Improved Sampling for Biological Molecules Using Shadow Hybrid Monte Carlo

Scott S. Hampton and Jesús A. Izaguirre

University of Notre Dame, Notre Dame IN 46556, USA

Abstract. Shadow Hybrid Monte Carlo (SHMC) is a new method for sampling the phase space of large biological molecules. It improves sampling by allowing larger time steps and system sizes in the molecular dynamics (MD) step of Hybrid Monte Carlo (HMC). This is achieved by sampling from high order approximations to the modified Hamiltonian, which is exactly integrated by a symplectic MD integrator. SHMC requires extra storage, modest computational overhead, and a reweighting step to obtain averages from the canonical ensemble. Numerical experiments are performed on biological molecules, ranging from a small peptide with 66 atoms to a large solvated protein with 14281 atoms. Experimentally, SHMC achieves an order magnitude speedup in sampling efficiency for medium sized proteins.

1 Introduction

The sampling of the configuration space of complex biological molecules is an important and formidable problem. One major difficulty is the high dimensionality of this space, roughly $3N$, with the number of atoms N typically in the thousands. Other difficulties include the presence of multiple time and length scales, and the rugged energy hyper-surfaces that make trapping in local minima common, cf. [1]. This paper introduces Shadow Hybrid Monte Carlo (SHMC), a propagator through phase space that enhances the scaling of hybrid Monte Carlo (HMC) with space dimensionality.

The problem of sampling can be thought of as estimating expectation values for a function $A(\Gamma)$ with respect to a probability distribution function (p.d.f.) $\rho(\Gamma)$, where $\Gamma = [\mathbf{x}^{\mathrm{T}}, \mathbf{p}^{\mathrm{T}}]^{\mathrm{T}}$, and $\mathbf{x}^{\mathrm{T}}$ and $\mathbf{p}^{\mathrm{T}}$ are the vectors of collective positions and momenta. For the case of continuous components of Γ,

$$\langle A(\Gamma) \rangle_{\rho} = \int A(\Gamma)\rho(\Gamma)\mathrm{d}\Gamma \,. \tag{1}$$

Examples of observables A are potential energy, pressure, free energy, and distribution of solvent molecules in vacancies [2,3].

Sampling of configuration space can be done with Markov chain Monte Carlo methods (MC) or using molecular dynamics (MD). MC methods are rigorous sampling techniques. However, their application for sampling large biological molecules is limited because of the difficulty of specifying good moves for dense

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3037, pp. 268–275, 2004.

systems [4] and the large cost of computing the long range electrostatic energy, cf. [3, p. 261]. MD, on the other hand, can be readily applied as long as one has a "force field" description of all the atoms and interactions among atoms in a molecule. Additionally, MD enables relatively large steps in phase space as well as global updates of all the positions and momenta in the system. MD finds changes over time in conformations of a molecule, where a conformation is defined to be a semi-stable geometric configuration. Nevertheless, the numerical implementation of MD introduces a bias due to finite step size in the numerical integrator of the equations of motion.

MD typically solves Newton's equations of motion, a Hamiltonian system of equations,

$$\dot{\Gamma}(t) = J\mathcal{H}_\Gamma(\Gamma(t)), \qquad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}, \tag{2}$$

with a Hamiltonian

$$\mathcal{H}(\mathbf{x}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathrm{T}}\mathrm{M}^{-1}\mathbf{p} + U(\mathbf{x}), \tag{3}$$

where M is a diagonal matrix of masses, $U(\mathbf{x})$ is the potential energy of the system, and $\mathbf{p} = M\dot{\mathbf{x}}$ are the momenta. Eq. (2) can also be written as

$$\dot{\mathbf{x}}(t) = \mathrm{M}^{-1}\mathbf{p}(t), \qquad \dot{\mathbf{p}}(t) = F(\mathbf{x}(t)), \tag{4}$$

where the conservative forces $F(\mathbf{x}(t)) = -\nabla U(\mathbf{x}(t))$.

