JOURNAL OF COMPUTATIONAL PHYSICS 22, 245–268 (1976)

Efficient Estimation of Free Energy Differences from Monte Carlo Data

CHARLES H. BENNETT

IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

Received February 13, 1976; accepted May 3, 1976

Near-optimal strategies are developed for estimating the free energy difference between two canonical ensembles, given a Metropolis-type Monte Carlo program for sampling each one. The estimation strategy depends on the extent of overlap between the two ensembles, on the smoothness of the density-of-states as a function of the difference potential, and on the relative Monte Carlo sampling costs, per statistically independent data point. The best estimate of the free energy difference is usually obtained by dividing the available computer time approximately equally between the two ensembles; its efficiency (variance × computer time)$^{-1}$ is never less, and may be several orders of magnitude greater, than that obtained by sampling only one ensemble, as is done in perturbation theory.

I. INTRODUCTION

A well-known deficiency of the Monte Carlo [1, 2] and molecular dynamics [3] methods, commonly used to study the thermodynamic properties of classical systems having 10^2 to 10^4 degrees of freedom, is their inability to calculate quantities such as the entropy or free energy, which cannot be expressed as canonical or microcanonical ensemble averages. In general, the free energy of a Monte Carlo (MC) or molecular dynamics (MD) system can be determined only by a procedure analogous to calorimetry, i.e., by establishing a reversible path between the system of interest and some reference system of known free energy. "Computer calorimetry" has a considerable advantage over laboratory calorimetry in that the reference system may differ from the system of interest not only in its thermodynamic state variables but also in its Hamiltonian, thereby making possible a much wider variety of reference systems and reversible paths. Often the path between an analytically tractable reference system and the system of ultimate physical interest will include one or more intermediate systems. These may be interesting in their own right (e.g., the hard sphere fluid), or they may be special systems, important only as calorimetric stepping stones, whose Hamiltonians contain artificial terms designed to stabilize the system against phase transitions [4, 5], induce favorable importance weighting [6, 7], or otherwise enhance the system's efficiency as a computational tool [8–10].

Whether the calorimetric path has one step or many, one eventually faces the statistical problem of extracting from the available data the best estimate of the free energy differences between consecutive systems. Specializing the question somewhat, one might inquire what is the best estimate one can make of the free energy difference between two MC systems (i.e., two canonical ensembles on the same configuration space), given a finite sample of each ensemble. Section II of this paper derives the "acceptance ratio estimator," a near-optimal solution of this estimation problem, based only on the data in the two ensemble samples (a special case is the estimation of the free energy difference between two ensembles using data from only one of them; however, it will be argued that it is usually preferable to gather data from both ensembles). The efficiency of the acceptance ratio estimator is proportional to the degree of overlap between the two ensembles.

Section III presents a related method, the "interpolation method," which yields an improved free energy estimate under an additional assumption that is often physically justified, namely, that the density of states in each ensemble is a smooth function of the difference potential. When this assumption is justified, the interpolation method can yield a good free energy estimate even when the overlap between the two ensembles is negligible. On the other hand, when the two ensembles neither overlap nor satisfy this smoothness assumption, no method of statistical analysis can yield a good estimate of the free energy difference, and one must collect additional MC data from one or more ensembles intermediate between the two originally considered.

Section IV compares the present methods with older methods of MC free energy estimation, viz, numerical integration of a derivative of the free energy, perturbation theory, and previous overlap methods. This section also discusses some of the problems of designing and sampling intermediate ensembles.

II. The Acceptance Ratio Method

IIa. *Acceptance Probabilities and Configurational Integrals*

In this section the acceptance ratio method, to be developed more rigorously in Sections IIb and IIc, will be discussed from a physical and qualitative point of view.

In most classical systems of interest the kinetic part of the canonical partition function is trivially calculable; hence the problem of finding the free energy of a given (N, T, V) macrostate reduces to that of evaluating the canonical configurational integral

$$Q = \int \exp[-U(q_1 \cdots q_N)]\, dq_1 \cdots dq_N . \tag{1}$$

Here $U = \Phi/kT$ is the temperature-scaled potential energy, a function of the system's N configurational degrees of freedom, $q_1, q_2, \ldots, q_N$. It is convenient to allow U sometimes to take on the "value" plus infinity (but never minus infinity) so that external constraints, such as those that define the system's volume and shape, may be incorporated directly in the potential function and Q may be defined, as above, by an unbounded integral. With these conventions *any* nonsingular probability density $\rho(\mathbf{q})$ may be viewed as a canonical ensemble density, determined by a potential of the form $U(\mathbf{q}) = \text{const} - \ln \rho(\mathbf{q})$.

Existing methods are incapable, in general, of evaluating integrals of the form (1), because the dominant contribution typically comes from a small but intricately shaped portion of configuration space; however it is not difficult to derive useful formulas for the *ratio* between two such integrals, defined by two *different* potential functions, U_0 and U_1, acting on the *same* configuration space $\{(q_1 \cdots q_N)\}$. Equation (4), for example, to be derived presently, expresses the ratio Q_0/Q_1 as a ratio of canonical averages involving the "Metropolis" function, $M(x) = \min\{1, \exp(-x)\}$. The Metropolis function, because it has the property $M(x)/M(-x) = \exp(-x)$, is used in the standard Monte Carlo algorithm [1, 2] to assign Boltzmann-weighted acceptance probabilities to trial moves, a move that would change the (temperature-scaled) energy by ΔU being accepted with probability $M(\Delta U)$. Here, however, we consider an unorthodox kind of trial move—one that keeps the same configuration $(q_1 \cdots q_N)$, but switches the potential function from U_0 to U_1 or vice-versa. For each configuration, the acceptance probabilities for such a pair of trial moves must satisfy the relation

$$M(U_1 - U_0)\exp(-U_0) = M(U_0 - U_1)\exp(-U_1). \tag{2}$$

Integrating this identity over all of configuration space and multiplying by the trivial factors Q_0/Q_0 and Q_1/Q_1, one obtains:

$$Q_0 \frac{\int M(U_1 - U_0)\exp(-U_0)\, dq_1 \cdots dq_N}{Q_0}$$

$$= Q_1 \frac{\int M(U_0 - U_1)\exp(-U_1)\, dq_1 \cdots dq_N}{Q_1}. \tag{3}$$

The quotients on both sides can be recognized as canonical averages, i.e., quantities that can be measured during ordinary MC runs on systems 0 and 1 respectively. Representing these averages by the conventional angle brackets, one obtains the desired result:

$$\frac{Q_0}{Q_1} = \frac{\langle M(U_0 - U_1)\rangle_1}{\langle M(U_1 - U_0)\rangle_0}. \tag{4}$$

The physical meaning of this formula is that a Monte Carlo calculation that included potential-switching trial moves (in a fixed ratio to ordinary, configuration-changing trial moves) would distribute configurations between the unknown U_1 and the reference U_0 system in the ratio of their configurational integrals. The potential-switching moves need not actually be carried out, however, since the desired ratio can be estimated more accurately simply by taking the indicated averages over separately-generated samples of the U_0 and U_1 ensembles.

Before proceeding further, a few general remarks on the scope and limitations of the acceptance ratio method are in order. The requirement that the two systems be defined by potentials acting on the same configuration space is not a serious limitation, since, for most pairs of macrostates one might care to compare, a rather trivial transformation of the coordinates (e.g., a dilation or shear) suffices to make the two configuration spaces congruent. It is possible even to compare systems with a different number of degrees of freedom (as in the MC simulation of a grand canonical ensemble); the lower-order system is simply given one or more dummy coordinates, whose contribution to Q can later be factored out and computed analytically.

