6

Optimization of Variational Wave Functions

6.1 Introduction

In the first work that used the variational Monte Carlo approach to study a many-body system, a relatively simple wave function has been considered, containing only a couple of parameters; thus, it was possible to obtain their optimal values by a simple fitting procedure of few energy values (McMillan, 1965). The possibility to include several variational parameters in a Monte Carlo framework is a relatively recent achievement. Indeed, until the beginning of the nineties, people limited the optimization to very few (e.g., at most two or three) variational parameters, which were determined with trial and error calculations of energies or by fitting methods. Nowadays, there is a growing interest in attaining fast and accurate schemes to optimize a large number of parameters to describe correlated many-body wave functions.

Hereafter, we denote by the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ the set of all p variational parameters describing the wave function:

$$\Psi_\alpha(x) = \langle x|\Psi_\alpha\rangle, \tag{6.1}$$

where $\{|x\rangle\}$ denotes a complete basis set in the Hilbert space (which is taken to be orthogonal and normalized). The first breakthrough in the optimization has been done by realizing that, once a sampling for $\Psi_\alpha(x)$ has been obtained, the reweighting technique can be used to compute the energy or the variance for a wave function with different parameters, i.e., $\Psi_{\alpha+\delta\alpha}(x)$ (Umrigar et al., 1988). This method was shown to be very efficient and stable for a small number of particles (e.g., for few atoms and molecules), especially when the variance minimization was employed, while it has considerable drawbacks for large number of particles. Few years after the reweighting technique was applied, a further progress in the stochastic optimization techniques of several parameters was achieved. The basic idea is to compute energy differences for an arbitrary small variation of the parameters, namely

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

evaluating the energy derivatives. It turns out that these quantities correspond to well-defined estimators with variance that increases only polynomially (e.g., at most quadratically) with the number of particles. The simplest approach is based upon the steepest-descent minimization (Harju et al., 1997). More elaborated algorithms include the so-called *stochastic reconfiguration* (Sorella, 1998, 2001), approximated versions of the Newton-Raphson approach (Umrigar and Filippi, 2005; Sorella, 2005), and the so-called *linear method* (Nightingale and Melik-Alaverdian, 2001; Umrigar and Filippi, 2005; Umrigar et al., 2007; Toulouse and Umrigar, 2007). The key point in all these techniques is to devise a scheme that is stable even in the presence of the statistical noise and may converge quickly to the global minimum. These iterative schemes are based on energy derivatives, which must be calculated as accurately as possible within a given Monte Carlo simulation.

We emphasize that a robust and efficient method for the wave function optimization will also improve the convergence of projection techniques, which crucially depend on the quality of the guiding wave function that is used (see Chapters 8 and 9). In particular, the fixed-node approach benefits from an optimized wave function, since this approximation becomes exact by systematically improving the trial state (see Chapter 10).

6.2 Reweighting Techniques for the Optimization of Wave Functions

Here, we briefly discuss how the reweighting technique, described in section 3.2, can be used to get, with a single simulation, the energy or variance for several sets of variational parameters $\boldsymbol{\alpha}$. In particular, we consider the method based on the variance optimization, which has been widely used for several years after being introduced by Umrigar et al. (1988). Suppose that a set of configurations $\{x_i\}$ with $i = 1, \ldots, N$ have been sampled from the square of the variational state $\Psi_\alpha(x)$. Of course, the energy and variance corresponding to this "initial" guess with $\boldsymbol{\alpha}$ are given by:

$$E_\alpha = \frac{\langle \Psi_\alpha | \mathcal{H} | \Psi_\alpha \rangle}{\langle \Psi_\alpha | \Psi_\alpha \rangle} \approx \frac{1}{N} \sum_{i=1}^{N} e_{L,\alpha}(x_i), \tag{6.2}$$

$$\sigma^2_{\Psi_\alpha} = \frac{\langle \Psi_\alpha | (\mathcal{H} - E_\alpha)^2 | \Psi_\alpha \rangle}{\langle \Psi_\alpha | \Psi_\alpha \rangle} \approx \frac{1}{N} \sum_{i=1}^{N} \left[e_{L,\alpha}(x_i) - E_\alpha \right]^2, \tag{6.3}$$

where the local energy is given by, see Eq. (5.7):

$$e_{L,\alpha}(x_i) = \frac{\langle x_i | \mathcal{H} | \Psi_\alpha \rangle}{\langle x_i | \Psi_\alpha \rangle}. \tag{6.4}$$

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

Then, by using Eq. (3.10), with:

$$\mathcal{R}_{\alpha+\delta\alpha_k}(x) = \left| \frac{\Psi_{\alpha+\delta\alpha_k}(x)}{\Psi_\alpha(x)} \right|^2, \tag{6.5}$$

we can estimate (with the same set of configurations $\{x_i\}$) also the energy and the variance that correspond to a wave function with (slightly) different variational parameters $\boldsymbol{\alpha} + \boldsymbol{\delta\alpha}$:

$$E_{\alpha+\delta\alpha} = \frac{\sum_{i=1}^{N} e_{L,\alpha+\delta\alpha}(x_i)\mathcal{R}_{\alpha+\delta\alpha}(x_i)}{\sum_{i=1}^{N} \mathcal{R}_{\alpha+\delta\alpha}(x_i)}, \tag{6.6}$$

$$\sigma^2_{\Psi_{\alpha+\delta\alpha}} = \frac{\sum_{i=1}^{N} \left[e_{L,\alpha+\delta\alpha}(x_i) - E_{\alpha+\delta\alpha}\right]^2 \mathcal{R}_{\alpha+\delta\alpha}(x_i)}{\sum_{i=1}^{N} \mathcal{R}_{\alpha+\delta\alpha}(x_i)}. \tag{6.7}$$

Therefore, by performing a single simulation with variational parameters $\boldsymbol{\alpha}$, it would be possible to "reconstruct" the energy and variance also for different values in the neighborhood of $\boldsymbol{\alpha}$. In order to find the optimal set of parameters, we can minimize either the energy of Eq. (6.6) or the variance of Eq. (6.7) with respect to $\boldsymbol{\delta\alpha}$, by applying the steepest-descent approach (for the given set of configurations $\{x_i\}$, these are non-linear functions of the variational parameters). Within the reweighting method, the advantage of the variance minimization is that this is a sum of positive terms and, therefore, a minimum is always present (in other words, the variance is always a positive quantity, even for a finite number of samples). Instead, by performing the energy minimization, it is not guaranteed to find a result that is bounded from below; indeed, it may happen that a configuration x_i, which has been sampled according to $|\Psi_\alpha(x)|^2$ (corresponding to a finite local energy $e_{L,\alpha}(x_i)$), gives a very large and negative local energy for the parameters $\boldsymbol{\alpha} + \boldsymbol{\delta\alpha}$; for example, x_i may be arbitrary close to the nodal surface of $\Psi_{\alpha+\delta\alpha}(x)$, such that $e_{L,\alpha+\delta\alpha}(x_i) \to -\infty$.

The reweighting technique is unstable when the number of particles is large. Indeed, let us consider the case where only one component α_k of the vector $\boldsymbol{\alpha}$ has been incremented by $\delta\alpha_k$; then, the reweighting factors are determined by the ratio of two many-body wave functions that are exponentially large or small, leading to:

$$\mathcal{R}_{\alpha+\delta\alpha_k}(x) \approx \exp(\pm\delta\alpha_k N_p), \tag{6.8}$$

where N_p is the number of particles. Therefore, for large N_p, even a tiny change of the variational parameters would imply an exponential increase of the statistical noise. This fact makes it extremely hard (if not impossible) to find out the optimal set of parameters in a system with several particles, because the statistical noise in the variation of the energy or variance is very large. Therefore, it is necessary to

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

consider alternative optimization approaches: one possibility is to consider methods that are based upon the calculation of energy derivatives. Indeed, although approaches based upon the variance minimization are quite satisfactory, there are good motivations to develop energy-based minimization methods: for example, an energy-optimized wave function usually provides better expectation values with respect to variance-optimized ones, since the former approach is more sensitive to low-energy variation than the latter one.

