Analysis of the statistical error in umbrella sampling simulations by umbrella integration

Johannes Kästner\textsuperscript{a} and Walter Thiel\textsuperscript{b}

Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany

(Received 20 February 2006; accepted 28 April 2006; published online 20 June 2006)

Umbrella sampling simulations, or biased molecular dynamics, can be used to calculate the free-energy change of a chemical reaction. We investigate the sources of different sampling errors and derive approximate expressions for the statistical errors when using harmonic restraints and umbrella integration analysis. This leads to generally applicable rules for the choice of the bias potential and the sampling parameters. Numerical results for simulations on an analytical model potential are presented for validation. While the derivations are based on umbrella integration analysis, the final error estimate is evaluated from the raw simulation data, and it may therefore be generally applicable as indicated by tests using the weighted histogram analysis method. © 2006 American Institute of Physics. [DOI: 10.1063/1.2206775]

I. INTRODUCTION

Free energy is a fundamental concept in thermodynamics and physical chemistry. The most accurate methods to calculate free-energy differences include extensive sampling of the system.\textsuperscript{1,2} Applications range from the solid state over catalytic reactions and biochemical processes to rational drug design.

Two methods commonly applied to calculate the free energy change are biased molecular dynamics,\textsuperscript{3} also known as umbrella sampling, and thermodynamic integration.\textsuperscript{4} We have shown that thermodynamic integration along a reaction coordinate, which is a function $\xi(r)$ of atom positions only, can be regarded as a limiting case of umbrella sampling.\textsuperscript{3}

In umbrella sampling, the change of the free energy $A$ along the reaction coordinate $\xi$ is calculated. The reaction is divided into windows with a specific center $\xi_i$ assigned to each window. An additional energy term, a $\xi$-dependent bias, keeps the system in the vicinity of $\xi_i$. We consider only bias potentials in the form of harmonic restraints,

$$w_i(\xi) = \frac{1}{2}K(\xi - \xi_i)^2. \quad (1)$$

A simulation provides the distribution $P_i^b(\xi)$ of the biased system along the reaction coordinate in window $i$. In the original formulation of umbrella sampling, the unbiased free energy $A_i^u$ for each window is calculated directly from $P_i^b(\xi)$.

The different $A_i^u$ may be combined using the weighted histogram analysis method (WHAM).\textsuperscript{5,6} We found it advantageous\textsuperscript{5} to calculate the unbiased mean forces for each window:

$$\frac{\partial A_i^u}{\partial \xi} = -\frac{1}{\beta} \frac{\partial \ln P_i^b(\xi)}{\partial \xi} \frac{d\xi_i}{d\xi}, \quad (2)$$

which can then directly be averaged and integrated to yield $A(\xi)$. This method of analysis is called umbrella integration (UI).\textsuperscript{5}

We have shown and rationalized that it is well justified to approximate $P_i^b(\xi)$ by a normal distribution.\textsuperscript{5} Thus, only the mean $\xi_i^b$ and the variance $(\sigma_i^b)^2$ of $\xi$ have to be sampled in each window:

$$\frac{\partial A_i^u}{\partial \xi} = \frac{1}{\beta} \frac{1}{(\sigma_i^b)^2} - K(\xi - \xi_i^b). \quad (3)$$

The approximation of $P_i^b(\xi)$ by a normal distribution is equivalent to the truncation of a power series of $A(\xi)$ after the quadratic term in $\xi$.\textsuperscript{5} It is worth noting that the linear term of this power series is solely determined by $\xi_i^b$ while $(\sigma_i^b)^2$ defines the quadratic term. This can best be illustrated by integrating Eq. (3) for one window, and centering it around $\xi_i^b$, with $\xi' = \xi - \xi_i^b$:

$$A_i^u(\xi') = \frac{1}{2} \left( \frac{1}{\beta(\sigma_i^b)^2} - K \right) \xi'^2 + K(\xi_i^b - \xi_i^b)\xi' + C_i. \quad (4)$$

A further approximation that neglects the quadratic contribution has been proposed\textsuperscript{8-10} but will not be discussed here.