Most special ensembles used in Monte-Carlo work can be expressed as canonical ensembles by appropriate definition of the potential function U. The (N, T, P) ensemble, for example, can be represented [2] by making the volume a coordinate and the pressure a parameter of U. Importance-weighted ensembles [6, 7] can be viewed as canonical ensembles defined by U functions containing additive terms designed to concentrate the probability density in desired portions of configuration space.

The only important practical limitation on the method is that *both* mean acceptance probabilities (i.e., both averages in Eq. (4)) must be large enough to be determined with reasonable statistical accuracy in a Monte-Carlo run of reasonable duration. If only one of the acceptance probabilities is too small, it can be increased, at the expense of the other, by shifting the origin of one of the potential functions by an additive constant. Simultaneous smallness of both probabilities indicates that there is insufficient overlap between the U_0 and U_1 ensembles, and, in order to obtain a good estimate of Q_1, one must either:

1. Find a new reference potential which exhibits greater overlap with U_1.

2. Perform additional MC calculations under one or more intermediate potentials, so as to form an overlapping chain between U_0 and U_1.

3. Use curve-fitting methods, to be discussed in Section III, to interpolate between the U_0 and U_1 ensembles, thereby obtaining a good estimate of the free energy difference in spite of the lack of overlap.

Although Eq. (4) is not strictly correct for the (N, E, V) or $(N, E, V,$ linear

momentum) ensembles sampled by molecular dynamics, in practice it often can be used with molecular dynamics data, owing to the close similarity (except near phase transitions) of the configurational distributions in the various ensembles, for systems having more than a few degrees of freedom. Since temperature is not an independent variable in a constant-energy ensemble, the temperatures used in defining the temperature-scaled potentials U_0 and U_1 would have to be taken from time averages of the kinetic energy.

An exact microcanonical analog of Eq. (4) exists for the somewhat special case of two systems at the same energy E whose Hamiltonians, H_0 and H_1, have equal "soft" parts, i.e., H_0 and H_1 are equal wherever neither is infinite. For such a pair of systems the ratio of the microcanonical phase integrals is given by

$$\frac{\int \delta(H_0 - E)\, dq^N\, dp^N}{\int \delta(H_1 - E)\, dq^N\, dp^N} = \exp(S_0 - S_1)/k = \frac{[M(H_0 - H_1)]_1}{[M(H_1 - H_0)]_0}, \tag{5}$$

with square brackets here denoting microcanonical phase averages. The numerator and denominator of Eq. (5) have a very simple interpretation: e.g., $[M(H_1 - H_0)]_0$ is the fraction of points on the H_0 energy surface that also lie on the H_1 energy surface. The formulation of a more general microcanonical analog of Eq. (4) is frustrated by the fact that, for a general pair of Hamiltonians, H_0 and H_1, the two energy surfaces would have an intersection of zero measure.

Returning to the canonical ensemble, it may be noted that Eq. (4) is not the most general formula for Q_0/Q_1 as a ratio of canonical averages. A more general formula results if one includes in both the numerator and denominator an arbitrary weighting function. Let $W(q_1 \cdots q_N)$ be any everywhere-finite function of the coordinates. It then follows easily that

$$\frac{Q_0}{Q_1} = \frac{Q_0 \int W \exp(-U_0 - U_1)\, dq^N}{Q_1 \int W \exp(-U_1 - U_0)\, dq^N} = \frac{\langle W \exp(-U_0)\rangle_1}{\langle W \exp(-U_1)\rangle_0}. \tag{6}$$

Note that configurations having infinite energy under either U_0 or U_1 or both make no contribution to Eq. (6) so long as W is finite; henceforth, W will by convention be set equal to zero for all such configurations.

Most previous direct or overlap methods for estimating free energy (to be reviewed in Section IV) can be viewed as special cases of Eq. (6), with particular forms of the potentials U_0 and U_1 and the weight function W. Equation (4), for example, corresponds to the choice $W = \exp(+\min\{U_0, U_1\})$. The next section (IIb) shows, by some rather lengthy statistical arguments, that the optimized estimator of Q_0/Q_1 as a ratio of canonical averages differs from Eq. (4) in two respects: (1) the Fermi function, $f(x) = 1/(1 + \exp(+x))$, is used instead of the Metropolis function; and (2) the origin of one of the potential functions is shifted so as to (roughly) equalize the two acceptance probabilities.

It is also shown that, when one is free to vary the amount of computer time spent sampling the two ensembles, roughly equal time should be devoted to each.

IIb. *Optimized Acceptance Ratio Estimator—Large Sample Regime*

Optimization of the free energy estimate is most easily carried out in the limit of large sample sizes. Let the available data consist of n_0 statistically independent configurations from the U_0 ensemble and n_1 from the U_1 ensemble, and let this data be used in Eq. (6) to obtain a finite-sample estimate of the reduced free energy difference $\Delta A = A_1 - A_0 = \ln(Q_0/Q_1)$. For sufficiently large sample sizes the error of this estimate will be nearly Gaussian, and its expected squre will be

$$\begin{aligned}
&\text{Expectation of } (\Delta A\text{est} - \Delta A)^2 \\
&\approx \frac{\langle W^2 \exp(-2U_1)\rangle_0}{n_0[\langle W \exp(-U_1)\rangle_0]^2} + \frac{\langle W^2 \exp(-2U_0)\rangle_1}{n_1[\langle W \exp(-U_0)\rangle_1]^2} - \frac{1}{n_0} - \frac{1}{n_1} \\
&= \frac{\int ((Q_0/n_0)\exp(-U_1) + (Q_1/n_1)\exp(-U_0))\, W^2 \exp(-U_0 - U_1)\, dq^N}{[\int W \exp(-U_0 - U_1)\, dq^N]^2} \\
&\quad - (1/n_0) - (1/n_1). \qquad (7)
\end{aligned}$$

By making the integral in the numerator stationary with respect to a variation of W at constant value of the integral in the denominator, the optimum W function is found:

$$W(q_1 \cdots q_N) = \text{const} \times \left(\frac{Q_0}{n_0}\exp(-U_1) + \frac{Q_1}{n_1}\exp(-U_0)\right)^{-1}. \qquad (8)$$

Substituting this into Eq. (6) yields

$$\frac{Q_0}{Q_1} = \frac{\langle f(U_0 - U_1 + C)\rangle_1}{\langle f(U_1 - U_0 - C)\rangle_0}\exp(+C), \qquad (9a)$$

where

$$C = \ln\frac{Q_0 n_1}{Q_1 n_0} \qquad (9b)$$

and f denotes the Fermi function $f(x) = 1/(1 + \exp(+x))$. Equation (9a) is true for any value of the shift constant C, but the particular value specified by Eq. (9b) minimizes the expected square error (Eq. (7)) when the canonical averages are evaluated by finite sample means, with sample sizes n_0 and n_1.