6.3 Energy Derivatives

In this section, we discuss the basics ingredients that are necessary to compute the derivatives of the variational energy with respect to a given variational parameter α_k:

$$f_k = -\frac{\partial E_\alpha}{\partial \alpha_k} = -\frac{\partial}{\partial \alpha_k} \frac{\langle \Psi_\alpha | \mathcal{H} | \Psi_\alpha \rangle}{\langle \Psi_\alpha | \Psi_\alpha \rangle}. \tag{6.9}$$

The dependence of E_α on the variational parameters is just a consequence of the fact that the wave function $|\Psi_\alpha\rangle$ depends upon $\boldsymbol{\alpha}$. Thus, in order to differentiate E_α, it is convenient to expand $|\Psi_\alpha\rangle$ for small changes $\alpha_k \rightarrow \alpha_k + \delta\alpha_k$. For a given configuration $|x\rangle$, where $\Psi_\alpha(x)$ is a complex number (in case of complex parameters, we can assume that all the α_k are real, once we consider their real and imaginary parts separately), we have that:

$$\Psi_{\alpha+\delta\alpha_k}(x) = \Psi_\alpha(x) + \delta\alpha_k \frac{\partial \Psi_\alpha(x)}{\partial \alpha_k} + O(\delta\alpha_k^2), \tag{6.10}$$

where the notation $\Psi_{\alpha+\delta\alpha_k}(x)$ means that only the component α_k of the vector $\boldsymbol{\alpha}$ has been incremented by $\delta\alpha_k$. In the following, for simplicity, we assume that $\Psi_\alpha(x) \neq 0$ for all the configurations. For fermionic systems in the continuous space, the nodal region $\Psi_\alpha(x) = 0$ represents a negligible (i.e., with zero measure) integration domain. On the lattice, accidental configurations with $\Psi_\alpha(x) = 0$ can be removed by considering a tiny perturbation of the variational *Ansatz* (e.g., by adding a small noisy part) and considering the limit of vanishing perturbation. Then, Eq. (6.10) can be formally written in terms of a local operator $\mathcal{O}_k$, corresponding to the parameter α_k and defined by diagonal matrix elements $\mathcal{O}_k(x)$:

$$\langle x | \mathcal{O}_k | x' \rangle = \delta_{x,x'} \mathcal{O}_k(x), \tag{6.11}$$

$$\mathcal{O}_k(x) = \frac{\partial \ln \Psi_\alpha(x)}{\partial \alpha_k} = \frac{1}{\Psi_\alpha(x)} \frac{\partial \Psi_\alpha(x)}{\partial \alpha_k}; \tag{6.12}$$

here, in principle, $\mathcal{O}_k(x)$ may depend upon the variational parameters $\boldsymbol{\alpha}$, however, to keep the notation simple, we prefer not to put the label α in the local operators.

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

The important point is that $\mathcal{O}_k(x)$ can be usually computed for the given *Ansatz* of the variational state (see section 6.7 for the Jastrow-Slater wave functions). In this way, we can write a formal expansion of the many-body state as:

$$|\Psi_{\alpha+\delta\alpha_k}\rangle = (1 + \delta\alpha_k \mathcal{O}_k)|\Psi_\alpha\rangle, \tag{6.13}$$

which can be readily verified by taking the overlap of both sides of the above equation with $|x\rangle$ and using Eqs. (6.11) and (6.12). Notice that the diagonal operator $\mathcal{O}_k$ is not necessarily Hermitian, as its diagonal elements are not necessarily real, for a generic complex case.

Let us now show how to obtain the explicit form of the energy derivative with respect to a given variational parameter. It is clear that the variational energy E_α (as well as any other correlation function) does not depend on the overall normalization (and global phase) of the wave function. In other words, by scaling the wave function by an arbitrary complex constant c, i.e., $|\Psi_\alpha\rangle \to c|\Psi_\alpha\rangle$, E_α remains unchanged. In order to exploit this property, it is better to consider explicitly normalized wave functions. First of all we define:

$$|v_{0,\alpha}\rangle \equiv \frac{|\Psi_\alpha\rangle}{||\Psi_\alpha||}, \tag{6.14}$$

where $||\Psi_\alpha||$ indicates the norm of the state $|\Psi_\alpha\rangle$. Then, we define a set of states (one for each value of $k = 1, \ldots, p$):

$$|v_{k,\alpha}\rangle \equiv \left(\mathcal{O}_k - \overline{\mathcal{O}}_k\right)|v_{0,\alpha}\rangle, \tag{6.15}$$

where

$$\overline{\mathcal{O}}_k = \langle v_{0,\alpha}|\mathcal{O}_k|v_{0,\alpha}\rangle = \frac{\langle\Psi_\alpha|\mathcal{O}_k|\Psi_\alpha\rangle}{\langle\Psi_\alpha|\Psi_\alpha\rangle}. \tag{6.16}$$

The states $|v_{k,\alpha}\rangle$ are orthogonal to $|v_{0,\alpha}\rangle$, as easily verified when using Eq. (6.16); however, they are neither normalized nor orthogonal to each other, i.e., in general $\langle v_{k,\alpha}|v_{k',\alpha}\rangle \neq \delta_{k,k'}$ for $k, k' \neq 0$. Therefore, the set of states $|v_{0,\alpha}\rangle$ and $\{|v_{k,\alpha}\rangle\}$ defines a *semi-orthogonal* basis.

In order to compute the normalized wave function when the parameter α_k is changed, we first compute the norm of $|\Psi_{\alpha+\delta\alpha_k}\rangle$:

$$\begin{aligned} ||\Psi_{\alpha+\delta\alpha_k}||^2 &= \langle\Psi_\alpha|(1 + \delta\alpha_k\mathcal{O}_k)^*(1 + \delta\alpha_k\mathcal{O}_k)|\Psi_\alpha\rangle \\ &= ||\Psi_\alpha||^2\left[1 + 2\Re(\delta\alpha_k\overline{\mathcal{O}}_k) + O(\delta\alpha_k^2)\right]. \end{aligned} \tag{6.17}$$

Then, we have that:

$$\begin{aligned} |v_{0,\alpha+\delta\alpha_k}\rangle &= \frac{|\Psi_{\alpha+\delta\alpha_k}\rangle}{||\Psi_{\alpha+\delta\alpha_k}||} = |v_{0,\alpha}\rangle + \left[\delta\alpha_k\mathcal{O}_k - \Re(\delta\alpha_k\overline{\mathcal{O}}_k)\right]|v_{0,\alpha}\rangle + O(\delta\alpha_k^2) \\ &= \left[1 + i\Im(\delta\alpha_k\overline{\mathcal{O}}_k)\right]|v_{0,\alpha}\rangle + \delta\alpha_k|v_{k,\alpha}\rangle + O(\delta\alpha_k^2), \end{aligned} \tag{6.18}$$

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

which can be finally recast as:

$$|v_{0,\alpha+\delta\alpha_k}\rangle = \exp(i\delta\phi)\left[|v_{0,\alpha}\rangle + \delta\alpha_k|v_{k,\alpha}\rangle\right] + O(\delta\alpha_k^2), \tag{6.19}$$

where $\delta\phi = \Im(\delta\alpha_k\overline{\mathcal{O}}_k)$.

By using the above expression, it is immediate to work out the derivative of the variational energy E_α with respect to a given variational parameter α_k:

$$\begin{aligned}\frac{\partial E_\alpha}{\partial\alpha_k} &= \lim_{\delta\alpha_k\to 0}\frac{\langle v_{0,\alpha+\delta\alpha_k}|\mathcal{H}|v_{0,\alpha+\delta\alpha_k}\rangle - \langle v_{0,\alpha}|\mathcal{H}|v_{0,\alpha}\rangle}{\delta\alpha_k} \\ &= \langle v_{k,\alpha}|\mathcal{H}|v_{0,\alpha}\rangle + \langle v_{0,\alpha}|\mathcal{H}|v_{k,\alpha}\rangle = 2\Re\left[\frac{\langle\Psi_\alpha|\mathcal{H}(\mathcal{O}_k - \overline{\mathcal{O}}_k)|\Psi_\alpha\rangle}{\langle\Psi_\alpha|\Psi_\alpha\rangle}\right].\end{aligned} \tag{6.20}$$

Notice also that, as expected, the phase factor $\delta\phi$ does not enter in the above expression.

In order to evaluate the force f_k by a standard Monte Carlo sampling, we introduce a completeness relation to have:

$$\begin{aligned}f_k &= -2\Re\left[\frac{\sum_x\langle\Psi_\alpha|\mathcal{H}|x\rangle\langle x|\left(\mathcal{O}_k - \overline{\mathcal{O}}_k\right)|\Psi_\alpha\rangle}{\sum_x\langle\Psi_\alpha|x\rangle\langle x|\Psi_\alpha\rangle}\right] \\ &= -2\Re\left[\sum_x\frac{e_L^*(x)\left(\mathcal{O}_k(x) - \overline{\mathcal{O}}_k\right)|\Psi_\alpha(x)|^2}{\sum_x|\Psi_\alpha(x)|^2}\right],\end{aligned} \tag{6.21}$$

where $e_L^*(x)$ is the complex conjugate of the local energy (here, we omitted the index α, in harmony with the notation adopted for the local operators $\mathcal{O}_k$), as it is generally a complex-valued function. Then, f_k can be evaluated by considering:

$$f_k \approx -2\Re\left[\frac{1}{N}\sum_{i=1}^N e_L^*(x_i)\left(\mathcal{O}_k(x_i) - \overline{\mathcal{O}}_k\right)\right], \tag{6.22}$$

$$\overline{\mathcal{O}}_k \approx \frac{1}{N}\sum_{i=1}^N \mathcal{O}_k(x_i). \tag{6.23}$$

We remark that for the evaluation of the errorbars, it is important to take advantage of the correlation between the expectation value $\overline{\mathcal{O}}_k$ and the actual derivatives f_k, which must be computed using the same sample $\{x_i\}$.

Finally, we emphasize two main properties of the above expression of derivatives, which represent the basis of any efficient stochastic minimization method:

1. From Eq. (6.22), we have that if the wave function is an exact eigenstate of $\mathcal{H}$, the local energy coincides with the corresponding exact eigenvalue E, regardless the sample $\{x_i\}$. This implies that f_k identically vanishes without statistical fluctuations and thus we recover the zero-variance property for energy derivatives.

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

2. In the general case, the variance is always bounded by the square of the number of particles N_p^2, because the local energy is an extensive random variable that is multiplied by a derivative estimator, i.e., $\mathcal{O}_k(x)$, that is also at most extensive. Thus the variance of the product of two extensive quantity is at most $O(N_p^2)$. This implies that these expressions can be used to optimize wave functions with a large number of particles, since, at variance with methods based on the reweighting technique, the statistical fluctuations to estimate energy derivatives increases at most polynomially with the number of particles.

