). In the paper [37], it was suggested to update the componsents σ_j consecutively. For each component j , a fixed number (say n_{sc}) of subcomponents σ_{ji} is changed to generate a trial move for the Metropolis algorithm. Note that N_{sc} has to be odd. Otherwise the Markov chain is not irreducible: the chain will never go from an initial configuration σ to a configuration σ' that differs from σ in an odd

number of subcomponents σ_{ij}. The number n_{sc} is choosen such that about half of the trial moves is accepted (as said before, a somewhat lower acceptance rate might be more appropriate). This way of generating trial moves is suited for sampling schemes based on rank-one updates, as discussed in 4.2.1. If the whole matrix U_σ is recalculated for every configuration σ', this way of generating trial moves can be improved in a few ways.

- Choose n_{sc} randomly between 1 and a maximum value N_{sc} at each Markov step. N_{sc} can be somewhat larger than the fixed value for n_{sc} mentioned above. This procedure will lead to approximately the same acceptance rate for the Markov chain, but it connects a configuration σ to much more trial configurations σ'. Therefore the underlying transition kernel of the Markov chain will connect more configurations. As discussed in section 3.2.6, this leads to shorter autocorrelations and a faster convergence.

- If U_σ is recalculated completely for every trial configuration σ, it is advantageous to pick the n_{sc} subcomponents not from one inverse-temperature slice, but randomly from all subcomponents of σ. Again, the underlying transition kernel of the Markov chain will connect more configurations.

As discussed in section 2.1, the Boltzmann operator $e^{-\beta\hat{H}}$ has to be cut in N_t inverse-temperature slices $e^{-\frac{\beta}{N_t}\hat{H}}$ in order to avoid errors in its decomposition originating from the non-commutativity of parts of the Hamiltonian H. To get a good accuracy with SDQMC, N_t sometimes has to be quite large, of the order of a few hundred. In such cases, a lot of matrix multiplications are needed to build up the matrix U_σ for just one configuration σ. Because matrix multiplications require $\mathcal{O}\left(N_S^3\right)$ flops, this will be the bottleneck of the SDQMC calculation. Recalculating U_σ completely for every trial configuration σ is then not very efficient. The Markov-chain Monte-Carlo sampling scheme can be arranged in such a way that only a fraction of the matrix multiplications has to be repeated for the evaluation of a new configuration σ'. This can lead to a considerable reduction in computation time.
The N_t inverse-temperature slices are split into two parts, the first N_{t1} slices versus the last $N_{t2} = N_t - N_{t1}$ slices. The matrix U_σ is computed in two steps. Using the notation 4.157, the matrix $T_{N_{t1}}$ is calculated first. It is stored in the computer memory. Then it is multiplied with the last N_{t2} slices to obtain U_σ. Trial moves for the Metropolis sampling are generated by changing only the last N_{t2} components of σ. In other words, only the last N_{t2} slices are changed. $T_{N_{t1}}$ remains unaltered. Therefore, the first N_{t1} matrix multiplications have not to be repeated for the evaluation of σ'. This saves a lot of computer time. Of course, one should change the other components of σ too in order to get an irreducible Markov chain. Therefore, after a number of Markov steps (say N_c) where only the last N_{t2} components were changed, the components are shifted cyclically over a number of slices (say N_{shift}). Because of the permutational symmetry of the (grand) canonical trace, the next N_c Markov steps amount to a sampling of N_{t2} other components of σ. After such a shift, $T_{N_{t1}}$ has to be recalculated. However, because this happens only every N_c Markov steps, this has no big effect on the needed amount of computer time. What numbers should be taken for N_{t2}, N_c and N_{shift} ?

- If N_{t2} is too large, then too many matrix multiplications will still have to be done to obtain the $U_{\sigma'}$ and the calculation will be slow. If N_{t2} is too small, the Markov chain stays too long in the same region of the configuration space (the first N_{t1} slices remain the same) and the autocorrelations will be long. No general rule for N_{t2} can be given. For most applications, we obtained good results when N_{t2} was taken such that the amount of computer time needed for the N_{t2} matrix multiplications to build $U_{\sigma'}$ from $T_{N_{t1}}$ was of the same order as the amount of computer time needed for the evaluation of the trace of $\hat{U}_{\sigma'}$.

- If N_c is too small, then the computation of $T_{N_{t1}}$ will dominate the calculation, so it will be slow. If N_c is too long, the Markov chain stays too long in the same region of the configuration space (the first N_{t1} remain the same) and the autocorrelations will be long. No general rule for N_c can be given. For most applications we obtained good results with N_c such that the amount of computer time needed for the N_c Markov steps was somewhat larger than the amount of computer time needed to compute $T_{N_{t1}}$

- Inspired by the Gibbs sampler (see section 3.4.2), one could suggest to cut the N_t inverse-temperature slices in N_t/N_c parts of equal length and to shift after every N_c Markov steps a new part to last N_{t2} positions. This amounts to taking N_{shift} a multiple of N_t/N_c (provided that N_t/N_c is an integer number). Taking $N_{\text{shift}} = N_t/N_c$ leads to a deterministic-scan Gibbs sampler. Taking $N_{\text{shift}} = mN_t/N_c$ with m a random integer number between 1 and N_c leads to a random-scan Gibbs sampler. However, shorter autocorrelations are observed when N_{shift} is drawn randomly between 1 and N_t. This can be understood from the fact that the transition kernel for such a sampling scheme couples a configuration σ to much more other configurations σ'. As explained in section 3.2.6, this leads to a faster convergence and shorter autocorrelations for the Markov chain.

The N_c local updates (local, because only the N_{t2} last slices are updated) and the shift of σ over N_{shift} can be taken together and considered as one transition kernel for a *hybrid* Markov-chain Monte-Carlo method, as described in section 3.4.

Applications

Overview

SDQMC results for the 4×4 Hubbard model are presented in chapter 5. Special attention is given to a comparision between results obtained in the canonical and the grand canonical ensemble. SDQMC calculations for the nuclear pairing Hamiltonian are discussed in chapter 6. Results were obtained for a model with pairing in a degenerate shell and for a mean-field plus pairing model for nuclei in the Fe region. An outlook for SDQMC calculations of neutrino-nucleus scattering cross-sections is given in chapter 7.

5

The Hubbard model

For this work, we studied the two-dimensional Hubbard model mainly because it is a good test case for quantum Monte-Carlo methods. It has been studied extensively using SDQMC in the grand canonical ensemble and with ground-state projection [7, 14, 35], as well as using other quantum Monte-Carlo techniques (e.g. Green-function Monte-Carlo, worldline Monte-Carlo [7]). Our aim was to develop SDQMC algorithms for the canonical ensemble, with the application to atomic nuclei as the final goal. Therefore we did not systematically study the physical properties of the Hubbard model, nor did we calculate magnetic susceptibilities, spin-spin correlation functions, etc. However, we calculated thermodynamical quantities such as the internal energy and the specific heat. Our results are the first ones for the Hubbard model obtained within the canonical ensemble (with fixed numbers of spin-up and spin-down particles).

5.1 The Hubbard Hamiltonian

To introduce the Hubbard model, we follow Yosida[47]. When atoms are put together in a crystal lattice, the influence of the neighbouring atoms is felt by the valence electrons of the atoms in the lattice. The orbitals of the outermost valence electrons, which are responsible for the cohesive energy of the crystal, are modified compared to the orbitals in free atoms. In metals these electrons can move in the crystal. They become conduction electrons which are described by the Bloch function.
Let us represent the electron orbitals in the incompletely filled shell of the ions in the crystal lattice ($3d$ orbitals in iron-group elements or $4f$ orbitals in rare-earth elements) by a nondegenerate localized orbital $\phi_j(r)$, where the index j indicates a site on the lattice. The Hamiltonian that describes the interactions of the valence electrons can be expressed as

$$\hat{H} = \sum_{i,j} t_{ij} \left(\hat{a}^\dagger_{\uparrow i} \hat{a}_{\uparrow j} + \hat{a}^\dagger_{\downarrow i} \hat{a}_{\downarrow j} \right) + U \sum_i \hat{a}^\dagger_{\uparrow i} \hat{a}_{\uparrow i} \hat{a}^\dagger_{\downarrow i} \hat{a}_{\downarrow i}. \tag{5.1}$$

This Hamiltonian is called the 'Hubbard Hamiltonian'. The operator $\hat{a}^\dagger_{\uparrow i}$ ($\hat{a}^\dagger_{\downarrow i}$) creates an electron with spin up (spin down) in the localized orbital $\phi_i(r)$ on lattice site i. If these localized orbitals are mutually orthogonal, the above operators and their Hermitian conjugates fulfill the well-known commutation relations for fermion operators. If the localized orbitals are not mutually orthogonal, the $\phi_i(r)$ should be interpreted as Wannier functions, which are orthogonalized linear combinations of Bloch functions. We restrict the model to a single conduction band, i.e. we assume that there are only two orbitals per lattice site: one spin-up and one spin-down orbital.
The first term in 5.1 represents the 'hopping' of electrons throughout the lattice.

$$\hat{H}_1 = \sum_{i,j} t_{ij} \left(\hat{a}^\dagger_{\uparrow i} \hat{a}_{\uparrow j} + \hat{a}^\dagger_{\downarrow i} \hat{a}_{\downarrow j} \right). \tag{5.2}$$

This term transfers electrons from site i to site j with a strength t_{ij}. For numerical calculations one assumes a finite, periodic lattice. Because of the periodicity of the t_{ij}, $\hat{H}_1$ can be diagonalized by a Fourier transformation. If we indicate the sites on the lattice with a vector notation, e.g. for a two-dimensional $L \times L$ square lattice $j = (j_1, j_2)$, then we can define the operators

$$\hat{c}^\dagger_{sn} = \frac{1}{L} \sum_{n=(n_1,n_2)} e^{i\frac{2\pi}{N}(n_1 j_1 + n_2 j_2)} \hat{a}^\dagger_{sj}, \tag{5.3}$$

where the index s denotes the spin ($\uparrow$ or $\downarrow$). The kinetic term $\hat{H}_1$ is then given by

$$\hat{H}_1 = \sum_{ns} e_n \hat{c}^\dagger_{sn} \hat{c}_{sn}, \tag{5.4}$$

with e_n the single-particle energy related to a state $\hat{c}^\dagger_{sn}|\rangle$. For simplicity it is often assumed that t_{ij} is equal to a constant value t for sites i and j next to one another, and zero otherwise. The term then expresses 'nearest neighbour hopping'. If one retains t_{ij} only for sites i and j close to one another, and puts $t_{ij} = 0$ otherwise, one assumes a 'tight binding limit'. This limit is justified by the fact that the orbitals ϕ_i and ϕ_j will have an extremely small overlap for sites i and j that are not close to one another on the lattice [48]. For our calculations, we limited the t_{ij} to 'nearest neighbour hopping' only. The single-particle energies are then given by

$$e_n = 2t \left[\cos\left(\frac{2\pi}{N} n_1\right) + \cos\left(\frac{2\pi}{N} n_2\right) \right]. \tag{5.5}$$

The second term in the Hamiltonian 5.1 describes the Coulomb repulsion between two electrons on the same lattice site. Note that the Pauli principle forces two electrons in the same orbital to have opposite spins. Electrons in two different orbitals will interact electromagnetically too, but these interactions are neglected. For some systems this omission is justified by the screening of the interactions between electrons on different sites and by the small overlap between the Wannier orbitals [48].
The restriction to a single conduction band, nearest-neighbour hopping and on-site repulsion only, leads to the 'minimal Hubbard model'. The Hamiltonian is then given by

$$\hat{H} = t \sum_{\langle i,j \rangle} \left(\hat{a}^\dagger_{\uparrow i} \hat{a}_{\uparrow j} + \hat{a}^\dagger_{\downarrow i} \hat{a}_{\downarrow j} \right) + U \sum_i \hat{n}_{\uparrow i} \hat{n}_{\downarrow i}, \tag{5.6}$$

where the notation $\langle i,j \rangle$ indicates that the summation for the one-body part runs over pairs of neighbouring sites, and

$$\hat{n}_{\uparrow i} = \hat{a}^{\dagger}_{\uparrow i}\hat{a}_{\uparrow i} \tag{5.7}$$

$$\hat{n}_{\downarrow i} = \hat{a}^{\dagger}_{\downarrow i}\hat{a}_{\downarrow i} \tag{5.8}$$

According to the dimensionalty of the underlying lattice, the model is called the one-, two- ,three- or even infinite-dimensional Hubbard model. Though the Hamiltonian looks very simple, it already leads to a complicated many-body problem. Because the one- and two-body terms do not commute, the system will have a highly correlated ground state. The many-body problem for the one-dimensional Hubbard model can be solved exactly using the so called 'Bethe Ansatz' [47]. While the one- and infinitely dimensional models are more of theoretical interest, the two- and three-dimensional models are related to interesting physical systems. Over recent years, the two-dimensional model has drawn considerable attention in connection to high-temperature superconductivity. The electronic properties of high-temperature superconductors are believed to originate from electrons moving in planes of cupper and oxygen ions, isolated by layers of non-conducting ions. It was suggested by Anderson [49] that electron correlations described by the Hubbard model might give a qualitative picture of high-temperature superconductivity. Up to now, calculations have failed to give conclusive evidence for this conjecture.
So far, we assumed that the two-body interaction strength U in 5.1 was positive, i.e. that the interaction was repulsive. This is obvious if the two-body term describes a Coulomb repulsion. However, couplings to lattice deformations (phonons) or other collective degrees of freedom, can lead to short-range *attractive* correlations between electrons. This is realized in an effective way in the attractive Hubbard model. It has been studied in connection to superconductivity, because, by design, it yields superconductivity in its ground state [7].
An interesting symmetry exists between the repulsive and the attractive Hubbard model. The asymmetric particle-hole transformation

$$\hat{a}_{\uparrow i} \rightarrow (-1)^{i_1+i_2}\hat{a}_{\uparrow i}, \tag{5.9}$$

$$\hat{a}_{\downarrow i} \rightarrow (-1)^{i_1+i_2}\hat{a}^{\dagger}_{\downarrow i}, \tag{5.10}$$

with i_1 and i_2 indicating the lattice 'position' $i = (i_1, i_2)$, transforms the Hamiltonian 5.6 to

$$\hat{H} = t\sum_{\langle i,j \rangle}\left(\hat{a}^{\dagger}_{\uparrow i}\hat{a}_{\uparrow j} + \hat{a}^{\dagger}_{\downarrow i}\hat{a}_{\downarrow j}\right) - U\sum_{i}\hat{n}_{\uparrow i}\hat{n}_{\downarrow i} + U\hat{N}_{\uparrow}, \tag{5.11}$$

where $\hat{N}_{\uparrow} = \sum_i \hat{n}_{\uparrow i}$ is the number operator for the spin-up particles. The repulsive Hubbard model for $N_{\uparrow}$ spin-up and $N_{\downarrow}$ spin-down particles is transformed into an attractive Hubabrd model for $N_{\uparrow}$ spin-up and $N_{\mathcal{S}} - N_{\downarrow}$ spin-down particles, with $N_{\mathcal{S}}$ the number of lattice sites (because of the symmetry between spin-up and spin-down states, we define $N_{\mathcal{S}}$ here such that it equals half the number of single-particle states). Therefore, the energy of a repulsive Hubbard model with intercation strength U and with $N_{\uparrow}$ spin-up and $N_{\downarrow}$ spin-down electrons is equal to $UN_{\uparrow}$ plus the energy of an attractive Hubbard model with intercation strength $-U$ and with $N_{\uparrow}$ spin-up and $N_{\mathcal{S}} - N_{\downarrow}$ spin-down electrons. This

property can be extended to any lattice that has two types of sites, say A and B, such that the electrons hop from A sites to B sites and from B sites to A sites, but never from A to A sites nor from B to B sites. Such lattices are called 'bipartite' lattices.
Applying this relation twice, once with a transformation of the spin-up particles and once with a transformation of the spin-down particles, leads to a relation between the energy $E_{N_\uparrow,N_\downarrow}$ of a repulsive $(N_\uparrow, N_\downarrow)$ system and the energy $E_{N_\mathcal{S}-N_\uparrow,N_\mathcal{S}-N_\downarrow}$ of a repulsive $(N_\mathcal{S} - N_\uparrow, N_\mathcal{S} - N_\downarrow)$ system:

$$E_{N_\mathcal{S}-N_\uparrow,N_\mathcal{S}-N_\downarrow} = E_{N_\uparrow,N_\downarrow} + U\left(N_\mathcal{S} - N_\uparrow - N_\downarrow\right). \tag{5.12}$$

Because we used the Hubbard model mainly as a benchmark for testing SDQMC algorithms and because of limitations in computer power, we restricted our calculations to small lattices. Most calculations were done with a periodic, two-dimensional square lattice of 4×4 sites. This leads to 32 single-particle states, and matrices of dimension $N_\mathcal{S} = 16$. To reduce the influence of finite-size effects, which is necessary to make contact with real physical systems, larger lattice sizes are needed, with 10×10 sites or more. Parametrizations used to describe real crystals, have values for t of the order of $1eV$ and values for U of the order of 1 to $10eV$ [50]. In the rest of this chapter, we work in appropriate energy units such that $t = 1$.

