2.2 Local Error Decomposition for Trapezoidal Integration

Estimating the uncertainty of a calculated quantity is critically important in many applications.[74] The calculation of ΔG using integration based methods such as TI is no exception. In such calculations, the value of the ensemble average of the derivative of the potential energy with respect to a predefined pathway ($<\partial V/\partial \lambda>_\lambda$) is numerically integrated to produce a value of ΔG. Section 2.1 outlined a method based on the KS–statistic ($\mathrm{KS_{SE}}$ given by Eq. 14) which can be used to estimate the uncertainty in an ensemble average obtained by MD simulation. In this section we outline a general method for estimating the uncertainty of the Trapezoidal method (section 2.2.3) by considering both uncertainties in the points (section 2.2.1) as well as truncation error due to linear interpolation (section 2.2.2).

The definite integral of a function $f(x)$ represented by a vector of discrete points $\boldsymbol{y}$ can be estimated using the Trapezoidal rule:

$$\int_{x_0}^{x_N} f(x)dx \approx \frac{1}{2}\sum_{i=1}^{N}(y_i + y_{i-1})(x_i - x_{i-1}) \tag{21}$$

The Trapezoidal rule approximates the area under the curve between consecutive points using linear interpolation. Alternative numerical integration algorithms differ in the method of interpolation between points, Simpson's rule for example the interpolation is based on a 3rd degree polynomial. All numerical integration algorithms contain an inherent element of uncertainty associated with the method of interpolation. This is discussed in more detail in section 2.2.2. If the discrete points y_i are exact representations of $f(x_i)$, and these points are appropriately spaced in order to reproduce the highest non-zero derivatives, higher-order interpolation methods will yield more accurate results. If, however, $f(x_i)$ is only approximated by y_i, any uncertainty within the points is amplified by the use of higher-order methods. It is for this reason we chose to use the Trapezoidal rule for numerical integration calculations rather than high-order methods such as Simpson's rule. Note that as discussed in section 2.2.3, if truncation errors are allowed to cancel during the summation of local error contributions this effectively incorporates 2nd order contributions for functions that contain both concave and convex regions.

2.2.1 Trapezoidal Rule Point Uncertainty Propagation (Known Unknowns)

For the calculation of ΔG using integration based methods we have the case where $f(x)$ is approximated by a series of points y_i which have associated uncertainties σ_{yi}. We therefore would like to analytically calculate how these errors propagate through the Trapezoidal rule and contribute to the uncertainty of the integral. If we assume that the errors corresponding to each of the points (y_i ± σ_{yi}) are independent and normally distributed, we can then apply the simplified form of the general Gaussian error propagation formula[75]:

$$f(x \pm \sigma_x, y \pm \sigma_y, \ldots) = f(x, y, \ldots) \pm \sqrt{\left(\frac{\partial f}{\partial x}\sigma_x\right)^2 + \left(\frac{\partial f}{\partial y}\sigma_y\right)^2 + \ldots} \tag{22}$$

where f is an arbitrary function of variable ($x, y, \ldots$) with associated independent and normally distributed errors ($\sigma_x, \sigma_y, \ldots$). The error due to uncertainty in the points of applying the Trapezoidal rule can be obtained by applying Eq. 22 to Eq. 21. Which results in:

$$Error = \frac{1}{2}\sqrt{(x_1 - x_0)^2\sigma_{y_0}^2 + \sum_{i=1}^{N-1}(x_{i+1} - x_{i-1})^2\sigma_{y_i}^2 + (x_N - x_{N-1})^2\sigma_{y_N}^2} \tag{23}$$

Note that this expression for the total propagation error is a sum of contributions from each y_i point, it is therefore trivial to identify which points contribute most to the overall error.

2.2.2 Trapezoidal Rule Truncation Error (Unknown Unknowns)

When applying the Trapezoidal rule, the errors associated with the points $\boldsymbol{y}$ are a well-defined source of uncertainty. They contain specific information about the value of the function f at a given point x. Another source of error in all numerical integration methods results from implicit assumptions made regarding the form of the function between the evaluated data points i.e. the interpolation method. This source of error, often referred to as truncation error, is not so well-defined and therefore associated with an addition degree of uncertainty. This additional uncertainty is due to the assumption that the discrete points are sufficiently close together to accurately reproduce the highest non-zero derivative of the form of the function being integrated. However, in many applications the reason for using numerical integration is precisely because the form of the underlying function is unknown. Truncation error is therefore much more difficult to quantify accurately than the propagation of uncertainties in the points.

