Variational Monte Carlo Notes for Boulder Summer School 2010

Bryan Clark

July 21, 2010

VMC

The Variational Monte Carlo method helps to accomplish two things:

- Given a wave function, Ψ_T , extract information out of it (i.e. energy, spin-spin correlation functions, etc.)
- Given a class of wave functions, $\Psi_T[\alpha]$ parameterized by α , find the α that produces the "best" wavefunction.

We will start with the first of these. Taking a wave function and extracting observables can be done by computing integrals of the form

$$\frac{\int |\Psi_T(c)|^2 \, O(c;\Psi_T)\, dc}{\int |\Psi_T(c)|^2 \, dc}$$

by sampling configurations c with probability proportional to $|\Psi_T(c)|^2$ and computing the average $\langle O(c)\rangle_{|\Psi_T|^2}$. For example, one might want to compute the spin-spin correlation functions.

Another, observable of interest might be the energy

$$\langle \Psi_T|H|\Psi_T\rangle = \frac{\int dc\, \Psi_T^*(c) H \Psi_T(c)}{\int dc\, \Psi_T^* \Psi_T} = \frac{\int dc\, \Psi_T^*(c)\Psi_T(c)\,\frac{H\Psi_T(c)}{\Psi_T(c)}}{\int dc\, \Psi_T^*(c)\Psi_T(c)}$$

We define the local energy to be

$$E_L \equiv \frac{H\Psi_T(c)}{\Psi_T(c)}$$

Why VMC

1. Although VMC is a biased method, there are situations where it is useful even in the face of unbiased techniques. For example, unbiased QMC

methods often can give correlation functions but not the nature of the wave-function. It is sometimes useful to (even approximately) know this latter information.

2. Fermions are hard I: In many cases, there aren't any unbiased (low scaling) methods that compute properties of fermions. This is largely because of the sign problem. VMC does not have a sign problem (unless you happen to be working in a non-orthogonal basis.) Although only an approximate method, it allows access to "answers" that are not accessible in other ways.

3. Fermions are hard II: There are other(even better) approximate methods (like fixed node diffusion Monte Carlo which we will talk about next lecture) that build on top of good trial wave-functions. Consequently, we can use the output of VMC as the input to these better methods.

MCMC

When using the variational monte carlo method, one sample a configuration c with probability $|\Psi_T(c)|^2$ using Markov chain Monte Carlo. We have already talked about the use of MCMC in previous talks so we will not spend much time with that here. It should be pointed out that to accomplish this, it is important to select some basis to work in. Then a snapshot of your Monte Carlo simulation is the configuration c.

MCMC algorithm

1. Choose arbitrary c

2. Choose configuration c' with probability $T(c \to c')$ (often taken to be a uniform "box" or gaussian around the current position c)

3. Accept configuration c' with metropolis acceptance probablity

$$min\left(1, \frac{T(c' \to c)}{T(c \to c'} \frac{|\Psi_T(c')|^2}{|\Psi_T(c)|^2}\right)$$

4. Compute observables $O(c; \Psi_T)$ ocassionally.

5. Loop back to (2).

There are not typically ergodic problems with VMC simulations so one does not usually have to be concerned with sophisticated Monte Carlo moves, etc.

Selecting the “best” wave-functions

There are a number of possible metrics that we might use to say that we have selected the “best” wave-function from a class of wave-functions. They include

- minimizing the energy. We know that the true ground state has the lowest energy minimized over all wave-functions Ψ_T (where we are restricted to wave-functions with the right symmetry).
- minimizing the variance of E_L. For the true ground state, the local energy $E_L = \frac{[H\Psi_0](c)}{\Psi_0(c)} = \frac{E\Psi_0(c)}{\Psi_0(c)} = E_0$ is independent of the given configuration c. This means that the variance of E_L,

$$\sigma(E_L) = \langle E_L^2 \rangle - \langle E_L \rangle^2$$

  is equal to 0.
- maximizing the overlap with the true ground state $|\langle \Psi_T | \Psi_0 \rangle|$.