6.4 The Stochastic Reconfiguration

The knowledge of energy derivatives of Eq. (6.9) allows us to employ the steepest-descent method (Press et al., 2007) to change the variational parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$, even when p is very large:

$$\alpha_k' = \alpha_k + \delta\alpha_k, \tag{6.24}$$

$$\delta\alpha_k = \Delta f_k, \tag{6.25}$$

where Δ is an arbitrary (small) constant. In principle, its value can be optimized to reach the lowest possible energy at each iteration; however, in most applications, it is a common practice to keep Δ constant along the minimization procedure. Then, the variational parameters are iteratively improved along a Markov chain procedure. In absence of noise, the steepest-descent method always converges to a minimum, where the Euler conditions $f_k = 0$ are satisfied. Indeed, let us suppose that $f_k \neq 0$, then the energy for $\boldsymbol{\alpha}'$ is given by a Taylor expansion to linear order in Δ:

$$E_{\alpha'} = E_\alpha + \sum_k \frac{\partial E_\alpha}{\partial \alpha_k}\delta\alpha_k + O(\Delta^2) = E_\alpha - \Delta \sum_k f_k^2 + O(\Delta^2), \tag{6.26}$$

where we used that $\partial E_\alpha / \partial \alpha_k = -f_k$ and $\delta\alpha_k = \Delta f_k$. Therefore, for small Δ, when the linear truncation is accurate enough in the Taylor expansion, we obtain that:

$$\Delta E \equiv E_{\alpha'} - E_\alpha = -\Delta \sum_k f_k^2 \leq 0; \tag{6.27}$$

here, the equality sign holds only when $f_k = 0$. Thus, the method converges to a minimum for a large number of iterations just because the energy monotonically decreases with the number of iterations. Within the steepest-descent method only the first derivative of the energy is computed and it is certain that a small change of the parameters $\boldsymbol{\delta\alpha} = (\delta\alpha_1, \ldots, \delta\alpha_p)$ parallel to the force $\mathbf{f} = (f_1, \ldots, f_p)$ will decrease the energy; the only issue concerns the size of Δ, which must be taken sufficiently small to make the quadratic term in Eq. (6.26) negligible.

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

Let us discuss a simple case where all the parameters are "equivalent" and consider a wave function that is a linear combination of p orthonormal Slater determinants $|D_k\rangle$ (i.e., $\langle D_k|D_{k'}\rangle = \delta_{k,k'}$):

$$|\Psi_\alpha\rangle = \sum_k \alpha_k |D_k\rangle; \tag{6.28}$$

then, it is reasonable to make a search with the condition that

$$\delta s^2 = \sum_k \delta\alpha_k^2 \tag{6.29}$$

is small enough along the optimization procedure, where it is assumed that for small δs the Taylor expansion of the energy to first order in $\boldsymbol{\delta\alpha}$ is accurate enough. In order to enforce the constraint of Eq. (6.29), the usual scheme is to introduce a Lagrange multiplier μ and minimize the quadratic form:

$$\Delta E + \mu\delta s^2 = \sum_k \left(-f_k\delta\alpha_k + \mu\delta\alpha_k^2\right); \tag{6.30}$$

this approach yields a minimum condition:

$$\delta\alpha_k = \frac{f_k}{2\mu}, \tag{6.31}$$

which is compatible with the steepest-descent approach of Eq. (6.25) with $\Delta = 1/2\mu$. Although the above argument makes sense for simple wave functions, as for example the ones with linear dependence upon the parameters, serious difficulties arise in strongly-correlated states, where the dependence on the variational parameters is highly non-linear (e.g., in the Jastrow factors). Here, a small change of a given variational parameter can produce very different wave functions and physical quantities, whereas another parameter may weakly affect the wave function. In order to overcome these difficulties, it is important to introduce a more appropriate metric δs^2 that is used to estimate the "proximity" of two normalized (complex) wave functions $|v_{0,\alpha}\rangle$ and $|v_{0,\alpha+\delta\alpha}\rangle$:

$$\delta s^2 = \mathrm{Min}_{\delta\theta}||\exp(-i\delta\theta)v_{0,\alpha+\delta\alpha} - v_{0,\alpha}||^2. \tag{6.32}$$

In the previous definition of δs^2, the minimization on the phase factor $\delta\theta$ is necessary because we do not want to distinguish between two wave functions that differ only by an overall phase factor, as they produce the same correlation functions. In other words, we want to define a distance δs^2 that vanishes when we have physically equivalent wave functions. Then, we replace in Eq. (6.32) the

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

expression for $|v_{0,\alpha+\delta\alpha}\rangle$ that is obtained by generalizing Eq. (6.19) to the case where several parameters are changed:

$$|v_{0,\alpha+\delta\alpha}\rangle = \exp(i\delta\phi)\left[|v_{0,\alpha}\rangle + \sum_k \delta\alpha_k |v_{k,\alpha}\rangle\right] + O(|\delta\alpha|^2), \tag{6.33}$$

where $\delta\phi = \sum_k \Im(\delta\alpha_k \overline{\mathcal{O}}_k)$. Now, the minimization over $\delta\theta$ gives $\delta\theta = \delta\phi$, thus leading to:

$$\delta s^2 = \sum_{k,k'} \langle v_{k,\alpha}|v_{k',\alpha}\rangle \delta\alpha_k \delta\alpha_{k'} + o\left(|\delta\alpha|^2\right). \tag{6.34}$$

Since all increments $\delta\alpha_k$ are assumed real (as discussed previously, here we assume that all parameters are real), we can symmetrize the previous expression with respect to the indices k and k' and neglect the terms that are $o(|\delta\alpha|^2)$, obtaining:

$$\delta s^2 = \sum_{k,k'} \left(\frac{\langle v_{k,\alpha}|v_{k',\alpha}\rangle + \langle v_{k',\alpha}|v_{k,\alpha}\rangle}{2}\right) \delta\alpha_k \delta\alpha_{k'}. \tag{6.35}$$

In this way, we can finally identify a matrix $\mathbf{S}$ that fully determines the metric in the space of normalized wave functions:

$$S_{k,k'} = \Re\left(\langle v_{k,\alpha}|v_{k',\alpha}\rangle\right), \tag{6.36}$$

which implies that the distance between two wave functions reads:

$$\delta s^2 = \sum_{k,k'} S_{k,k'} \delta\alpha_k \delta\alpha_{k'}. \tag{6.37}$$

At this point, it is natural to improve the steepest-descent method by using the metric given by $\mathbf{S}$. The minimization of $\Delta E + \mu\delta s^2$ with the metric δs^2 given in Eq. (6.37) improves the convergence to the minimum of the variational energy with respect to the simple steepest-descend approach, as non-equivalent parameters can be appropriately changed with a different scale. This approach is called *stochastic reconfiguration* (Sorella, 1998 and 2001). The minimization of $\Delta E + \mu\delta s^2$ gives:

$$\sum_{k'} S_{k,k'} \delta\alpha_{k'} = \frac{f_k}{2\mu}, \tag{6.38}$$

which is a set of linear equations for the unknown vector $\boldsymbol{\delta\alpha}$. After having solved this linear system, we can update the variational parameters until convergence is reached; as in the steepest-descent method, we can set $\Delta = 1/(2\mu)$ small enough, which may be kept fixed during the optimization. We would like to stress the fact

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

that, since the matrix $\mathbf{S}$ is strictly positive definite, the energy is monotonically decreasing along the optimization as:

$$\Delta E = -\Delta \sum_{k,k'} S^{-1}_{k,k'} f_k f_{k'} < 0. \tag{6.39}$$

Within a Monte Carlo procedure, the matrix $\mathbf{S}$ is evaluated by a finite sampling of N configurations $\{x_i\}$ as:

$$S_{k,k'} \approx \Re \left[\frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{O}_k(x_i) - \overline{\mathcal{O}}_k \right) \left(\mathcal{O}_{k'}(x_i) - \overline{\mathcal{O}}_{k'} \right) \right]; \tag{6.40}$$

the forces are also computed in a similar way, see Eq. (6.22). Therefore, the solution of the linear system (6.38) is affected by statistical errors, yielding statistical fluctuations of the final variational parameters $\{\alpha_k\}$, even when convergence has been reached. In this regime, it is convenient to perform several iterations, in order to obtain accurate values for the variational parameters by averaging them after equilibration.