Statistical methods\textsuperscript{11} provide a well-defined error bar for the quantities $\xi_i^b$ and $(\sigma_i^b)^2$ that are needed in Eq. (3) as input data for umbrella integration. The purpose of this work is to show how this error is propagated through the analysis and thus to calculate the sampling error associated with this approach. Error estimation could also be done directly from the sampling data using the bootstrap algorithm,\textsuperscript{12} which would, however, give only one number per simulation without yielding any insight into the cause of the error. Our derivations will allow us to gain such insight and to choose the parameters of umbrella simulations such that the sampling error is minimized at a reasonable computational cost.
II. THEORY

In this section, we derive approximate expressions for the statistical error of umbrella sampling simulations that are analyzed by umbrella integration.\(^5\) While aiming at applicability to general systems, a simple analytic example potential is used to test and to validate our approximations. This system can be solved analytically and thousands of independent simulations can easily be run to calculate the mean statistical error as well as the systematic error with respect to the analytic solution. In the following, we first describe the example potential and then discuss methods for the evaluation of the sampling error.

A. Analytic example

We used the same analytic two-dimensional potential energy function as in our previous work:\(^5\) \(E(\xi, \phi) = f(\xi) + k(\xi) \phi^2\), with \(f(\xi) = -c \xi^2 + (c^2/4b) \xi^4\) and \(k(\xi) = k_{\text{min}} + 2db + c\sqrt{18d^2b}/c \xi + d \xi^2\). The reaction coordinate is \(\xi\).

This function has two minima with \(E = 0\) at \(\xi_{\text{min}} = \pm \sqrt{2b}/c\) and a barrier of height \(b\) at \(\xi = 0\). All stationary points are at \(\phi = 0\). In \(\phi\) direction, \(E(\xi, \phi)\) is broad at one minimum and narrow at the other minimum. Therefore, the minima have the same potential energy but a different free energy, \(A(\xi) = f(\xi) + \ln(k(\xi))/2\beta\).

We use atomic units (a. u.) throughout the paper unless explicitly specified otherwise. The parameters are assigned the following values. The barrier is \(b = 0.01\). Using \(c = 0.005\) places the minima at \(\xi_{\text{min}} = \pm 2\). \(d = 0.01\) and \(k_{\text{min}} = 0.01\) define the width of \(E\) in the direction of \(\phi\).

Biased Metropolis Monte Carlo\(^13\) (MC) simulations at \(T = 300\) K were performed on this system with a step size of 0.1 which led to typically 30\%-50\% accepted MC steps. After calculating 20000 steps to ensure that the trajectory is independent of the starting conditions, only every fifth step was used for analysis to reduce correlation between the steps.

Throughout this section we used 50 000 independent simulations with 40 windows in each simulation to average the sampling error with respect to the analytic solution.

B. Data collection in each window

The simulation of the biased system provides a trajectory of \(\xi_i\) whose mean value \(\xi_i\) and variance \((\sigma_i^2)\) need to be determined. There are various tests\(^11\) to decide if the system is well equilibrated with respect to \(\xi\). To check for a lack of trend in both the mean and the variance, we apply the Mann-Kendall test for trend\(^14\) which requires uncorrelated input data. Since the data from a molecular dynamics (MD) or MC trajectory are intrinsically correlated, we decorrelate them by collecting them into segments of a length that exceeds the correlation length.\(^11\) We usually start from 100 steps per segment and compute the mean and the variance of \(\xi\) in each segment. The Mann-Kendall test is applied on these segments. Whenever one of the tests (mean or variance) fails for a given trajectory, more and more steps from the beginning of the trajectory are dropped until the remaining data pass both tests. Only this part of the trajectory is used for the subsequent analysis.

Once a trend-free part of the trajectory is found, the Shapiro-Wilk test for normality\(^15\) as well as the von Neumann test for serial correlation\(^16\) are applied on the segmented values.\(^11\) If one test fails, the segment width is increased until all four tests pass. If this decreases the number of segments below 24, the tests are not powerful enough\(^11\) and we consider the system not to be equilibrated. Further simulations are necessary in this case.