Numerical integrators for MD generate a solution $\Gamma^n \approx \Gamma(n\delta t)$, where the step size or time step used in the discretization is δt. Typical integrators can be expressed as

$$\Gamma^{n+1} = \Psi(\Gamma^n), \tag{5}$$

where Ψ represents a propagator through phase space. Any time reversible and volume preserving integrator can be used for HMC. SHMC requires in addition that the integrator be symplectic (cf. [5, p. 69]). An integrator is symplectic if $\partial_\Gamma \Psi(\Gamma)^{\mathrm{T}} J \partial_\Gamma \Psi(\Gamma) \equiv J$. In this work, both implementations use the Verlet/Leapfrog discretization [6], which satisfies the constraints for both propagators.

HMC, introduced in [7], uses MD to generate a global MC move and then uses the Metropolis criterion to accept or reject the move. HMC rigorously samples the canonical distribution and eliminates the bias of MD due to finite step size. Unfortunately, the acceptance rate of HMC decreases exponentially with increasing system size N or time step δt. This is due to discretization errors introduced by the numerical integrator and cause an extremely high rejection rate. The cost of HMC as a function of system size N and time step δt has been investigated in [8,9]. These errors can be reduced by using higher order integrators for the MD step as in [10]. However, higher order integrators are not an efficient alternative for MD for two reasons. First, the evaluation of the force is very expensive, and these integrators typically require more than one force evaluation per step. Second, the higher accuracy in the trajectories is not needed in MD, where statistical errors and errors in the force evaluation are very large.

2 Shadow HMC

SHMC is a biased variation on HMC. It uses a smooth approximation to the modified Hamiltonian to sample more efficiently through phase space. The modified Hamiltonian is exactly conserved by the numerical integrator and a cheap, arbitrarily accurate, approximation called a shadow Hamiltonian has been proposed in [11]. SHMC samples a non-canonical distribution defined by high order approximations to the modified Hamiltonian, which greatly increases the acceptance rate of the method. A reweighting of the observables is performed in order to obtain proper canonical averages, thus eliminating the bias introduced by the shadow Hamiltonian. The overhead introduced by the method is modest in terms of time, involving only dot products of the history of positions and momenta generated by the integrator. There is moderate extra storage to keep this history. In this generalization of HMC, sampling is in all of phase space rather than configuration space alone.

Let $\tilde{\rho}(\mathbf{x}, \mathbf{p})$ be the target density of SHMC, where

$$\tilde{\rho}(\mathbf{x}, \mathbf{p}) \propto \exp\left(-\beta \tilde{\mathcal{H}}(\mathbf{x}, \mathbf{p})\right), \tag{6}$$

$$\tilde{\mathcal{H}}(\mathbf{x}, \mathbf{p}) = \max\left\{\mathcal{H}(\mathbf{x}, \mathbf{p}), \mathcal{H}_{[2k]}(\mathbf{x}, \mathbf{p}) - c\right\}. \tag{7}$$

Here, $\mathcal{H}_{[2k]}(\mathbf{x}, \mathbf{p})$ is the much smoother shadow Hamiltonian, defined in Section 3, and c is an arbitrary constant that limits the amount by which $\mathcal{H}_{[2k]}$ is allowed to depart from $\mathcal{H}(\mathbf{x}, \mathbf{p})$.

Algorithm 1 lists the steps for calculating SHMC. The first step is to generate a set of momenta, $\mathbf{p}'$, usually chosen proportional to a Gaussian distribution. $\mathbf{p}'$ is accepted based on a Metropolis criterion step proportional to the difference of the total and shadow energies. This step is repeated until a set of momenta are accepted. Next, the system is integrated using MD and accepted with probability proportional to Eq. (6). Finally, in order to calculate unbiased values, the observables are reweighted.