Probably maximizing the overlap with the true ground state is the best thing to do, but no one knows how this can actually be done (although optimizing the overlap with the fixed node wave-function can be done (with some fair amount of difficulty) and might be a reasonable proxy for this). The general wisdom, then, is that the best metric is to select the wave funtion that has the lowest energy E. In practice whether this works well depends partially on the sort of properties in which you are interested. For example, one can get very good energies and still get the far off diagonal terms of your 1-body density matrix incorrect as the energy is often not very sensitive to this.

Optimization of Wave Functions

Optimization is hard and still somewhat akin to black magic. In fact, even finding local minima is difficult (to say nothing of global minima) Even though a lot of recent work has gone into procedures for better optimizing wave functions, it is still an active and important research area.

A method to optimize poorly

In this section we start by describing a bad, but naively appealing way to optimize our wave function. We will then see how to improve this method so that it works.

Local optimization of deterministic functions is something that has been worked on for a long time. VMC, though, is a stochastic process and returns the energy $E \pm \delta$ with some error bar δ. As a start, we would like to write down an approximation for the energy that is a deterministic function. Then, we could naively use a typical black box optimization method.

To accomplish this, imagine we have 1000 configuration $\{c_1...c_{1000}\}$ that are all sampled from the distribution $|\Psi_T[\alpha_0]|^2$. We know that an approximation to the energy $E[\alpha_0]$ is

$$\langle E(\alpha_0)\rangle \approx \sum_{i=1}^{1000} E_L(c_{i;}\alpha_0)$$

It is also possible, though, that from these *same* 1000 configurations, we can also estimate the energy of $E(\alpha)$ where α is a different set of parameters then α_0. Notice, that the configurations we want to use have not been sampled from $\Psi_T(\alpha)$. Nonetheless, we can reweight them so they look as if they've been sampled from this other distribution. To accomplish this we will use the function

$$\langle E(\alpha)\rangle \approx \frac{\sum_{i=1}^{1000} E_L(c_{i;}\alpha)\frac{\Psi_T(c_i,\alpha)}{\Psi_T(c_i,\alpha_0)}}{\sum_{i=1}^{1000} \frac{\Psi_T(c_i,\alpha)}{\Psi_T(c_i,\alpha_0)}}$$

as our estimate for the energy of our system with parameters α.

Notice that $\langle E(\alpha)\rangle$ is a deterministic objective function. Optimizing deterministic objective functions is a long and well studied subject. At this point, one might be tempted to stick this objective function into a black-box optimizer and get our results.

There are two problems with this approach though:

1. If the wave functions $\Psi_T(\alpha)$ and $\Psi_T(\alpha_0)$ differ significantly then there will be a very small number of effective points and our average energy will be very unreliable. This is not a major problem. When the number of effective points becomes small, you can always refresh the configurations ocassionally. Also, it is empirically the case that if you just optimize without reweighting the configurations (i.e. assume $\Psi_T(c_i;\alpha)/\Psi_T(c_i,\alpha_0) \approx 1$) this works and is often more stable (although a couple refreshes will still be necessary).

2. Imagine the limit where we only have one configuration. Almost certainly we will find parameters α that are good for this configuration but bad in general. This problem persists even in the limit of 1000's of configurations. One of the fundamental reasons for this problem is that the energy is not bounded from below. Consequently, their are parameter regions that can (for a finite set of configurations) severly underestimate the energy.

The (now somewhat deprecated) approach to dealing with this problem is that instead of optimizing the energy of your system, optimize the variance of E_L. Because this is bounded below by 0, it is less sensitive to undersampling problems. The main downside to this is that minimizing the energy is likely a better proxy for the "best" wave-function then minimizing the variance. (Often you can also get away with minimizing some linear combination of the variance and the energy.)

Suppose we really want to optimize the energy though. Is their anything we can do?

Generic Approach I: Optimize the infinite sample.

Fundamentally our problem was that we were minimizing with respect to a finite sample and not an infinite sample. So, if we had a finite number of samples and walked toward the parameters that minimize the finite sample energy we get something that is too low. Instead let's figure out how we can bias it to walk toward the true ground state.