This procedure ensures that the analyzed part of the trajectory of each window has no drift in \(\xi\) and is well equilibrated.\(^11\) It provides a correlation-independent estimate of the sampling errors associated with \(\xi_i\) and \((\sigma_i^2)\). The variance of the mean values of the segments gives an estimate of the variance of \(\xi_i\) caused by the sampling, var\((\xi_i)\). Likewise, the variance of \((\sigma_i^2)\) values of each segment provides an estimate of var\((\sigma_i^2)\).

Error propagation in Eq. (3) leads to

\[
\text{var}(\Delta A^u) = \frac{1}{\beta^2(\sigma_i^2)^4} \left( \text{var}(\xi_i^2) + \frac{(\xi_i^2 - \bar{\xi_i}^2)^2}{(\sigma_i^2)^4} \text{var}(\sigma_i^2) \right). \tag{5}
\]

This expression for the variance of the unbiased mean force in a given window can be evaluated using the segment-based estimates of the variances var\((\xi_i^2)\) and var\((\sigma_i^2)\) (see above).

A simpler alternative is to ignore the correlation between the steps, i.e., to use \(\text{var}(\xi_i^2) = (\sigma_i^2)^2/N_i\), with \(N_i\) being the number of steps in the stationary region of window \(i\). Assuming a normal distribution, one obtains\(^17,18\) \(\text{var}(\sigma_i^2) = 2(\sigma_i^2)^4/N_i\). With these simplifications, Eq. (5) becomes

\[
\text{var}(\Delta A^u) = \frac{2(\xi_i^2 - \bar{\xi_i}^2)^2}{N_i \beta^2(\sigma_i^2)^4}. \tag{6}
\]

Figure 1 compares the results from these two approximate expressions to the actually sampled variance for the window centered at \(\xi_i = -0.0897\). It is obvious that the sampled variance (solid black curve) is better reproduced by Eq. (5) than by Eq. (6) which underestimate the actual error typically by a factor of 2 or more. Equation (5) describes the

![Figure 1: Different methods to calculate var(ΔA^u)/δA](https://example.com/figure1.png)
sampled error in $\partial A_n^\mu/\partial \xi$ reasonably well near the mean $(\xi^2)$ of the distribution, while it deviates significantly at the tails of the distribution. This deviation has only a minor effect on the final error estimate, as the tails of the distribution are assigned a low weight when combining the windows (see below). The larger deviations far from $\xi^2$ ($|\xi| > 0.4$ in Fig. 1) may be caused by a breakdown of the assumption that a power series of $A_n(\xi)$ can be truncated after the quadratic term in $\xi$ which should be valid only for a limited range of $\xi$ around $\xi^2$.

C. Combining the windows

According to Eq. (3) the unbiased mean forces $\partial A_n^\mu/\partial \xi$ are linear functions of $\xi$ in each window. They are combined and averaged to a global solution $\partial A/\partial \xi$ on a grid in $\xi$, with the normalized weight
\begin{equation}
\frac{\partial A}{\partial \xi} = \sum_i p_i(\xi) \left( \frac{\partial A_n^\mu}{\partial \xi} \right),
\end{equation}

with $p_i(\xi) = N_i P_i^b(\xi) / \sum_i N_i P_i^b(\xi)$. $N_i$ is the number of steps used for window $i$. Thus, $\partial A_n^\mu/\partial \xi$ as well as the weights $p_i$ depend on the sampled distributions. In principle one would have to apply error propagation to both of these terms. However, the variations $\partial A_i^\mu/\partial \xi$, which are given by Eq. (5), are much larger than the variations in the weights. Thus we neglect the latter and obtain
\begin{equation}
\text{var} \left( \frac{\partial A}{\partial \xi} \right) = \sum_i p_i^2 \text{var} \left( \frac{\partial A_n^\mu}{\partial \xi} \right).
\end{equation}