The purpose of the constant c is to minimize the difference in the energies so that the reweighted observables of $\mathcal{H}_{[2k]}$ are unbiased. Let $\Delta\mathcal{H} = \mathcal{H}_{[2k]} - \mathcal{H}$. Experiments suggest that $\Delta\mathcal{H}$ is predominantly positive in MD simulations. This is most likely due to the fact that the shadow Hamiltonian is designed to exactly conserve energy of the numerical solution of quadratic Hamiltonians such as those used in MD[11]. Currently, c is chosen proportional to the expected value of the discretization error, $\langle\Delta\mathcal{H}\rangle$. This value is obtained after running a sufficient number of steps and monitoring $\Delta\mathcal{H}$ at each step.

3 Shadow Hamiltonian

The modified equations of a system of differential equations are exactly satisfied by the approximate discrete solution of the numerical integrator used to solve them. These equations are usually defined as an asymptotic expansion in powers

Algorithm 1 Shadow Hybrid Monte Carlo (SHMC)

1. **MC Step:** Given $\mathbf{x}$, generate $\mathbf{p}'$ with p.d.f. $\tilde{\rho}(\mathbf{x}, \mathbf{p})$, using the acceptance-rejection method:
   a) Generate $\mathbf{p}'$ having p.d.f. $\rho_{\mathrm{p}}(\mathbf{p})$
   b) Accept with probability
   $$\min\left\{1, \frac{\exp(-\beta(\mathcal{H}_{[2k]}(\mathbf{x}, \mathbf{p}') - c))}{\exp(-\beta\mathcal{H}(\mathbf{x}, \mathbf{p}'))}\right\}$$
   c) Repeat (1a) - (1b) until P is accepted.
2. **MD Step:** Given Γ :
   a) $\Gamma' = R\Psi(\Gamma)$ (where Ψ nearly conserves $\mathcal{H}_{[2k]}$)
   b) Accept Γ' with probability
   $$\min\left\{1, \frac{\tilde{\rho}(\Gamma')}{\tilde{\rho}(\Gamma)}\right\}$$
   c) If rejected, choose Γ.
3. **Reweighting Step:** Given $\{A, \Gamma\}$, reweight observable A using $\rho(\Gamma)/\tilde{\rho}(\Gamma)$ before computing averages.

of the discretization time step. If the expansion is truncated, there is excellent agreement between the modified equations and the discrete solution [12].

In the case of a Hamiltonian system, Eq. (2), symplectic integrators conserve exactly (within roundoff errors) a modified Hamiltonian $\mathcal{H}^{\delta \mathrm{t}}$. For short MD simulations (such as in HMC) $\mathcal{H}^{\delta \mathrm{t}}$ stays close to the true Hamiltonian, cf. [5, p. 129–136]. Work by Skeel and Hardy [11] shows how to compute an arbitrarily accurate approximation to the modified Hamiltonian integrated by symplectic integrators based on splitting. The idea is to compute

$$\mathcal{H}_{[2k]}(x, p) = \mathcal{H}^{\delta \mathrm{t}}(x, p) + O(\delta t^{2k}) \,. \tag{8}$$

$\mathcal{H}_{[2k]}$ is the shadow Hamiltonian of order $2k$. It follows from centered finite difference approximations to derivative terms in the expansion of $\mathcal{H}^{\delta \mathrm{t}}$, and from interpolation to the evaluation points. It is a combination of trajectory information, that is, k copies of available positions and momenta generated by the MD integration, and an extra degree of freedom β that is propagated along with the momenta. By construction, $\mathcal{H}_{[2k]}$ is exact for quadratic Hamiltonians, which are very common in MD. Details can be found in the original reference.