5.2 Decompositions for SDQMC

As discussed in chapter 2, several ways of decomposing the Boltzmann operator $e^{-\beta\hat{H}}$ for application in SDQMC calculations exist. For the Hubbard model, we tried several expansions.
First of all, there is Hirsch's discrete Hubbard-Stratonovich decomposition discussed in section 2.2.2. The one- and two-body part of the Hamiltonian are separated using the Suzuki-Trotter formula 2.12. The decomposition 2.72 for the two-body interaction leads to diagonal matrices, that can be multiplied quickly ($2N_\mathcal{S}^2$ flops per matrix multiplication). The matrix representation of the operator $e^{-\beta\hat{H}_1}$ leads to a dense matrix. The specific eigenstructure 5.3 however, allows fast multiplications with this matrix using a fast Fourier transform. This requires $6\ _2\log(N_\mathcal{S})N_\mathcal{S}^2$ flops per matrix multiplication [12], which is fast compared to the $N_\mathcal{S}^3$ flops needed for a general matrix multiplication For small lattice dimensions this number of flops can be optimized even further; for a 4×4 lattice to $6N_\mathcal{S}^2$ flops, for a 8×8 lattice to $11N_\mathcal{S}^2$ flops. Another approach for the matrix $e^{-\beta H_1}$ is to use the series expansion

$$e^{-\beta H_1} = 1 - \beta H_1 + \frac{\beta^2}{4}H_1^2 + \frac{\beta^3}{6}H_1^3 + \frac{\beta^4}{24}H_1^4. \tag{5.13}$$

Because of the sparse structure of the matrix H_1, the matrix multiplications with H_1 can be performed quickly. The expansion should at least be of fourth order, in order to make the error small compared to the error originating from the Suzuki-Trotter breakup 2.12. This approach requires $20N_\mathcal{S}^2$ flops per matrix multiplication, so it is advantageous only

for large lattices. It has the disadvantage that it is not a completely exact representation of $e^{-\beta\hat{H}_1}$.

To improve the speed of the matrix multiplications for the one-body part even more, a somewhat different decomposition was used. It is based on the operators $\hat{H}_{[i]}$ defined by

$$\hat{H}_{[i]} = t\sum_j \left(\hat{a}^\dagger_{\uparrow i}\hat{a}_{\uparrow j} + \hat{a}^\dagger_{\downarrow i}\hat{a}_{\downarrow j}\right) + U\hat{n}_{\uparrow i}\hat{n}_{\downarrow i}, \tag{5.14}$$

where the summation over j runs over neighbouringh sites of site i only. Obviously, $\hat{H} = \sum_i \hat{H}_{[i]}$. The operators $\hat{H}_{[i]}$ have the particular property that

$$\left[\hat{H}_{[i]}, \hat{H}_{[j]}\right] = 0, \tag{5.15}$$

if $(i_1 + i_2)$ and $(j_1 + j_2)$ are both even or both odd. We split $\hat{H}$ in two parts, $\hat{H} = \hat{H}_o + \hat{H}_e$, with

$$\hat{H}_o = \sum_{i,\ (i_1+i_2 \text{odd})} \hat{H}_{[i]}, \tag{5.16}$$

$$\hat{H}_e = \sum_{i,\ (i_1+i_2 \text{even})} \hat{H}_{[i]}. \tag{5.17}$$

The Boltzman operator $e^{-\beta\left(\hat{H}_o+\hat{H}_e\right)}$ is split into factors $e^{-\beta\hat{H}_o}$ and $e^{-\beta\hat{H}_e}$ using the Suzuki-Trotter formula 2.12. Because of the commutation relation 5.15, each of these factors can be split into single factors $e^{-\beta\hat{H}_{[i]}}$ without approximations. The operators $e^{-\beta\hat{H}_{[i]}}$ are rank-two operators (see section 2.2.2). They can be approximated with errors of order β^2 using 2 terms of the form 2.53. An exact representation can be obtained using 3 terms, but we used only the form with 2 terms. The error of the decomposition of the Boltzmann operator is now related to the commutators between $\hat{H}_o$ and $\hat{H}_e$, instead of the commutators between $\hat{H}_1$ and $\hat{H}_2$. The error is comparable in both cases. The big advantage of this decomposition is that the matrix multiplications for one inverse-temperature slice now require only $6N_S^2$ flops in total, one-body and two-body interaction included, for any lattice size. Furthermore, we found that this $\hat{H}_o$-$\hat{H}_e$ decomposition needed considerably less inverse temperature slices to converge than the first decomposition.

A third decomposition that we used, is based on a Suzuki-Trotter separation of the one- and two-body parts of the Hamiltonian, just like the first method. The two-body part is handled using a decomposition of the form 2.94. The second-order expansion 2.97 was used, but it required too many inverse temperature slices to converge (of the order of $N_t = 1000$ for a 4×4 system with $U = 4$, $\beta = 8$.) The fourth-order expansion 2.99 converged faster. The use of these decompositions was motivated by the fact that the matrices involved in the decomposition of the two-body interaction had very few non-zero elements, so that matrix multiplications could be performed extremely fast. A second reason was that these decompositions lead to configuration spaces of smaller dimensions than the first two decompositions. Therefore, the MCMC sampling could be expected to converge faster and to have shorter autocorrelations. However, because a large number of inverse temperature slices were needed compared to the first two methods, the advantages were cancelled. Furthermore, it was difficult to set up the Markov chain such that reasonable acceptance

rates were obtained: the weight distribution of the configurations was not smooth enough to allow an efficient generation of trial moves.
The results presented in the next section were all obtained using the first or the second decomposition.
A way to improve the MCMC sampling is obtained using guided sampling (see section 3.4.5). By neglecting the one-body part of the Hamiltonian, the matrices representing the terms $\hat{U}_\sigma$ in the decomposition 4.1 for the Boltzmann operator all become diagonal. Their N-body trace can be evaluated quickly, in $\mathcal{O}(N N_{\mathcal{S}})$ operations. Therefore, their weight is useful as a guiding weight for the sampling of the exact terms. Especially at low values for β ($\leq 2/|t|$) and at strong interaction strengths ($U \simeq 8$), the efficiency of the Markov chain is considerably improved.

5.3 Thermodynamic properties of the 4×4 Hubbard model

All results presented in this section pertain to a minimal, repulsive Hubbard model on a periodic 4 lattice, in the canonical ensemble. We did a few calculations for larger lattices too, (6×6 and 8×8). They required much more computation time and therefore we did not perform systematic studies of the influence of the temperature, the number of particles nor the interaction strength on the thermodynamic properties of such large systems.
All calculations were done for a fixed number of spin-up particles ($N_\uparrow$) and spin-down particles ($N_\downarrow$). This puts a further restriction on the canonical ensemble, which normally contains all states with $N_\uparrow + N_\downarrow = N$. To obtain the full canonical ensemble, one should add up the results for all particle numbers $N_\uparrow$ and $N_\downarrow$ that satisfy $N_\uparrow + N_\downarrow = N$. This can easily be done by adding one line to the algorithm 4.97. Because our aim was not to obtain a complete description of the Hubbard model, but to develop a SDQMC algorithm, we did not make this summation. Calculations for a fixed set of particle numbers $(N_\uparrow, N_\downarrow)$, have the advantage that the ground state is reached at lower temperatures of β, while the algorithm 4.97 still allows to take into account the full complexity of the ground state. To indicate a system with $N_\uparrow$ spin-up and $N_\downarrow$ spin-down particles we the notion '$(N_\uparrow, N_\downarrow)$ filling' is used.
In order to avoid confusion, the symbol E in this chapter is used for the internal energy of the system, while the symbol U is reserved for the interaction strength parameter in the Hamiltonian 5.6. All error limits indicate 95%-confidence intervals. The error bars were omitted if they were smaller than the markers of the data points. The lines in the plots are ment to guide the eye, they are no fits nor theoretical predictions.
For a clear interpretation of the results, it is interesting to take a closer look at the spectrum of the one-body part of the Hamiltonian, i.e. the Hubbard model for $U = 0$. The single-particle energies are given by expression 5.5. A schematic picture of this spectrum for the 4×4 lattice is given in figure 5.1. Configurations with 5 spin-up or 5 spin-down electrons correspond to closed shells. The $(5 \uparrow 5 \downarrow)$ system has a ground state energy of -24.

Figure 5.1: *Single-particle spectrum of the* 4×4 *Hubbard model at* $U = 0$.

5.3.1 Results at $(7 \uparrow 7 \downarrow)$ filling.

To illustrate how thermodynamical quantities can be calculated using SDQMC calculations for the Hubbard model, we discuss the calculations at $(7 \uparrow 7 \downarrow)$ filling in some more detail. A first issue, is to fix the number of inverse temperature intervals, N_t. A number of SDQMC runs were done with different values for N_t. Using Hirsch's discrete Hubbard-Stratonovich transform 2.72, a good convergence of the results was obtained at $N_t = 20\beta$ for an interaction strength of $U = 4$, while at interaction strength $U = 8$, we had to take $N_t = 40\beta$. Thus we took $N_t = 5\beta U$. We observed that less inverse temperature slices are needed to obtain convergence with the fast decomposition discussed in 5.2. However, for consistency, we took $N_t = 5\beta U$ for these calculations too.
The sign problem (see section 4.5) spoiled the calculations above a certain value for β. Figure 5.2 shows the evolution of the average sign as a function of β. The average sign $\bar{s}$ decreases faster for $U = 8$ than for $U = 4$. As discussed in section 4.5.4, $\bar{s}$ is related to the non-Hermitian part of the operators $\hat{S}_\sigma$ that show up in the decomposition of the Boltzmann operator 4.1. For the Hubbard model, these terms originate mainly from the non-commutativity of the one-body Hamiltonian $\hat{H}_1$ with the operators $\hat{n}_{si}$ used in the discrete Hubbard-Stratonovich transform 2.72. Formally, the operator $\beta \hat{H}_{t,U}$ is equivalent with the operator $(2\beta)\hat{H}_{t/2,U/2}$. A system with Hamiltonian $\hat{H}_{t/2,U}$ will lead to a larger average sign $\bar{s}$ at any value of β than a system with Hamiltonian $\hat{H}_{t,U}$. Therefore, for $t = 1$, the average sign at an inverse temperature β for $U = 8$ will be smaller than the average sign at inverse temperature β for $U = 4$, but at least as large as the average sign at inverse temperature 2β for $U = 4$. This can be seen in figure 5.2.
As discussed in section 4.5.1, the errors on the Monte-Carlo results are inversely propor-

Figure 5.2: *Average sign $\bar{s}$ as a function of the inverse temperature β for the 4×4 Hubbard model at $(7 \uparrow 7 \downarrow)$ filling, for $U = 4$ and $U = 8$. The dotted curve shows the average sign for the system with $U = 4$ at inverse temperature 2β.*

tional to $\bar{s}$. Figure 5.3 illustrates this for the errors on the internal energy E and on C/β^2. (The error on C/β^2 is plotted instead of the error on the specific heat C, because the latter is calculated by multiplying $\langle \hat{H}^2 \rangle_C - \langle \hat{H} \rangle_C^2$ with β^2, see expression 4.113). Though the error depends on a number of factors (length of the Markov chain, autocorrelations), figure 5.3 clearly shows that for small $\bar{s}$, the errors diverge like $1/\bar{s}$ (this is the dotted curve in figure 5.3).

Results for the internal energy E and the specific heat C are shown in figure 5.4. At high values of β, the internal energy E could be calculated much more accurately than the specific heat C. One reason for this is the factor β^2 in expression 4.113. For $U = 4$ the results for E became inaccurate around $\beta = 12$, the results for C became inaccurate around $\beta = 5$. For $U = 8$ the results for E became inaccurate around $\beta = 5$, the results for C became inaccurate around $\beta = 3$. Peaks in the specific-heat curve generally are related to changes in the internal structure of the system as the temperature increases or decreases. A clear peak in the specific heat is a signature of a phase transition. At $U = 4$, one peak in the specific heat, around $\beta = 1$, can be observed, while at $U = 8$, it looks like there could be two peaks, one around $\beta \simeq 0.5$, and maybe one at values of $\beta \geq 2$. This indicates a qualitative difference between the system for $U = 4$ and $U = 8$. Note that these calculations were restricted to fixed values of $N_\uparrow$ and $N_\downarrow$. So phase transitions related to a change in particle number or a change in the difference between the number of spin-up and spin-down particles cannot be observed here. Though β is the natural variable for the computations, a presentation of the results as a function of temperature is equivalent. Results for the internal energy E and the specific heat C as a function of

Figure 5.3: *Size of the errror limits on the internal energy E and the quantity C/β^2 for the 4×4 Hubbard model at $(7 \uparrow 7 \downarrow)$ filling, for $U = 8$. The dotted curve is $1/\bar{s}$*

temperature are shown in figure 5.5.
By integrating E numerically from $\beta = 0$ to $\beta = 1/T$, the logarithm of the canonical partition function Z was obtained as a function of temperature. From relation 4.12, the entropy S was obtained. These results are shown in figure 5.6

5.3.2 Results at various fillings

.
We performed calculations for the minimal repulsive Hubbard model at various fillings. Some results for $U = 4$ are shown in figures 5.7 to 5.12. For the 4×4 model at $U = 4$, we observe three types of curves for the specific heat. The $(5 \uparrow 5 \downarrow)$, $(5 \uparrow 6 \downarrow)$ and $(4 \uparrow 5 \downarrow)$ systems have a strong peak in the specific heat curve around $\beta \simeq 2$ to 2.5. They also have a low ground state energy. The $(5 \uparrow 5 \downarrow)$ has no sign problems. Clearly, these properties are related to the closed-shell structure at $N_\uparrow = 5$ or $N_\downarrow = 5$ (see figure 5.1). The $(8 \uparrow 8 \downarrow)$ and $(7 \uparrow 9 \downarrow)$ systems exhibit a nearly flat plateau in the specific heat curve from $\beta \simeq 1$ to $\beta \simeq 5$. This indicates that these systems have a lot of low-lying excited states. Because of the half filling ($N_\uparrow + N_\downarrow = 16$), there are no sign problems for these systems. All the other systems that we have studied show a maximum of $C \simeq 6$ around $\beta \simeq 1$ to 1.5. They cool to their ground states faster than the half-filled or closed-shell systems. This indicates that the non-half-filled open-shell systems have a larger gap between the ground state and the first-excited state than the half-filled or closed-shell systems.
We also notice very little difference between the results for some systems that have the same total number of particles but a different distribution over spin-up and spin-down