Figure 2 illustrates that Eq. (9) in combination with Eq. (5) reproduces the sampled error in the mean force quite well. Even though the estimates of $\text{var}(\partial A/\partial \xi)$ from the individual simulations (green lines) contain some noise themselves, their magnitude agrees with the solid black reference curve in general. This is sufficient, as one will see in Eq. (15) that only an average of $\text{var}(\partial A/\partial \xi)$ over $\xi$ is required for the final error bar. The maxima between the windows and minima at the window means are also reproduced well, at least in the middle of the range of $\xi$ shown. For $|\xi| > 2.7$ both variances increase steeply and the sampled variance in Fig. 2 deviates significantly from our estimate. However, this occurs in the region outside of the sampling range and is thus irrelevant for our purposes.

D. Integration

The free-energy change $\Delta_{\xi_{\text{start}}-\xi_{\text{end}}} A$ is obtained by integrating $\partial A/\partial \xi$ over the interval $[\xi_{\text{start}}, \xi_{\text{end}}]$ on a grid provided by the bins, $(\partial A/\partial \xi)_k = (\partial A/\partial \xi)_{\xi_k}$, where $k$ numbers the bins. The total number of bins in the interval is $N_{\text{bins}}$. The integral according to Simpson’s rule can be written as
\begin{equation}
\Delta A = \int_{\xi_{\text{start}}}^{\xi_{\text{end}}} \frac{\partial A}{\partial \xi} d\xi = \frac{h}{3} \sum_{k=1}^{N_{\text{bins}}} C_k \left( \frac{\partial A}{\partial \xi} \right)_k,
\end{equation}

where $h$ is the bin width, $h = (\xi_{\text{end}} - \xi_{\text{start}})/(N_{\text{bins}} - 1)$, and $C_k = 1$ for $k=1$ and $k = N_{\text{bins}}$, otherwise $C_k = 2$ for odd $k$ and $C_k = 4$ for even $k$. $N_{\text{bins}}$ has to be odd.

The error contribution to $\Delta A$ caused by the uncertainty in $\partial A/\partial \xi$, see Eq. (9), can be calculated by error propagation. Neglecting a correlation between the bins, the variance of $\Delta A$ would be estimated by
\begin{equation}
\text{var}(\Delta A) = \text{var} \left( \frac{\partial A}{\partial \xi} \right) \frac{\langle \xi_b - \xi_k \rangle}{h},
\end{equation}

with $\text{var}(\partial A/\partial \xi)$ being the average of $\text{var}(\partial A/\partial \xi)$ over the interval $[\xi_{\text{start}}, \xi_{\text{end}}]$. Thus there would be a linear relation between the bin width $h$ and the variance of $\Delta A$.

Since there is a correlation between $\partial A/\partial \xi$ of different bins, the covariances expressing this correlation have to be taken into account:
\begin{equation}
\text{var}(\Delta A) = h^2 \sum_{k=1}^{N_{\text{bins}}} \sum_{l=1}^{N_{\text{bins}}} \text{cov} \left( \frac{\partial A}{\partial \xi} \right)_{kl},
\end{equation}

with the covariance defined as
\begin{equation}
\text{cov}(f_{kl}) = \langle f_k f_l \rangle - \langle f_k \rangle \langle f_l \rangle,
\end{equation}

where $f_k$ represents $\partial A/\partial \xi$ at $\xi = \xi_k$ and $\langle \cdot \rangle$ represents an average over the different simulations. The covariance can be accurately sampled using many independent simulations (Fig. 3).

Correlations between the errors in $\partial A/\partial \xi$ are caused by the broadness of the distributions of each window. All bins in a given window show a correlated error. The covariance was simulated for the analytic example, Fig. 3. It can be well approximated by a Gaussian with a width $\sigma_k^2$ of its nearest window:
\begin{equation}
\text{cov} \left( \frac{\partial A}{\partial \xi} \right)_{kl} = \text{var} \left( \frac{\partial A}{\partial \xi} \right)_k \exp \left[ -\frac{1}{2} \left( \frac{\xi_l - \xi_k}{\sigma_k} \right)^2 \right].
\end{equation}