The magnitude, σ^2, of this minimum square error can be found by taking the variance of Eq. (9a), or by substituting Eq. (8) into (7); σ^2 can be conveniently

expressed in terms of n_0, n_1, and the normalized configuration-space density functions ρ_0 and ρ_1:

$$\sigma^2 = \frac{\langle f^2\rangle_0 - \langle f\rangle_0^2}{n_0\langle f\rangle_0^2} + \frac{\langle f^2\rangle_1 - \langle f\rangle_1^2}{n_0\langle f\rangle_0^2} \tag{10a}$$

$$= \left(\int \frac{n_0 n_1 \rho_0 \rho_1}{n_0\rho_0 + n_1\rho_1}\, dq^N\right)^{-1} - \frac{n_0 + n_1}{n_0 n_1}. \tag{10b}$$

In Eq. (10a) the argument of f is understood to be $(U_1 - U_0 - C)$ or $(U_0 - U_1 + C)$ in the 0 and 1 expectations, respectively, with $C = \ln(Q_0 n_1/Q_1 n_0)$ as specified by Eq. (9b). In Eq. (10b), $\rho(q_1 \cdots q_N)$ denotes the density $(1/Q)\exp[-U(q_1 \cdots q_N)]$. Since σ^2 is a monotonically decreasing function of both n_0 and n_1, it follows that for some $\bar{n}$, lying between n_0 and n_1,

$$\sigma^2 = \frac{2}{\bar{n}}\left[\left(\int \frac{2\rho_0\rho_1}{\rho_0 + \rho_1}\, dq^N\right)^{-1} - 1\right]. \tag{11}$$

The integral in this equation is clearly a measure of the "overlap" between the two densities in configuration space. Equation (11) thus says that Q_0/Q_1 can be determined accurately as a ratio of canonical averages by (and only by) sampling a number of configurations greater than the reciprocal of the overlap between ρ_0 and ρ_1.

The optimized formula for Q_0/Q_1 (Eq. (9a)) differs from that derived earlier (Eq. (4)) only in the use of the Fermi function in place of the Metropolis function, and in the shifting of the origin of one of the potentials by an additive constant C. Figure 1 shows both the Fermi and Metropolis functions along with a typical probability density for values of their argument x, the change in energy accompanying a potential-switching move (i.e., $x = U_0 - U_1 + C$ under U_1, and

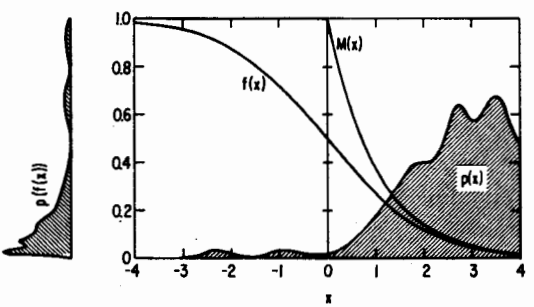

FIG. 1. The Fermi function, $f(x) = 1/[1 + \exp(+x)]$, and the Metropolis function, $M(x) = \min\{1, \exp(-x)\}$, are shown along with a typical probability density, $p(x)$, for their argument. Plotted on the left is a typical probability density for *values* of the Fermi function.

$U_1 - U_0 - C$ under U_0). When the shift constant is properly chosen (Eq. (9b)), most potential-switching moves (like most trial moves in an ordinary MC calculation) will result in an increase in energy and hence will lie on the positive "tail" of the f (or M) function. The advantage of the Fermi function in estimating free energy differences lies in its having a softer shoulder than the Metropolis function. This narrows the distribution of acceptance probabilities $p(f(x))$, and makes possible a more accurate estimation of the ensemble average acceptance probability from a given body of data. The Metropolis function, on the other hand, is the better acceptance function to use in the ordinary MC algorithm for *generating* new configurations, because here one seeks to maximize the acceptance probability itself, without regard to its variance. The shifting of the energy origin by C serves to maximize the number of configurations falling near the soft shoulder of the f function, while minimizing the number falling far out on its tail. It should perhaps be pointed out that in the special case of two potentials whose soft parts are identical, Eq. (9a) becomes equivalent to Eq. (4) and yields no better estimate of Q_0/Q_1.

In practice, of course, one cannot determine the optimum shift constant C exactly, because it depends on the unknown quantity Q_0/Q_1; however a value sufficiently close to the optimum can be found by adjusting C until Eqs. (9a) and (9b) become self-consistent for the given body of data. This estimation procedure can be expressed conveniently as a pair of simultaneous equations in $\Delta A\text{est}$ and C:

$$\Delta A\text{est} = \ln \frac{\sum_1 \{f(U_0 - U_1 + C)\}}{\sum_0 \{f(U_1 - U_0 - C)\}} + C - \ln(n_1/n_0) \tag{12a}$$

$$\Delta A\text{est} = C - \ln(n_1/n_0). \tag{12b}$$

Equation (12a) (the finite-sample analog of Eq. (9a)) estimates the free energy difference in terms of explicit sums over the n_1 configurations comprising the U_1 ensemble sample and the n_0 comprising the U_0 ensemble sample; Eq. (12b) (or equivalently $\sum_1 = \sum_0$) is the self-consistency criterion for selecting C.

The large-sample regime assumed in Eqs. (9)–(12) may now be expressed as a condition on the sums $\sum_0$ and $\sum_1$: namely, that for some range of C-values about the true C of Eq. (9b), these sums differ relatively little from their respective expectations, $n_0\langle\ \rangle_0$ and $n_1\langle\ \rangle_1$. Under this condition the self-consistent procedure yields essentially an optimum estimate of ΔA, differing from the true value by a quantity of order σ. This follows from the fact that the first term on the right-side of Eq. (12a) is a monotonically decreasing function of C with a slope of nearly -1 and a value, for the correct C of Eq. (9b), within about σ of zero. One may be sure of being in the large-sample regime whenever both $\sum_1$ and $\sum_0$ are large compared to unity, because the terms comprising the sums are statistically independent and all lie between zero and one (the large sample condition is thus equivalent to the condition

discussed in connection with Eqs. (10) and (11), viz, n_0 and n_1 must be great enough to adequately sample the region of overlap between ρ_0 and ρ_1).

Figure 2 shows a representative graphical solution of Eqs. (12a) and (12b) for four sets of simulated Monte-Carlo data drawn from the same pair of ensembles (the sample sizes $n_0 = n_1 = 10^6$, lay in the large sample regime, with $\sum_0 \approx \sum_1 \approx 200$). Note that the straight line of Eq. (12b) cuts through a region where the four curves of (12a) differ least from each other and from the true value of ΔA. In the large sample regime the standard error of the estimate ΔA est will be less than ± 1, and can be computed in the usual manner by solving Eqs. (12a) and (12b) for several large, independent bodies of data, as was done in Fig. 2.

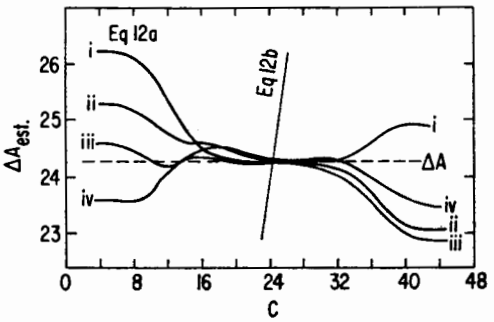

FIG. 2. Acceptance ratio estimate of the free energy difference between two MC ensembles in the large-sample regime by simultaneous graphical solution of Eqs. (12a) and (12b). For a description of the two ensembles, and the method of sampling them, see the Appendix. The four curves (i–iv) plot the right side of Eq. (12a) as a function of the shift constant C for four independent sets of data, each consisting of 10^6 points randomly chosen from the 0 ensemble and 10^6 from the 1 ensemble. The slanting straight line is Eq. (12b), while the dashed horizontal line gives the true free energy difference ΔA. The mean and standard error of the four finite-sample estimates are 24.290 ± 0.017; these are in satisfactory agreement with the true free energy difference, 24.268, and the error $\sigma = \pm 0.021$ predicted by Eq. (10b) for sample sizes $n_0 = n_1 = 4 \times 10^6$.