A shadow Hamiltonian of order $2k$, k even, is constructed as a linear combination of centered differences of the position and momenta of the system. The formulae for the 4^{th} and 8^{th} order shadows, $k = 2$ and $k = 4$ respectively, follow:

$$\mathcal{H}_{[4]} = \frac{1}{2\delta t}\left(\mathrm{A}_{10} - \frac{1}{6}\mathrm{A}_{12}\right) , \tag{9}$$

$$\mathcal{H}_{[8]} = \frac{1}{2\delta t}\left(210\mathrm{A}_{10} - \frac{2}{7}\mathrm{A}_{12} - \frac{19}{210}\mathrm{A}_{14} + \frac{5}{42}\mathrm{A}_{30} + \frac{13}{105}\mathrm{A}_{32} - 315\mathrm{A}_{34}\right). \quad (10)$$

Define the i^{th} centered difference formula to be $\delta\omega^{[i]}$. So, for example, $\delta\mathbf{x}^{[2]}$ would represent the 2^{nd} centered difference of the positions:

$$\delta\mathbf{x}^{[2]} = \mathbf{x}^{n+1} - 2\mathbf{x}^{n} + \mathbf{x}^{n-1}$$

Now define A_{ij}:

$$\mathrm{A}_{ij} = \begin{cases} \delta\mathbf{x}^{[i]} \cdot \delta\mathbf{p}^{[j]}\mathrm{M} - \delta\mathbf{x}^{[j]} \cdot \delta\mathbf{p}^{[i]}\mathrm{M} - \delta\beta^{[i]} & : \quad j = 0 \\ \delta\mathbf{x}^{[i]} \cdot \delta\mathbf{p}^{[j]}\mathrm{M} - \delta\mathbf{x}^{[j]} \cdot \delta\mathbf{p}^{[i]}\mathrm{M} & : \quad j \neq 0 \end{cases} \quad (11)$$

Finally, the β term propagated by Leapfrog is:

$$\beta = -\delta t(\mathbf{x}^{n} \cdot F^{n} + 2U(\mathbf{x}^{n})), \quad (12)$$

where the forces $\mathbf{F}$, the positions $\mathbf{x}$, and the momenta $\mathbf{p}$, are vectors of length $3N$, and N is the number of atoms in the system. M is a diagonal matrix containing the mass of each atom.

4 Numerical Tests

SHMC was tested with a 66 atom Decalanine, and a more complex solvated protein, BPTI, with 14281 atoms. The methods and example systems are available by obtaining ProtoMol [13] from our website[1]. Simulations were run on a Linux cluster administered by the Department of Computer Science and Engineering at the University of Notre Dame. Each node contains 2, 2.4 GHz Xeon processors and 1 GB RDRAM.

The performance of HMC and SHMC is dependent upon the input parameters of time step δt and trajectory length L. Here, L is amount of simulated time for one MC step. L should be long enough so that the longest correlation times of interest are sampled during an MD step, thus avoiding the random walk behavior of MC. SHMC also needs a tuning parameter c to indicate allowed divergence between the shadow and total energy.

Several techniques have been used to compare SHMC and HMC. The efficiency of sampling is measured by computing the cost to generate a new geometric conformation. The statistical error is measured by computing the potential energy and its standard deviation.

Statistical Correctness. In order to test the statistical correctness of the reweighted values of SHMC, the potential energies (PE) and their standard deviations were computed. Table 1 shows the average potential energy (PE) for Decalanine. Looking through the values, there is little difference statistically speaking. All of the reweighted values are within at least one standard deviation of the unweighted HMC values. Additionally, the reweighted standard deviation is acceptable in all cases.

[1] http://protomol.sourceforge.net

Table 1. Average potential energy (kcal/mol) and standard deviation for Decalanine for HMC and SHMC using an 8^{th} order shadow Hamiltonian.

	Time step (fs)					
Method	0.5	0.75	1.0	1.25	1.5	2
HMC	97.5 ± 6.5	97.4 ± 6.9	100 ± 6.6	99.8 ± 6.7	98.1 ± 7.1	97.4 ± 9.1
SHMC	103 ± 6.7	102 ± 7	96.8 ± 7.2	98.9 ± 6.8	97.3 ± 8	99.7 ± 8.4
c	0.4	0.4	0.6	1.2	1.2	2.8

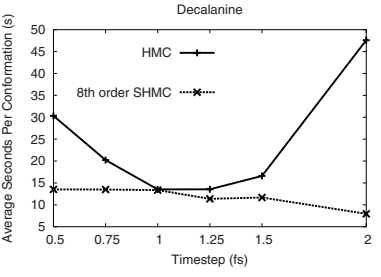

Fig. 1. Average computer time per discovered conformation for 66-atom Decalanine.