We now assume that the width of the covariances is significantly smaller than the integration range $[\xi_{\text{start}}, \xi_{\text{end}}]$, implying that no covariance reaches out of the integration range at both sides. This will always be satisfied in practice, as the
integration has to be spread over many windows in order to reduce the sampling error. For a sufficiently large number of bins, the sums in Eq. (12) can be replaced by integrals, which leads to a particularly simple formula for the final error in $\Delta A$:

$$\text{var}(\Delta A) = \text{var} \left( \frac{\partial A}{\partial \xi} \right) \left[ (\xi_b - \xi_b) \sigma^b \sqrt{2\pi} - 2(\sigma^h)^2 \right],$$

with $\sigma^b$ being the average of $\sigma^h$ over the windows contributing to the interval. As the distributions of all windows in our system have similar widths, we use an average of all $\sigma^h$ for $\sigma^b$.

The error propagation in the integration step is dominated by the correlation within each window. When this correlation is considered, the error in $\Delta A$ is independent of the bin width ($h$) but depends on the broadness (width) of the distributions of the windows ($\sigma^h$). With typical values of $\sigma^h=0.1$ and $h=0.02$, Eq. (11) underestimates the error by about an order of magnitude compared with Eq. (15).

The integral, Eq. (10), is taken between the limits $\xi_a$ and $\xi_b$. As one is usually interested in free-energy differences between stationary points, these limits must satisfy $\partial A / \partial \xi |_{\xi_a} = 0$. The error in $\Delta A$ will in principle also depend on the accuracy of $\xi_a$ and $\xi_b$. The statistical error in the determination of $\xi_a$ and $\xi_b$ is, however, negligible for a reasonable number of bins. The roots of $\partial A / \partial \xi$ can easily be found as the function is given on a grid in $\xi$.

The error of Simpson’s rule $|e_i|$ is given by

$$|e_i| \leq \frac{(\xi_b - \xi_a)h^4}{180} |f^{(4)}(\xi^*)|,$$

where $|f^{(4)}(\xi^*)|$ is an upper limit of the absolute value of the fourth derivative of $\partial A / \partial \xi$ on the interval [$\xi_a, \xi_b$]. Numerical tests show that this error contribution is also negligibly small.

To summarize the error estimation proposed in this work, we use Eqs. (5), (9), and (15) and give results at the 95% confidence level in the form $\Delta A \pm 1.96 \sqrt{\text{var}(\Delta A)}$.

### III. DISCUSSION

In this section we address the dependence of the statistical errors in $\Delta A$ on the umbrella sampling parameters. We again employ the analytic potential function (Sec. II A) for numerical tests (5000 independent simulations for each data point, 40 windows, and $K=6\kappa$ unless noted otherwise, see text below). Since the analytic solution is available for our analytic example potential, we can discuss not only statistical but also systematic errors, which is not possible in general. Throughout this section, the estimated statistical error was obtained from Eqs. (5), (9), and (15); the sampled systematic error is $\Delta A - \Delta A_{\text{analytic}}$, the sampled statistical error is given as $\sqrt{(\Delta A)^2 - (\Delta A_{\text{analytic}})^2}$, and the sampled total error is $\sqrt{(\Delta A)^2 - (\Delta A_{\text{analytic}})^2}$.

#### A. Choice of parameters for umbrella sampling

A number of parameters must be defined for umbrella sampling simulations. Besides the form of the reaction coordinate, these are the strength of the bias ($K$), the number of windows to be sampled, the position of these windows ($\xi_i$), and the number of steps to be sampled in each window. For the analysis, one needs to define the number of bins, which are used for integration in U1, and for building the histograms, weighting and averaging in WHAM. Specific for umbrella integration is the segment width of the coarse-graining procedure described in Sec. II B. Knowledge of the sampling error can guide the choice of these parameters in order to minimize the sampling error at reasonable computational effort.