By considering the limit of $\Delta \to 0$, the optimization procedure reduces to a set of stochastic differential equations that are approximately described by a Langevin dynamics (see Chapter 4):

$$\frac{d\boldsymbol{\alpha}(t)}{dt} = \mathbf{S}^{-1}\mathbf{f}[\boldsymbol{\alpha}(t)] + \boldsymbol{\eta}(t), \tag{6.41}$$

where we have explicitly indicated that the forces depend upon the variational parameters. Here, we neglect the correlation between the force components in the statistical noise, since this aspect does not alter the qualitative discussion. The noise is described by:

$$\langle \eta_k(t) \rangle = 0, \tag{6.42}$$

$$\langle \eta_k(t) \eta_{k'}(t') \rangle = 2T_{\text{noise}} \delta_{k,k'} \delta(t - t'). \tag{6.43}$$

Within a Monte Carlo optimization, the parameter T_{noise} depends on the fact that both $\mathbf{f}$ and $\mathbf{S}$ are evaluated by using a finite sampling with N configurations. We would like to emphasize that the inversion of the matrix $\mathbf{S}$ introduces a controllable bias, which vanishes for $N \to \infty$. This is due to the fact that the matrix elements $S_{k,k'}$, obtained from Eq. (6.40), are random variables distributed around the corresponding exact values with fluctuations that are $O(1/\sqrt{N})$. Therefore, the noise can be tuned to zero by increasing N as $T_{\text{noise}} \propto 1/N$, for the central limit theorem. The number of iterations necessary to reach convergence is weakly dependent on the "temperature" T_{noise} (see Chapter 4); however, it crucially depends on the energy landscape of the system. The optimal value for N is the smallest one that provides the fluctuations of the parameters within the desired accuracy. It should be noted

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

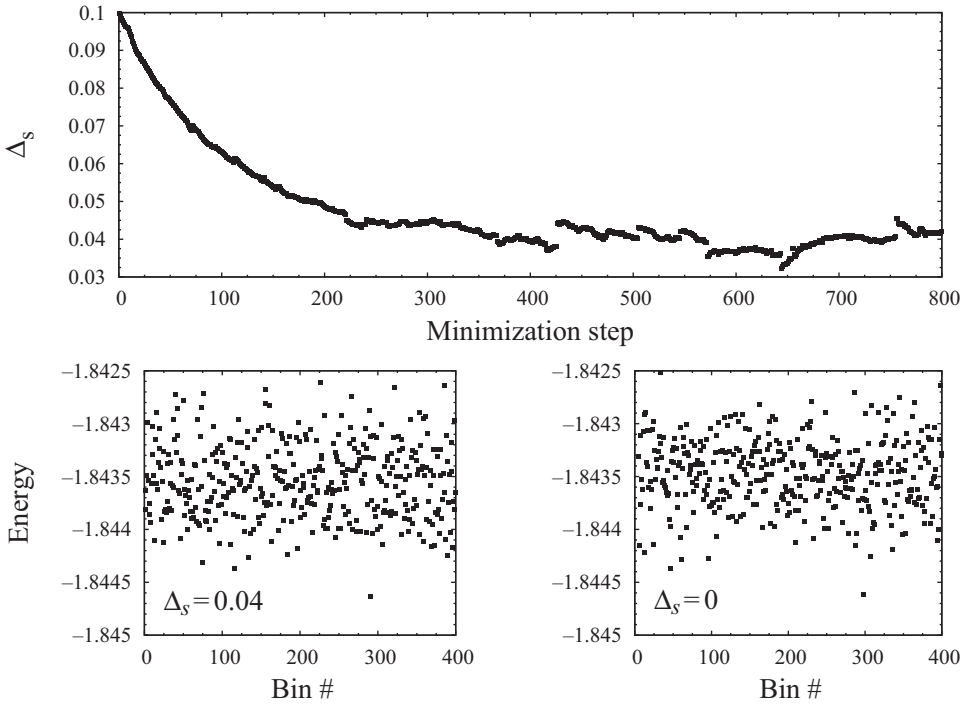

Figure 6.1 Results for the attractive Hubbard model with $U/t = -1$ for $L = 162$ sites at half filling (i.e., $N_e = L$). The wave function is obtained by applying a density-density Jastrow factor (1.65) to a BCS wave function (1.91) with on-site s-wave pairing $\Delta_k = \Delta_s$; the number of Jastrow pseudo-potentials is 29. Upper panel: by using the stochastic reconfiguration, the optimization of Δ_s as a function of the Monte Carlo steps, here $N = 5000$. Lower panels: the variational energy calculated by fixing all the variational parameters, for the optimal value $\Delta_s = 0.04$ (left panel) and for $\Delta_s = 0$ (right panel).

that the variational parameters, averaged over the Langevin simulation, are close to the true energy minimum, but may be affected by a bias that scales to zero with T_{noise}, due to the presence of non-quadratic terms in the energy landscape. Therefore, the best strategy to obtain accurate results is to average the variational parameters by systematically increasing the value of N.

As an example for the convergence of the stochastic reconfiguration, we report in Fig. 6.1 the case of the attractive Hubbard model with $U/t = -1$ at half filling, with $N_e = L = 162$. The variational wave function is given by the BCS state (1.91) with on-site s-wave pairing $\Delta_k = \Delta_s$, supplemented with the density-density Jastrow factor (1.65). For this value of the electron-electron interaction, a small value of Δ_s is stabilized, i.e., $\Delta_s \approx 0.04$. We would like to emphasize that the energy gain due to this term is $\Delta E = 0.00007(2)$, which is much smaller than the energy fluctuations that are present in the simulation; nevertheless, an accurate determination of this

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

variational parameter is possible by using the stochastic optimization, demonstrating the power of such Monte Carlo methods.

6.4.1 Covariance Property

One important property of the stochastic reconfiguration optimization is that it has the same efficiency when the parametrization of the many-body state is changed without changing its form. This fact represents an important improvement with respect to the simple steepest-descent method. For example, let us consider the case in which the wave function contains the Gutzwiller factor g and a fugacity term f (suitable whenever the number of fermions is not fixed along the simulation); for the fermionic case, in which $n_i = n_{i,\uparrow} + n_{i,\downarrow}$ and $n_{i\sigma}^2 = n_{i\sigma}$, the correlation term can be written in two equivalent ways:

$$\exp\left(-\sum_i \frac{g}{2} n_i^2 - \sum_i f n_i\right) = \exp\left(-\sum_i f' n_i\right) \prod_i \left[1 + (g' - 1) n_{i\uparrow} n_{i\downarrow}\right], \quad (6.44)$$

where $g' = \exp(-g)$ and $f' = f + g/2$. Choosing one parametrization (with g and f) or the other (with g' and f') does not necessarily give the same efficiency in the optimization procedure (while it must give the same energy at equilibrium, apart from problems related to local minima). Here, we show that, within the stochastic reconfiguration technique, the minimization procedure has the same effectiveness when changing the way of parameterizing the quantum state.

In the general case, let us consider an arbitrary transformation of the variational parameters:

$$\alpha_k' = \alpha_k'(\alpha_1, \ldots, \alpha_p), \quad (6.45)$$

which gives the same quantum state, i.e., $\Psi_{\alpha'}(x) = \Psi_\alpha(x)$. Then, we want to impose that, after one step of optimization, the same improvement is obtained when considering the parametrization with $\boldsymbol{\alpha}$ or $\boldsymbol{\alpha}'$. The condition $\Psi_{\alpha'+\delta\alpha'}(x) = \Psi_{\alpha+\delta\alpha}(x)$ would require that:

$$\delta\alpha_k' = \sum_{k'} \frac{\partial \alpha_k'}{\partial \alpha_{k'}} \delta\alpha_{k'} = (\mathbf{J}\boldsymbol{\delta\alpha})_k\,, \quad (6.46)$$

where we have defined the Jacobian matrix by $J_{k,k'} = \frac{\partial \alpha_k'}{\partial \alpha_{k'}}$. The previous relation is indeed satisfied within the stochastic reconfiguration technique. In fact, the local operator $\mathcal{O}_k'(x)$, which enters into the calculation of the energy derivative, is given by:

$$\mathcal{O}_k'(x) = \frac{1}{\Psi_{\alpha'}(x)} \frac{\partial \Psi_{\alpha'}(x)}{\partial \alpha_k'} = \sum_{k'} \frac{\partial \alpha_{k'}}{\partial \alpha_k'} \left(\frac{1}{\Psi_\alpha(x)} \frac{\partial \Psi_\alpha(x)}{\partial \alpha_{k'}}\right) = \sum_{k'} J_{k,k'}^{-1} O_{k'}(x), \quad (6.47)$$

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

where we have used the fact that $\Psi_{\alpha'}(x) = \Psi_\alpha(x)$ and the property that the Jacobian of the inverse transformation is just the inverse of the Jacobian of the direct transformation. In this way, the new forces $\mathbf{f}'$ and the new matrix $\mathbf{S}'$ are:

$$\mathbf{f}' = \mathbf{J}^{-1}\mathbf{f}, \tag{6.48}$$

$$\mathbf{S}' = \mathbf{J}^{-1}\mathbf{S}\mathbf{J}^{-1}. \tag{6.49}$$

Therefore, the iteration step with the new parametrization $\boldsymbol{\alpha}'$ is given by:

$$\delta\boldsymbol{\alpha}' = \Delta(\mathbf{S}')^{-1}\mathbf{f}' = \Delta\left(\mathbf{J}^{-1}\mathbf{S}\mathbf{J}^{-1}\right)^{-1}\mathbf{J}^{-1}\mathbf{f} = \mathbf{J}\delta\boldsymbol{\alpha}, \tag{6.50}$$

which is exactly what we required at the beginning. This fact can be named *covariance property* of the method, because (exactly as in general relativity) the matrix $\mathbf{S}$ represents the metric of our non-linear space (of wave functions) and the direction of the optimization is covariant and does not depend on the way in which the wave function is represented. This property is very important since it implies that there is no way to improve the method by performing any non-linear transformation among the parameters.