Figure 5.4: *internal energy E (figure a) and the specific heat C (figure b) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(7 \uparrow 7 \downarrow)$ filling, for $U = 4$ and $U = 8$.*

Figure 5.5: *internal energy E (figure a) and the specific heat C (figure b) as a function of the temperature T, for the 4×4 Hubbard model at $(7 \uparrow 7 \downarrow)$ filling, for $U = 4$ and $U = 8$.*

Figure 5.6: *Logarithm of of the canonical partition function Z and entropy S as a function of the temperature T, for the 4×4 Hubbard model at $(7 \uparrow 7 \downarrow)$ filling, for $U = 4$ and $U = 8$.*

states. This is the case for $(8 \uparrow 8 \downarrow)$ filling and $(7 \uparrow 9 \downarrow)$ filling, that have very similar energies for all values of β. The same holds for the $(7 \uparrow 7 \downarrow)$ filling and $(6 \uparrow 8 \downarrow)$ filling, and for the $(6 \uparrow 6 \downarrow)$ filling and $(5 \uparrow 7 \downarrow)$ filling. This means that for these systems there is actually very little energy needed to flip one spin. It also means that in a complete canonical or grand canonical ensemble, these states will separate only at very low temperature.
The internal energies thus obtained for various fillings are listed in tables A.1 and A.2. Ground state energies obtained by diagonalization methods and by SDQMC with ground-state projection are listed for comparison (taken from the review [7] and references therein).
We also studied some systems at interaction strength $U = 8$. At this strength, also the $(5 \uparrow 5 \downarrow)$ system became sensitive to the sign problem. The results are presented in figures 5.13 to 5.16.
The specific heat of the two-dimensional Hubbard model has been studied by Duffy and Moreo using a SDQMC method in the grand canonical ensemble [52]. They calculated the internal energy at several temperatures T, while tuning the chemical potential at each temperature such that $\langle \hat{N} \rangle$ remained constant. The obtained energies were fitted by a polynomial in T of order 4 or 6. The specific heat was then obtained by deriving this polynomial to T. Note that the specific heat curve is then given by a polynomial of order 3 or 5. Therefore, this procedure might lead to more pronounced or even artificial peaks in the specific-heat curve than a procedure based on the evaluation of the specific heat for each Monte-Carlo sample. Furthermore, it is hard to establish error limits with this procedure (no error limits are given in reference [52]). Duffy and Moreo find two peaks in the specific-heat curve for the half-filled 6×6 Hubbard model. These peaks are more pronounced at stronger interaction strengths U. A low temperature peak around $T = 0.25$ is ascribed to the spin degrees of freedom, while a high temperature peak around $T = 1$ to 3 is ascribed to the charge degrees of freedom. The former degrees of freedom correspond to fluctuations in $N_\uparrow$ and $N_\downarrow$, whith fixed $N = N_\uparrow + N_\downarrow$, while the latter correspond to fluctuations in N. However, we observe similar features in the specific-heat curve for the 4×4 Hubabrd model at $(8 \uparrow, 8 \downarrow)$ filling with $U = 8$ (see figure 5.16). Because $N_\uparrow$ and $N_\downarrow$ are fixed for this system, this might suggest a different interpretation of the peaks. Further research is needed to settle this question.

5.4 Some remarks concerning the canonical and grand canonical ensemble

Using the results for the 4×4 Hubbard model at $U = 4$, we can simulate a grand canonical ensemble. We could calculate the values for the grand canonical ensemble directly using the SDQMC method discussed in section 4.2. What we want to study here, however, is the contribution of each $(N_\uparrow N_\downarrow)$ system to the grand canonical ensemble in order to compare the merits of the grand canonical and the canonical SDQMC methods.
The canonical partition function is calculated for all fillings for which results are presented in tables A.1 and A.2, and for the fillings whose energies can be obtained from these results using relation 5.12. The values for the partition functions are multiplied with a

Figure 5.7: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(5 \uparrow 5 \downarrow)$ filling, for $U = 4$.*

Figure 5.8: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(6 \uparrow 6 \downarrow)$ filling, for $U = 4$.*

Figure 5.9: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(8 \uparrow 8 \downarrow)$ filling, for $U = 4$.*

Figure 5.10: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(5 \uparrow 7 \downarrow)$ filling, for $U = 4$.*

Figure 5.11: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(6 \uparrow 8 \downarrow)$ filling, for $U = 4$.*

Figure 5.12: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(7 \uparrow 9 \downarrow)$ filling, for $U = 4$.*

Figure 5.13: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(4 \uparrow 4 \downarrow)$ filling, for $U = 8$.*

Figure 5.14: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(5 \uparrow 5 \downarrow)$ filling, for $U = 8$.*

Figure 5.15: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(6 \uparrow 6 \downarrow)$ filling, for $U = 8$.*

Figure 5.16: *Internal energy E (figure a), the specific heat C (figure b) and average sign $\bar{s}$ (figure c) as a function of the inverse temperature β, for the 4×4 Hubbard model at $(8 \uparrow 8 \downarrow)$ filling, for $U = 8$.*

Figure 5.17: *The function $F(N_\uparrow, N_\downarrow)$ as defined by 5.19 for several values of the chemical potential μ for states with particle number $N = N_\uparrow + N_\downarrow$.*

factor $e^{-\beta\mu(N_\uparrow+N_\downarrow)}$ to obtain the weight of each system in the grand canonical ensemble. Not all possible fillings are included, so we only have an approximation for the grand canonical ensemble. But the included fillings are the ones with the lowest ground-state energies. Therefore, we obtain a good approximation to the grand canonical ensemble at lower temperatures. The grand canonical results presented hereafter are all based on this 'simulated' grand canonical ensemble.
First of all, we consider very low temperatures, i.e. very high values of β. Contributions of excited states to the canonical partition fuction can be neglected, so that it is given by

$$Z_{N_\uparrow N_\downarrow} = e^{-\beta E_0\left(N_\uparrow, N_\downarrow\right)}. \tag{5.18}$$

As an estimate for the ground state energy $E_0(N_\uparrow, N_\downarrow)$ we took the values for the internal energy at $\beta = 5$ as listed in tables A.1 and A.2. The weight of the $(N_\uparrow, N_\downarrow)$ system in the grand canonical ensemble is be proportional to the factor $e^{-\beta F\left(N_\uparrow, N_\downarrow\right)}$, with $F(N_\uparrow, N_\downarrow)$ given by

$$F(N_\uparrow, N_\downarrow) = E_0(N_\uparrow, N_\downarrow) + \mu(N_\uparrow + N_\downarrow). \tag{5.19}$$

This function $F(N_\uparrow, N_\downarrow)$ is plotted for several values of the chemical potential μ in figure 5.17 At $\mu = 2$ the curve is symmetric around $N = 8$ and the half filled systems dominate. This is a consequence of relation 5.12. At $\mu = 0$ the $(5\uparrow 5\downarrow)$ system dominates. At $\mu = 1$ the curve is almost flat. At this chemical potential the systems with $N = 5, 6, 7$ and 8 have almost equal weights. This illustrates the point we want to make in this section: it could happen that the grand canonical ensemble at low temperature is dominated by systems with $N = 5 + 5$ or $N = 8 + 8$, for any value of the chemical potential between

Figure 5.18: *Contribution of systems with various fillings to the grand canonical ensemble as a function of the chemical potential μ, for the 4×4 Hubbard model at $U = 4$ and $\beta = 5$.*

$\mu = 0$ and $\mu = 2$, and that systems with $N = 6 + 6$ or $N = 7 + 7$ have only very small contributions. This would happen if the points for $N = 6 + 6$ and $N = 7 + 7$ would lie a little bit below the curve that connects the points for $N = 5 + 5$ and $N = 8 + 8$. Our results for the energies at low temperatures are not precise enough to settle this point. But this raises the question: can information on the $(6 \uparrow 6 \downarrow)$ system or the $(7 \uparrow 7 \downarrow)$ be obtained from grand canonical results ?
Depending on the chemical potential μ, systems with different fillings dominate the grand canonical ensemble. This is illustrated for $\beta = 5$ in figure 5.18 The contribution of the $N = 12$ systems (the $(6 \uparrow 6 \downarrow)$ and $(5 \uparrow 7 \downarrow)$ systems), is maximal at $\mu \simeq 0.8$. But even then they only contribute 40%. Furthermore they are always dominated by the $(5 \uparrow 6 \downarrow)$ system or the $(6 \uparrow 7 \downarrow)$ system. The same holds for the $(7 \uparrow 7 \downarrow)$ system and the $(6 \uparrow 8 \downarrow)$ system, that are always dominated by the $(6 \uparrow 7 \downarrow)$ system or the $(7 \uparrow 8 \downarrow)$ system. To see the evolution of this effect with varying values of β, the average particle number $\langle \hat{N} \rangle_{GC}$ is calculated as a function of the chemical potential μ. The result is shown in figure 5.19 for several values of β. For a range of values of β, the value of μ is determined for which $\langle \hat{N} \rangle_{GC}$ is equal to 12 or 14. At these chemical potentials, the contribution of the systems with various fillings is calculated, so that a curve is obtained for these contributions as a function of β. These curves are plotted in figure 5.20 for $N = 12$ and 5.21 for $N = 14$.
It seems like the canonical systems with $N = 12$ ($N = 14$) will make up at most 50% of the grand canonical system with $\langle \hat{N} \rangle_{GC} = 12$ ($\langle \hat{N} \rangle_{GC} = 14$).
At present, the question if the ground state of the repulsive Hubbard model can be superconductive, is still unanswered and a subject of intensive research [51]. Monte-Carlo calculations with ground-state projection (see section 4.4) have been performed

Figure 5.19: *Average particle number $\langle \hat{N} \rangle_{GC}$ as a function of the chemical potential μ for the 4×4 Hubbard model at $U = 4$ and $\beta = 5$.*

Figure 5.20: *Contribution of systems with various fillings to the grand canonical ensemble as a function of the inverse temperature β, for the 4×4 Hubbard model at $U = 4$ and $\langle \hat{N} \rangle_{GC}$=12*

Figure 5.21: *Contribution of systems with various fillings to the grand canonical ensemble as a function of the inverse temperature β, for the 4×4 Hubbard model at $U = 4$ and $\langle \hat{N} \rangle_{GC}$=14*

mainly for closed-shell or half-filled systems, because of the sign problem. However, as mentioned in the previous section, the systems that are more difficult to study with this algorithm seem to have qualitatively different properties. A larger gap between the ground state and the first-excited state might be one of these properties. As indicated in this section, also the grand canonical algorithm is limited for the study of these systems. In order to make the repulsive Hubbard model superconductive, it has been suggested that one might have to include next-to-nearest-neighbour hopping in the one-body part of the Hamiltonian. Inclusion of such a term will leave the wavefunctions of the single-particle eigenstates unchanged, but it will change their single-particle energies. This can lead to a different shell structure for the one-body part of the Hamiltonian. At strong interaction strengths U, the main effect of these shifts in single-particle energies will be to shift the ground-state energies of systems with fixed $N_\uparrow$ and $N_\downarrow$. The correlated structure of the ground state might remain qualitatively unaltered. If all $(N_\uparrow, N_\downarrow)$ fillings are taken into account (at a given value for the chemical potential μ), there is only one state with definite filling $(N_{\uparrow 0}, N_{\downarrow 0})$ that has the lowest energy, the absolute ground state. An absolute ground state of a Hubbard model with next-to-nearest-neighbour hopping that exhibits superconductivity, might not be the absolute ground state of the minimal Hubbard model, but it might still be the ground state for a fixed filling $(N_{\uparrow 0}, N_{\downarrow 0})$. These arguments all motivate a systematic study of the minimal Hubbard model using SDQMC in the canonical ensemble, with fixed $(N_\uparrow, N_\downarrow)$.

6

The nuclear pairing Hamiltonian

6.1 The nuclear pairing Hamiltonian and nuclear many-body theory

The atomic nucleus is a complicated many-body system (with $A \simeq 20$ to 200 particles). The problem of describing a large number of protons and neutrons, all strongly interacting, is a challenge for theoretical nuclear physics. In a way, nuclear theory is the art of sympli-fying the many-body problem in such a way that accurate predicitions about the properties of real nuclei can be made, using models that allow precise numerical calculations.
A first, crude approximation to the nuclear many-body problem, is the use of a mean-field potential. Instead of describing the interactions between the nucleons in the system, one introduces a potential that reflects the average interaction of a nucleon with all other nucleons. This 'mean-field' potential can then be treated as if it referred to an external field. In this way, the many-body problem of N nucleons interacting with one another is reduced to N one-body problems of a nucleon moving in a mean field. A commonly used parametrization for this mean field is the Woods-Saxon potential [54]. A mean-field potential can be derived from effective nucleon-nucleon forces (e.g. forces of the Skyrme-type [55]) in a self-consistent way using the Hartree-Fock method [53].
In order to improve on the mean field approximation, one has to take into account the 'residual interactions', i.e. the interactions that are not accounted for within the mean-field potential. Let $\hat{H}$ denote the full Hamiltonian of the many-body system and $\hat{H}_{\rm mf}$ the Hamiltonian of the mean-field potential, then the Hamiltonian $\hat{H}_{\rm res}$ for the residual interactions is trivially given by

$$\hat{H}_{\rm res} = \hat{H} - \hat{H}_{\rm mf}\,. \tag{6.1}$$

In practice, one uses effective parametrizations for the residual Hamiltonian $\hat{H}_{\rm res}$. Among many others, the Landau-Migdal force is a commonly used parametrization [53]. In order to obtain an accurate description of the atomic nucleus, one should diagonalize the

Hamiltonian $\hat{H}_{\rm mf} + \hat{H}_{\rm res}$ in the space of all many-body states. However, this space is extremely large. If we would allow each particle to be in any of $N_{\mathcal{S}}$ single-particle states, the many-body space for N fermions would contain $\binom{N}{N_{\mathcal{S}}}$ states. This space becomes too large to allow an exact diagonalization of the Hamiltonian. Therefore the space has to be truncated in some way. A truncation with a clear physical motivation, is given by the nuclear shell model [56]. A core for the atomic nucleus is built by filling the lowest shells of eigenstates of the mean-field potential. Configurations where the lowest shells are completely filled constitute 'magic' nuclei. In experiments it is found that nuclei whose particle numbers correspond to these 'magic' fillings, are strikingly stable and more strongly bound than others. Non-magic nuclei can be considered to consist of a 'magic' core plus a few 'valence' nucleons in the lowest open shells of the mean-field potential. For these valence nucleons, many-body states with definite symmetry properties are considered. In such way a restricted many-body space is constructed, whose dimension is small enough to allow diagonalization of the Hamiltonian. Note that the trunctation of the many-body space has to be taken into account in the effective parametrization of the residual interaction. For the shell model, this can be done using the Brueckner G-matrix [53]. For medium heavy to heavy nuclei ($A \geq 50$), except for configurations with only a few valence nucleons, the valence shells become too large to allow diagonalization. SDQMC is a promising method to study these systems [6]: it allows to study much larger model spaces than diagonalization methods. Calculations where one considers multiple shells for the valence nucleons or where one considers more nucleons as valence nucleons ('no-core' calculations) could be possible. This requires the elimination of spurious states related to center-of-mass motion, a problem that is not yet satisfactory solved for SDQMC. Apart from a truncation of the model space, one also has to simplify the form of the interaction in order to make calculations feasible. A simple form that accounts for the short-range correlations induced by the residual interaction, is the nuclear pairing Hamiltonian $\hat{H}_P$ [53], that takes the form

$$\hat{H}_P = -\sum_{t=p,n} G_t \sum_{k,k'>0} \hat{a}^\dagger_{k't} \hat{a}^\dagger_{\bar{k}'t} \hat{a}_{\bar{k}t} \hat{a}_{kt}, \tag{6.2}$$

where the operators $\hat{a}^\dagger_{kt}$ create a particle in the corresponding single-particle eigenstates of the mean-field Hamiltonian in the valence shell, and where the index t indicates proton or neutron states. The notation $\bar{k}$ indicates the time-reversed state of the state k. In the so-called 'BCS' phase convention [53], time reversal has a simple form:

$$|\overline{nljmt}\rangle = |nlj{-}mt\rangle, \tag{6.3}$$

$$|\overline{nlj{-}mt}\rangle = -|nljmt\rangle. \tag{6.4}$$

Note that $|\bar{\bar{k}}t\rangle = -|kt\rangle$. The notation $k, k' > 0$ in 6.2 indicates that if state k is included in the summation, then $\bar{k}$ should not be included. A simple way to state this is that the summation for k and k' should run over states with $m > 0$ only. The interaction strength G_t depends on the model space and the system under study. A parametrization suggested by Bes and Sorensen for an atomic nuleus with A nucleons is [57]

$$G = \frac{20MeV}{A}. \tag{6.5}$$

A parametrization with different strengths for protons and neutrons is given in [58].
The combination of a mean-field potential and the pairing Hamiltonian leads to a Hamiltonian $\hat{H} = \hat{H}_{\rm mf} + \hat{H}_P$. Though this Hamiltonian looks simple, it already leads to a complicated many-body problem. An often used technique to tackle this problem, is the Bardeen-Cooper-Shrieffer (BCS) theory [53]. It leads to equations that can be handled easily in a numerical way and it has a clear interpretation in terms of quasiparticles. The disadvantage is that it gives only an approximate solution, and that leads to many-body states with changing numbers of particles. For some systems, exact solutions can be found (see section 6.2.1 and the paper by Richardson and Sherman [61]). A general, accurate solution for this many-body problem is even at present a topic of intensive research [59, 60, 63]. We have found that SDQMC is a very useful method for the study of the ground-state and finite-temperature properties of the nuclear pairing Hamiltonian.