Equations (12a) and (12b) represent an estimation strategy optimized with respect to a given pair of large samples, with fixed sizes n_0 and n_1. We now consider the allocation of computer time *between* the two ensembles if the sample sizes are not fixed beforehand, but are free to be chosen so as to minimize σ^2 with respect to n_1/n_0 at a constant total cost in computer time. Let us assume that the time required to compute each (statistically-independent) data point is t_0 in the 0 ensemble and t_1 in the 1 ensemble, so that the total computing time is $n_0 t_0 + n_1 t_1$. A crude but effective rule for choosing n_1/n_0 is simply to allocate equal time to the two ensembles, i.e.,

$$n_1/n_0 = t_0/t_1 . \tag{13}$$

The estimation efficiency, $1/[(n_0\ell_0 + n_1\ell_1)\,\sigma^2]$, resulting from this equal-time allocation is at least half as great as that resulting from any other allocation. This follows from the fact that σ^2 is a monotonically decreasing function of both n_0 and n_1; (i.e., even the best allocation of 1 hour of computer time *between* the two ensembles yields no better estimate than would obtained by devoting a full hour to *each* ensemble). In the special case $\ell_0 = \ell_1 = 1$, the equal-time estimation efficiency is approximately one fourth the overlap integral (cf. Eq. (11)).

The equal time rule gives a sufficiently good n_1/n_0 ratio for most practical situations; however, for the sake of elegance, the true optimum ratio can be expressed in terms of the variance of the Fermi functions by solving the variational equation

$$\ell_1(d\sigma^2/dn_0) = \ell_0(d\sigma^2/dn_1). \tag{14}$$

Explicitly differentiating Eq. (10a) with respect to n_0 and n_1, one obtains

$$-\frac{\ell_1(\langle f^2\rangle_0 - \langle f\rangle_0^2)}{n_0^2\langle f\rangle_0^2} = -\frac{\ell_0(\langle f^2\rangle_1 - \langle f\rangle_1^2)}{n_1^2\langle f\rangle_1^2}, \tag{15}$$

with the argument of f understood to be $U_1 - U_0 - C$ in the 0 expectations and $U_0 - U_1 + C$ in the 1 expectations. One might worry about the implicit n-dependence that the various Fermi expectations have by virtue of the n-dependent shift constant C, defined in Eq. (9b). However, these implicit n-dependences have no effect on the *derivatives* of σ^2, because Eq. (9b) is itself the solution to the variational condition $\partial\sigma^2/\partial C = 0$. Equations (9a) and (9b) also cause the denominators on the two sides of Eq. (15) to be equal. Thus Eq. (15) reduces to an equation in one unknown, defining an optimum value for the shift constant C:

$$\ell_1 \operatorname{Var}_0[f(U_1 - U_0 - C)] = \ell_0 \operatorname{Var}_1[f(U_0 - U_1 + C)], \tag{16}$$

with Var_0 and Var_1 denoting the absolute variances. These variances, of course, cannot be estimated with much precision from finite samples of the ensembles, but by adjusting the sample sizes until Eqs. (12a) and (12b) can be solved self-consistently for the same value of C as Eq. (16), one might obtain some improvement over the equal time strategy, particularly in cases where the optimum time ratio is far from 1: 1.

As noted earlier, the efficiency of estimating ΔA is at least half-optimal, and not very sensitive to the n_1/n_0 ratio, in the neighborhood of $n_1/n_0 = \ell_0/\ell_1$ (the equal time rule). This can be seen in Fig. 3, which plots the log estimation efficiency versus the log sampling ratio for the pair of model ensembles considered earlier. Although the optimum n_1/n_0 ratio is 1.8, the estimation efficiency at $n_1/n_0 = 1$ is almost (99.7%) as good. (At first it might appear that by making the costs ℓ_0 and ℓ_1 very disparate, the optimum time ratio could be displaced far from unity;

however this is not so. If, for example ϕ_1/ϕ_0 is changed from 1 to 10^{-4}, the optimum n_1/n_0 ratio is indeed greatly increased as one would expect; but the optimum *time* ratio, $n_1\phi_1/n_0\phi_0$, changes only from 1.8 to 1.3)

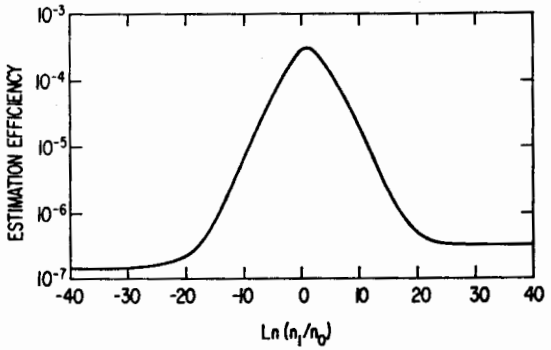

FIG. 3. Dependence of estimation efficiency, $1/[(n_0 + n_1)\sigma^2]$, on the n_1/n_0 ratio for the model ensembles described in the Appendix, assuming equal sampling costs ($\phi_1 = \phi_0$) in the two ensembles. The horizontal wings of the curve indicate the rather poor efficiency with which $A_1 - A_0$ can be estimated when only one ensemble is sampled, as in infinite-order perturbation theory. The efficiency curve was calculated by exact evaluation of Eq. (10b) over the rather trivial configuration space of the model ensembles.

Figure 3 also shows that the estimation efficiency can become very bad if one flouts the equal time rule by sampling only one ensemble. The poor efficiency results from the fact that, when only one ensemble is sampled (say the 0 ensemble), the acceptance ratio estimator reduces to the average of a pure exponential (cf. Eq. (22)), whose variance can be expressed in terms of the densities ρ_0 and ρ_1 as $[\int (\rho_1^2/\rho_0)\, dq^N - 1]/n_0$. This expression is less transparent than the formula (Eq. (11)) relating the variance of the two-ensemble estimate to the overlap integral; however, its qualitative meaning is that an accurate one-ensemble estimate requires that the sampled ensemble include *all* important configurations of the other ensemble. A good two-ensemble estimate, on the other hand, requires only that each ensemble include *some* important configurations of the other ensemble.

IIc. *Acceptance Ratio Estimates in the Small Sample Regime*

The treatment so far has been limited to the large sample regime, in which both sums in Eq. (12a) can be made simultaneously greater than unity. Unfortunately, in many cases of interest, the overlap between the two ensembles is so slight that even with the largest practical n_0 and n_1 this condition cannot be met. We shall now show that even in this small sample regime Eq. (12a) can yield a useful estimate of ΔA, though the error bounds will be greater than ± 1 and can no longer be

estimated from the spread among independent estimates. When either sum in (12a) (say $\sum_1$) is small compared to unity, its most serious source of statistical error becomes the possibility that some important class of configurations, whose total ensemble probability is low ($1/n_1$ or less) but whose f values are high (near unity at worst), may not be sampled at all. Such failures to sample could cause either sum to *underestimate* its expectation by a quantity of order unity (corresponding errors due to *over*-sampling of high-f configurations could also occur, but, owing to the convexity of the log function, their effect on Eq. (12a) would be much less). In order to find bounds on the possible failure-to-sample errors, and hence on ΔA, we take advantage of the fact that $\sum_1$ is a monotonically decreasing function of C, while $\sum_0$ is monotonically increasing. Therefore, by decreasing C to that value, C_1, for which $\sum_1$ becomes equal to unity, we can make all the failure-to-sample errors appear in the denominator of Eq. (12a) and obtain a value, $\Delta A\text{est}+$, which may overestimate, but is unlikely to seriously under-