#### 1. Umbrella potential

The strength of the umbrella potential has to be decided before any window is simulated, preferably even before the system is equilibrated. In any form of umbrella sampling, the bias potential has to be strong enough to flatten the transition state in order to enable sampling of this region. That is, $A^\kappa$ should not exhibit maxima for any window. This can be fulfilled by choosing $K$ larger than $-\min(\partial^2 A / \partial \xi^2)$. In general, $-\min(\partial^2 A / \partial \xi^2)$ is often close to $-\partial E / \partial \xi_{TS}$ at the transition state. We introduce the auxiliary quantity $\kappa = -\partial^2 E / \partial \xi_{TS}^2$. $K > \kappa$ ensures proper sampling around the transition state. However, $\kappa$ is unknown initially. An estimate of $\kappa$ can be derived from the assumption of a sinusoidal $E(\xi)$:

$$\kappa = \frac{\pi^2}{2} \frac{E^\kappa}{(\Delta \xi)^2} \approx \frac{5}{2} \frac{E^\kappa}{(\Delta \xi)^2}.$$  

Here, $E^\kappa$ is the height of the (potential) energy barrier and $\Delta \xi$ is the difference in $\xi$ between the minimum and the transition state. This rough estimate may be used for any system as long as approximate values for $E^\kappa$ and $\Delta \xi$ are known. Otherwise they may be guessed, and preliminary umbrella sampling simulations around the transition state will then provide a more accurate value of $\kappa$. 

FIG. 3. The variance var($\partial A / \partial \xi$) over the whole reaction range (dots, top curve). The solid curve marked with circles shows the covariances of $\partial A / \partial \xi$ between the bin at $\xi = -2.0$ (labeled $k$) and a bin running over the range of $\xi$ (labeled $l$). The two other solid curves show the corresponding covariances $\text{cov}(\partial A / \partial \xi)_w$ with $\xi_a = 0.0$ and $\xi_b = 2.0$, respectively. The dashed lines are the approximations of $\text{cov}(\partial A / \partial \xi)_w$ according to Eq. (14).
The errors in $\Delta A$ for different values of $K$ in units of $\kappa$ are shown in Fig. 4. The estimated statistical error and the sampled statistical error agree quite well and are fairly constant over the range of $K$ shown in Fig. 4. For small $K$, $\Delta A$ is underestimated by the sampling which results in a negative systematic error. For $K=2\kappa$, the systematic error and the statistical error are similar in their absolute value and thus contribute equally to the total error. For $K\gg 5\kappa$, the systematic error becomes negligible. The sampled total error is the quadratic deviation of the sampled result from the analytic value. It is dominated by the systematic error for small $K$ and by the statistical error for large $K$. As a rule of thumb, choosing $K\approx 2\kappa$ seems advisable for umbrella integration. It should, however, not be chosen too high, as a higher $K$ requires more windows to be sampled, see below.

A global histogram (sum of all $P^i$) can be used to check the proper choice of $K$, Fig. 5. It is obvious that too small values of $K$ ($K<\kappa$) impede sampling around the transition state. On the other hand, too large values of $K$ result in nonoverlapping distributions of the individual windows. Choosing $K$ too large also increases the noise in $(a_i^b)^2$, and thus in the slopes of the individual $\partial A_i/\partial \xi_i$. This may be caused by a decoupling of $\xi$ from the other vibrational modes and thus by poor equilibration of $\xi$. WHAM analysis relies more on the overlap between the windows than umbrella integration, so that a smaller $K$ may generally be more favorable in WHAM.

It is possible to use different values of $K$ for different windows. For example, it can be advantageous to use larger values of $K$ around the transition state while using smaller values near minima. We will show in the next section that this also allows for fewer windows to be sampled in the region of the minima.

2. Number and range of windows

Overlap between the distributions of the windows is not strictly required for umbrella integration analysis, in contrast to WHAM. One can derive a free-energy profile of nonoverlapping windows by interpolating $\partial A_i/\partial \xi_i$. However, overlapping windows are advantageous in UI as the error decreases with increasing overlap. The width of the distribution of a window with constant or linear $\Lambda_i^b(\xi)$ is $1/\sqrt{\beta K}$ which thus provides an estimate for the average width of $a_i^b$. When simulating our example system with different barrier heights (parameter $b$) and different $K$, we found that the error significantly increases whenever the distance between the windows, $\Delta \xi = \xi_{i+1}^{\text{min}} - \xi_i^{\text{max}}$, becomes larger than $3/\sqrt{\beta K}$. This is illustrated in Fig. 6. Thus

$$\Delta \xi < 3/\sqrt{\beta K}$$

should be fulfilled, which is the case above around 20 windows in Fig. 6. Our expression for the statistical error underestimates the sampled total error when Eq. (18) is violated.