6.2 Some properties of the nuclear pairing Hamiltonian

The Hamiltonian we study in this chapter has the form

$$\hat{H} = \sum_{t=p,n} \sum_{k>0} e_{kt} \left(\hat{a}^\dagger_{kt} \hat{a}_{kt} + \hat{a}^\dagger_{\bar{k}t} \hat{a}_{\bar{k}t} \right) - \sum_{t=p,n} G_t \sum_{k,k'>0} \hat{a}^\dagger_{k't} \hat{a}^\dagger_{\bar{k}'t} \hat{a}_{\bar{k}t} \hat{a}_{kt}. \tag{6.6}$$

The e_{kt} are the single-particle energies of the mean-field potential. From the structure of the Hamiltonian 6.6 we can make a few observations.

- Protons and neutrons decouple. The interactions between protons and neutrons are only represented by the mean field, there is no proton-neutron term in the residual interaction. This is of course a crude simplification of the real situation, but it allows us to separate the many-body problem for this model in two smaller ones, one for protons and one for neutrons. Therefore, in what follows, we will suppress the index t.

- There is a symmetry between states $k > 0$ and their time-reversed states $\bar{k}$. The number operators $\hat{N}_+$ and $\hat{N}_-$ for the $k > 0$ and $k < 0$ states respectively, both commute with the Hamiltonian. Hence the number of particles N_+ in states $k > 0$ and the number of particles N_- in states $k < 0$ are both conserved quantities.

- If a state k is occupied and its time-reversed state $\bar{k}$ is not, then the particle in state k only feels the mean field. It does not interact with the other particles and it remains in its state k . It does have an effect on the other particles in the system because it blocks both states k and $\bar{k}$ for the other particles, due to the Pauli principle.

- If a pair of particles occupies a state k and $\bar{k}$, it can be scattered to another pair of states k' and $\bar{k}'$, but the particles will allways remain 'accompagnied'. Burglin and Rowley [60] define two particles to be 'accompagnied' if they occupy a pair of states k and $\bar{k}$. A particle in a state k whose time-reversed state $\bar{k}$ is not occupied, is called 'unaccompagnied'. This notion should not be confused with the notion of 'paired' particles: two particles in a single j-shell can couple to a state with angular

momentum $J = 0$. In such a state, these particles are said to be 'paired'. In the two-particle state $\sum_m \hat{a}^\dagger_{jm}\hat{a}^\dagger_{\bar{j}m}|\rangle$ the particles are paired, while in the two-particle states $\hat{a}^\dagger_{jm}\hat{a}^\dagger_{\bar{j}m}|\rangle$, for any m, they are accompagnied. Burglin and Rowley also introduce the operator [60]

$$\hat{V} = \sum_{k>0}\left(\hat{a}^\dagger_{kt}\hat{a}_{kt} + \hat{a}^\dagger_{\bar{k}t}\hat{a}_{\bar{k}t}\right) - 2\sum_{k>0}\left(\hat{a}^\dagger_{kt}\hat{a}^\dagger_{\bar{k}t}\hat{a}_{\bar{k}t}\hat{a}_{kt}\right). \tag{6.7}$$

This operator gives the number of unaccompagnied particles. Though its physical meaning is completely different, we note the formal analogy between the operator $\hat{V}$ and the two-body term of the Hubbard Hamiltonian 5.1. The operator $\hat{V}$ commutes with the Hamiltonian. Therefore, the number of unaccompagnied particles is a conserved quantity. Furthermore, the many-body space can be split up in subspaces, each having a definite number of accompagnied and unaccompagnied particles. Burglin and Rowley use this splitting in order to obtain many-body spaces with small enough dimensions to allow diagonalization techniques such as the Lanczos algorithm [11] for the study of the lowest eigenstates of the system. Note that the subspace where all particles are accompagnied, contains only $J_z = 0$ states. The $\hat{V}$-subspaces are not invariant under rotations.

- Because of the attractive nature of the pairing, the ground state of a system with an even number of particles (even-even systems if both protons and neutrons are considered) will belong to the subspace where all particles are accompagnied.

6.2.1 Pairing in a degenerate shell

If all single-particle levels of the many-body system are degenerate, i.e. all e_k are equal, then the many-body problem for the Hamiltonian 6.6 can be solved exactly. Without loss of generality we can assume here that $e_k = 0$ for all k. Now we consider all possible Slater determinants that can be obtained by placing N identical particles in the $N_\mathcal{S}$ single-particle states. For every pair of states $(k, \bar{k})$ there are three possibilities: either none of the states is occupied, both states are occupied or only one of the states is occupied. In the latter case, the particle in the state k or $\bar{k}$ will not interact with the other particles and the Slater determinant will not couple to Slater determinants for which k and $\bar{k}$ are both occupied or both empty. A half-filled couple of states $(k, \bar{k})$ is sterile for the rest of the many-body problem. Therefore, one can reduce the many-body problem of N particles in the $N_\mathcal{S}$ single-particle states, from which exactly one occupies a state k or $\bar{k}$, to a many-body problem of $N-1$ particles in $N_\mathcal{S}-2$ single-particle states (the states k and $\bar{k}$ are left out). By repeated application of this reduction, one comes to a many-body problem of a system where all particles are accompagnied.
We discuss this for N accompagnied particles in $N_\mathcal{S}$ single-particle states. If all particles are accompagnied, then one can assign to every state $k > 0$ a quaspin $s^{(k)}$:

- $s^{(k)} = +\frac{1}{2}$ if both k and $\bar{k}$ are occupied,
- $s^{(k)} = -\frac{1}{2}$ if both k and $\bar{k}$ are empty.

One can define angular-momentum like operators for this quasispin in the following way [53] (for $k > 0$):

$$\hat{s}_+^{(k)} = \hat{a}_k^\dagger \hat{a}_{\bar{k}}^\dagger, \tag{6.8}$$

$$\hat{s}_-^{(k)} = \hat{a}_k \hat{a}_{\bar{k}}, \tag{6.9}$$

$$\hat{s}_0^{(k)} = \frac{1}{2}\left(\hat{a}_k^\dagger \hat{a}_k + \hat{a}_{\bar{k}}^\dagger \hat{a}_{\bar{k}} - 1\right). \tag{6.10}$$

This defines the quasispin vector $\hat{\mathbf{s}}^{(k)}$ These operators satisfy the well known commutation relations for angular-momentum operators:

$$\left[\hat{s}_+^{(k)}, \hat{s}_-^{(k)}\right] = 2\hat{s}_0^{(k)}, \tag{6.11}$$

$$\left[\hat{s}_0^{(k)}, \hat{s}_+^{(k)}\right] = \hat{s}_+^{(k)}, \tag{6.12}$$

$$\left[\hat{s}_0^{(k)}, \hat{s}_-^{(k)}\right] = -\hat{s}_-^{(k)}. \tag{6.13}$$

A set of values for the quasispins $s^{(1)}, s^{(2)}, \ldots s^{(\Omega)}$, where $\Omega = N_{\mathcal{S}}/2$ is the number of pairs $(k, \bar{k})$, specifies a fully paired Slater determinant. Just like angular momenta, the quasispins can be added together to obtain a total quasispin vector $\hat{\mathbf{S}}$:

$$\hat{\mathbf{S}} = \sum_{k>0} \hat{\mathbf{s}}^{(k)}. \tag{6.14}$$

The component $\hat{S}_0$ is related to the number operator by

$$\hat{S}_0 = \frac{\hat{N} - \Omega}{2}. \tag{6.15}$$

The interesting point about quasispin is that the pairing Hamiltonian 6.2 can be rewritten as

$$\hat{H}_P = -G\hat{S}_+\hat{S}_- \tag{6.16}$$

$$= -G\left(\hat{\mathbf{S}}\cdot\hat{\mathbf{S}} - \hat{S}_0^2 + \hat{S}_0\right). \tag{6.17}$$

This structure shows that the eigenstates of the Hamiltonian are given by the eigenstates of the total quasispin $\hat{\mathbf{S}}$. The eigenstates with definite quasispin quantum numbers S and S_0 can be constructed from the states with definite $s^{(k)}$, $k = 1, \ldots, \Omega$, using the well-known techniques for angular-momentum coupling [56]. Note that N-particle states must have $S_0 = (N - \Omega)/2$. Therefore the total quasispin for the N-particle eigenstates must be at least $|N - \Omega|/2$. The energy of an N-particle eigenstate with quasispin S is given by

$$E(S) = -G\left[S(S+1) - S_0^2 + S_0\right] \tag{6.18}$$

$$= -G\left[S(S+1) - \frac{1}{4}(N-\Omega)^2 + \frac{1}{2}(N-\Omega)\right]. \tag{6.19}$$

The maximum value for S is $\Omega/2$. This value of S gives the ground-state energy. The vacuum $|\rangle$ corresponds to a state with $S = \Omega/2$, $S_0 = \Omega/2$. The ground states for

energy level (MeV)	fully accompagnied subspace	complete space
-12	1	1
-6	5	65
-2	9	429
0	5	429

Table 6.1: *Comparision of degeneracies of energy levels for the $(h_{11/2})^6$ configuration in the fully accompagnied subspace and the complete many-body space, $(G = 1MeV)$.*

particle-numbers $N = 0, 2, \ldots, 2\Omega$ form a multiplet for which $S = \Omega/2$. Therefore, the N-particle ground state can be constructed by applying the raising operator $\hat{S}_+$ $N/2$ times to the vacuum $|\rangle$:

$$|\mathrm{GS}_N\rangle = \left(\hat{S}_+\right)^{N/2} |\rangle. \tag{6.20}$$

This shows that a system of N particles in a single j shell, with N even, has a ground state where all particles are paired.
Instead of the quasispin S one often uses the 'seniority' quantum number s defined by

$$s = \Omega - 2S. \tag{6.21}$$

It can take on the values $s = 0, 2, \ldots, \min(N, N_{\mathcal{S}} - N)$. The energy as a function of seniority is given by

$$E(N, s) = -\frac{G}{4}(N - s)(2\Omega - s - N + 2). \tag{6.22}$$

The degeneracy $d(s)$ of the states with energy $E(N, s)$ is given by [53]

$$d(0) = 1 \tag{6.23}$$

$$d(s) = \binom{\Omega}{s/2} - \binom{\Omega}{s/2 - 1} \text{ for s} > 0. \tag{6.24}$$

Because it is sometimes overlooked [53, 62], we remark here that these degeneracies only apply to the subspace in which all particles are accompagnied. They sum up to a total of $\binom{\Omega}{N/2}$ states. The complete many-body space contains much more states, $\binom{2\Omega}{N}$ in total. Taking all these states into account, one finds the same energy levels because the unaccompagnied particles do not interact and hence do not contribute to the energy. But the degeneracies change dramatically for the excited levels. Table 6.1 compares the degeneracies for a system of 6 particles in a shell with 12 single-particle states (e.g. a $(h_{11/2})^6$ configuration) in the fully accompagnied and the complete space.

6.3 Decomposition scheme for SDQMC

In this section we present a decomposition scheme for the Boltzmann operator $e^{-\beta\hat{H}}$, with $\hat{H}$ of the form 6.6. As mentioned before, this Hamiltonian does not couple protons with

neutrons. Therefore, a first and important step for the application of SDQMC to this system, is to split it in two systems, a proton and a neutron system, such that separate SDQMC calculations can be performed for each of them. A next step is to split up the inverse temperature β in N_t inverse temperature intervals (see section 2.1). In each inverse temperature interval, the operator $e^{-\frac{\beta}{N_t}\hat{H}}$ is split up further as follows:

$$\begin{aligned} e^{-\frac{\beta}{N_t}\hat{H}} &= e^{-\frac{\beta}{4N_t}\hat{H}_1} e^{\frac{\beta G}{2N_t}\hat{P}_1} e^{\frac{\beta G}{2N_t}\hat{P}_2} \cdots e^{\frac{\beta G}{2N_t}\hat{P}_{\Omega-1}} e^{\frac{\beta G}{2N_t}\hat{P}_{\Omega}} e^{-\frac{\beta}{2N_t}\hat{H}_1} \\ &\quad \times e^{\frac{\beta G}{2N_t}\hat{P}_{\Omega}} e^{\frac{\beta G}{2N_t}\hat{P}_{\Omega-1}} \cdots e^{\frac{\beta G}{2N_t}\hat{P}_2} e^{\frac{\beta G}{2N_t}\hat{P}_1} e^{-\frac{\beta}{4N_t}\hat{H}_1}, \end{aligned} \tag{6.25}$$

with

$$\hat{H}_1 = \sum_{k>0} e_k \left(\hat{a}_k^\dagger \hat{a}_k + \hat{a}_{\bar{k}}^\dagger \hat{a}_{\bar{k}} \right), \tag{6.26}$$

$$\begin{aligned} \hat{P}_k &= \left(\sum_{k'>0} \hat{a}_{k'}^\dagger \hat{a}_{\bar{k}'}^\dagger \right) \hat{a}_{\bar{k}} \hat{a}_k \\ &= \hat{S}_+ \hat{s}_-^{(k)}, \end{aligned} \tag{6.27}$$

where we used the notation of the quasispin operators introduced in the previous section. The ascending-descending ordering of the factors $e^{\frac{\beta G}{2N_t}\hat{P}_k}$ in expression 6.25 is necessary to reduce the error originating from the non-commutativity of the exponents to order $\left(\frac{\beta}{N_t}\right)^3$. The operators $\hat{P}_k$ have the property that $\hat{P}_k^2 = \hat{P}_k$. Therefore, their exponential can be written as

$$e^{\frac{\beta G}{2N_t}\hat{P}_k} = 1 + \left(e^{\frac{\beta G}{2N_t}} - 1 \right) \hat{P}_k. \tag{6.28}$$

This form allows an exact decomposition using rank-two operators of the type discussed in section 2.2.2.