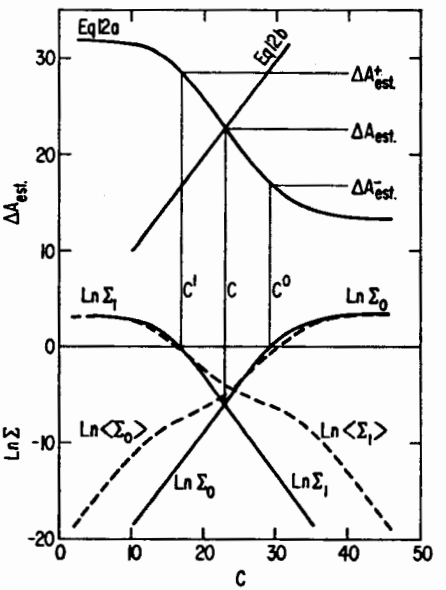

FIG. 4. Acceptance ratio estimate of the free energy difference in the small-sample regime. The upper pair of curves show the construction of the optimum estimate $\Delta A\text{est}$ by graphical solution of Eqs. (12a) and (12b), from two small samples ($n_0 \approx n_1 \approx 20$) of the pair of ensembles described in the Appendix. The lower pair of solid curves show the construction of the upper and lower estimates, $\Delta A\text{est}^+$ and $\Delta A\text{est}-$, by Eqs. (17a) and (17b). The dashed curves show the log expectations, $\ln\langle\Sigma_0\rangle = \ln[n_0\langle f(U_1 - U_0 - C)\rangle_0]$ and $\ln\langle\Sigma_1\rangle = \ln[n_1\langle f(U_0 - U_1 + C)\rangle_1]$, which are closely approximated by the observed log sums as long as the latter are greater than unity, but are poorly approximated at C values for which the log sums are less than unity.

estimate, the true free energy difference ΔA. The construction of this upper estimate and the corresponding lower estimate, ΔAest−, is shown in Fig. 4. The computation of error bounds in the small sample regime may be summarized:

$$\Delta A\text{est}- = \text{R12a}(C_0) \lesssim \Delta A \lesssim \text{R12a}(C_1) = \Delta A\text{est}+, \tag{17a}$$

where R12a(C) denotes the right side of Eq. (12a), and C_0 and C_1 are defined by

$$\sum_0 \{f(U_1 - U_0 - C_0)\} = 1 = \sum_1 \{f(U_0 - U_1 + C_1)\}. \tag{17b}$$

Parenthetically it is interesting to note that the right side of Eq. (12a), which would be independent of C in the large-sample limit, here decreases with a slope of about -1 throughout the range where both sums are much smaller than unity. This is necessarily so because, under these conditions, neither sum can receive contributions except from the nearly exponential positive tail of the f function. The self-consistent estimate ΔAest obtained by solving Eq. (12a) with (12b) therefore lies about midway between the upper and lower bounds computed by Eq. (17). It is still the best estimate of ΔA in the sense that it equalizes the damage that would be done by a unit failure-to-sample error in the numerator or the denominator of Eq. (12a).

IId. *Practical Considerations in Using Acceptance Ratio Methods*

In using Eqs. (12) and (17) with real MC data account must be taken of the fact that successively generated configurations of a Markov chain are not statistically independent, but on the contrary highly correlated. Each of the sums, $\sum_0$ and $\sum_1$, and the numbers n_0 and n_1, must therefore refer to a subset of configurations, chosen from the chain so infrequently as to be uncorrelated, or else be defined in terms of the whole chain as follows:

$$\sum = \tau^{-1} \sum_{\text{wc}}; \qquad n = \tau^{-1} n_{\text{wc}}. \tag{18}$$

Here $\sum_{\text{wc}}$ denotes a sum over the whole chain, having n_{wc} configurations, and τ is an empirically estimated autocorrelation time of the Markov chain with respect to values of the f function (This autocorrelation time can be defined as the large k limit of the quantity $k \cdot \text{Var}[f^{(k)}]/\text{Var}[f]$, where $f^{(k)}$ denotes the mean of k consecutive f values. Clearly τ can be estimated accurately only if it is considerably shorter than the total chain length n_{wc}). The cost, $\mathfrak{t}$, of a statistically independent data point, discussed in connection with Eqs. (13)–(16), would be similarly defined as τ times the computer time required to make one MC move.

Another practical note: although the evaluation of the sums in Eq. (12a) could in principle be done after the MC data (typically a Markov chain of several million configurations) had been generated, it is inconvenient to store all this data. A

better approach would be during the run to accumulate values of Σ_0 and Σ_1, using a mesh of C values sufficiently fine to permit accurate graphical solution of Eqs. (12a) and (12b) at the end of the run. Alternatively, one could store a pair of histograms, $h_0(\Delta U)$ and $h_1(\Delta U)$, of the values of the difference potential $\Delta U = U_1 - U_0$, observed while sampling the 0 and 1 ensembles, respectively. The interval width of the histograms need be no smaller than the desired precision of estimating ΔA, and they can be summed over easily at the end of the runs to evaluate Eq. (12a). Such histograms are also useful in their own right, in the interpolative method for estimating ΔA to be discussed in the next section.

The acceptance ratio method is at its best when the overlap between the two ensembles, as defined by the integral in Eq. (11), is not too small, e.g., in solid state vacancy calculations [8, 10] where the difference potential, $U_1 - U_0$, depends strongly on only a few atomic coordinates. It is more common for the difference potential to depend strongly on all the coordinates, resulting in an overlap many orders of magnitude less than unity. The two data histograms, $h_1(\Delta U)$ and $h_0(\Delta U)$ will then be separated by a wide gap (cf. Fig. 5) that cannot be filled in by any reasonable amount of additional sampling of either ensemble, and the acceptance ratio method (Eq. (17) in particular) will yield only the rather crude conclusion that the true free energy difference ΔA is somewhere between $\max\{\Delta U\}_1$ and $\min\{\Delta U\}_0$.

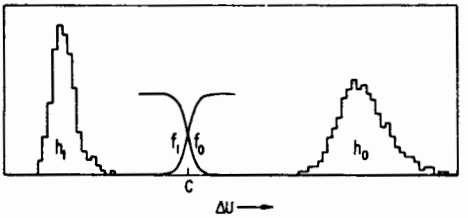

FIG. 5. Histograms of values of $\Delta U = U_1 - U_0$, h_0 representing a typical set of ΔU values sampled from the 0 ensemble, and h_1 a typical set sampled from the 1 ensemble. In the gap between the two histograms are a pair of complementary Fermi functions $f_0 = f(\Delta U - C)$ and $f_1 = f(C - \Delta U)$ whose origin can be shifted to the left or right, but whose widths are insufficient to achieve good overlap of f_0 with h_0, and f_1 with h_1, simultaneously.

The estimate of ΔA can be considerably improved if, as is often the case, the two histograms are sufficiently smooth to justify extrapolating them into the gap region of the ΔU spectrum, from which no data have been collected. The following section will deal with this method of estimating ΔA, which is no longer a pure acceptance ratio method, because it is based on the additional assumption that the ΔU spectrum is smooth even where no data have been collected.

When the smoothness assumption is not justified, i.e., when the data histograms

are ragged as well as widely separated, an improved estimate of ΔA can be obtained by using the acceptance ratio method in a multistage manner, with data collected from a chain of intermediate ensembles extending from U_0 to U_1.

III. Interpolation or Curve-Fitting Method

This section concerns the estimation of ΔA from histograms of ΔU values, $h_0(\Delta U)$ and $h_1(\Delta U)$, which are smooth but may be separated by a gap wide compared to kT. This situation is likely to arise when the difference potential depends strongly but more or less equally on many coordinates, e.g., when U_0 and U_1 represent a condensed system of many identical atoms interacting via one pair potential in the U_0 system and another in the U_1 system. If one is willing to infer from the histograms' smoothness that the reduced density-of-states functions $p_0(\Delta U)$ and $p_1(\Delta U)$, which the histograms approximate, extend smoothly into the gap region, one can obtain a much better estimate of ΔA than could be obtained from the acceptance ratio method alone.