The steep increase of the total error once $\Delta \xi$ becomes larger than $3/\sqrt{\beta K}$ is caused by an increase of the systematic error which is not covered by our estimate and which probably arises to a large extent from the truncation of the power series in Eq. (3). While no conclusive assessment of systematic errors can be given in general, it can be expected that our
procedure will underestimate the sampling error whenever less windows than required by Eq. (18) are used.

Equation (18) relates the widths of the distributions (about $1/\sqrt{\beta K}$), causing the overlap between the windows, to the distance of the centers of the windows $\Delta \xi$ as these are chosen before the simulations are performed. The overlap, however, is more directly influenced by the distance of the window means $\bar{\xi}_{i+1}^b - \bar{\xi}_{i}^b$. Especially in the region of the transition state, $\frac{\Delta \xi}{\xi_{i+1} - \xi_{i}}$ is larger than $\Delta \xi$ as the system proceeds away from the transition state. To ensure sufficient overlap, more windows than required by Eq. (18) should be used whenever possible. One can, alternatively, follow a previous proposal and simulate the windows consecutively, using the width of the distribution of the previous window to define $\xi_{i+1}^b$ for the next window. For a constant number of total steps simulated in the production runs, the error decreases with increasing number of windows (and therefore less steps sampled per window).

In practice, we usually equilibrate the whole system either in one minimum without restraint or approximately at the transition state while applying a restraint. Then each window only needs a short equilibration with the restraint $w_i$ specific for that window. The tests described in Sec. II B are used to test the equilibration. Compared to the simulation time needed for equilibration, the production time is generally rather short.

The range of window centers $\xi_i^c$ is determined by the start and the end of the process under study. The range of $\xi_i^c$ should exceed the interval of the stationary points as this ensures accurate results for the latter. As particular interest rests on the stationary points it is advisable to place one window directly at the position of each such point that is known in advance, especially when the number of windows is small. If the range of $\xi_i^c$ turns out to be too small, additional windows can always be added.

3. Other parameters

The number of steps sampled per window can always be increased once a simulation is analyzed. The sampling error decreases approximately with the square root of the number of steps.

The segment width of the coarse graining procedure described in Sec. II B does not enter $\Delta A$. Its effect on the error bar is negligible: The final error in $\Delta A$ varies by less than 5% if the segment width is changed between 50 and 1000.

The integration in Sec. II D becomes more accurate with an increasing number of bins. Convergence is usually reached around 200–400 bins. To be on the safe side, we recommend the use of 400 bins between each stationary point of interest. In contrast to umbrella integration, WHAM does not converge with the number of bins: Too few bins result in increasing systematic error, while too many bins result in increasing statistical error.20,3

In Sec. II B we discussed the effect of the truncation of the power series of $A(\xi)$. This may become an issue if the windows span very broad ranges of $\xi$, i.e., for very few windows or very weak restraints. These cases, however, should be avoided anyway as the total error will then increase significantly, as shown in the last section. Truncation errors may also be estimated by comparing the results of umbrella integration with those of WHAM analysis of the same simulation data. The truncation of the power series in UI is expected to be acceptable for systems with smooth energy surfaces. Smooth curves $A(\xi)$ can well be described by a series of polynomials, which is effectively done in UI. Any steps or hystereses, i.e., nonsmooth energy surfaces, are usually a sign for a reaction coordinate that is not properly chosen. In this case, a different definition of the reaction coordinate should be considered.

