

Supplementary material (Ytreberg and Zuckerman)

Here we attempt to make precise the approximations made in applying black-box re-weighting (BBRW) when the observed density P^{obs} is estimated using a subset of the full set of coordinates. Note that, if the density is estimated using all coordinates, then no approximation is made in the sense that the final weights will be exact in a suitable limit – e.g., dense sampling, even if biased, when using nearest-neighbor distances. As will be implicit in the discussion below, a subset of *collective* coordinates may also be used so long as they reproduce the density corresponding to the Cartesian-measure ($dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 \dots$). We also suggest that avenues for systematic improvement of our work are possible.

We will assume that the desired distribution, P , is given by the Boltzmann distribution, i.e., $P(\mathbf{r}) \propto \exp[-\beta U(\mathbf{r})]$, where \mathbf{r} is the full set of coordinates.

Density estimation using a subset of coordinates can be understood in the context of a potential of mean force (PMF). Specifically, we will write the PMF in terms a subset of coordinates, \mathbf{R} , deduced from the full set of coordinates $\mathbf{r} = (\mathbf{R}, \mathbf{x})$. Presumably \mathbf{R} constitutes the most important, or the set of “slow” coordinates, but such an assumption is not formally necessary. We consider a slightly modified PMF to emphasize the \mathbf{x} integrals and their dimensionality, as reflected in “effective configuration-space volumes,” v ,

The Boltzmann factor of the PMF, as usual, gives the probability density p of non-integrated coordinates (\mathbf{R} , in this case), and one has

$$p(\mathbf{R}) \propto \int d\mathbf{x} e^{-\beta U(\mathbf{R}, \mathbf{x})} \equiv v_x^{\text{ref}} e^{-\beta \text{PMF}(\mathbf{R})}. \quad [\text{S1}]$$

We are using lower-case p for probability here to draw attention to the reduced set of coordinates in the argument. The volumetric factor v_x^{ref} with dimensions of \mathbf{x} is a constant for all \mathbf{R} . Below we will use Eq. [S1] to determine the nature of the approximations associated with the two strategies employed in the paper: binning and nearest-neighbors.

When the observed density has been estimated in the reduced \mathbf{R} space, one requires a modified version of the fundamental black-box relation (Eq. [1] in the paper). Specifically, for a configuration j , i.e., $(\mathbf{R}_j, \mathbf{x}_j)$, one has

$$W^{\text{bb}}(j) = \frac{p(\mathbf{R}_j)}{p^{\text{obs}}(\mathbf{R}_j)}. \quad [\text{S2}]$$

To see that this expression is correct, assume that n_j configurations have been generated (i.e., are “observed”) in an small volume $d\mathbf{R}$ around the point \mathbf{R}_j . Then, we estimate $p^{\text{obs}}(\mathbf{R}_j) \approx n_j/d\mathbf{R}$ for each configuration in that volume. Further, the PMF is approximately constant over the volume and $p(\mathbf{R}) \approx p(\mathbf{R}_j)$ for all the configurations.

Therefore each weight, W^{bb} , is approximately constant within the volume $d\mathbf{R}$. Now consider the total weight in the volume; this is the sum over all n_j of the W^{bb} values which

leads to the cancellation of n_j . The total weight in the volume is therefore proportional to $p(\mathbf{R}_j)$, as demanded by statistical mechanics. Note that we have assumed the limiting case of dense sampling, which is the correct limit for checking the mathematics. The use of the PMF in the numerator implies that weights indeed depend, in principle, on the excluded coordinates \mathbf{x} .

Because Eq. [S2] is a formally exact expression, it should be used regardless of whether density estimation (in \mathbf{R} space) is performed based on binning, nearest neighbors, or some other method. Further the motivation for using the \mathbf{R} -space relation is identical to that presented in the paper: the \mathbf{R} coordinates will *not* be assumed to be distributed properly, and therefore BBRW will be used to deduce (i.e., to correct) the distribution over \mathbf{R} .

Here we want to understand the precise nature of the approximations made in the paper – i.e., those made by using regular Boltzmann factors in Eq. [S2] (i.e., by assuming $p(j) \approx P(j) = \exp[-\beta U(j)]$) rather than Boltzmann factors of the PMF, as required by substituting Eq. [S1] in [S2].