The prior problem was mainly one of undersampling. Another way to say this is that if you followed the gradient of the finite set of configurations, you were travelling in the wrong direction. There are new sophisticated ways of solving this problem. Instead of describing them all in complete detail, here I would like to describe some key tricks that help resolve this. This first trick is important for optimization, but it's really much more general. It's a really a lesson on variance reduction and how it's important to think hard about it when doing Monte Carlo. You should think of variance reduction in the same bag of tricks as optimization ensemble and loop/cluster moves.

Trick I: Variance reduction or how to add 0 and get much better answers.

So we want to walk in the direction of the gradient.

Let's take the derivative of

$$\begin{aligned}\partial_\alpha \langle E\rangle &= \partial_\alpha\left[\frac{\int |\Psi|^2 \frac{H\Psi}{\Psi}}{\int |\Psi|^2}\right]\\ &= \frac{\int \left(2\frac{\Psi_\alpha}{\Psi}\frac{H\Psi}{\Psi} + \frac{H\Psi_\alpha}{\Psi} - \frac{H\Psi}{\Psi}\frac{\Psi_\alpha}{\Psi}\right)|\Psi|^2}{\int |\Psi|^2} - \frac{\int \frac{H\Psi}{\Psi}|\Psi|^2}{\int |\Psi|^2}\frac{\int 2\frac{\Psi_\alpha}{\Psi}|\Psi|^2}{\int |\Psi|^2}\\ &= \int |\Psi|^2\left(\frac{\Psi_\alpha}{\Psi}E_L + \frac{H\Psi_\alpha}{\Psi} - 2\langle E\rangle\frac{\Psi_\alpha}{\Psi}\right) / \int |\Psi|^2\end{aligned}$$

where we define $\Psi_\alpha \equiv \frac{\partial}{\partial\alpha}\Psi$.

Let us now look more closely at the term

$$\int \frac{\Psi^*\Psi\left[H\Psi_\alpha\right]}{\Psi} = \int \frac{\Psi^*\left[\Psi H\right]\Psi_\alpha}{\Psi} = \int \frac{\Psi_\alpha}{\Psi}\frac{H\Psi}{\Psi}|\Psi|^2$$

where the first equality is allowed by the hermiticity of H. In other words, we have learned that

$$0 = \int dc\,|\Psi(c)|^2\left[\frac{\left[H\Psi_\alpha\right](c)}{\Psi(c)} - E_L(c)\frac{\Psi_\alpha(c)}{\Psi(c)}\right]$$

but it is not true for all c that

$$|\Psi(c)|^2\left[\frac{\left[H\Psi_\alpha\right](c)}{\Psi(c)} - E_L(c)\frac{\Psi_\alpha(c)}{\Psi(c)}\right]$$

is equal to 0. We can then add 0 to $\partial_\alpha \langle E \rangle$ giving us

$$\partial_\alpha \langle E \rangle = \int |\Psi|^2 \, \frac{\Psi_\alpha}{\Psi} \left(E_L - \langle E \rangle \right) / \int |\Psi|^2$$

If you use this bottom equation as a gradient to minimize with respect to, you will get much better answers then the earlier equation. At first glance, this seems like a bit of a paradox. All we have done is remove a term that is equal to 0. What is going on, though, is the following. These two equations are the same in the limit of an infinite number of configuration. They are not the same on a finite sample. In fact, the bottom equation will not be the true gradient of the finite sample (so if you're checking your code with finite differences, it will naively look wrong). Consequently, the set of parameters where the bottom gradient is 0 is not going to be the minimum of the finite sample. By using a property of the infinite sample (i.e. the hermiticity of H) we have a situation where we are better optimizing for the infinite system even on a finite sample.

To further understand what is going on, let's look at the limit where we have the exact ground state $\Psi_T = \Psi_0$. We know that in the exact ground state the gradient should be 0. The first equation, though, has configuration by configuration fluctuations. If we are summing an infinite number of configurations, it will add up to 0 but for any finite sample it will push you away even from the true ground state. Since $E_L = \langle E \rangle$, when you have the true ground state wavefunction, the second equation is 0 configuration by configuration. Even summing over a single configuration would give a useful gradient.