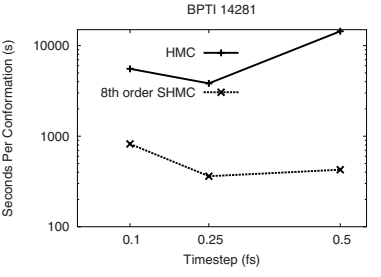

Fig. 2. Average computer time per discovered conformation for 14281-atom BPTI.

Sampling Efficiency. The number of molecular conformations visited by HMC and SHMC is determined using a method suggested in [14]. The sampling efficiency of HMC and SHMC is defined as the computational cost per new conformation. This value is calculated by dividing the running time of the simulation by the number of conformations discovered. This is a fair metric when comparing different sampling methods, since it takes care of the overhead of more sophisticated trial moves, and any other effects on the quality (or lack thereof, e.g., correlation) of samples produced by different sampling techniques.

Figure 1 shows the number of conformations per second as a function of the time step for Decalanine. At its best, HMC is only as good as SHMC for one time step, $\delta t = 1$. In terms of efficiency, SHMC shows a greater than two-fold speedup over HMC when the optimal values for both methods are used. Figure 2 shows even more dramatic results for BPTI with 14281 atoms. The speedup in this case is a factor of 10. This is expected, since the speedup increases asymptotically as $\mathcal{O}(N^{1/4})$[15].

The following graphs demonstrate how c affects simulations. Figure 3 shows a plot of the standard deviation of the potential energy as a function of the value chosen for c. The system is Decalanine, with a time step of 2 fs. Figure 4 shows that the probability of accepting the MD move also decreases as c increases. In the first case, a large c is desirable and in the second case a small c is best.

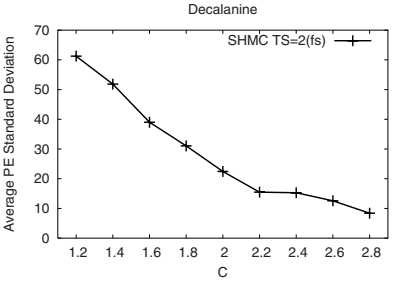

Fig. 3. The effect of c on the standard deviation of the potential energy.

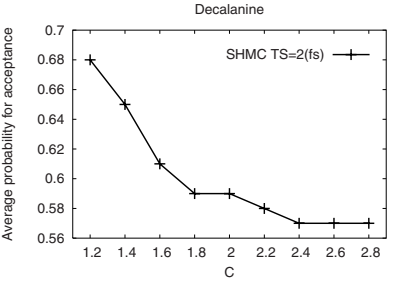

Fig. 4. The effect of c on the probability of accepting the MD step.

5 Discussion

SHMC is a rigorous sampling method [15] that samples a p.d.f. induced by a modified Hamiltonian. Because this modified Hamiltonian is more accurate than the true Hamiltonian, it is possible to increase the efficiency of sampling. Since the modified Hamiltonian is by construction close to the true Hamiltonian, the reweighting does not damage the variance. The additional parameter, c, of SHMC, measures the amount by which the modified and the true Hamiltonian can depart. Different regions of phase space may need different optimal parameters. Here, c is chosen to satisfy both bounds on the statistical error of sampling and an acceptable performance. A rule of thumb is that it should be close to the difference between the true and the modified Hamiltonian. Other criteria are possible, and it would be desirable to provide "optimal" choices.