6.4.2 A Scale-Invariant Regularization

A fundamental ingredient along the stochastic optimization is the inversion of the matrix $\mathbf{S}$, which is a non-negative matrix, even when considering a finite sampling of N configurations. However, it may happen that it has eigenvalues that differ by several order of magnitudes and, therefore, it is always worth doing a *pre-conditioning* that stabilizes the inversion, namely considering the scaling:

$$S^{\mathrm{pc}}_{k,k'} = \frac{S_{k,k'}}{\sqrt{S_{k,k}S_{k',k'}}}. \tag{6.51}$$

In this way, we first find the solutions of the linear equations:

$$\sum_{k'} S^{\mathrm{pc}}_{k,k'}\delta\alpha^{\mathrm{pc}}_{k'} = \frac{f^{\mathrm{pc}}_k}{2\mu}, \tag{6.52}$$

where the scaled unknowns are $\delta\alpha^{\mathrm{pc}}_{k'} = \sqrt{S_{k',k'}}\delta\alpha_{k'}$ and the scaled forces are $f^{\mathrm{pc}}_k = f_k/\sqrt{S_{k,k}}$; then, we obtain the original $\boldsymbol{\delta\alpha}$ by a simple rescaling $\delta\alpha_k = \delta\alpha^{\mathrm{pc}}_k/\sqrt{S_{k,k}}$.

A most serious problem arises whenever the matrix $\mathbf{S}^{\mathrm{pc}}$ acquires very small and even negligible eigenvalues, implying that the application of $(\mathbf{S}^{\mathrm{pc}})^{-1}$ to the force $\mathbf{f}^{\mathrm{pc}}$ can be unstable, especially if we consider that both quantities are evaluated within a stochastic procedure. For example, if there is a redundancy in the parametrization

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

of the wave function, one particular linear combination of the forces will vanish, as well as the matrix $\mathbf{S}^{\mathrm{pc}}$ will be singular. For example, by diagonalizing $\mathbf{S}^{\mathrm{pc}}$, we have that:

$$\lambda_i \, (\delta\alpha^{\mathrm{pc}})_i^U = (f^{\mathrm{pc}})_i^U, \tag{6.53}$$

where λ_i are the eigenvalues of $\mathbf{S}^{\mathrm{pc}} = \mathbf{U}^\dagger \Lambda \mathbf{U}$, $(\boldsymbol{\delta\alpha}^{\mathrm{pc}})^U = \mathbf{U}\boldsymbol{\delta\alpha}^{\mathrm{pc}}$, and $(\mathbf{f}^{\mathrm{pc}})^U = \mathbf{U}\mathbf{f}^{\mathrm{pc}}$. For a redundant parameter, $(f^{\mathrm{pc}})_k^U = \lambda_k = 0$, which makes the solution of the problem undetermined. Therefore, it is clear that some form of regularization is necessary. A simple but efficient possibility is to modify the diagonal elements only:

$$S_{k,k'}^{\mathrm{pc},\epsilon} = S_{k,k'}^{\mathrm{pc}} + \epsilon \delta_{k,k'}. \tag{6.54}$$

This defines a regular matrix that can be safely inverted for any $\epsilon > 0$; in this way, we get rid of a possible dependence among the variational parameters. Indeed, Eq. (6.53) is modified into:

$$(\lambda_i + \epsilon)(\delta\alpha^{\mathrm{pc}})_i^U = (f^{\mathrm{pc}})_i^U, \tag{6.55}$$

which implies that $(\delta\alpha^{\mathrm{pc}})_i^U \approx 0$ for the cases in which $(f^{\mathrm{pc}})_i^U \approx 0$ and $\lambda_i \ll \epsilon$. Therefore, the effect of introducing a finite ϵ is to keep fixed the parameter direction along which $(f^{\mathrm{pc}})_i^U \approx 0$. Since $(f^{\mathrm{pc}})_i^U \approx \sqrt{\lambda_i}$, neglecting this direction implies an error in the optimization of the order of $\sqrt{\epsilon}$.

Although this regularization does not maintain the covariance property discussed above, it preserves a subset of transformations obtained when scaling the parameters by arbitrary constants ξ_k:

$$\alpha_k' = \xi_k \alpha_k; \tag{6.56}$$

indeed, we can immediately verify this property by using that, in this case, $\mathbf{J}$ is diagonal and, therefore, $\mathbf{S}^{\mathrm{pc}}$ of Eq. (6.51) remains unchanged when performing the transformation (6.49) on the matrix $\mathbf{S}$. Therefore, such a regularization is *scale invariant*. The smaller is the value of ϵ the faster will be the approach to the lowest variational energy, but the probability of an instability will be higher (if the matrix $\mathbf{S}$ is singular); a common practice is to use $\epsilon \approx 10^{-3}$, for a reasonable efficiency and accuracy.

6.4.3 Optimization Using the Signal to Noise Ratio

At this point, we want to discuss a very simple argument that explains the reason why the method described above represents an efficient stochastic optimization procedure. The generalized force component f_k, defined in Eq. (6.22) is a random variable with a corresponding noise, whose value can be estimated statistically, as

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

well as its correlation with the other components. Indeed, the covariance matrix can be evaluated by using jackknife or bootstrap methods described in section 3.10:

$$\sigma^2_{k,k'} \approx \langle\langle f_k f_{k'}\rangle\rangle - \langle\langle f_k\rangle\rangle\langle\langle f_{k'}\rangle\rangle. \tag{6.57}$$

Let us now consider a particular direction in the parameter space that is described by taking $\boldsymbol{\alpha}(\Delta) = \boldsymbol{\alpha}_0 + \Delta\boldsymbol{\tau}$, where Δ parametrizes a line change of the parameters in the direction of the vector $\boldsymbol{\tau}$, starting from $\boldsymbol{\alpha}_0$. This implies that the generalized force f_τ along the direction $\boldsymbol{\tau}$ is given by:

$$f_\tau = -\frac{\partial E_\alpha}{\partial \Delta} = -\sum_k \frac{\partial E_\alpha}{\partial \alpha_k}\frac{\partial \alpha_k}{\partial \Delta} = \mathbf{f}\cdot\boldsymbol{\tau}. \tag{6.58}$$

Its standard deviation is:

$$\sigma_\tau = \sqrt{\sum_{k,k'} \tau_k \sigma^2_{k,k'} \tau_{k'}}. \tag{6.59}$$

Now, we can look at the direction where the signal (i.e., the actual value of the derivative) is the largest possible compared to its standard deviation. A change of parameters along this direction guarantees a lowering of the energy if the squared signal to noise ratio

$$\Sigma^2(\boldsymbol{\tau}) = \frac{\sum_{k,k'} \tau_k f_k f_{k'} \tau_{k'}}{\sum_{k,k'} \tau_k \sigma^2_{k,k'} \tau_{k'}} \tag{6.60}$$

is much larger than one (e.g., if the derivative along a given direction is non-zero by more than 3 standard deviations). In this case, the energy will be lowered by changing $\boldsymbol{\alpha}_0 \to \boldsymbol{\alpha}_0 + \Delta\boldsymbol{\tau}$, with Δ sufficiently small. In this sense, a simple standard maximization yields that the "best direction" is given by:

$$\tau_k = \Delta_\tau \sum_{k'} \left(\sigma^2\right)^{-1}_{k,k'} f_{k'}, \tag{6.61}$$

where Δ_τ is an arbitrary constant, as the maximum of $\Sigma^2(\boldsymbol{\tau})$ does not change when scaling the solution by an arbitrary constant. In practice, for accurate variational wave functions, the covariance matrix is expected to be close to the matrix $\mathbf{S}$, since the fluctuations of the force component f_k are mainly given by the fluctuations of $(\mathcal{O}_k - \overline{\mathcal{O}}_k)$, and the fluctuations of the local energy can be neglected, leading to:

$$\sigma^2_{k,k'} \simeq S_{k,k'}. \tag{6.62}$$

Therefore, Eqs. (6.61) and (6.62) explain the reason why the stochastic reconfiguration approach gives a particularly efficient way to perform Monte Carlo

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

optimizations. In fact, the minimization according to Eq. (6.38) follows (approximately) the one in which there is a maximal signal to noise ratio, which is fundamental for a stable stochastic optimization.

6.5 Stochastic Reconfiguration as a Projection Technique

Here, we show that the stochastic reconfiguration method is equivalent, in the case of real wave functions, to an imaginary-time projection technique, which is restricted in the subspace defined by $|v_{0,\alpha}\rangle$ and $\{|v_{k,\alpha}\rangle\}$, as defined in Eqs. (6.14) and (6.15). In order to prove this statement, let us define a wave function that is obtained from the variational one $|\Psi_\alpha\rangle$ by applying one step of the power method (see section 1.7):

$$|\Psi_\Delta\rangle = (1 - \Delta\mathcal{H})|\Psi_\alpha\rangle, \tag{6.63}$$

where Δ is a small constant, such to improve the energy with respect to $|\Psi_\alpha\rangle$. Here, we consider an imaginary-time evolution (i.e., real Δ), but all the results are also valid for a real-time propagation (i.e., $\Delta \to i\Delta$), which allows the basis for simulating quantum dynamics within variational Monte Carlo (Carleo et al., 2012). However, in this case, we cannot assume that the variational parameters are real because they will be turned to complex already after the first steps of dynamics. Therefore, we are led to consider general complex quantum states, that are analytic functions of the variational parameters . Then, it is easy to show that Eqs. (6.14), (6.15), and (6.19) remain unchanged, even when considering complex variations $\{\delta\alpha_k\}$. In general, for any finite value of Δ, the state $|\Psi_\Delta\rangle$ will be different from any state that can be parametrized by using the original variational form, namely $|\Psi_\Delta\rangle \neq |\Psi_{\alpha'}\rangle$ for any choice of the parameters $\boldsymbol{\alpha}'$. Nevertheless, for small enough Δ, $|\Psi_\Delta\rangle$ is very close to $|\Psi_\alpha\rangle$ and we may ask whether a suitable parametrization with $\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{\delta\alpha}$ can give an accurate representation of $|\Psi_\Delta\rangle$. In particular, we can minimize the "distance" between $|\Psi_\Delta\rangle$ and $e^{i\delta\theta}|\Psi_{\alpha+\delta\alpha}\rangle$ (where $\delta\theta$ is a generic angle that has to be taken into account for the real-time propagation). By using the notations of Eqs (6.14) and (6.15), we have:

$$|\Psi_\Delta\rangle = ||\Psi_\alpha||(1 - \Delta\mathcal{H})|v_{0,\alpha}\rangle, \tag{6.64}$$

$$|\Psi_{\alpha+\delta\alpha}\rangle = e^{i\delta\phi}||\Psi_{\alpha+\delta\alpha}||\left[|v_{0,\alpha}\rangle + \sum_k \delta\alpha_k|v_{k,\alpha}\rangle\right], \tag{6.65}$$

where $||\Psi_{\alpha+\delta\alpha}|| = C||\Psi_\alpha||$ with $C \approx 1$, see Eq. (6.17). The minimization of the distance can be achieved by projecting $|\Psi_\Delta\rangle$ into the subspace spanned by the semi-orthogonal basis set:

$$\langle v_{0,\alpha}|\Psi_\Delta\rangle = e^{i\delta\theta}\langle v_{0,\alpha}|\Psi_{\alpha+\delta\alpha}\rangle, \tag{6.66}$$

$$\langle v_{k,\alpha}|\Psi_\Delta\rangle = e^{i\delta\theta}\langle v_{k,\alpha}|\Psi_{\alpha+\delta\alpha}\rangle, \tag{6.67}$$

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

which lead to the following set of linear equations for the $\delta\alpha_k$'s:

$$1 - \Delta\langle v_{0,\alpha}|\mathcal{H}|v_{0,\alpha}\rangle = Ce^{i(\delta\theta+\delta\phi)}, \tag{6.68}$$

$$-\Delta\langle v_{k,\alpha}|\mathcal{H}|v_{0,\alpha}\rangle = Ce^{i(\delta\theta+\delta\phi)}\sum_{k'}\langle v_{k,\alpha}|v_{k',\alpha}\rangle\delta\alpha_{k'}. \tag{6.69}$$

Here, the first equation defines the normalization constant $C = 1 + O(\Delta)$, which is irrelevant, ($\delta\theta = -\delta\phi$ for real Δ), while the second one can be recast as:

$$\sum_{k'}\tilde{S}_{k,k'}\delta\alpha_{k'} = \frac{\tilde{f}_k}{2\mu}, \tag{6.70}$$

where $\mu = Ce^{i\theta}/\Delta$ and both the force component $\tilde{f}_k$ and the matrix $\tilde{S}_{k,k'}$ are generally complex and given by:

$$\tilde{f}_k = -2\langle v_{k,\alpha}|\mathcal{H}|v_{0,\alpha}\rangle = -2\frac{\langle\Psi_\alpha|\left(\mathcal{O}_k - \overline{\mathcal{O}}_k\right)^*\mathcal{H}|\Psi_\alpha\rangle}{\langle\Psi_\alpha|\Psi_\alpha\rangle}, \tag{6.71}$$

$$\tilde{S}_{k,k'} = \langle v_{k,\alpha}|v_{k',\alpha}\rangle = \frac{\langle\Psi_\alpha|\left(\mathcal{O}_k - \overline{\mathcal{O}}_k\right)^*\left(\mathcal{O}_{k'} - \overline{\mathcal{O}}_{k'}\right)|\Psi_\alpha\rangle}{\langle\Psi_\alpha|\Psi_\alpha\rangle}. \tag{6.72}$$

We would like to emphasize that the projection approach of Eq. (6.70) coincides with the minimization algorithm we have derived in Eq. (6.38) only when the variational wave function $\Psi_\alpha(x)$ is real.

We finally remark that, also within the projection technique, the matrix $\tilde{\mathbf{S}}$ requires some regularization to avoid too small eigenvalues in its inversion. In the case where we want to integrate the short time dynamics with an accuracy Δ^2, then the regularization cutoff has to be scaled according to $\epsilon \propto \Delta^4$. This is not a problem in principle, but makes prohibitive a highly accurate time integration, since we have to increase the number of samples $N \propto 1/\epsilon^2$, in order to obtain a statistical accuracy that is $O(\epsilon)$; otherwise, the error is larger than the target accuracy. Fortunately, this issue is not a problem if we are interested only to the optimization of the variational wave function and not to the accurate description of the exact time propagation, as in real time quantum dynamics.

6.6 The Linear Method

The stochastic reconfiguration method performs very well whenever there are not very different energy scales in the optimization problem, as in the case of lattice models or in *ab-initio* electronic calculations where the high-energy scales of the core electrons are eliminated by using pseudo-potentials. However, in particular cases, it may happen that the speed of convergence to the minimum remains

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

extremely slow, also within the stochastic reconfiguration scheme. An established way to speed-up any optimization scheme would be to take into account the exact second derivatives of the energy with respect to the variational parameters; this can be done by using the Newton-Raphson scheme, which is extremely efficient, since it requires very few steps to converge to an energy minimum. Unfortunately, the exact evaluation of the Hessian matrix:

$$H_{k,k'} = \frac{\partial^2 E_\alpha}{\partial \alpha_k \partial \alpha_{k'}} \tag{6.73}$$

is quite cumbersome in practice. Indeed, this would require to compute the Taylor expansion of the variational wave function up to the second order in the variation of the parameters $\boldsymbol{\alpha}$:

$$\Psi_{\alpha+\delta\alpha}(x) \approx \Psi_\alpha(x) + \sum_k \delta\alpha_k \frac{\partial \Psi_\alpha(x)}{\partial \alpha_k} + \frac{1}{2}\sum_{k,k'} \delta\alpha_k \delta\alpha_{k'} \frac{\partial^2 \Psi_\alpha(x)}{\partial \alpha_k \partial \alpha_{k'}}; \tag{6.74}$$

as before, see Eqs. (6.11) and (6.12), we can express the second derivative in terms of local operators $\mathcal{O}_{k,k'}$:

$$\langle x|\mathcal{O}_{k,k'}|x'\rangle = \delta_{x,x'}\mathcal{O}_{k,k'}(x), \tag{6.75}$$

$$\mathcal{O}_{k,k'}(x) = \frac{1}{\Psi_\alpha(x)} \frac{\partial^2 \Psi_\alpha(x)}{\partial \alpha_k \partial \alpha_{k'}}; \tag{6.76}$$

however, the actual evaluation of these operators is not easy in general (while it can be afforded for Jastrow parameters).

Instead of using the full Hessian approach, a great simplification arises when using the expansion of the wave function up to the linear order (i.e., by considering the first derivatives only) and then compute the expectation value of the Hamiltonian, even beyond the linear regime:

$$E_{\alpha+\delta\alpha} = \frac{\sum_{k,k'=0}^{p} z_k^* z_{k'} \langle v_{k,\alpha}|\mathcal{H}|v_{k',\alpha}\rangle}{\sum_{k,k'=0}^{p} z_k^* z_{k'} \langle v_{k,\alpha}|v_{k',\alpha}\rangle}, \tag{6.77}$$

where we have explicitly indicated that the sum includes the terms with $k = 0$ and $k' = 0$; then $z_k = \delta\alpha_k$ for $k > 0$ and $z_0 = 1$. This approach has been dubbed as *linear method* (Umrigar et al., 2007; Toulouse and Umrigar, 2007). Within this simplified approach, we have to compute the matrix elements:

$$H_{k,k'} = \langle v_{k,\alpha}|\mathcal{H}|\, v_{k',\alpha}\rangle, \tag{6.78}$$

$$\overline{S}_{k,k'} = \langle v_{k,\alpha}|v_{k',\alpha}\rangle, \tag{6.79}$$

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

where $\overline{S}_{0,0} = 1$, $\overline{S}_{k,0} = \overline{S}_{0,k} = 0$ for $k > 0$, and $\overline{S}_{k,k'} = \langle v_{k,\alpha} | v_{k',\alpha} \rangle$ for k and $k' > 0$. Then, the change $\delta\alpha_k$ can be obtained by minimizing the energy with respect to each z_k. This procedure leads to a generalized eigenvalue equation (since the states $\{|v_k\rangle\}$ are not orthogonal to each other):

$$\sum_{k'=0}^{p} H_{k,k'} z_{k'} = E \sum_{k'=0}^{p} \overline{S}_{k,k'} z_{k'}. \tag{6.80}$$

Among all possible right eigenvectors of the above generalized eigenvalue equation, it is not always convenient to take the one corresponding to the lowest (real) eigenvalue E. Indeed, whenever the matrices are particularly large, or the simulation is particularly noisy, it is better to take the eigenvector with maximum overlap $|z_0|^2$ with $|v_{0,\alpha}\rangle$; in this way, the correction $\delta\alpha_k = z_k/z_0$ is small and the expansion of $|\Psi_{\alpha+\delta\alpha}\rangle$ remains under control.