Trick II: Linearize

Another useful trick is the following. These implicitly uses the idea of variance reduction from above (although not in as clear and obvious a way as enumerated above) Given a wave function $\Psi[\alpha]$ you can taylor exapnd this Hamiltonian and write

$$\Psi[\alpha] = \Psi[\alpha_0] + \sum_i (\alpha - \alpha_i) \frac{\partial \Psi}{\partial \alpha_i}$$

where α_0 is some initial set of parameters. Then, write

$$H_{ij} = \left\langle \frac{\Psi_i}{\Psi_0} \frac{H\Psi_j}{\Psi_0} \right\rangle_{|\Psi_0|^2}$$

and

$$S_{ij} = \left\langle \frac{\Psi_i}{\Psi_0} \frac{\Psi_j}{\Psi_0} \right\rangle_{|\Psi_0|^2}$$

Then if we solve the generalized eigenvalue equation

$$H \Delta\alpha = E S \Delta\alpha$$

we retreive a set of changed parameters $\Delta\alpha$. In principle, one could then update the set of parameters as $\alpha \leftarrow \alpha + \Delta\alpha$. Although a reasonable approach, in the case where your parameter weren't linear to begin with, this is not the best way to update the parameters. (see prl 98, 110201 for a better approach).

Generic Approach II: Optimize Stochastically

There is another (reasonably new) approach for this optimization problem. This is essentially to stochastically follow the "gradient" by "refreshing" the configurations you are using to calculate the gradients frequently. One might imagine that this avoids the problem of undersampling because as you're heading to energies that are too low, you switch to new configurations which have different energies.

More concretely, one imagines that the following could happen:

- Accumulate a few configurations from the markov chain, compute the gradients
$$g_{approx} = \sum_c \left[\nabla_\alpha \langle E \rangle\right](c)$$
and walk some small distance in the direction of the gradient τg_{approx} .

If this gradient involved only E_L and not Ψ_T, then one can show, from the linearity of expectations, that the expected gradient equals the true gradient. In this (not satisfied) limit this optimization method would be formally correct (to find local minima) and one can argue that in the long term limit, this would sample parameters whose average should be the true parameters. The problem with this idea are that the

- the gradient does involve Ψ_T and
- often these stochastic gradients can have extremely large values. This then throws you into crazy parameter regimes that are hard to return from.

To resolve these difficulties, for each parameter, one walks a fixed distance $r\delta sign[\partial_\alpha \langle E \rangle]$ where δ is some small constant and r is a parameter that starts at 1 and can be ramped down as the optimization progresses. The fact that this seems to work well in practice is somewhat mysterious. The advantages of this method is that it seems reasonably robust and is fairly simple to implement. We should note that this primarily works because as you move around in parameter space, you don't change the markov chain much.

A comment on global minima

These approaches have been mainly concerned with finding a good local minima. (In practice, these methods often do a bit better then this because they are stochastic and tend not to get stuck in shallow local minima). Often this is good enough. Many systems seem not to have too many local minima and we can typically start somehere close to the right answer (or try a number of initial starting points).

In cases, where this is not good enough, there are other potential approaches. For example, you can do some form of stochastic simulated annealing.

Wave Function Zoo

Ideally we want to write down a wave-function that is a good representation of the ground state. Properties of useful wavefunctions are those that

- can be represented compactly compactly (with a polynomial number of parameters)
- are easy to manipulate. (i..e. averages like the energy of the wave function can be computed quickly). Usually you want to be able to calculate these averages exactly although the recent interest in PEPS is somehow a set of interesting wavefunctions that contrasts with this.