The efficiency of Monte Carlo methods can be improved using other variance reduction techniques. For example, [16] improves the acceptance rate of HMC by using "reject" and "accept" windows. It accepts whether to move to the accept window or to remain in the reject window based on the ratio of the sum of the probabilities of the states in the accept and the reject windows. SHMC is akin to importance sampling using the modified Hamiltonian. The method of control variates [17] could also be used in SHMC.

Conformational dynamics [18,19] is an application that might benefit from SHMC. It performs many short HMC simulations in order to compute the stochastic matrix of a Markov Chain. Then it identifies almost invariant sets of configurations, thereby allowing a reduction of the number of degrees of freedom in the system.

Acknowledgments. This work was partially supported by an NSF Career Award ACI-0135195. Scott Hampton was supported through an Arthur J. Schmitt fellowship. The authors would like to thank Robert Skeel, David Hardy, Edward Maginn, Gary Huber and Hong Hu for helpful discussions.

References

1. Berne, B.J., Straub, J.E.: Novel methods of sampling phase space in the simulation of biological systems. Curr. Topics in Struct. Biol. **7** (1997) 181–189
2. Leach, A.R.: Molecular Modelling: Principles and Applications. Addison-Wesley, Reading, Massachusetts (1996)
3. Schlick, T.: Molecular Modeling and Simulation - An Interdisciplinary Guide. Springer-Verlag, New York, NY (2002)
4. Brass, A., Pendleton, B.J., Chen, Y., Robson, B.: Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. Biopolymers **33** (1993) 1307–1315
5. Sanz-Serna, J.M., Calvo, M.P.: Numerical Hamiltonian Problems. Chapman and Hall, London (1994)
6. Hockney, R.W., Eastwood, J.W.: Computer Simulation Using Particles. McGraw-Hill, New York (1981)
7. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195** (1987) 216–222
8. Creutz, M.: Global Monte Carlo algorithms for many-fermion systems. Phys. Rev. D **38** (1988) 1228–1238
9. Mehlig, B., Heermann, D.W., Forrest, B.M.: Hybrid Monte Carlo method for condensed-matter systems. Phys. Rev. B **45** (1992) 679–685
10. Creutz, M., Gocksch, A.: Higher-order hybrid monte carlo algorithms. Phys. Rev. Lett. **63** (1989) 9–12
11. Skeel, R.D., Hardy, D.J.: Practical construction of modified Hamiltonians. SIAM J. Sci. Comput. **23** (2001) 1172–1188
12. Hairer, E., Lubich, C.: Asymptotic expansions and backward analysis for numerical integrators. In: Dynamics of Algorithms, New York, IMA Vol. Math. Appl 118, Springer-Verlag (2000) 91–106
13. Matthey, T., Cickovski, T., Hampton, S., Ko, A., Ma, Q., Slabach, T., Izaguirre, J.A.: ProtoMol: an object-oriented framework for prototyping novel algorithms for molecular dynamics. Submitted to ACM Trans. Math. Softw. (2003)
14. Kirchhoff, P.D., Bass, M.B., Hanks, B.A., Briggs, J., Collet, A., McCammon, J.A.: Structural fluctuations of a cryptophane host: A molecular dynamics simulation. J. Am. Chem. Soc. **118** (1996) 3237–3246
15. Hampton, S.: Improved sampling of configuration space of biomolecules using shadow hybrid monte carlo. Master's thesis, University of Notre Dame, Notre Dame, Indiana, USA (2004)
16. Neal, R.M.: An improved acceptance procedure for the hybrid Monte Carlo algorithm. J. Comput. Phys. **111** (1994) 194–203
17. Lavenberg, S.S., Welch, P.D.: A perspective on the use of control variables to increase the efficiency of monte carlo simulations. Management Science **27** (1981) 322–335
18. Schüette, C., Fischer, A., Huisinga, W., Deuflhard, P.: A direct approach to conformational dynamics based on hybrid Monte Carlo. J. Comput. Phys **151** (1999) 146–168
19. Schüette, C.: Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules. Technical report, Konrad-Zuse-Zentrum für Informationstechnik Berlin (1999) SC 99-18.