Approximation used in estimating density by binning on a subset of coordinates

First, how much weight should be assigned to a bin in \mathbf{R} space, based on the exact PMF formalism? The weight should be proportional to the integral of Eq. [S1] over the volume of bin i , which yields the local partition function, z_i , namely,

$$z_i \equiv \int_{\text{bin } i} d\mathbf{R} \int d\mathbf{x} e^{-\beta U(\mathbf{R}, \mathbf{x})} \equiv v_x^{\text{ref}} \int_{\text{bin } i} d\mathbf{R} e^{-\beta \text{PMF}(\mathbf{R})}. \quad [\text{S3}]$$

Equation [S3] gives the total weight in a bin using our exact PMF formalism.

We now describe the approximation used in the paper for binning (Eq. [5]) by considering the total amount of weight assigned to a given bin by the BBRW procedure. (Note that, ultimately, the sum of these bin totals determines the weight of the whole state and the associated relative free energy.) If one substitutes (in the paper) Eq. [5] into Eq. [1] and sums over all weights in a given bin i , it is seen that the total weight is just \bar{P}_i , the average Boltzmann factor for configurations in the bin. Therefore, as stated in the paper, the approximation for the binning approach is that z_i is assumed proportional to \bar{P}_i .

Such an estimate based solely on Boltzmann factors does not account for the entropy *within* a bin (although it is critical to note that BBRW automatically accounts for entropy in \mathbf{R} space). While this may be a poor approximation for individual \mathbf{R} values, our results suggest it may be quite reasonable when many configurations are averaged over to produce a free-energy difference estimate for two large states. That is, the average entropy in a well-chosen set of fast coordinates may be quite similar from one state to another. Nevertheless, we believe that including the local entropy in the fast coordinates is important for the future, and can be handled as sketched below in the ‘cleaner formalism.’

Approximation used in the nearest-neighbors strategy.

When density is estimated by distances to near neighbors, as in Eq. [6] of the paper, the net result is quite similar to that for the binning approach. In essence, Boltzmann factors are again averaged, instead of a rigorous estimation of the PMF. This can be seen by considering a small volume in \mathbf{R} space. As above, the observed density estimates will be similar for each structure, but the numerators of Eq. [S2] are approximated by ordinary Boltzmann factors of the potential, U . Hence the total weight in such a volume will be approximately equal to the average Boltzmann factor of nearby structures. The fundamental similarity to the binning case suggests that entropy in the fast coordinates is not properly accounted for in the nearest-neighbors strategy..

Thus, for both the nearest-neighbors and binning strategies, the approximation made is that the local partition function is given by \overline{P} , rather than by Eq. [S3]. Qualitatively, variations in local entropy are ignored.

A cleaner formalism for future work.

A difficulty of the approximations described above is that there is no obvious way to improve them systematically. We therefore describe a slightly different approach, which employs the Boltzmann factor of the average energy (as opposed to the average Boltzmann factor). Here we make the reasonable assumption that configurations can be properly sampled *locally* – i.e., that local canonical averages can be estimated in the usual way by averaging over configurations without any kind of weighting.

Consider re-writing the basic PMF relation [S1] by separating off the locally averaged energy:

$$v_x^{\text{ref}} e^{-\beta \text{PMF}(\mathbf{R})} = \int d\mathbf{x} e^{-\beta U(\mathbf{R}, \mathbf{x})} \equiv v_x^{\text{eff}}(\mathbf{R}) \cdot e^{-\beta \langle U(\mathbf{R}, \mathbf{x}) \rangle_{\text{fixed } \mathbf{R}}} . \quad [\text{S4}]$$

This *defines* an \mathbf{R} -dependent effective volume in \mathbf{x} space, $v_x^{\text{eff}}(\mathbf{R})$. However, because the average energy has been separated off in a precise way, and also because the PMF is a free energy, the local entropy, $S_x(\mathbf{R})$ can be defined via the relation $v_x^{\text{eff}}(\mathbf{R}) \propto e^{+S_x(\mathbf{R})/k_B}$. Equation [S4] suggests two approximation schemes: (i) crudely ignore the local entropy by assuming $v_x^{\text{eff}}(\mathbf{R}) = \text{const}$, and hence estimate the PMF by the average energy; and (ii) employ an estimator for $S_x(\mathbf{R})$ based on the local ensemble.

We anticipate scheme (ii) will work particularly well if “fast” coordinates \mathbf{x} are chosen on an automated basis – i.e., via principal-components or a similar analysis. Further, we have already tried scheme (i) with binning, and it works well, but in a smaller region of bin sizes than the average Boltzmann factor approach reported in the paper.