Projected Gutzweiller

One reasonable place to start for writing down good wave-functions is with reasonably simple limits. For example, let's take a non-interacting system (potentially living in some external potential) We can think of this as coming from a free fermion system, some tight binding model, Hartree-Fock or density functional theory. We know that the ground state for such a wave function can the be represented as

$$\Psi_T = \det M$$

where

$$M_{ij} = \Phi_i(r_j)$$

where Φ_i is a d-dimensional function that takes a position/configuration of a single electron and returns a value. The parameters α of this wave function is selecting the values of these n d-dimensional functions. One should note that their is a redundancy here. Because the determinant is invariant with respect to any unitary transformation of M, there are many different choices for the same orbitals that give the same wavefunctions.

Now, we can ask how we can intorduce some correlations into these wave functions. Notice that so far the spin up and spin down electrons don't even notice the existence of each other. One approach is, instead of writing our wave function as a single determinant, we can instead write it is a sum of many determinants. For example

$$\Psi_T = \sum_k \det\left[M_k\right]$$

where for each M_k we select n orbitals ϕ_i (different matrices can have the same orbitals although obviously it isn't useful to have any two matrices have exactly the same set of orbitals). We know that we can write a complete basis as a sum of slater determinants. Therefore, in the limit of a large number of slater determinants this is always a good representation. This is essentially what quantum chemists do when doing a full CI calculation (and for small molecules it works reasonably well). The basic problem with this approach is that it is not size extensive. The number of determinants required will generically scale exponentially with the size of your system making it not particularly effective for condensed matter systems. Instead, we would like to add correlation in a more size-extensive way. To accomplish this, let's consider for a moment the opposite limit of free fermions: a hubbard model with an extremely large U. Because of the large energy cost to having double occupancy, we want to include in our wavefunction some term that forbids them. One approach to this is to add a projection that explicitly projects out double occupancy. This is typically called gutzweiller projection which we will write as

$$\Psi_T = P_G \det M$$

where P_G is the projection that forbids double occupancy. We should note that this constraint is easy to implement in real space in a variational Monte Carlo simulation (and very hard to implement when our representation is not in real space). We can simply reject any Monte Carlo moves where two particles end up on the same point.

Slater-Jastrow

Of course, at intermediate U/t it is probably a bit extreme to disallow all double occupancy. Instead, let us soften this projection by replacing it with a Jastrow Factor. We can write a jastrow factor as

$$J = \exp\left[-\sum_i u_1(r_i) - \sum_{ij} u_2(r_i, r_j) - \sum_{ijk} u_3(r_i, r_j, r_k) + ...\right]$$

where u is a real function. Often we will only work with 1 and 2 body terms and let the two body term u_2 be a function only of the magnitude $|r_i - r_j|$. The parameters in these systems include those of projected gutzweiller and the functions u_1, u_2, u_3

A few comments about Jastrow factors:

- We can see that if we let $u_2(0) \to \infty$ this reduces back to the projected gutzweiller function.
- In a translationally invariant function, it might seem a bit strange to include a one body function. One way to think about its important is the following: Often the slater matrix ansatz will have a pretty good distribution for the density (especially if you happen to be pulling it from Density Functional Theory). By putting in the 2-body terms you've added some correlation into the system (important) but you've screwed up the density. The one body term u_1 will fix the density back up.
- In the continuum (i.e. when our Hamiltonian is $\nabla^2 + 1/r$) we can actually put certain constraints on u because we can solve the two body problem as the particles come close to each other. This is called the cusp condition.

- Imagine we are working with a bosonic system. Then this expansion is a comlete basis for the ground state of bosonic wave functions. Jastrow's were first used by Mcmillan for calculating ground state energies of Helium-4. We can see what we've really done for our fermionic wave function is write it as
  $$\Psi = \Psi_{fermion}\Psi_{boson}$$
  a product of a fermionic and bosonic wave function
- If we were to generalize our jastrow wave functions to allow complex factors we can get quantum hall states, etc. We will talk about this a bit more when we talk about Huse-Elser/CPS states.