This approach is less risky than it might first appear, and in fact is more akin to interpolation than to extrapolation, because p_0 and p_1 are not independent:

$$\frac{p_1(x)}{p_0(x)} \equiv \frac{\langle \delta(U_1 - U_0 - x)\rangle_1}{\langle \delta(U_1 - U_0 - x)\rangle_0} = \exp(\Delta A - x); \tag{19}$$

thus it is a matter of finding a single function (p_0, say) which fits both histograms, while satisfying the normalization constraints

$$\int_{-\infty}^{\infty} p_0(\Delta U)\, d\,\Delta U = 1, \tag{20a}$$

and

$$\int_{-\infty}^{\infty} p_1(\Delta U)\, d\,\Delta U \equiv \int_{-\infty}^{\infty} p_0(\Delta U) \exp(\Delta A - \Delta U)\, d\,\Delta U = 1. \tag{20b}$$

This can be done conveniently by expanding $\ln p_0$ as a polynomial in ΔU and performing a least-squares fit of the expansion coefficients, along with ΔA, to the histogram data, subject to the normalization constraints. The adequacy of the polynomial approximation, as well as the range of plausible ΔA values, can be judged by chi-square tests.

Estimation of ΔA also can be performed graphically (cf. Fig. 6), by plotting the two functions

$$-\tfrac{1}{2}\Delta U \qquad +\ln p_0, \tag{21a}$$

and

$$+\tfrac{1}{2}\Delta U \qquad +\ln p_1 \tag{21b}$$

versus ΔU on the same graph. Each function is plotted (solid curves) over the range of ΔU values for which it can be accurately estimated from the histogram data. By virtue of Eq. (19), the two functions are parallel, differing only by the unknown additive constant ΔA; hence, in order to estimate ΔA, one need only find a plausible parallel extrapolation (dashed curves) of the two functions into the range of ΔU values corresponding to the gap between the two histograms. Probably the easiest way to do this is to cut the graph in half vertically and slide the right half up or down until it can be smoothly joined onto the left half by some plausible extrapolation. The range of vertical shifts for which this can be done is then the range of plausible ΔA values. In Fig. 6 this range can be seen to be about $\Delta A = -85.5 \pm 2.5$, which is considerably narrower than the gap between the two histograms.

Clearly the interpolation method is at its best when the data histograms are smooth and significantly broader than kT (kT = 1 in the reduced units of ΔU),

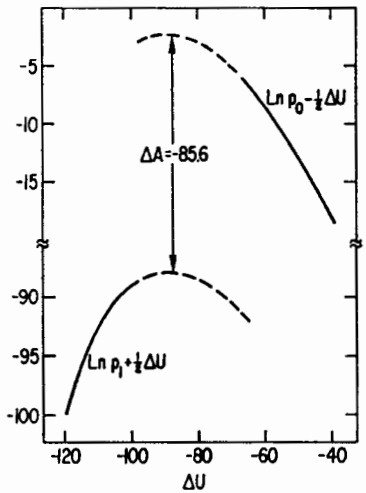

FIG. 6. The curve-fitting method of free energy estimation (cf. Eqs. (21a) and (21b)), applied to data of Valleau and Card [17]. The U_1 system here consisted of 200 charged (100+ and 100−) hard spheres at a fixed temperature and volume; the U_0 system was the same, except that the spheres were uncharged. The left and right solid curves in the present figure were obtained, respectively, from the left and right probability density curves of Valleau and Card [17, Fig. 2] (omitting the tails where the density was less than one tenth maximum) by: scaling the energy to units of kT, taking the logarithm, and adding $+\frac{1}{2}\Delta U$ and $-\frac{1}{2}\Delta U$, respectively. Their figure also includes a middle curve, corresponding to an intermediate ensemble which they used to bridge the gap between the two outer curves. With the help of this intermediate ensemble, they estimated ΔA to be -85.60 ± 0.58 (cf. [17, Table I]. As the present figure suggests, the outer curves are smooth enough to yield a fairly good estimate of ΔA (about -85.5 ± 2.5) by interpolation across the gap, without help from the intermediate ensemble.

and separated by a gap not much broader than the histograms themselves. Under these conditions much information about the shape of p_0 and p_1 in the overlap region can be inferred from data points lying far outside that region, data points which contribute hardly anything to the acceptance probabilities on which the method of the previous section is based. From this it can be seen that when two smooth histograms overlap very slightly (i.e., overlap integral in Eq. (11) of order $1/\bar{n}$), the interpolation method will still be an improvement over the straight acceptance ratio estimator. However, the more the two histograms overlap, the less important it becomes to guess the shape of the reduced density-of-states functions, p_0 and p_1, and, in the large-overlap limit (overlap integral ≈ 1), interpolation does not improve the estimate at all.

On the other hand, when the gap between h_0 and h_1 is excessively wide compared to the width of the histograms themselves, accurate interpolation becomes difficult, and it is best to use the interpolation method in a multistage manner, using MC data collected under one or more intermediate potentials, e.g., $U_\lambda = U_0 + \lambda(U_1 - U_0)$; $0 < \lambda < 1$. This linear form of the intermediate potential is convenient because it allows all the $\ln p_\lambda(\Delta U)$ data to be plotted on the same graph and fitted to the same polynomial, but it may be inferior to the more general intermediate potentials discussed in Section IVc.

IV. Discussion

In this section the acceptance ratio and interpolation methods of Sections II and III are compared with other methods of free energy estimation, viz, perturbation theory, numerical integration of a derivative of the free energy, and previous overlap methods.

IVa. *Perturbation Theory*

Perturbation theory [11, 12], which estimates the free energy of the U_1 system by extrapolation from the U_0 system, can be viewed as the limiting case of the acceptance ratio method in the absence of any data from the U_1 system. In this limit (i.e., $n_1 \to 0$) Eqs. (9a) and (9b) reduce to

$$A_1 - A_0 = -\ln\langle \exp(U_0 - U_1)\rangle_0 , \tag{22}$$

an infinite-order perturbation formula [13] which is exact, provided there are no configurations for which U_0 is infinite but U_1 is finite. To obtain finite-order formulas, one assumes the potential U to depend on a continuous parameter

λ in such a way that as λ is varied from 0 to 1, U passes smoothly from U_0 to U_1. The potential U_1 in Eq. (22) then can be expanded in a Taylor series about U_0:

$$\begin{aligned} A_1 - A_0 &= -\ln\langle\exp[-(\partial U/\partial\lambda) - (\partial^2 U/\partial\lambda^2)/2 - \cdots]\rangle_0 \\ &= \langle(\partial U/\partial\lambda)\rangle_0 \\ &\quad + \tfrac{1}{2}[\langle\partial^2 U/\partial\lambda^2\rangle_0 - \langle(\partial U/\partial\lambda)^2\rangle_0 + (\langle\partial U/\partial\lambda\rangle_0)^2] \\ &\quad + \cdots \\ &= \partial A/\partial\lambda + \tfrac{1}{2}\,\partial^2 A/\partial\lambda^2 + \cdots, \end{aligned} \tag{23}$$

where all the derivatives with respect to λ are evaluated in the U_0 ensemble, at $\lambda = 0$. The expansion is usually truncated at second order because the statistical uncertainty in measuring the higher derivatives, by a MC run of reasonable duration, is typically so great that they contribute only noise to the infinite-order formula (Eq. (22)). The simplest perturbation formula results by taking $U_\lambda = U_0 + \lambda\,\Delta U$, in which case $\partial^2 U/\partial\lambda^2 = 0$ and $\partial U/\partial\lambda = \Delta U$; Eq. (23) then expresses ΔA in terms of the mean and moments of ΔU, as measured in the reference system.