$$\begin{aligned} e^{\frac{\beta G}{2N_t}\hat{P}_k} &= \frac{1}{2\Omega} \sum_{k'>0} \Big[\left(1 + \gamma \hat{a}_{k'}^\dagger \hat{a}_k\right) \left(1 + \gamma \hat{a}_{\bar{k}'}^\dagger \hat{a}_{\bar{k}}\right) \\ &\quad + \left(1 - \gamma \hat{a}_{k'}^\dagger \hat{a}_k\right) \left(1 - \gamma \hat{a}_{\bar{k}'}^\dagger \hat{a}_{\bar{k}}\right) \Big], \end{aligned} \tag{6.29}$$

with γ given by

$$\gamma^2 = \Omega \left(e^{\frac{\beta G}{2N_t}} - 1 \right). \tag{6.30}$$

In the decomposition for the Boltzmann operator that is obtained in this way, the operators $\hat{U}_\sigma$ from expression 4.1 can be split up in two parts, $\hat{U}_\sigma = \hat{U}_{\sigma+}\hat{U}_{\sigma-}$, with a part $\hat{U}_{\sigma+}$ for the states $k > 0$ and a formally equal part $\hat{U}_{\sigma-}$ for the time-reversed states. The corresponding matrix U_σ has a structure

$$U_\sigma = \begin{pmatrix} U_{\sigma+} & 0 \\ 0 & U_{\sigma-} \end{pmatrix}, \tag{6.31}$$

where furthermore the submatrices are equal, $U_{\sigma-} = U_{\sigma+}$. As discussed in section 4.5.3, such matrices lead to good sign characteristics for even-even systems. Note that for the canonical trace of $\hat{U}_\sigma$ one has to sum over all fillings (N_+, N_-) for which $N_+ + N_- = N$:

$$\hat{\mathrm{Tr}}_N \left(\hat{U}_\sigma \right) = \sum_{N_+ + N_- = N} \hat{\mathrm{Tr}}_{N_+} \left(\hat{U}_{\sigma+} \right) \hat{\mathrm{Tr}}_{N_-} \left(\hat{U}_{\sigma-} \right). \tag{6.32}$$

Further advantages of this decomposition are that its error is only proportional to the commutators of the operators $\hat{H}_1, \hat{P}_1, \hat{P}_2, \ldots, \hat{P}_\Omega$, The leading error term is proportional to $\frac{\beta^3}{N_t^2}$. The matrices that represent the operators $1 + \gamma \hat{a}^\dagger_{k'} \hat{a}_k$ are extremely simple (the unity matrix with the element (k', k) set to γ). Therefore the matrix multiplications that are needed to build up U_σ are simple and fast. Multiplying a matrix with the matrix representation of one inverse-temperature slice requires 2 multiplications with a diagonal matrix and 2Ω multiplications with matrices of the from $1 + \gamma a^\dagger_{k'} a_k$ (a_k denotes the unit row vector representing $\hat{a}_k$). For one slice this requires $6\Omega^2$ flops. The decomposition can easily be extended for Hamiltonians of the form

$$\hat{H} = -\sum_{k,k'>0} G_{k\,k'} \hat{a}^\dagger_{k'} \hat{a}^\dagger_{\bar{k}'} \hat{a}_{\bar{k}} \hat{a}_k. \tag{6.33}$$

To give an idea of the computer time needed for these calculations, we calculated the neutron internal energy for a system with 30 neutrons distributed over 70 single-particle states (see section 6.4.2), at a temperature $T = 0.25MeV$ and an interaction strength $G = 16MeV/56$. This is a very large system at a low temperature. Most calculations were performed at higher temperatures and in much smaller model spaces, so they required much less computer time. For the decomposition of the Boltzmann operator, $N_t = 160$ inverse temperature slices were used. The QR-stabilization technique discussed in section 4.6.1 was applied every 20 slices. The updating scheme of section 4.6.2 was used with $N_{sc} = 195$, $N_{t2} = 40$, $N_c = 9$. 20 independent Markov chains were run of 9000 thermalization and 45000 sampling steps each. Observables were evaluated every 45 Markov steps. For the internal energy a value $U_n = -630.79MeV$ was obtained, with a statistical error of $0.10MeV$ (at 95%-confidence level). On a PC with a 200-MHz pentium-pro processor, running a Linux operating system, this calculation took 19 hours.

6.4 Thermodynamical properties of the nuclear pairing model

6.4.1 Pairing in a degenerate shell

Because a system with pairing in a degenerate shell can be solved analytically, it was an ideal test case for the SDQMC method. Furthermore, recently, N. J. Cerf [62] presented a quantum Monte-Carlo method for the study of thermodynamic properties of nuclear many-body systems using a monopole pairing interaction in the canonical ensemble. He presented finite temperature results for a model with a $(h_{11/2})^6$ configuration. This seemed to us an ideal point of comparision for our finite temperature calculations in the canonical ensemble. To our surprise, we found quite different results.
The space of configurations taken into account by Cerf is too limited to calculate properties at finite temperature. The peak in the specific heat versus temperature curve around 1 MeV that is mentioned in [62] is, in our opinion, an artefact of a too small many-body space. With a Monte-Carlo calculation in the complete N-particle many-body space, we observe a much sharper peak in the specific heat around 1.25 MeV, originating from nucleon pair

breakup.
The method of Cerf is based on a path-integral over chains of configurations $C(t_1)$, $C(t_2)$, ..., $C(t_D)$ that are periodic in time. The $C(t_i)$ are states in the Fock space of many-body states. They are sampled with a Metropolis random walk. Though no details are given on how the many-body states are sampled, it appears to us that Cerf only considers states where all particles are accompagnied, i.e. if a state $|jlm\rangle$ is occupied, then the state $|jl-m\rangle$ is also occupied [60]. This set of configurations is sufficient for the study of the ground state of the system [63], where all nucleons are accompagnied. At finite temperature, however, the pairs of accompagnied particles can be broken up by thermal fluctuations. Then also configurations with unaccompagnied nucleons have to be taken into account.
In reference [62], Monte-Carlo results starting from a model with a $(h_{11/2})^6$ configuration and with a constant pairing strength of $G = 1MeV$, were compared to exact results obtained in the quasispin formalism. The energy levels and degeneracies for this model are listed in table 6.1. We calculated the thermodynamical properties of this system using the SDQMC in the canonical ensemble (see section 4.3), based on the decomposition presented in the previous section. Figure 6.1 shows the internal energy of the system as a function of temperature. Our Monte-Carlo results are in excellent agreement with the exact results that are based on the degeneracies given in the third column of table 6.1. They differ clearly from the results for the fully accompagnied states only, with which the Monte-Carlo results of Cerf coincide. It is observed that the internal energy starts to rise at temperatures near to 1 MeV. This leads to a distinct peak in the specific heat around 1.25 MeV, as is shown in figure 6.2. The curve for the fully accompagnied states shows a lower and broader peak at temperatures around 2 MeV. This peak was associated by Cerf with the vanishing of nucleon pair correlations. However, the breakup of the pairs starts already at lower temperatures. This results in the stronger peak that we observe around 1.25 MeV.
Though the Monte-Carlo method presented in [62] offers an interesting way to study pair correlations in nuclei, we emphasize that a full treatment of the complete many-body space is required to study properties at finite temperature.

6.4.2 Thermodynamical properties of a model with pairing for Fe nuclei.

Thermodynamical properties of nuclei in the Fe region were studied in a model with a Hamiltonian of the form 6.6. For the mean-field potential, a Woods-Saxon potential $U(r)$ is used, given by [64]

$$U(r) = V_c - Vf(x) + \left(\frac{\hbar}{m_\pi c}\right)^2 V_{so}\,(\sigma\cdot\mathbf{l})\,\frac{1}{r}\frac{d}{dr}f(x_{so}), \tag{6.34}$$

where

$$\begin{aligned} V_c &= Ze^2/r, \quad r \geq R_c, \\ &= \left[Ze^2/(2R_c)\right](3 - r^2/R_c^2), \quad r \leq R_c, \end{aligned} \tag{6.35}$$

$$R_c = r_c A^{1/3},$$

Figure 6.1: *Internal energy U versus temperature T. Error bars on our Monte-Carlo data were omitted because they are smaller than the symbols marking the data points. Cerf's Monte-Carlo results are not shown. They coincide with the dashed curve.*

Figure 6.2: *Specific heat C versus temperature T. Error bars on the Monte-Carlo data represent 95%-confidence intervals. Cerf's Monte-Carlo results are not shown. They coincide with the dashed curve.*

$$f(x) = (1+e^{x})^{-1} \text{ with } x = \left(r - r_0 A^{1/3}\right)/a, \tag{6.36}$$

$$\left(\frac{\hbar}{m_\pi c}\right)^2 = 2.000\ fm^2. \tag{6.37}$$

Here, A is the number of nucleons, Z the number of protons. The other parameters are taken as in [64]

$$\begin{aligned} V &= 53.3 + 27(A-2Z)/A - 0.4Z/A^{1/3}\ MeV, \\ r_0 &= 1.25\ fm, \\ a &= 0.65\ fm, \\ V_{so} &= 7.5\ MeV, \\ r_{so} &= r_0, \\ a_{so} &= 0.47 fm. \end{aligned}$$

To calculate the mean field and its eigenfunctions, we use the parameters for the nucleus $^{56}_{26}\mathrm{Fe}_{30}$. This mean field is used for all nuclei in this particular mass region. For every set of angular momentum quantum numbers l and j, the Woods-Saxon potential is diagonalized in a basis of the lowest 60 harmonic-oscillator eigenfunctions with the appropriate symmetry. In this way, the single-particle eigenstates and their energies listed in table 6.2, are obtained. Also a number of unbound states (with energy > 0) are obtained. In fact, the Woods-Saxon potential exhibits a continuum of unbound eigenstates. Due to the expansion in a finite number of basis functions, discrete unbound energy levels are obtained. These can be seen as a discrete approximation to the continuum of unbound states.
The $1s_{\frac{1}{2}}$, $1p_{\frac{3}{2}}$, $1p_{\frac{1}{2}}$,$1d_{\frac{5}{2}}$, $1d_{\frac{3}{2}}$ and $2s_{\frac{1}{2}}$ orbitals are considered to be compeletely filled. They form an inert core for the many-body problem. The $1f_{\frac{7}{2}}$, $2p_{\frac{3}{2}}$,$2p_{\frac{1}{2}}$ and $1f_{\frac{5}{2}}$ orbitals constitute the valence shell.
For the strength of the pairing interaction we took $G = 20MeV/56$, in accordance with expression 6.5. The same strength was used for protons and neutrons, and for all nuclei in the Fe mass region.
The lines that connect the data points on the figures in this section are ment to guide the eye. They do not correspond to analytical results or fitted curves. Error limits represent 95% confidence intervals. If no error limits are shown, this means that they are smaller than the markers of the data points, unless it is stated that no error limits were determined.

Proton and neutron contributions

Some thermodynamical properties of the pairing model for $^{56}_{26}\mathrm{Fe}_{30}$ were studied using SDQMC. Because the proton and neutron systems are not coupled to one another, separate results for both particle types are obtained. The internal energy of the total system and the contributions of the proton and neutron subsystems are shown as a function of temperature in figure 6.3. The same is done for the specific heat in figure 6.4 The neutrons contribute more to the internal energy than the protons, because there are more valence neutrons than valence protons. This also leads to a slightly stronger peak in the specific-heat

orbital	single-particle energies (MeV)	
	protons	neutrons
$1s_{\frac{1}{2}}$	-34.7106	-42.0333
$1p_{\frac{3}{2}}$	-25.3351	-32.2120
$1p_{\frac{1}{2}}$	-24.0715	-31.1979
$1d_{\frac{5}{2}}$	-15.0034	-21.5607
$1d_{\frac{3}{2}}$	-12.7911	-19.6359
$2s_{\frac{1}{2}}$	-12.3511	-19.1840
$1f_{\frac{7}{2}}$	-4.1205	-10.4576
$2p_{\frac{3}{2}}$	-2.0360	-8.4804
$2p_{\frac{1}{2}}$	-1.2334	-7.6512
$1f_{\frac{5}{2}}$	-1.2159	-7.7025
$3s_{\frac{1}{2}}$	4.7316	-0.3861
$2d_{\frac{5}{2}}$	5.6562	0.2225
$2d_{\frac{3}{2}}$	6.1324	0.9907
$1g_{\frac{9}{2}}$	6.6572	0.5631
$3p_{\frac{3}{2}}$	6.6663	2.5931
$3p_{\frac{1}{2}}$	6.7469	2.6915
$4s_{\frac{1}{2}}$	8.9016	4.4706
$1g_{\frac{7}{2}}$	9.1386	3.5488

Table 6.2: *Single-particles eigenstates of the Woods-Saxon potential.*

Figure 6.3: *Internal energy U as a function of temperature T for a system with 6 protons and 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, $G = 20MeV/56$. The energy scale is adapted such that the total, proton and neutron internal energies all tend to 0 at low temperature.*

Figure 6.4: *Specific heat C as a function of temperature T for a system with 6 protons and 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, $G = 20MeV/56$.*

curve for neutrons than for protons. Qualitatively, there is no big difference between the thermodynamical properties of both subsystems. This is not the case at lower values of the interaction strength G.

Dependence on the pairing interaction strength G

We have studied the pairing model for $^{56}_{26}\mathrm{Fe}_{30}$ for several values of the pairing interaction strength. SDQMC calculations were performed for 10 neutrons in a shell with 20 valence states ($1f_{\frac{7}{2}}$, $2p_{\frac{3}{2}}$,$2p_{\frac{1}{2}}$ and $1f_{\frac{5}{2}}$ orbitals) and for 6 protons in the same shell.
The neutron energy as a function of temperature is shown in figure 6.5. The energy scale was chosen such that the inert core had zero energy. The fact that the energy does not go to much higher values as the temperature increases, is due to the limited size of the model space: not enough high-lying states are included. As we shall discuss later on, the results for $T \geq 1.5MeV$ are not physical anymore. For larger values of G, the system is more strongly bound. Furthermore, when raising the temperature, the system stays in its ground state longer than for smaller values of G. This indicates that there is an energy gap between the ground state and the first excited state proportional to G, as is expected from BCS theory. The neutron specific heat as a function of temperature is shown in figure 6.6. To give a clear picture for the values at low temperature (high β), the neutron specific heat is also shown as a function of β, in figure 6.7. The dotted line indicates the results for $G = 0$, thus for a pure mean field. With increasing strength G, the peak in the specific heat curve shifts to a slightly higher temperature and becomes more pronounced. In general, peaks in the specific heat can be interpreted as signs of a phase transition. We see here that the pairing correlations, for $G \geq 20MeV/56$, seem to induce a phase transition in the system.
Analogous calculations were done for protons. The proton energy as a function of temperature is shown in figure 6.8. The proton specific heat is shown as a function of temperature in figure 6.9, and also as a function of β, in figure 6.10. The same discussion as for the neutron results, applies here. There is, however, a striking difference in the specific-heat curve for low values of G: a second peak develops around $\beta = 5MeV^{-1}$ for $G = 10MeV/56$. At this value of the pairing strength, the first peak in the specific-heat curve, around $\beta = 1.5MeV^{-1}$, coincides with the peak in the the specific-heat curve for a pure mean field. This means that this peak is related to the condensation of the valence particles in the lowest energy levels of the valence shell (the $1f_{\frac{7}{2}}$ orbital). The second peak is entirely due to pair correlations, that develop among the 6 particles in the $1f_{\frac{7}{2}}$ orbital. In figure 6.11, the number of accompagnied pairs and the expectation value of the pairing operator $\hat{S}_+\hat{S}_-$ are shown as a function of β. While the system with $G = 20MeV/56$ becomes completely accompagnied and reaches full pairing strength at values of $\beta \geq 3MeV^{-1}$, the system with $G = 10MeV/56$ comes to this regime only at values of $\beta \geq 6MeV^{-1}$. In figure 6.12, the number particles in the $1f_{\frac{7}{2}}$ orbital and the number of particles in the other orbitals are shown. For the system with $G = 10MeV/56$, it is observed that approximately all 6 particles occupy states in the $1f_{\frac{7}{2}}$ orbital for values of $\beta \geq 3MeV^{-1}$. The fact that the pairing correlations reach their maximum for this system only at values of $\beta \geq 6MeV^{-1}$, means that the system passes through two phases

Figure 6.5: *Neutron energy U_n as a function of temperature T for a system with 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, for various values of the pairing strength G*

Figure 6.6: *Neutron specific heat C_n as a function of temperature T for a system with 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, for various values of the pairing strength G*

Figure 6.7: *Neutron specific heat C_n as a function of inverse temperature temperature β for a system with 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, for various values of the pairing strength G*

Figure 6.8: *Proton energy U_p as a function of temperature T for a system with 6 protons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, for various values of the pairing strength G*

Figure 6.9: *Proton specific heat C_p as a function of temperature T for a system with 6 protons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, for various values of the pairing strength G*

Figure 6.10: *Proton specific heat C_p as a function of inverse temperature temperature β for a system with 6 protons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction, for various values of the pairing strength G*

Figure 6.11: *The expectation value of the pairing operator $\hat{S}_+\hat{S}_-$ (dotted lines) and the number of accompagnied pairs (full lines) as a function of the inverse temperature β for systems with pairing strength $G = 10MeV/56$ and $G = 20MeV/56$ and 6 protons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction. No error limits were determined.*

as it is cooled: first, the 6 valence protons condense into the $1f_{\frac{7}{2}}$ orbital. At $\beta \simeq 3MeV^{-1}$, this stage is completed. If the temperature is lowered further, pair correlations among these particles can develop. At values of $\beta \geq 6MeV^{-1}$ the system is almost completely cooled to its ground state. For the system with $G = 20MeV/56$, the occupation of the $1f_{\frac{7}{2}}$ orbital reaches a maximum of about 5.3. The particles always remain spread over all the valence shell orbitals, because the pairing interaction is now strong enough to scatter them out of the $1f_{\frac{7}{2}}$ orbital, even in the ground state.