The matrix elements $\overline{S}_{k,k'}$ and $H_{k,k'}$ can be evaluated by a Monte Carlo sampling:

$$\overline{S}_{k,k'} \approx \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{O}_k(x_i) - \overline{\mathcal{O}}_k\right)^* \left(\mathcal{O}_{k'}(x_i) - \overline{\mathcal{O}}_{k'}\right), \tag{6.81}$$

$$H_{k,k'} \approx \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{O}_k(x_i) - \overline{\mathcal{O}}_k\right)^* \frac{\langle x_i | \mathcal{H} \left(\mathcal{O}_{k'} - \overline{\mathcal{O}}_{k'}\right) | \Psi_\alpha \rangle}{\langle x_i | \Psi_\alpha \rangle}, \tag{6.82}$$

in which we defined $\mathcal{O}_0 = \mathbb{I}$ and $\overline{\mathcal{O}}_0 = 0$ in order to include the cases where $k = 0$ or $k' = 0$. The factor in the r.h.s. of Eq. (6.82) can be expressed in terms of derivatives of the local energy, which can be efficiently computed within the variational Monte Carlo technique. Notice that the Hamiltonian matrix $\mathbf{H}$ is Hermitian only on average, since its Hermitian character is spoiled by statistical fluctuations that are present in samples with a finite number of configurations. Thus, it would be tempting to symmetrize the matrix to restore its Hermitian character, e.g., by considering $(H_{k,k'} + H^*_{k',k})/2$. Instead, our experience indicates that it is convenient to solve the above equations without performing such symmetrization, in order to maintain the so-called "strong" zero-variance property. Indeed, if a particular linear combination of states in the semi-orthogonal basis

$$|\varphi\rangle = \sum_{k=0}^{p} z_k \left(\mathcal{O}_k - \overline{\mathcal{O}}_k\right) |\Psi_\alpha\rangle \tag{6.83}$$

is a right eigenstate of the system (6.80) with eigenvalue E_φ, we obtain that:

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

$$\sum_{k'=0}^{p} H_{k,k'} z_{k'} \approx \frac{1}{N} \sum_{i=1}^{N} (\mathcal{O}_k(x_i) - \overline{\mathcal{O}}_k)^* \frac{\langle x_i | \mathcal{H} \sum_{k'=0}^{p} z_{k'} \left(\mathcal{O}_{k'} - \overline{\mathcal{O}}_{k'} \right) | \Psi_\alpha \rangle}{\langle x_i | \Psi_\alpha \rangle}$$

$$= E_\varphi \frac{1}{N} \sum_{i=1}^{N} (\mathcal{O}_k(x_i) - \overline{\mathcal{O}}_k)^* \frac{\langle x_i | \sum_{k'=0}^{p} z_{k'} (\mathcal{O}_{k'} - \overline{\mathcal{O}}_{k'}) | \Psi_\alpha \rangle}{\langle x_i | \Psi_\alpha \rangle}$$

$$\approx E_\varphi \sum_{k'=0}^{p} \overline{S}_{k,k'} z_{k'}, \tag{6.84}$$

which implies that Eq. (6.80) is satisfied with zero variance: by estimating both $H_{k,k'}$ and $\overline{S}_{k,k'}$ with the *same* configurations, we obtain a right eigenvector of the linear system corresponding to the exact eigenstate of $\mathcal{H}$ given by the assumption of Eq. (6.83). This is a "strong" zero-variance principle because it can be satisfied even when $|\Psi_\alpha\rangle$ is not an exact eigenstate of the Hamiltonian.

We would like to make some final considerations on the linear method, which may lead, in several cases, to a substantial improvement with respect to the stochastic reconfiguration approach. Even if its implementation is more difficult than the one of the stochastic reconfiguration technique and requires some extra computational time for each iteration, the total number of iterations needed to reach a converged result are usually drastically reduced as compared with the stochastic reconfiguration technique. Unfortunately, it is not possible to quantify the actual gain in computational time that is expected when using the linear method. In fact, in quantum Monte Carlo simulations, a stable, accurate, and fast optimization can be obtained even with a slowly convergent method, since each iteration may require a small number of samples; in this way, performing thousands of iterations with the stochastic reconfiguration technique does not represent a serious problem, especially when the number of variational parameters is small enough.

In all these optimization techniques, the number of samples that are necessary to achieve a meaningful accuracy of the matrices involved in the optimization should be much larger than the leading dimension of the matrix (i.e., the number of parameters), otherwise the sampling will produce rank-deficient singular (i.e., "dirty") matrices. A practical rule of thumb is to take the number of samples N such that:

$$N \geq 10 \times p, \tag{6.85}$$

where p is the number of parameters that must be optimized.

As an example of the linear method, we show the case of the repulsive Hubbard model with $U/t = 4$, on a cluster with $L = 98$ at half filling ($N_e = L$). In this case, the variational state is given by a Jastrow-Slater wave function, where the Jastrow term contains density-density terms, see Eq. (1.65), and the Slater determinant is constructed from free-electron orbitals. The variational parameters to be optimized are the Jastrow pseudo-potentials, for each independent distances

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

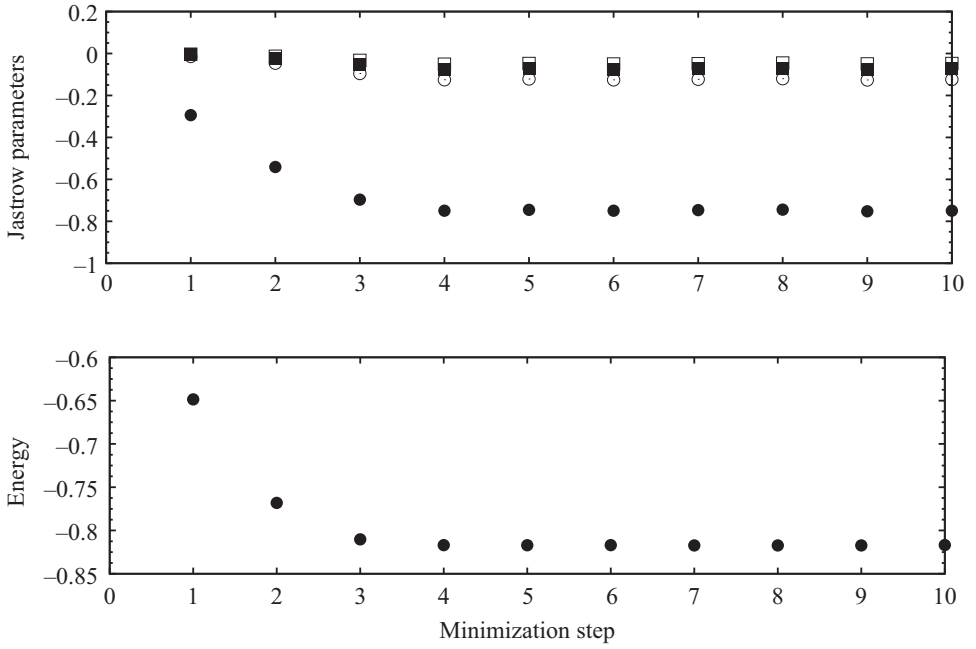

Figure 6.2 Results for the repulsive Hubbard model with $U/t = 4$ for $L = 98$ sites at half filling (i.e., $N_e = L$). The wave function is obtained by applying a density-density Jastrow factor (1.65) to a Slater determinant obtained by taking free-electron orbitals; the total number of Jastrow pseudo-potentials is 19 (including the on-site Gutzwiller term). By using the linear method, the optimization of the Jastrow pseudo-potentials (Gutzwiller term and the ones corresponding to the smallest 3 distances) is reported (upper panel). The variational energy along the minimization technique is also shown (lower panel). The number of independent Monte Carlo attempts for each optimization step is $N = 80000$.

(including the on-site Gutzwiller factor). In Fig. 6.2, we report the energy and few Jastrow parameters (the Gutzwiller term and the ones corresponding to the smallest four distances) for the first ten iterations (each one corresponding to $N = 80000$ steps). The initial parameters are set to zero. We would like to emphasize that, with the linear method, a very small number of steps $N \approx 4$ is sufficient to converge to the equilibrated values, also for the Jastrow terms that correspond to long-range distances.

Finally, we list a number of points that are still a subject of research, which have not been solved within the linear method yet:

- The "strong" zero-variance property does not hold whenever some parameters are restricted to be real (for example, usually we want to work with a real Jastrow factor). Instead, this method does not have this drawback if the calculation is restricted to a real *Ansatz* and a real Hamiltonian, since, in this case, both $H_{k,k'}$ and $\overline{S}_{k,k'}$ are real and the corresponding eigenvectors are also real.

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

- Far from the minimum, the change of the variational parameters given by $\delta\alpha_k = z_k/z_0$ can be rather large and sometimes may even increase the energy (because the change may exceed the regime where the linear expansion is valid). In this case, it is probably better to start the optimization with the stochastic reconfiguration technique and then apply the linear method only close to the minimum. Notice that this problem can be more and more relevant when increasing the system size.

6.7 Calculations of Derivatives in the Jastrow-Slater Case

Here, we show how derivatives of the wave function, see Eq. (6.12), can be computed in the case of a Jastrow-Slater *Ansatz*:

$$|\Psi_J\rangle = \mathcal{J}|\Phi_0\rangle, \tag{6.86}$$

where

$$\mathcal{J} = \exp\left(-\frac{1}{2}\sum_{i,j} v_{i,j} n_i n_j\right), \tag{6.87}$$

and $|\Phi_0\rangle$ is the ground state of a non-interacting (BCS) Hamiltonian:

$$\mathcal{H}_0 = \sum_{I,J} t_{I,J} d_I^\dagger d_J. \tag{6.88}$$

In the following, we consider the translationally invariant case, where both the Jastrow pseudo-potential and the parameters in $\mathcal{H}_0$ only depends upon the distance between the sites, i.e., $v_{i,j} \to v_k$ and $t_{I,J} \to t_k$. The generalization to a case where translational symmetry is broken is straightforward. A different approach can be devised for the Slater part if $|\Phi_0\rangle$ is defined without passing through an auxiliary Hamiltonian, namely defined by given single-particle orbitals that are suitably parametrized.