An error analysis of the Trapezoidal rule using a Taylor series expansion—which can be found in most numerical methods text books—results in a truncation error estimate for the interval between two point *a* and *b* given by:

$$Error \approx -\frac{(b-a)^3}{12} f''(\xi) \tag{24}$$

where $f''(\xi)$ is the second derivative over the interval [*a, b*]. The second derivative over this interval can be calculated numerically with either the forward, backward or central difference methods, and thus the 2nd order truncation error for each individual interval can be estimated. The primary assumption of this error estimate is that the calculation of the 2nd order term from the discrete points is accurate i.e. the calculation of $f''(\xi)$ by numerical differentiation. Which in turn is dependent on whether the discrete points being integrated are sufficiently close together to reproduce the underlying function.

2.2.3 Combined Trapezoidal Error Estimate

The combined estimate of the integration error associated with applying the Trapezoidal rule on a series of points which themselves contain errors is given by the sum of the results from the previous two sections, specifically:

$$\begin{aligned} Error =& \frac{1}{2}\sqrt{(x_1-x_0)^2\sigma_{y_0}^2 + \sum_{i=1}^{N-1}(x_{i+1}-x_{i-1})^2\sigma_{y_i}^2 + (x_N-x_{N-1})^2\sigma_{y_N}^2} \\ &+ \left|\sum_{i=0}^{N-1} -\frac{(x_{i+1}-x_i)^3}{12} f''(\xi)\right| \end{aligned} \tag{25}$$

Since the accuracy of the forward, backward and central difference methods for estimating $f''(\xi)$ over the interval $[x_{i+1}, x_i]$ are the same, the choice of method should have no systematic effect. However, unless the data points are perfectly symmetric they will give slightly different results. In practice, we calculate the sum of truncation errors with both forward and backward difference methods and take the larger of the two.

Note that the truncation error (given by Eq. 24) is not entirely independent of the point uncertainty error (Eq. 23) since the points themselves are used to estimate $f''(\xi)$ over the interval $[x_{i+1}, x_i]$. Using the technique applied in section 2.2.1, it is straight forward to propagate point uncertainties to obtain an error estimate in $f''(\xi)$. However, since some of the same points are used to estimate the uncertainty in neighbouring values of $f''(\xi)$, correlations between the errors of neighbouring intervals need to be accounted for. This becomes progressively more complex and the practical utility of considering such effects is unclear.

Summing the truncation errors and then taking the absolute value (Eq. 25) allows for cancelation of errors and therefore implicitly accounts for 2nd order contributions in some cases i.e. in cases where the function contain both concave and convex regions. The drawback of this approach however, is that given the uncertainty of the truncation error predications themselves, the errors can inappropriately cancel purely by chance and lead to significant underestimation of the total error. An alternative approach would be to add the absolute values of the individual interval truncation errors—which is the default behaviour for purely concave or convex functions since there can be no cancelation of errors—but this leads to significant over estimation of the error in typical cases. An initial analysis of the performance of Eq. 25 for estimating the error in solvation energy calculations using TI (details provided in the following section) found that in approximately 8% of cases the error was underestimated. While a failure of 8% may be adequate in many cases, in some applications, for example when parameterising force fields, a more robust estimate of the uncertainty is required.

2.2.4 Additional Uncertainty Heuristic

The addition of a simple heuristic to the integration uncertainty given by Eq. 25 dramatically reduced the underestimate rate without the significant increase in computational cost associated with not allowing for cancelation of errors i.e. summing the absolute value of the truncation error. The heuristic correction consists of adding the maximum interval truncation error found with either the forward or backward difference methods:

$$Error\ Heuristic = \max\{Er_i^{forward}, Er_i^{backward}\}_{i=0}^{N-1} \tag{26}$$

where $Eri^{forward}$ and $Eri^{\ backward}$ are the interval truncation errors calculated using Eq. 24 where $f''(\xi)$ over the interval $[x_{i+1}, x_i]$ is estimated using the forward and backward difference methods respectively. Thus a more robust Trapezoidal rule error estimate is given by:

$$
\begin{aligned}
Error =& \frac{1}{2}\sqrt{(x_1 - x_0)^2\sigma_{y_0}^2 + \sum_{i=1}^{N-1}(x_{i+1} - x_{i-1})^2\sigma_{y_i}^2 + (x_N - x_{N-1})^2\sigma_{y_N}^2} \\
&+ \max\left\{\left|\sum_{i=1}^{N-1} Er_i^{forward}\right|, \left|\sum_{i=1}^{N-1} Er_i^{backward}\right|\right\} \\
&+ \max\{Er_i^{forward}, Er_i^{backward}\}_{i=0}^{N-1}
\end{aligned}
\tag{27}
$$

An analysis on the relative performance of Eq. 25 and Eq. 27 is provided as a part of the following section. Eq. 27 was used for all the calculations of ΔG^{solv} by TI presented in Chapter 3, Chapter 6 and Chapter 7.