Alternatively, the acceptance ratio and curve fitting methods of Sections II and III may be viewed as double-ended, interpolative counterparts of ordinary extrapolative perturbation methods of infinite (Eq. (22)) and finite (Eq. 23)) order, respectively. The assumption of smoothness of the density-of-states function, on which the curve-fitting method is based, is then a less restrictive counterpart of the assumption that higher-order terms in Eq. (23) are negligible. Double-ended methods have the advantage of being able to set both upper and lower bounds on the free energy difference $A_1 - A_0$. The crudest of these bounds is the so-called Gibbs-Bogoliubov inequality [14],

$$\langle U_1 - U_0\rangle_1 \leqslant A_1 - A_0 \leqslant \langle U_1 - U_0\rangle_0, \tag{24}$$

which follows, via the convexity of the log function, from Eq. (22) and its analogue with U_0 and U_1 interchanged; more subtle bounds, e.g., Eq. (17), are discussed in Sections II and III.

In an ordinary single-ended perturbation treatment, on the other hand, no data is collected from the U_1 ensemble and only the right half of Eq. (24) can be used. One therefore cannot rule out the possibility of seriously overestimating $A_1 - A_0$ due to a failure to sample important configurations in the U_1 ensemble. Assurance that this kind of error has not occurred must come from specific knowledge of the potentials, or from independent confirmation of the properties of the U_1 system. The notably successful perturbation theory of liquids [12, 15, 16], whose reference system is the hard sphere fluid, was confirmed in this manner.

The perturbation theory of liquids also illustrates the chief strength of single-ended perturbation methods, namely, the possibility of using a single reference system to compute the properties of many different U_1 systems, without having to collect MC data on each of these separately. It should be noted, however, that when there is any doubt about the rapid convergence of an extrapolative perturbation from U_0 data, the estimate of $A_1 - A_0$ can be considerably improved, usually without much cost in computer time, by collecting a small amount of MC data on the U_1 system, then using a double-ended method. Indeed, whenever the second-order perturbation term differs significantly from zero, the two ensembles being compared are probably sufficiently different to warrant sampling both of them. The gain in estimation efficiency made possible by sampling both ensembles instead of only one is suggested in Fig. 3.

IVb. *Numerical Integration*

This method estimates the free energy difference $A_1 - A_0$ by numerically integrating the derivative $\partial A/\partial\lambda = \langle \partial U/\partial\lambda \rangle$, which is measured by equilibrium MC calculations at a mesh of values of the parameter λ between 0 and 1. The most commonly performed integration is of pressure versus volume [4, 5]; however, the method can be used to compute the free energy change attending any continuous deformation [8] of the potential, boundary conditions, or other parameters defining the MC macrostate. The integration method as ordinarily practiced is less than optimal because it ignores the information which each MC run provides about higher derivatives of the free energy (e.g., the isothermal compressibility); however, this information may be of poor statistical quality. Ideally each estimated derivative of A, at each mesh point, should be given a weight inversely proportional to its estimated standard error. Probably the easiest way to achieve this correct weighting of information is to use the acceptance ratio or interpolation methods in a multistage manner, to estimate $\int (\partial A/\partial\lambda)\, d\lambda$ between consecutive mesh points.

When information on higher derivatives is included it is clear that perturbation theory and the interpolation method of Section III are special cases of integration, with one or two mesh points, respectively. The number of mesh points actually needed depends on the smoothness of $\partial A/\partial\lambda$ as a function of λ, and on the ease and precision with which the derivatives can be estimated at each mesh point. In typical applications, where five to ten mesh points are used, numerical integration can determine $A_1 - A_0$ when the unknown and reference ensembles are too different to be compared by any method not using intermediate ensembles. On the other hand, when the 0 and 1 systems are similar enough to be compared directly, the generation of many intermediate ensembles, each of which must be allowed to equilibrate before representative data can be collected, is tedious and may be wasteful of computer time.

IVc. *Previous Overlap Methods*

Most previous overlap methods for determining free energy differences can be regarded as special cases of the acceptance ratio method, with particular forms of the potentials U_0 and U_1 and of the weight function W in Eq. (6). Perhaps the most common special case is the comparison of a pair of systems, one of which is restricted to a subset of the configurations accessible to the other. In other words, the difference potential $\Delta U = U_1 - U_0$ is a hard function, taking on only the values zero and plus infinity. With such a pair of potentials, one of the acceptance probabilities in Eq. (4) becomes identically unity, while the other is simply the fraction of microstates in the less restricted ensemble belonging to the more restricted ensemble. In what was probably the first application of an overlap method to a realistic system, McDonald and Singer [9] sampled a nested set of ensembles, defined by a decreasing sequence of upper bounds on the total energy of a gaseous Lennard–Jones system, and obtained the unnormalized density of states as a function of energy over a wide range of energies, from which they were able to compute the thermodynamic properties of the gas over a wide range of temperature.

Such nested sets of hard constraints can be used to restrict a MC system to any desired region of configuration space, and to determine the spontaneous probability of occupancy of that region in the absence of the constraints. In a molecular dynamics calculation, analogous constraint terms in the Hamiltonian can be used to sample trajectories passing through an arbitrary region of phase space, and to estimate the spontaneous frequency of such passages in an unconstrained system [10].

Apparently the first overlap calculation of a free energy difference between two systems whose potentials had differing "soft" parts was that of Valleau and Card [17]. These authors, interested in determining the thermodynamic properties of a fluid of charged hard spheres as a function of temperature, compared systems whose U differed by a constant factor, i.e., systems having the same unscaled potential but different temperatures. Their somewhat complicated procedure for for estimating ΔA (cf. [17, Appendix]) can be recognized as accomplishing the same result, with somewhat less statistical efficiency, as the Fermi-function weighting used in the acceptance ratio method of Section II. By emphasizing the importance of the density-of-states functions, $p_0(\Delta U)$ and $p_1(\Delta U)$, these authors adumbrated the interpolation method of Section III.

In the same paper Valleau and Card pointed out the possibility of using a specially tailored bridging ensemble, designed to have significant overlap with both the unknown and reference ensembles, in place of the many intermediate ensembles ordinarily used in the numerical integration method. In principle it is always possible to define such a bridging ensemble, no matter how different U_0 and U_1 may be. This may be done, for example, by defining the bridging potential

U_B, as an appropriately weighted log mean exponential of a sequence of overlapping potentials U_λ, extending between U_0 and U_1,

$$U_B = -\ln \sum_{\lambda=0}^{1} w_\lambda \exp(-U_\lambda), \tag{25}$$

the discrete weights w_λ being chosen to approximate $\exp(A_\lambda)$. By using a continuous weight function $w(\lambda)$ one can even define a bridging ensemble whose density of states is perfectly flat over the entire relevant interval of the ΔU spectrum, but to guess such a weight function would be tantamount to guessing the function $p_0(\Delta U)$ over the same interval. Torrie, Valleau, and Bain [6] used discretely weighted bridging ensembles to pass between an unconstrained hard sphere fluid and a single occupancy fluid (no two particles allowed in the same Wigner–Seitz cell) in a few sampling stages, thereby avoiding the long pressure–volume integration by which the communal entropy is usually estimated. Unfortunately, the bridging ensembles exhibited such long autocorrelation times (cf. Eq. (18)) that the hoped-for gain in computational efficiency was not realized. The bridging system, in other words, diffused much too slowly between the part of configuration space overlapping with U_0 and the part overlapping with U_1. In later work, Torrie and Valleau [7] obtained much better performance using continuously weighted bridging ensembles to estimate the free energy of a 32 particle Lennard–Jones fluid relative to that of the corresponding purely repulsive inverse twelfth power system. The difference in diffusion rates obtained in these two experiments probably is due in part to the superiority of a continuous weighting, but may also reflect the many-body character of the structural rearrangements involved in accommodating to the single-occupancy constraint. In difficult cases like the communal entropy calculation, some improvement in diffusion through a given bridging ensemble might be obtained by modifying the MC transition algorithm used to sample it (infinitely many transition algorithms, with different rates of diffusion through configuration space, can be used to sample the same canonical ensemble). So far almost all MC calculations have used simple transition algorithms in which only one particle is moved at a time, and trial moves are symmetrically distributed in direction. More efficient diffusion in the desired direction might be obtained by making trial moves preferentially parallel and antiparallel to the local gradient of the difference potential.