Dependence on the size of the model space

For the description of the high-temperature properties of the system, the model space given by the fp shell is too small. At temperatures of a few MeV, valence particles can be excited to higher-lying single-particle states, or core particles can be excited into the valence orbitals or higher energy states. In order to know up to what temperatures the results that we obtained in the fp shell are valid, we performed a number of calculations in larger model spaces. First, the $3s_{\frac{1}{2}}$, $2d_{\frac{5}{2}}$, $2d_{\frac{3}{2}}$ and $1g_{\frac{9}{2}}$ orbitals are added to the single-particle space. This leads to a many-body problem of 6 and 10 particles in 42 single-particle states. In a second extended model the core states are considered as valence states too. Therefore the $1s_{\frac{1}{2}}$, $1p_{\frac{3}{2}}$, $1p_{\frac{1}{2}}$, $1d_{\frac{5}{2}}$, $1d_{\frac{3}{2}}$ and $2s_{\frac{1}{2}}$ orbitals are added. Furthermore, also the $3p_{\frac{3}{2}}$, $3p_{\frac{1}{2}}$ and $4s_{\frac{1}{2}}$ orbitals are taken into account. This leads to a many-body problem of 26 and 30 particles in 70 single-particle states.

Figure 6.12: *The number particles in the $1f_{\frac{7}{2}}$ orbital (full line) and the number of particles in the other orbitals of the valence shell (dotted line) as a function of the inverse temperature β for systems with pairing strength $G = 10MeV/56$ and $G = 20MeV/56$ and 6 protons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction. No error limits were determined.*

Because multiple shells are used, the model space of the extended systems contains spurious excitations related to center-of-mass motion. Therefore, care has to be taken when relating high temperature results to internal excitations of the system. For the second extended model space (without core), these center-of-mass motions can be interpreted as thermal excitations of the collective degrees of freedom. This picture would be physically meaningful in the absence of a mean-field potential The fact that the mean-field potential is localized in space, breaks the translational invariance of the model. Therefore, one cannot separate the center-of-mass motion from the intrinsic excitations in a clean way [65]. A consistent treatment of spurious states is a topic for further research.
The results for the internal energy and the specific heat obtained using these model spaces are shown in figure 6.13 to 6.16. If the value for the pairing interaction strength G is not changed, then a system with a larger model space will have a lower ground-state energy because the larger model space allows stronger pair correlations. In order to obtain a comparable pairing energy, a reduced pairing interaction strength of $G = 16MeV/56$ is used for the extended shells. For the no-core system, the energy is shifted such that the ground-state energy coincides with the ground-state energy of the fp shell system.
In the second extended model space, at high temperatures ($T \geq 2MeV$), the specific-heat curve coincides with the specific-heat curve for $G = 0$. In this temperature region, the proton and neutron internal energy are some $5MeV$ lower than in the $G = 0$ case. Apart from this shift, the internal-energy curves are similar to the $G = 0$ case. This indicates that, at high temperatures, the pairing Hamiltonian enhances the binding energy but has

Figure 6.13: *Neutron energy U_n as a function of temperature T for a system with 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell (a), for a system with 10 neutrons in the first extended model space (b) and for a system with 30 neutrons in the second extended model space (c). The dashed line gives the result for the second extended model space without pairing ($G = 0$).*

Figure 6.14: *Neutron specific heat C_n as a function of temperature T for a system with 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell (a), for a system with 10 neutrons in the first extended model space (b) and for a system with 30 neutrons in the second extended model space (c). The dashed line gives the result for the second extended model without pairing ($G = 0$).*

Figure 6.15: *Proton energy U_p as a function of temperature T for a system with 6 protons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell (a), for a system with 6 protons in the first extended model space (b) and for a system with 26 protons in the second extended model space (c). The dashed line gives the result for the second extended model without pairing ($G = 0$).*

Figure 6.16: *Proton specific heat C_p as a function of temperature T for a system with 6 protons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell (a), for a system with 6 protons in the first extended model space (b) and for a system with 26 protons in the second extended model space (c). The dashed line gives the result for the second extended model without pairing ($G = 0$).*

no effect on the internal structure.
Because the internal energy is related to the derivative of the logarithm of the partition function Z_β (expression 4.10), and because Z_β is the Laplace transform of the level density $g(E)$ of excited states (expression 4.14), the constant shift in the internal energy at high temperature can be related to a shift in $g(E)$. Let U_1 and $g_1(E)$ denote the internal energy and the level density for the system with pairing, and U_0 and $g_0(E)$ the internal energy and the level density for the system without pairing. At high temperature (low β), we have $U_1 \simeq U_0 - 5MeV$. By integrating U_0 and U_1 form $\beta = 0$ to $\beta = 1/T$, the logarithms of the partition functions are obtained ($\ln(Z_{\beta 0})$ and $\ln(Z_{\beta 1})$ respectively).

$$\begin{aligned} \ln(Z_{\beta 1}) &= \ln(Z_{\beta 0}) + 5MeV\,\beta \qquad (6.38) \\ &\Downarrow \\ Z_{\beta 1} &= e^{5MeV\beta} Z_{\beta 0} \\ &\Downarrow \\ g_1(E) &= g_0(E + 5MeV) \qquad (6.39) \end{aligned}$$

The main effect of the pairing Hamiltonian on the level density of the excited states is a shift of the curve $g_0(E)$ to lower energies. At lower temperatures, the specific heat curve deviates from the curve for $G = 0$, because pairing correlations develop. This will have an effect on the tail of the partition function $Z_{\beta 1}$ for high β and on the level density $g_1(E)$ at low energies. By comparing the results for the fp shell and the first extended model space, we see that the fp shell is too small to describe the system at temperatures $T \geq 1.3MeV$. In order to compare with the results for the second extended model space around temperatures of $1MeV$, the pairing interaction strength G ought to be reduced somewhat more for the latter model space. The vanishing of pair correlations with increasing temperature, starting from $T \simeq 1MeV$, was also observed in shell-model quantum Monte-Carlo calculations for ${}^{54}_{26}\mathrm{Fe}_{28}$ based on more realistic interactions [66, 67]. The interesting topic of proton-neutron pairing in $N = Z$ nuclei [68], could of course not be adressed with the schematic mean-field plus pairing Hamiltonian 6.6.

Dependence on the number of particles

We studied systems with various numbers of neutrons in the fp shell: ${}^{54}_{26}\mathrm{Fe}_{28}$, ${}^{55}_{26}\mathrm{Fe}_{29}$, ${}^{56}_{26}\mathrm{Fe}_{30}$ and ${}^{57}_{26}\mathrm{Fe}_{31}$ were modelled by considering 8, 9, 10 and 11 neutrons in the fp valence shell, respectively. For the systems with 9 and 11 neutrons, the sign rule discussed in section 4.5.3, that guaranteed good sign characteristics for even-even systems, does not apply. The average sign $\bar{s}$ for the various systems is shown as a function of the inverse temperature β in figure 6.17. For the odd systems, accurate calculations are possible up to values of $\beta \simeq 4MeV^{-1}$. This corresponds to a temperature of $T \simeq 0.25MeV$. This temperature is low enough to get a good approximation of the ground state. The neutron internal energy U_n for the various systems is shown as a function of temperature in figure 6.18. The proton internal energy is not shown because it is equal for all four systems and it is already given in figure 6.8. While at high temperature the internal energy curves are equidistantly spaced, with an interval of about $9MeV$, there is a shift to lower energies for the systems with 8 and 10 neutrons at low energy. This is because the pairing correlations

Figure 6.17: *Average sign $\bar{s}$ as a function of inverse temperature β for systems with 8,9,10 and 11 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction ($G = 20MeV/56$).*

Figure 6.18: *Neutron internal energy U_n as a function of temperature T for systems with 8,9,10 and 11 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction ($G = 20MeV/56$).*

Figure 6.19: *Neutron internal-energy shift ΔU_{10} as a function of temperature T for a systems with 10 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction ($G = 20MeV/56$). The dashed line indicates the experimental value of the ground-state energy shift for ${}^{56}_{26}Fe_{30}$.*

are stronger for the even systems than for the odd systems at temperatures below $1MeV$. The shift in the internal energy for the system with 10 neutrons can be quantified as

$$\Delta U_{10} = \frac{U_{n9} + U_{n11}}{2} - U_{n10}, \tag{6.40}$$

With U_{n9}, U_{n10}, U_{n11} the neutron internal energies for the systems with 9, 10 and 11 valence neutrons respectively. The quantity ΔU_{10} is shown as a function of temperature in figure 6.19. The ground-state energy shift was calculated analogously to expression 6.40, with the internal energies replaced by the mass excesses given in reference [69]. A value of $1.776MeV$ was obtained. The SDQMC results approach this value remarkably well at temperatures below $0.5MeV$. The qualitative difference between the odd and the even systems also shows up in the specific-heat curve presented in figure 6.20.

We conclude this chapter by stating that SDQMC offers a powerful tool for the study of the nuclear pairing model. We have put emphasis on the thermodynamical properties. Occupation numbers and the pairing gap can be calculated too using SDQMC. Main advantages of SDQMC over other methods are that many-body correlations are taken into account exactly, particle numbers are constant and finite temperature results can be obtained. The major disadvantage of the method is that spectroscopic information can only be obtained indirectly. Finally, we remark that our calculations indicate that pairing correlations are important only at low temperature (below $1MeV$) and at low excitation energies.

Figure 6.20: *Neutron specific heat C_n as a function of temperature T for systems with 8,9,10 and 11 neutrons in the $1f_{\frac{7}{2}}2p_{\frac{3}{2}}2p_{\frac{1}{2}}1f_{\frac{5}{2}}$ shell, with a mean field and a pairing interaction ($G = 20MeV/56$).*

7

Neutrino scattering

In this chapter arguments are given that the nuclear temperature might be an important parameter for the calculation of neutrino-nucleus scattering cross sections relevant for supernova processes. We describe how this influence can be studied using SDQMC.

7.1 Neutrino-nucleus scattering cross-sections and the nuclear temperature

Of late years, neutrino-nucleus scattering has caught a lot of attention in connection to astrophysical topics like the supernova explosion mechanism and supernova nucleosynthesis. Especially for the latter topic, neutrino-nucleus cross-sections are necessary ingredients for understanding the mechanisms at work. Though most of the nuclei are synthesised via the long known s- and r-processes (neutron capture and beta decay), the origin of some nuclei cannot be explained in this way. Neutrino induced reactions might play a central role in the synthesis mechanism of nuclei like ^{11}Be, ^{19}F, ^{180}Ta [71, 72]. Also in the supernova explosion mechanism neutrinos play an important role. The most abundant elements in the outer shells of a supernova are ^{4}He, ^{12}C, ^{16}O, ^{20}Ne and ^{28}Si. Here it is important to know how much energy the neutrinos can transfer from the core to the envelope of the supernova. A third way in which neutrino scattering on atomic nuclei might play a role in supernovae, is in the r-process nucleosynthesis that might possibly take place during supernova explosions. Here, charged-current electron-neutrino captures can compete with the normal β-decays during the r-proces [70]. This might shorten the time scale for the r-proces.
Up till now, calculations of these cross-sections only considered excitations out of the nuclear ground state. However, due to the high temperatures in supernova processes (10^9K or higher), part of the nuclei may be in an excited state before interacting with a neutrino. Although even at such high temperatures only a small fraction of the nuclei will be in the first excited state, this might have an important influence on the total cross-section

due to two effects that can enhance the cross-section for scattering out of excited states compared to scattering out of the ground state.
First, there is an *energy effect*: supernova neutrinos have a thermal energy distribution, which can be described by a Fermi-Dirac distribution with a temperature of 5 to 10 MeV [73]. Neutrinos from the tail of this distribution will dominate the cross-section, which is proportional to E_ν^2. If the scattering takes place on a nucleus in the first excited state, the neutrino does not need to have such a high energy in order to bring the nucleus into the same resonant state and to keep the same final energy as it would need to have for scattering the nucleus out of its ground state. Consequently, a larger part of the tail of the neutrino spectrum can contribute, and since this tail falls of exponentially this can lead to a considerable enhancement for the scattering out of the first excited state.
Secondly, there can be a *cross-section effect*: if one decomposes the cross-section formula into its various multipoles, one would expect the cross-section to be dominated by the lowest multipoles. However, due to selection rules, shell effects and Pauli blocking the monopole contribution can be strongly suppressed for excitation out of the ground state. Continuum-RPA calculations show that this is actually the case for nuclei like ^{12}C and ^{16}O [74], for which the 1^- and 2^- transitions give the largest contributions. We performed a self-consistent continuum-RPA calculation based an effective force of the Skyrme type (SkeII). A continuum-RPA code developed by Jan Ryckebusch [75] for the study of electon scattering on atomic nuclei, was adapted for the study of neutrino scattering. The semi-inclusive cross-section for the neutral-current scattering of a $50MeV$ neutrino on an ^16O nucleus in the ground-state is shown in figure 7.1 as a function of the excitation energy ω. Also the 1^-- and 2^--multipole contributions induced by the axial-vector current are shown. Together, these two contributions account almost for the total cross-section. In the above cases monopole transitions are probably not suppressed for excitations out of the first excited state, so that again their contribution to the total cross-section is enhanced. Note that the *energy effect* is restricted to supernova neutrinos, due to their thermal energy distribution, while the *cross-section effect* is more general. In order to know to what extent these simple arguments are valid in a supernova-environment, a more realistic study of the temperature dependence of neutrino-nucleus scattering cross-sections is necessary. This can be done using the shell-model Monte-Carlo method.