Let us start from the case of the Jastrow factor, which is very simple. Given the translational invariance, we have that:

$$\mathcal{J} = \exp\left(-\frac{1}{2}\sum_k v_k \sum_i n_i \sum_{s_i(k)} n_{s_i(k)}\right), \tag{6.89}$$

where $s_i(k)$ indicates all the sites at distance k from the site i, e.g., in one spatial dimension $s_i(k)$ includes the two sites on the left and right (for an even number of sites, the largest distance corresponds to a single site); in two dimensions, the number of connected sites depends upon the punctual group of the lattice, but also

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

on the distance (e.g., on the square lattice, it can be 8, 4, 2, and 1). From Eq. (6.12), we have that:

$$\mathcal{O}_k(x) = -\frac{1}{2}\frac{\langle x|\left(\sum_i n_i \sum_{s_i(k)} n_{s_i(k)}\right)|\Psi_J\rangle}{\langle x|\Psi_J\rangle} = -\frac{1}{2}\langle x|\left(\sum_i n_i \sum_{s_i(k)} n_{s_i(k)}\right)|x\rangle, \quad (6.90)$$

where the last equality is due to the fact that n_i is diagonal in the basis $\{|x\rangle\}$ where electrons have definite positions (and spin) in the lattice, see section 5.4. Therefore, the calculation of the derivatives corresponds to the evaluation of a simple set of diagonal correlation functions.

Then, we consider the derivatives of the variational *Ansatz* with respect to the parameters contained in the non-interacting state $|\Phi_0\rangle$ for the case with determinants (see section 5.6). Then, a small change $t_k \rightarrow t_k + \delta t_k$ corresponds to a modification of the auxiliary Hamiltonian, i.e., $\mathcal{H}_0 \rightarrow \mathcal{H}_0 + \delta t_k \mathcal{V}^{(k)}$, where $\mathcal{V}^{(k)}$ is a two-body operator that includes all the hopping terms associated to t_k:

$$\mathcal{V}^{(k)} = \sum_{I,J} V_{I,J}^{(k)} d_I^\dagger d_J, \quad (6.91)$$

where the matrix $V_{I,J}^{(k)}$ is equal to 1 (0) if the hopping t_k connects (does not connect) the sites I and J. In the following, we will assume that the ground state of $\mathcal{H}_0$ is non-degenerate, implying that all orbitals corresponding to eigenvalues $\varepsilon_\alpha \leq \varepsilon_F$ are occupied, while all the others are empty. Then, the change in the non-interacting wave function can be computed in perturbation theory:

$$|\Phi_0(t_k + \delta t_k)\rangle = |\Phi_0(t_k)\rangle - \frac{\delta t_k}{\mathcal{H}_0 - E_0}\left(\mathcal{V}^{(k)} - \langle \mathcal{V}^{(k)}\rangle_0\right)|\Phi_0(t_k)\rangle + O(\delta t_k^2), \quad (6.92)$$

where $\langle \mathcal{V}^{(k)}\rangle_0 = \langle \Phi_0(t_k)|\mathcal{V}^{(k)}|\Phi_0(t_k)\rangle$ and E_0 is the non-interacting ground-state energy. The derivative can be easily worked out by expressing the perturbation $\mathcal{V}^{(k)}$ in the basis that diagonalizes the Hamiltonian $\mathcal{H}_0$, see Eq. (5.42):

$$\mathcal{V}^{(k)} = \sum_{\alpha,\beta} \overline{V}_{\alpha,\beta}^{(k)} \phi_\alpha^\dagger \phi_\beta, \quad (6.93)$$

where

$$\overline{V}_{\alpha,\beta}^{(k)} = \sum_{I,J} U_{I,\alpha}^* V_{I,J}^{(k)} U_{J,\beta}, \quad (6.94)$$

and $\mathbf{U}$ is the unitary matrix used to diagonalize $\mathcal{H}_0$, see Eq. (5.42). Thus, we obtain that:

$$|\Phi_0(t_k + \delta t_k)\rangle = |\Phi_0(t_k)\rangle - \delta t_k \sum_{\alpha \neq \beta} \frac{\overline{V}_{\alpha,\beta}^{(k)}}{\varepsilon_\alpha - \varepsilon_\beta} \phi_\alpha^\dagger \phi_\beta |\Phi_0(t_k)\rangle + O(\delta t_k^2); \quad (6.95)$$

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

here, we have used the fact that the state $\phi_\alpha^\dagger \phi_\beta |\Phi_0(t_k)\rangle$ is an eigenstate of $\mathcal{H}_0$ with eigenvalue $E_{\alpha,\beta} = E_0 + \varepsilon_\alpha - \varepsilon_\beta$. Therefore, within perturbation theory:

$$\frac{|\Phi_0(t_k + \delta t_k)\rangle - |\Phi_0(t_k)\rangle}{\delta t_k} = -\sum_{\alpha \neq \beta} \frac{\overline{V}^{(k)}_{\alpha,\beta}}{\varepsilon_\alpha - \varepsilon_\beta} \phi_\alpha^\dagger \phi_\beta |\Phi_0(t_k)\rangle + O(\delta t_k); \tag{6.96}$$

the local operator $\mathcal{O}_k(x)$ is obtained by taking the limit of $\delta t_k \to 0$:

$$\mathcal{O}_k(x) = -\sum_{\alpha \neq \beta} \frac{\overline{V}^{(k)}_{\alpha,\beta}}{\varepsilon_\alpha - \varepsilon_\beta} \frac{\langle x | \phi_\alpha^\dagger \phi_\beta |\Phi_0(t_k)\rangle}{\langle x | \Phi_0(t_k)\rangle}; \tag{6.97}$$

the last term can be computed by re-expressing the operators $\phi_\alpha^\dagger$ (and ϕ_β) in terms of $d_I^\dagger$ (and d_J):

$$\mathcal{O}_k(x) = -\sum_{I,J} W^{(k)}_{I,J} \frac{\langle x | d_I^\dagger d_J |\Phi_0(t_k)\rangle}{\langle x | \Phi_0(t_k)\rangle}, \tag{6.98}$$

where

$$W^{(k)}_{I,J} = \sum_{\alpha,\beta}{}' U_{I,\alpha} \frac{\overline{V}^{(k)}_{\alpha,\beta}}{\varepsilon_\alpha - \varepsilon_\beta} U^*_{J,\beta}; \tag{6.99}$$

here, the primed sum is carried over occupied orbitals ($\varepsilon_\beta \leq \varepsilon_F$) and unoccupied ones ($\varepsilon_\alpha > \varepsilon_F$).

The case with Pfaffians, see Eq. (5.106), can be obtained by a straightforward generalization of the previous steps. Here, we have to include also perturbations containing two creation or annihilation operators:

$$\mathcal{V}^{(k)} = \sum_{I,J} W^{(k)}_{I,J} d_I^\dagger d_J^\dagger + h.c., \tag{6.100}$$

which is associated to a small change in the pairing term (for a translationally invariant case, $\Delta^{\tau,\sigma}_{j,i} \to \Delta^{\tau,\sigma}_{k}$). Then, by inverting the Bogoliubov transformation of Eq. (5.107):

$$d_I = \sum_\alpha \left(u^*_{I,\alpha} \Phi_\alpha + v_{I,\alpha} \Phi_\alpha^\dagger \right), \tag{6.101}$$

the perturbation will possess both standard and anomalous terms, i.e., $\Phi_\alpha^\dagger \Phi_\beta$, which are similar to the ones in Eq. (6.93), but also $\Phi_\alpha^\dagger \Phi_\beta^\dagger$ and $\Phi_\alpha \Phi_\beta$. The rest of the derivation will follow the one shown for the determinant case.

Finally, we show how it is possible to compute the derivatives of the variational wave function with respect to the parameters η_k that define the backflow correlations. In our formulation, the "correlated" orbitals depend linearly upon these

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press

parameters, e.g., see the simplest case of Eq. (5.98). The logarithmic derivative can be computed by using the general formula for a matrix $\mathbf{M}$:

$$\frac{\partial}{\partial \eta_k} \ln \det \mathbf{M} = \frac{\partial}{\partial \eta_k} \mathrm{Tr} \ln \mathbf{M} = \mathrm{Tr}\left(\mathbf{M}^{-1} \frac{\partial \mathbf{M}}{\partial \eta_k} \right), \tag{6.102}$$

where $\partial \mathbf{M}/\partial \eta_k$ can be easily computed by differentiating every element of the matrix $\mathbf{M}$ with respect to the parameter η_k. Then, the local operator $\mathcal{O}_k(x)$ is obtained by taking the matrix $\tilde{\mathbf{U}}^b$ of Eq. (5.99), which defines the state with backflow correlations, as $\mathbf{M}$ in Eq. (6.102). Notice that the inverse matrix does not cost an extra computation time, since it is already calculated for the Metropolis algorithm.

https://doi.org/10.1017/9781316417041.007 Published online by Cambridge University Press