Overlap methods using a bridging potential of the form of Eq. (25) bear a certain similarity to numerical integration. Under a bridging potential, the system diffuses freely back and forth between the U_0 and U_1 parts of configuration space; in numerical integration, a series of intermediate MC runs is made, and the system is forced to diffuse, by the increment in the integration parameter, as each successive run equilibrates. Given a particular MC transition algorithm, which limits the

diffusion rate, numerical integration and bridging ensembles may be equally efficient statistically, if the higher-derivative information provided by the integration runs is taken into account.

V. Conclusion

The problem of free energy estimation can be broken into three parts:

1. what reference and (possibly) intermediate ensembles to use;
2. what MC transition algorithms to use for sampling the ensembles; and
3. how best to estimate the free energy from the resulting data.

The acceptance ratio and interpolation methods (developed in Sections II and III, respectively) offer a fairly complete solution to the third subproblem, viz, optimally estimating the free energy difference between two canonical ensembles given a finite MC sample of each (or, more generally, given MC routines able to sample each at some fixed cost per statistically independent data point). A good estimate can be arrived at if the ensembles being compared

1. exhibit significant overlap, allowing the acceptance ratio method of Section II to be used; or

2. are sufficiently similar that the density of states in each ensemble is a smooth function of $\Delta U = U_1 - U_0$, allowing the interpolation method of Section III to be used.

In either case it is dangerous not to sample both ensembles, unless one is known to include all important configurations of the other. When the two ensembles neither overlap nor satisfy the above smoothness condition, an accurate estimate of the free energy cannot be made without gathering additional MC data from one or more intermediate ensembles.

The first two subproblems are much less well understood than the third. The choice of a reference ensemble is more a matter of physics than of statistics, and it is probably best made on an individual, empirical basis. On the other hand the problems that arise in designing efficient bridging ensembles and transition algorithms to sample them appears to be the manifestation, in Monte Carlo work, of a general difficulty in the numerical simulation of systems with many degrees of freedom—the problem of moving efficiently through a complicated, labyrinthine configuration space. The problem arises whether one wishes to study the system dynamically (where it makes itself felt as a disparity of time scale between the phenomena of interest and the time step needed to integrate the equations of motion [18]), or statistically (as in Monte Carlo work), or merely by seeking the global

energy minimum in a space filled with steep-sided curving valleys, saddle points, and spurious local minima [19, 20]. Judging from results in these other fields, the problem may be partly alleviated by transition algorithms that make intelligent use of local anisotropy information, but some complicated systems will remain intrinsically sluggish and hard to simulate.

APPENDIX: Model Ensembles

The acceptance ratio estimators of Sections IIb and IIc were tested using a pair of simple model ensembles, defined on a discrete configuration space having only 23 states. Because of the trivial configuration space the ensembles could be sampled by a simple Poisson routine (rather than by a Markov chain) and the free energy difference, Fermi expectations, and all other quantities of interest could be calculated exactly, for convenient comparison with the estimates under study. These estimates were obtained by applying the estimators (e.g., Eqs. (12a) and 12b)) to finite random samples of the two ensembles. The defining properties of the ensembles are given in Table I.

TABLE I

State	ΔU	$-\ln p_0$	$-\ln p_1$	State	ΔU	$-\ln p_0$	$-\ln p_1$
a	2	30.352	8.084	*m*	26	7.352	9.084
b	4	26.352	6.084	*n*	28	5.352	9.084
c	6	22.352	4.084	*o*	30	4.352	10.084
d	8	18.352	2.084	*p*	32	2.352	10.084
e	10	15.352	1.084	*q*	34	1.352	11.084
f	12	13.352	1.084	*r*	36	1.352	13.084
g	14	12.352	2.084	*s*	38	1.352	15.084
h	16	11.352	3.084	*t*	40	2.352	18.084
i	18	11.352	5.084	*u*	42	4.352	22.084
j	20	10.352	6.084	*v*	44	6.352	26.084
k	22	9.352	7.084	*w*	46	8.352	30.084
l	24	8.352	8.084				

$A_1 - A_0$ is 24.268; overlap (integral in Eq. (11)) is 1.2×10^{-8}.

Acknowledgments

I first learned of overlap methods, and of the statistical difficulty of estimating free energies from machine data, while working with Berni Alder and Mary Ann Mansigh in 1966–70. Discussions with Aneesur Rahman, John Barker, and Betty Flehinger, and particularly reading Valleau and Card's 1972 paper, stimulated my thinking.

REFERENCES

1. M. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER, *J. Chem. Phys.* **21** (1953), 1087.
2. W. W. WOOD, *in* "The Physics of Simple Liquids," (H. N. V. Temperley, J. S. Rowlinson, and G. S. Rushbrooke, Eds.), North-Holland, Amsterdam, 1968.
3. B. J. ALDER AND T. E. WAINWRIGHT, *J. Chem. Phys.* **31** (1959), 459; **33** (1960), 1439.
4. W. G. HOOVER AND F. H. REE, *J. Chem. Phys.* **47** (1967), 4873; **49** (1968), 3609.
5. J.-P. HANSEN AND L. VERLET, *Phys. Rev.* **184** (1969), 151.
6. G. TORRIE, J. P. VALLEAU, AND A. BAIN, *J. Chem. Phys.* **58** (1973), 5479.
7. G. TORRIE AND J. P. VALLEAU, *Chem. Phys. Lett.* **28** (1974), 578; *J. Comput. Phys.*, to be published.
8. D. R. SQUIRE AND W. G. HOOVER, *J. Chem. Phys.* **50** (1969), 701.
9. I. R. MCDONALD AND K. SINGER, *J. Chem. Phys.* **47** (1967), 4766.
10. C. H. BENNETT, *in* "Diffusion in Solids: Recent Developments" edited (A. S. Nowick and J. J. Burton, Eds.), Academic Press, New York, 1975.
11. R. W. ZWANZIG, *J. Chem. Phys.* **22** (1954), 1420.
12. W. R. SMITH, *in* "Statistical Mechanics 1" (K. Singer, Ed.), Chem. Soc. Publ., London, 1973.
13. Z. W. SALSBURG, J. D. JACOBSON, W. FICKETT, AND W. W. WOOD, *J. Chem. Phys.* **30** (1959), 65.
14. W. R. SMITH, *in* "Statistical Mechanics 1" (K. Singer, Ed.), pp. 98–99. Chem. Soc. Publ., London, 1973.
15. J. A. BARKER AND D. HENDERSON, *J. Chem. Phys.* **47** (1967), 2856 and 4714.
16. D. LEVESQUE AND L. VERLET, *Phys. Rev.* **182** (1969), 307.
17. J. P. VALLEAU AND D. N. CARD, *J. Chem. Phys.* **57** (1972), 5457.
18. C. H. BENNETT, *J. Comput. Phys.* **19** (1975), 267.
19. R. FLETCHER AND M. J. D. POWELL, *Comput. J.* **6** (1963), 163.
20. M. LEVITT AND A. WARSHEL, *Nature* **253** (1975), 694.