7.2 Calculation of neutrino-nucleus scattering cross-sections using SDQMC

The shell-model Monte-Carlo method (SMMC) [6] presents an interesting approach to study the nuclear many-body problem. It allows the calculation of exact results, up to controllable statistical and systematical errors, in much larger model spaces than the shell-model methods based on diagonalisation. Therefore it can be applied to a wide range of nuclei. It is based on a stochastic evalutation of the nuclear thermodynamical partition function

$$Z = \hat{\mathrm{Tr}}\left(e^{-\beta\hat{H}}\right), \tag{7.1}$$

Figure 7.1: *Neutral-current neutrino scattering cross-section on ^{16}O as a function of the excitation energy ω for an incoming-neutrino energy of $50MeV$, calculated using continuum RPA. Total cross-section (thick full line), 1^- axial-vector contribution (dashed line) and 2^- axial-vector contribution (thin full line).*

where the inverse temperature β is an input parameter. So it is a finite temperature method. The response function $R_{A,B}(\tau)$ for operators $\hat{A}$ and $\hat{B}$ can be calculated using the expression

$$R_{A,B}(\tau) = \frac{\hat{\text{Tr}}\left(e^{-(\beta-\tau)\hat{H}}\hat{A}e^{-\tau\hat{H}}\hat{B}\right)}{\hat{\text{Tr}}\left(e^{-\beta\hat{H}}\right)}. \tag{7.2}$$

Inserting complete sets of eigenstates of $\hat{H}$ ($\{|i\rangle, |f\rangle\}$ with energies $E_{i,f}$) shows that

$$R_{A,B}(\tau) = \frac{1}{Z}\sum_{i,f} e^{-\beta E_i} e^{-\tau\left(E_f - E_i\right)} \langle i|\hat{A}|f\rangle\langle f|\hat{B}|i\rangle. \tag{7.3}$$

The neutrino-nucleus scattering cross-section is then given by the expression

$$\begin{aligned}\frac{d\sigma^{[2]}}{d\Omega\,[d\omega]} &= \frac{G^2}{(2\pi)^2}\,\epsilon_f^2 \\ &\times \left|\langle f|\int e^{-i\vec{q}\cdot\vec{r}}\left[\sqrt{2}\cos(\theta/2)\,\hat{J}^0(\vec{r}) - \frac{(\vec{e}_i + \vec{e}_f - i\vec{e}_i\times\vec{e}_f)}{\sqrt{2}\cos(\theta/2)}\cdot\hat{\vec{J}}(\vec{r})\right]d\vec{r}|i\rangle\right|^2\end{aligned} \tag{7.4}$$

where θ is the neutrino scattering angle, ϵ_f is the outgoing neutrino energy, $\vec{q}$ is the momentum transfer, $\vec{e}_i$ and $\vec{e}_f$ are the unit vectors along the incoming and outgoing neutrino direction respectively and $\hat{J}^0(\vec{r})$ and $\hat{\vec{J}}(\vec{r})$ are the components of the hadronic

weak neutral current. Using equation 7.2, with

$$\hat{A}^{\dagger} = \hat{B} = \int e^{-i\vec{q}\cdot\vec{r}} \left[\sqrt{2}\cos\left(\theta/2\right) \hat{J}^{0}\left(\vec{r}\right) - \frac{\left(\vec{e}_i + \vec{e}_f - i\vec{e}_i \times \vec{e}_f\right)}{\sqrt{2}\cos\left(\theta/2\right)} \cdot \hat{\vec{J}}\left(\vec{r}\right) \right] d\vec{r}, \tag{7.5}$$

leads to wrong results because $\vec{q}$ depends on the energy transfer $\omega = E_f - E_i$, which also shows up in the factor $e^{-\tau(E_f - E_i)}$ in equation 7.3. In order to obtain an expression based upon the response functions for operators $\hat{A}$ and $\hat{B}$ that are independent of ω, one can perform a series expansion of $e^{-i\vec{q}\cdot\vec{r}}$ in orders of ω. This leads to the result

$$e^{-i\vec{q}\cdot\vec{r}} = e^{-i\epsilon_i 2\sin(\theta/2)z} \sum_j \left(\cos(\theta/2)x - \sin(\theta/2)z\right)^j \frac{\omega^j}{j!}, \tag{7.6}$$

where ϵ_i is the ingoing neutrino energy and where we have choosen the axes such that $\vec{e}_z$ is oriented along the direction of $\vec{e}_i - \vec{e}_f$ and $\vec{e}_y$ is along the direction of $\vec{e}_i \times \vec{e}_f$. Note that the expansion parameter is small so that only a few terms in the expansion have to be taken into account. This leads to an expression for the cross-section in a form that is suitable for calculation using the SMMC:

$$\frac{d\sigma^{[2]}}{d\Omega\,[d\omega]} = \frac{G^2}{(2\pi)^2} \left(E_i - \frac{d}{d\tau}\right)^2 \sum_{j,k} \frac{1}{j!k!} \left(\frac{d}{d\tau}\right)^{j+k} \frac{\hat{\mathrm{Tr}}\left(e^{-(\beta-\tau)\hat{H}} \hat{A}_j^{\dagger} e^{-\tau\hat{H}} \hat{A}_k\right)}{\hat{\mathrm{Tr}}\left(e^{-\beta\hat{H}}\right)}, \tag{7.7}$$

with

$$\hat{A}_j = \sqrt{2} \int e^{-i\epsilon_i 2\sin(\theta/2)z} i^j \left(\cos(\theta/2)x - \sin(\theta/2)z\right)^j \tag{7.8}$$

$$\left(\cos(\theta/2)\hat{J}^0\left(\vec{r}\right) - \hat{J}_x\left(\vec{r}\right) + i\sin(\theta/2)\hat{J}_y\left(\vec{r}\right)\right) d\vec{r}. \tag{7.9}$$

The derivatives have to be taken on each Monte-Carlo sample in order to obtain small enough errors on the final result. After taking the limit of $\tau \to 0$, the SMMC will yield the cross-section $\langle \frac{d\sigma}{d\Omega}(\theta, \epsilon_i) \rangle_\beta$ for given scattering angle θ and incoming neutrino energy ϵ_i, integrated over the energy transfer variable ω. By taking additional derivatives towards τ and β, also the average energy transfer $\langle \omega \frac{d\sigma}{d\Omega}(\theta, \epsilon_i) \rangle_\beta$ and the average excitation energy $\langle E_f \frac{d\sigma}{d\Omega}(\theta, \epsilon_i) \rangle_\beta$ can be obtained. By performing an inverse Laplace transform on τ, which is numerically somewhat more tricky, even the strength distribution can be calculated.

7.3 Conclusions and outlook

The nuclear temperature might be an important parameter in carrying out calculations of neutrino-nucleus scattering cross sections relevant for supernova processes. This can be taken into account within SDQMC. Detailed formulas for the calculation of neutrino-nucleus scattering cross-sections have been derived. Results for the reactions $^{12}\mathrm{C}\left(\nu,\nu'\right)\mathrm{C}^*$, $^{16}\mathrm{O}\left(\nu,\nu'\right)\mathrm{O}^*$, $^{20}\mathrm{Ne}\left(\nu,\nu'\right)\mathrm{Ne}^*$ and $^{56}\mathrm{Fe}\left(\nu,\nu'\right)\mathrm{Fe}^*$, will be studied.

A

Detailed SDQMC results for the Hubbard model

In this appendix detailed SDQMC results are listed for the internal energy of the 4×4 repulsive Hubbard model. Ground-state energies obtained by diagonalization (DIAG) and SDQMC with ground-state projection (PQMC), listed in table A.1, are taken from reference [7].

β	$4\uparrow 4\downarrow$	error	$5\uparrow 5\downarrow$	error	$6\uparrow 6\downarrow$	error	$7\uparrow 7\downarrow$	error	$8\uparrow 8\downarrow$	error
0.50	-8.87	0.04	-8.96	0.02	-8.13	0.01	-6.18	0.03	-2.90	0.03
1.00	-13.65	0.02	-14.18	0.01	-13.57	0.01	-11.86	0.02	-8.73	0.04
1.50	-15.65	0.01	-16.45	0.01	-15.62	0.01	-13.76	0.01	-10.79	0.03
2.00	-16.60	0.01	-17.85	0.01	-16.61	0.01	-14.58	0.01	-11.69	0.04
2.50	-17.04	0.01	-18.74	0.01	-17.13	0.02	-15.00	0.02	-12.29	0.05
3.00	-17.26	0.01	-19.21	0.01	-17.40	0.01	-15.23	0.02	-12.74	0.06
3.50	-17.35	0.01	-19.43	0.01	-17.54	0.01	-15.36	0.03	-12.93	0.04
4.00	-17.40	0.01	-19.53	0.01	-17.60	0.01	-15.46	0.03	-13.10	0.04
4.50			-19.57	0.01	-17.64	0.02	-15.53	0.03	-13.20	0.04
5.00	-17.44	0.01	-19.59	0.01	-17.67	0.01	-15.55	0.06	-13.32	0.04
8.00			-19.60	0.01	-17.73	0.26	-15.49	0.86	-13.55	0.05
10.00			-19.60	0.01	-17.61	0.44				
12.00			-19.60	0.01			-15.66	0.52		
16.00									-13.56	0.03
DIAG	17.53		-19.58		-17.73		-15.74		-13.62	
PQMC	17.3		-19.4				-15.7		-13.6	

Table A.1: *Internal energy E obtained using SDQMC with the canonical algorithm, at several values of the inverse temperature β for the 4×4 Hubbard model at $U = 4$ at various fillings $(N_\uparrow N_\downarrow)$. Error limits indicate 95%-confidence intervals (statistical errors only). Ground-state energies obtained by diagonalization (DIAG) and SDQMC with ground-state projection (PQMC) are given for comparison.*

β	$4\uparrow 5\downarrow$	error	$4\uparrow 6\downarrow$	error	$5\uparrow 6\downarrow$	error	$5\uparrow 7\downarrow$	error	$6\uparrow 7\downarrow$	error	$6\uparrow 8\downarrow$	error	$7\uparrow 8\downarrow$	error	$7\uparrow 9\downarrow$	error
0.50	-9.01	0.02	-8.86	0.02	-8.65	0.03	-8.06	0.03	-7.27	0.03	-6.12	0.04	-4.72	0.03	-2.83	0.04
1.00	-14.02	0.05	-13.98	0.05	-14.03	0.06	-13.47	0.08	-12.80	0.09	-11.70	0.09	-10.38	0.11	-8.50	0.11
1.50	-16.16	0.02	-16.04	0.02	-16.15	0.02	-15.44	0.02	-14.79	0.03	-13.64	0.03	-12.39	0.03	-10.63	0.04
2.00	-17.33	0.03	-17.06	0.03	-17.30	0.04	-16.40	0.04	-15.62	0.05	-14.53	0.06	-13.24	0.06	-11.58	0.07
2.50	-17.96	0.02	-17.54	0.02	-17.92	0.03	-16.96	0.03	-16.10	0.04	-14.90	0.04	-13.72	0.05	-12.25	0.06
3.00	-18.28	0.02	-17.80	0.02	-18.30	0.02	-17.27	0.02	-16.34	0.03	-15.19	0.03	-14.02	0.04	-12.54	0.05
3.50	-18.42	0.02	-17.95	0.02	-18.48	0.02	-17.41	0.03	-16.49	0.03	-15.30	0.03	-14.21	0.04	-12.82	0.04
4.00	-18.50	0.01	-18.01	0.02	-18.56	0.02	-17.53	0.02	-16.59	0.03	-15.39	0.04	-14.29	0.04	-13.00	0.04
4.50	-18.51	0.02	-18.05	0.02	-18.61	0.02	-17.59	0.03	-16.58	0.03	-15.50	0.04	-14.39	0.04	-13.06	0.04
5.00	-18.53	0.02	-18.07	0.02	-18.64	0.02	-17.61	0.04	-16.59	0.04	-15.49	0.05	-14.45	0.04	-13.23	0.04
6.00	-18.56	0.02	-18.11	0.03	-18.64	0.03	-17.66	0.10	-16.58	1.40	-15.52	0.20	-14.51	0.07	-13.33	0.04
7.00	-18.54	0.02	-18.12	0.03	-18.68	0.04	-17.64	0.21	-16.77	0.49	-15.52	0.87	-14.52	0.51	-13.36	0.03
8.00	-18.54	0.03	-18.16	0.27	-18.66	0.09	-17.69	0.59	-16.81	0.78	-15.58	0.53	-14.62	0.42	-13.42	0.03
10.00	-18.53	0.30	-18.12	0.25	-18.68	0.47	-17.69	0.71	-17.05	0.46	-15.79	0.52	-14.60	1.03	-13.41	0.03

Table A.2: *Internal energy E obtained using SDQMC with the canonical algorithm, at several values of the inverse temperature β for the 4×4 Hubbard model at $U = 4$ at various fillings $(N_\uparrow N_\downarrow)$. Error limits indicate 95%-confidence intervals (statistical errors only).*

β	$4\uparrow 4\downarrow$	error	$5\uparrow 5\downarrow$	error	$6\uparrow 6\downarrow$	error	$7\uparrow 7\downarrow$	error	$8\uparrow 8\downarrow$	error
0.50			-8.38	0.02	-7.30	0.03	-4.97	0.08	0.19	0.11
1.00	-12.71	0.16	-12.67	0.04	-11.31	0.06	-8.83	0.04	-4.92	0.05
1.50			-14.39	0.06	-12.78	0.06	-9.93	0.06	-5.82	0.06
2.00	-15.37	0.02	-15.51	0.04	-13.62	0.04	-10.48	0.04	-6.42	0.09
2.50	-15.86	0.03	-16.24	0.05	-14.02	0.05	-10.81	0.06	-6.90	0.06
3.00	-16.13	0.06	-16.85	0.05	-14.37	0.09	-11.05	0.09	-7.33	0.15
3.50	-16.24	0.05	-17.19	0.07	-14.46	0.14	-11.25	0.13	-7.65	0.14
4.00	-16.32	0.51	-17.41	0.07	-14.71	0.26	-11.24	0.35	-8.01	0.15
4.50	-16.36	0.08	-17.53	0.07	-14.45	0.70	-11.81	0.55	-8.03	0.14
5.00	-16.39	1.19	-17.48	0.08	-13.77	1.57	-11.93	1.39	-8.12	0.08
5.50									-8.44	0.13
6.00	-16.67	0.61	-17.61	0.12					-8.36	0.14
7.00	-16.46	1.25	-17.51	0.22					-8.45	0.11
8.00	-16.75	0.79	-17.55	0.68					-8.36	0.13
10.00									-8.50	0.10

Table A.3: *Internal energy E obtained using SDQMC with the canonical algorithm, at several values of the inverse temperature β for the 4×4 Hubbard model at $U = 8$ at various fillings $(N_\uparrow N_\downarrow)$. Error limits indicate 95%-confidence intervals (statistical errors only).*

B

Samenvatting

In dit proefschrift wordt een methode voorgesteld voor de studie van fermionische veeldeeltjessystemen. Het gaat om een statistische methode (een zogeheten 'Monte-Carlo' methode) die een exacte beschrijving geeft van het quantummechanische veeldeeltjessysteem, op een controleerbare statistische en systematische fout na. In een eerste deel worden de bouwstenen van de methode aangebracht en in detail uitgewerkt. Een tweede deel geeft een aantal toepassingen van de methode, in de eerste plaats ter illustratie van de mogelijkheden van de methode.