RVB

We would like to do better then slater-jastrow. One approach is to use a projected BCS wave function. A projected BCS wave function applies a gutzweiler projection to the mean field BCS hamiltonian. We can write this as

$$P_G \left|BCS\right\rangle \quad = \quad P_G \prod_k (u_k + v_k c^\dagger_{k\uparrow} c^\dagger_{-k\downarrow})$$

Converting this to real space we get

$$\Psi = P_G \left(\sum_{r<r'} f_{r-r'} \left(c^\dagger_{r\uparrow} c^\dagger_{r'\downarrow} + c^\dagger_{r'\uparrow} c^\dagger_{r\downarrow} \right) \right)^{N/2} \left|0\right\rangle$$

Notice that this can be represented as a determinant where each row is an up spin, each column is a down spin and the matrix element is $f_{r-r'}$. These $f_{r-r'}$ are the parameters α in these systems. We can generalize this state, by multiplying it by a Jastrow instead of projecting.

Backflow

Earlier we discussed how we could take elements of our slater matrix to be $\phi_i(r_j)$. One can do better by instead of working with r_j working with generalized coordinates $r_j = r_j + \zeta_j(R)$ (i.e. using quasiparticles). As long as $\zeta_j(R)$ is symmetric with respect to R this produces a legitimate antisymmetric function. One adds to the parameters α the function $\zeta_j(R)$.

MultiSlater Jastrow

We've already discussed the idea of doing better then slater-determinants by just using a sum of them. Even better is to take a sum of slater determinants multiplied by a Jastrow. Even though this shares many of the problems that we discussed above, it still can help improve the wave function.

Huse Elser/Correlator Product States

For spin systems (and some fermion systems), one can often do very well with Huse Elser/CSP.

We can write a general wave function as

$$\Psi = \sum_{\{n\}} \Psi_{n_1 n_2 n_3 ... n_N} |n_1, n_2, n_3 ... n_n\rangle$$

This obviously has a lot of states. Instead we will decompose our state

$$\Psi_{n_1 n_2 .. n_N} = \prod_i C(n_i) \prod_{ij} C(n_i, n_j) \prod_{ijk} C(n_i, n_j, n_k) ...$$

into a product over correlation functions. We allow the C to be general complex values.

So for example our wave function might be

$$\Psi = \sum_{\{n\}} C(n1, n2) C(n3, n4, n5) |n_1, n_2, n_3, n_4, n_5\rangle$$

Notice, that if you do large enough products (instead of just 2-body terms and 3-body terms, do up to n-body terms), then you have a complete basis. If we think the physics is somehow local, it makes sense to choose correlators that are local. Let us count the total number of parameters. For two body terms we have $4n^2$ points. For three body terms it is $2n^3$ terms. Therefore, for a small “bond” dimension, the number of parameters is reasonably compact.

Others:

Valence Bond Basis, Geminals, MPS, PEPS, etc.

Fast Updates

Let us take the specific example of doing Variational Monte Carlo using a projected gutzweiller wave-function. To perform Variational Monte Carlo one needs to evaluate the ratio of the wave functions between configuration c and c' where they differ by the location of the particle at position k . This operation needs to be done over and over again (for different postions k', etc). One can write this as

$$\left| \frac{\Psi[c']}{\Psi[c]} \right|^2 = \left[\frac{\det[M + e_k^T u]}{\det M} \right]^2$$

where e_k^T is the unit vector with a 1 at location k and 0 at all other positions and $u = \phi_k(s_i) - \phi_k(s_j)$. Evalauting determinant ratios of this type is a common operation that is required in a variety of different quantum Monte Carlo simulations (CTQMC, Hirsh-Fye, DQMC, VMC, DMC, etc.). The naive approach involves simply taking the determinant of the numerator and denominator separately. This is not only slow (an $O(n^3)$ algorithm), but also is not numerically stable in the limit of large matrices. A superior approach is to use

$$\frac{\det[M + e_k u^T]}{\det M} = 1 + u^T M^{-1} e_k$$

to evaluate this ratio. In order to accomplish this, we need to have access to the value M^{-1}. To avoid the calculation of the value M^{-1} at each step we can calculate it once and then update it. This update can be done by the Shermann Morrison formula,

$$(M + e_k u^T)^{-1} = M^{-1} - \frac{M^{-1} e_k u^T M^{-1}}{1 + u^T M^{-1} e_k}$$