B. General applicability of the error estimate

While our derivations are based on the umbrella integration analysis, the final error estimate is calculated from the sampled raw data alone. The same data can be used for WHAM analysis. We have performed WHAM analysis of simulations for different umbrella potentials, as shown in Fig. 4. WHAM iterations were converged until the free-energy constants changed by less than $10^{-5}E_h$ between two consecutive iterations. It is obvious from Fig. 4 that our UI-based error estimate also gives a realistic assessment of the statistical error in WHAM analysis, at least for the analytical model potential investigated. It does, however, not account for the increased statistical error in WHAM when the number of bins (currently 400) is increased too much.

C. Comparison with previous work

Previously, we used the same analytic potential as in this work. Sampling 40 windows and 20 000 steps in each window with $K=6\kappa$, we obtained errors in the barriers of 0.097 and $-0.136 \text{kJ/mol}$ compared with the analytic solution. Using Eqs. (5), (9), and (15) we can now estimate the associated error bars to be 0.31 and 0.32 $\text{kJ/mol}$ for the forward and backward barriers, respectively. For the more accurate sampling with 80 windows and 80 000 steps per window, we reported errors of $-0.013$ and $-0.035 \text{kJ/mol}$, while we now arrive at error bars of 0.11 $\text{kJ/mol}$ for both barriers. The current error bars for our previous results thus always cover the exact analytic solution and overestimate the actual error in all cases.

We have also applied the proposed error estimates to quantum mechanics/molecular mechanics (QM/MM) umbrella sampling calculations for an enzymatic reaction. As in our previous work, we consider the hydroxylation reaction in the enzyme $p$-hydroxybenzoate hydroxylase (PHBH). Details of the setup are described elsewhere. A confidence interval of $\Delta \Delta A=101.5\pm2 \text{kJ/mol}$ for the barrier derived from umbrella sampling matches well with the value of $101\pm2 \text{kJ/mol}$ obtained by thermodynamic integration with the same QM/MM methodology. For the exothermicity, we obtain $\Delta A=-208.1\pm3.4 \text{kJ/mol}$ with umbrella sampling, while thermodynamic integration gives $\Delta A=-212\pm2 \text{kJ/mol}$. This shows that our estimates for the error bar are consistent with those from thermodynamic integration.21

For illustration purposes, we have also carried out two umbrella integration analyses for the PHBH example using only every second of the 38 sampled windows (odd and
even, respectively). The number of windows is then smaller than required by Eq. (18), and we thus expect to underestimate the error bar, see Fig. 6. We obtain $\Delta^2 A = 100.7 \pm 3.7$ kJ/mol and $\Delta^2 A = 105.6 \pm 3.6$ kJ/mol for the barrier and $\Delta A = -206.7 \pm 6.3$ kJ/mol and $\Delta A = -207.7 \pm 6.4$ kJ/mol for the reaction energy. The error bars are almost twice as large as those obtained by integrating all windows (see above), but in one case ($\Delta^2 A = 105.6 \pm 3.6$ kJ/mol) the error bar is still too small to include the barrier determined from the full treatment ($\Delta^2 A = 101.5 \pm 2$ kJ/mol).

IV. CONCLUSION

Statistical tests ensure the equilibration of umbrella sampling simulations and provide error bars for the quantities that enter umbrella integration analysis. Equations (5), (9), and (15) give an estimate of the statistical error in the resulting free-energy difference. This allows us to find optimal parameters for umbrella sampling simulations. The force constant of the umbrella potential $K$ has to be larger than $-\frac{\partial^2 E}{\partial \xi^2}$ at the transition state. It should be at least twice this value. No clear upper limit for $K$ appears to be reasonable. Enough windows should be sampled to keep the distance between the windows $\Delta \xi < 3/\sqrt{8K}$, which requires more windows for higher $K$.

The accuracy of umbrella sampling simulations can thus be controlled by umbrella integration analysis. In view of this advantage and the stable convergence behavior, umbrella integration appears to be a promising tool for free-energy simulations.

ACKNOWLEDGMENT

This work was supported by the Deutsche Forschungsgemeinschaft, Grant No. SFB 663/4.

14 H. B. Mann, Econometrica 13, 245 (1945).