In hoofdstuk 1 wordt een matrixrepresentatie voor Slater-determinanten aangebracht. Een Slater-determinant-golffunctie voor een systeem met N identieke deeltjes verdeeld over $N_{\mathcal{S}}$ één-deeltjes-toestanden kan voorgesteld worden door een $N_{\mathcal{S}} \times N$ matrix. Gebruik makend van deze matrix-voorstelling, kan de exponentiële van eender welke één-deeltjes-operator voorgesteld worden door een $N_{\mathcal{S}} \times N_{\mathcal{S}}$ matrix. De overlap van twee Slater-determinanten, het inwerken van de exponentiële van een één-deeltjes-operator op een Slater-determinant en het canonisch en groot-canonisch spoor van zo een operator kunnen dan berekend worden met behulp van eenvoudige matrixbewerkingen.
Om deze technieken te kunnen toepassen op de Boltzmann-operator $e^{-\beta\hat{H}}$, die over het algemeen de exponentiële van een twee-deeltjes-operator is, wordt in hoofdstuk 2 uitgelegd hoe de Boltzmann-operator ontwikkeld kan worden in een som van exponentiëlen van één-deeltjes-operatoren. Met behulp van de Suzuki-Trotter-formule worden het één- en twee-deeltjes-stuk in de exponent gescheiden. Dan wordt de exponentiële van het twee-deeltjes stuk geschreven als een som van exponentiëlen van één-deeltjes-operatoren. Hiervoor kan men gebruik maken van de 'Hubbard-Stratonovich-transformatie', die de twee-deeltjes-interactie vervangt door een één-deeltjes-interactie met een aantal willekeurig fluctuerende 'auxiliaire velden'. Een alternatief voor de Hubbard-Stratonovich-transformatie wordt gegeven door decomposities gebaseerd op operatoren van rang één en twee. Gebruik makend van de matrix-representatie uit hoofdstuk 1, kunnen operatoren van rang één en twee voorgesteld worden door matrices van rang één en twee. Zulke operatoren kunnen

op verschillende manieren aangewend worden om de Boltzmann-operator te ontwikkelen in een som van hanteerbare termen.
Het aantal termen dat bekomen wordt in de decompositie van de Boltzmann operator is veel te groot om ze allemaal uit te rekenen en op te tellen. Om dit probleem te omzeilen wordt een steekproef genomen uit deze termen, aan de hand waarvan dan een statistiche schatting gemaakt wordt voor het gezochte resultaat. De statistiek leert ons dat het steekproefgemiddelde het exacte resultaat zal benaderen, als de steekproef maar groot genoeg is. Om op een efficiënte manier een representatieve steekproef te bekomen, wordt gebruik gemaakt van 'Markov-keten-Monte-Carlo-methodes'. In hoofdstuk 3 worden deze methodes voorgesteld, in een op zichzelf staande bespreking. De convergentie van deze methodes wordt aangetoond. Omdat vastgesteld werd dat in de praktijk vaak te lange sampling-intervallen en niet helemaal correcte foutenmarges gehanteerd worden, wordt in het bijzonder aandacht besteed aan het bepalen van het optimale sampling-interval en de foutenmarges voor de Monte-Carlo resultaten. Veelgebruikte Markov-keten-Monte-Carlo-methodes zoals het Metropolis algoritme, de Gibbs-sampler en de warmtebad-methode worden voorgesteld. Twee mogelijkheden om de efficiëntie van Markov-keten-Monte-Carlo-methodes te verhogen, namelijk variantie-reductie en 'geleide Markov-ketens', worden aangebracht.
De bouwstenen aangebracht in hoofdstukken 1 tot 3, worden samengebracht in hoofdstuk 4 tot de 'Slater-determinant-quantum-Monte-Carlo methode'. Met deze methode kunnen grondtoestandseigenschappen en thermodynamische eigenschappen van discrete, fermionische veeldeeltjessystemen bestudeerd worden in het canonisch en groot-canonisch ensemble. Bij de berekeningen voor het canonisch ensemble, moet de karakteristieke veelterm van een groot aantal matrices bepaald worden. Daarom werd een efficiënt en nauwkeurig algoritme ontwikkeld voor het berekenen van de coëfficiënten van de karakteristieke veelterm van een algemene vierkante matrix. Een algemeen probleem van quantum Monte-Carlo methodes voor fermionen vormt het zogenaamde 'tekenprobleem'. De oorzaak van dit probleem en mogelijke remedies worden besproken. Tenslotte wordt er in hoofdstuk 4 aandacht geschonken aan enkele praktische punten, zoals de stabilizatie van de methode bij lage temperaturen en de optimalizatie van de Markov-keten-Monte-Carlo-sampling met behulp van 'hybride' samplers.

Ter illustratie van de Slater-determinant-quantum-Monte-Carlo methode worden in een tweede deel een aantal toepassingen uitgewerkt.
Omdat het uitvoerig bestudeerd is met behulp van quantum-Monte-Carlo methodes, vormt het 'Hubbard-model' een ideale test-case voor de methode. In hoofdstuk 5 wordt een specifieke, efficiënte decompositie van de Boltzmann-operator voor dit model voorgesteld. Thermodynamische eigenschappen van het repulsieve 4×4-Hubbard-model werden berekend met de Slater-determinant-quantum-Monte-Carlo methode in het canonisch ensemble. Tot hier toe werden quantum-Monte-Carlo berekeningen voor het Hubbard-model steeds uitgevoerd in het groot-canonisch ensemble of met grondtoestandsprojectie. Wegens de beperkingen van deze methodes (o.a. het tekenprobleem), hebben de resultaten van deze berekeningen voornamelijk betrekking op geloten-schil-configuraties of half-gevulde modelruimtes. Onze berekeningen in het canonisch ensemble tonen aan dat open-schil-configuraties kwalitatief verschillende eigenschappen hebben. De gedetailleerde cijfers zijn weergegeven in appendix A.

In hoofdstuk 6 wordt de Slater-determinant-quantum-Monte-Carlo methode toegepast op een atoomkernmodel met een gemiddeld-veld potentiaal en een paarvormingskracht. Ook voor dit model wordt een specifieke, efficiënte decompositie van de Boltzmann-operator gegeven. Resultaten voor een model met één enkele, ontaarde schil worden vergeleken met de exacte resultaten bekomen in het quasispin-formalisme. Toepassing van het model op kernen in het $^{56}\mathrm{Fe}$-massagebied levert waarden op voor een aantal thermodynamische grootheden. Paarvormingscorrelaties spelen een belangrijke rol in de structuur van atoomkernen bij lage temperatuur. Bij hoge temperatuur ($T \geq 1MeV$), beperkt hun effect zich tot een extra bijdrage aan de bindingsenergie.
In hoofdstuk 7 tenslotte, wordt geargumenteerd dat een formalisme bij eindige temperatuur wenselijk is voor een accurate beschrijving van neutrino-interacties met atoomkernen in supernova's. Er wordt geschetst hoe de Slater-determinant-quantum-Monte-Carlo methode hierop toegepast kan worden.

BIBLIOGRAPHY

[1] Rombouts S and Heyde K, 1994 *J. Phys. A: Math. Gen.* 27 3293-3298

[2] Thouless D J, 1960 *Nucl. Phys.* **21** 225

[3] Knuth D E, *The art of computer programming.* **Vol. 2** 1969 (Reading, Massachusetts: Addison-Wesley Publishin Company)

[4] Suzuki M, *Quantum Monte-Carlo Methods.* 1986 Solid State Sciences, ed M. Suzuki, Springer Berlin **74** (KUL WNAT 530.1 (063)TANI 1987)

[5] Suzuki M, 1991 *Phys. Lett.* **A 146** 319

[6] Koonin S E, Dean D J and Langanke K, 1997 *Phys. Rep.* **278** 1-78

[7] von der Linden W, 1992 *Phys. Rep.* **220** 53

[8] Stratonovich R L, 1957 *Doklady Akad. Nauk S.S.S.R.* **115** 1097 (translation: 1958 *Soviet Phys. Doklady* **2** 416)

[9] Hubbard J, 1959 *Phys. Rev. Lett.* **3** 77

[10] Dean D, 1996 private communication.

[11] Golub G H and Van Loan C F, 1989 *Matrix Computations* (London: The Johns Hopkins University Press)

[12] Press W H, Teukolsky S A, Vetterling W T and Flannery B P, 1992 *Numerical Recipes in Fortran* (Cambridge University Press)

[13] Negele, J W and Orland H, 1988 *Quantum Many-Particle Systems* (Redwood City: Addison-Wesley Publishing Company)

[14] Hirsch J E, 1985 *Phys. Rev.* **B 31** 4403

[15] Tierney L, *Markov chains for exploring posterior distributions* 1994 *Ann. Stat.* **22** 1701-1762

[16] Baxter, R J, 1982 *Exactly solved models in statistical mechanics* (London: Academic Press)

[17] Metropolis, N, Rosenbluth, A W, Rosenbluth, M N, Teller, A H and Teller E, 1953 *J. Chemical Physics* **21** 1087-1091

[18] Hastings W K, *Monte-Carlo sampling methods using Markov chains an their applications.* 1970 *Biometrika* **57** 97-109

[19] Peskun P H, *Optimum Monte-Carlo sampling using Markov chains.* 1973 *Biometrika* **60** 607-612

[20] Roberts G O and Tweedie R L, *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms.* 1996 *Biometrika* **83** 95-110

[21] Gelman A, Roberts G O and Gilks W R *Efficient Metropolis Jumping Rules* 1995 in *Bayesian Statistics V,* ed. Bernardo J M et al. (Oxford University Press).

[22] Gelfland A E and Smith A F M, 1990 *J. Amer. Statist. Assoc.* **85** 398-409

[23] Roberts G O and Rosenthal J S, 1996 *Geometric Ergodicity and Hybrid Markov Chains* preprint available from http://www.stats.bris.ac.uk/MCMC/

[24] Roberts G O and Rosenthal J S, 1997 *Two Convergence Properties of Hybrid Samplers*

[25] Liu J S , 1996 *Statistics and Computing* **6** 113.

[26] Smith R L and Tierney L, 1996 *Exact Transition Probabilities for the Independence Metropolis Sampler*, preprint available from http://www.stats.bris.ac.uk/MCMC/

[27] Caracciolo S, Pelissetto A and Sokal A D, 1994 *Phys. Rev. Lett.* **72** 179.

[28] Brooks S P and Roberts G O, 1995, *Diagnosing Convergence of Markov-Chain Monte-Carlo Algorithms*, research report available from http://www.statslab.cam.ac.uk/Reports/1995/1995-12.html

[29] Liu J S , 1996 ,*Biometrika* **83** 681.

[30] Lifshitz E M and Pitaevskii L P, 1980 *Statistical Physics* **I** (Oxford: Pergamon Press)

[31] Finkelstein R J, 1969 *Thermodynamics and Statistical Physics* (San Fransisco: W. H. Freeman and Company)

[32] Deisz J J, von der Linden W, Preuss R and Hanke W, 1995 *Evaluation of dynamical spectra for T=0 quantum Monte-Carlo simulations: Hubbard lattices and continuous systems*in *Computer Simulations in Condensed Matter Physics VIII*, Eds. Landau D P, Mon K K and Schaettler H B (Heidelberg: Springer Verlag)

[33] von der Linden W, Preuss R and Hanke W, 1996, *J. Phys.: Condens. Matter* **8** 3881

[34] Löwdin P 1955 *Phys. Rev.* **97** 1471

[35] White S R, Scalapino D J, Sugar R L, Loh E Y, Gubernatis J E and Scalettar R T 1989 *Phys. Rev.* **B 40** 506

[36] Haake F, Kuś M, Sommers H-J, Schomerus H and Życzkowski K, 1996 *J.Phys. A: Math. Gen.* **29** 3641

[37] Lang G H, Johnson C W, Koonin S E and Ormand W E 1993 *Phys. Rev. C* **48** 1518

[38] Ormand W E, Dean D J, Johnson C W, Lang G H and Koonin S E 1994 *Phys. Rev. C* **49** 1422

[39] S. Rombouts and K. Heyde, submitted to J. Comp. Phys.

[40] http://www.netlib.org/netlib/lapack/

[41] *DEC Fortran User Manual for DEC OSF/1 AXP Systems* 1994 (Maynard: Digital Equipment Corporation)

[42] Fahy S and Hamann D R, 1991, *Phys. Rev. B* **43** 765

[43] Sorella S, Tosatti E, Baroni S, Car R and Parrinello M, 1988, *Int. J. Mod. Phys. B* **1** 993

[44] Sorella S, Baroni S, Car R and Parrinello M, 1989, *Europhys. Lett.* **8** 663

[45] Alhassid Y, Dean D J, Koonin S E, Lang G and Ormand W E, 1994 *Phys. Rev. Lett.* **72** 613

[46] Bohr A and Mottelson B R, 1969 *Nuclear Structure* (New York: W. A. Benjamin, Inc.)

[47] Yosida K, 1996 *Theory of magnetism, Springer series in solid-state sciences* **122**, (Berlin: Springer-Verlag)

[48] Auerbach A, 1994 *Interacting electrons and quantum magnetism*, (Berlin: Springer-Verlag)

[49] Anderson P W,1987 *Science* **235** 1196.

[50] López Sancho M P, Rubio J, Refolio M C and López Sancho J M, 1995 *J. Phys.: Condens. Matter* **7** L695

[51] Husslein T, Morgenstern I, Newns D M, Pattnaik P C, Singer J M and Matuttis H G, 1996 *Quantum Monte-Carlo evidence for d-wave pairing in the 2D Hubbard model at a van Hove singularity*, preprint available from *http://xxx.lanl.gov/abs/cond-mat/9608030*

[52] Duffy D and Moreo A, 1996 *Specific Heat of the 2D Hubbard Model*, preprint available from *http://xxx.lanl.gov/abs/cond-mat/9612132*

[53] P. Ring and P. Schuck, *The Nuclear Many-Body Problem* (Springer, New York, 1980)

[54] Woods R D and Saxon D S, 1954 *Phys. Rev.* **95** 577

[55] Waroquier M E L, 1981, *Een effectieve Skyrme-type interactie uniform voor kern-structuur berekeningen doorheen de ganse massatabel*, University Gent, hoger-aggregaats-thesis

[56] Heyde K L G, 1990 *The nuclear shell model* (Berlin: Springer-Verlag)

[57] Bes D R and Sorensen R A, 1969 *Adv. Nucl. Phys.* **2** 129

[58] Nilsson S G and Ragnarsson I, 1995 *Shapes and shells in nuclear structure* (Cambridge University Press)

[59] Molique H, 1996 *Etats exotiques à hauts spins et nouvelle méthode pour l'énergie d' appariement nucléaire*, University L. Pasteur, Strasbourg, Ph.D. thesis

[60] Burglin O and Rowley N, 1996 *Nucl. Phys. A* **602** 21; *erratum in* **609** 600

[61] Richardson R W and Sherman N,1964 *Nucl. Phys.* **52** 221

[62] Cerf N J,1996 *Phys. Rev. Lett.* **76**, 2420

[63] Cerf N J and O. Martin, 1993 *Phys. Rev. C* **47**, 2610

[64] Perey C M and Perey F G, 1972 *Nucl. Data Tables* **10** 540

[65] Lawson R D,1980 *Theory of the nuclear shell model* (Oxford: Clarendon Press)

[66] Dean D J, Koonin S E, Langanke K, Radha P B and Alhassid Y, 1995 *Phys. Rev. Lett.* **74** 2909

[67] Langanke K, Dean D J, Radha P B and Koonin S E, 1996 *Nucl. Phys. A* **602** 244

[68] Langanke K, Dean D J, Koonin S E and Radha P B, 1997 *Nucl. Phys. A* **613** 253

[69] Tuli J K, 1995 *Nuclear wallet cards* (The U.S. Nuclear Data Network, Brookhaven National Laboratory)

[70] Qian Y Z, Haxton W C, Langanke K and Vogel P, 1996 *Neutrino-induced neutron spallation and supernova r-process nucleosynthesis*, preprint available from http://xxx.lanl.gov/abs/nucl-th/9611010

[71] Woosley S E and Haxton W C, 1988 *Nature* **334** 45

[72] Haxton W C, 1993 *Nucl. Phys. A* **553** 397c

[73] Woosley S E, Hartmann D H, Hoffman R D and Haxton W C, 1990 *Astrophys. J.* **356** 272

[74] Kolbe E, Langanke K, Krewald S and Thielemann F, 1992 *Nucl. Phys. A* **540** 599

[75] Ryckebusch J, 1988, University Gent, Ph.D. thesis