Free Energies, Phase Space Overlap, and Dissipated Work

1 Free energies and overlaps

[**?**] The difference in free energies between two systems involves the relationship between their corresponding regions ("subspaces") of phase space, and to *measure* this free energy difference it is necessary to consider both. These two regions may be wildly different, which gives rise to the *overlap problem*: a single simulation ensemble needs to be generated in which both regions of phase space are adequately represented.

To understand the performance of a free-energy calculation method it is therefore necessary to know something about how much overlap there is between the important regions of phase space for the two systems. The free energy difference ΔF by itself says nothing about this: systems can have the same ΔF but radically different overlap properties.

In particular, density-of-states approaches and work-based approaches form the two main pillars of equilibrium free energy methods. Focusing on work-based approaches built on Jarzynski's nonequilibrium work (NEW) formalism (encompassing FEP and TI as special or limiting cases), the pitfall is that they are generally prone to systematic errors (bias) when phase space overlap is poor. Applying them usually means efficiency is sacrificed in order to gain a result that can be considered accurate and converged.

Consider two systems A and B with potential energies U_A and U_B. For a given configuration γ drawn from the equilibrium ensemble of A, the "instantaneous switching" work

$$W_{A\to B}(\gamma) = U_B(\gamma) - U_A(\gamma) \tag{1}$$

is the work done when we perturb that configuration from system A to system B. There is a reverse version $W_{B\to A} = U_A - U_B$ as well.

Jarzynski's identity states that the equilibrium free energy difference can be obtained from an exponential average over nonequilibrium work values:

$$e^{-\beta\Delta F_{A\to B}} = \left\langle e^{-\beta W_{A\to B}} \right\rangle_A \tag{2}$$

where the average is over trajectories (or instantaneous switches) initiated in equilibrium with system A. Equivalently,

$$\beta\Delta F_{A\to B} = -\ln \left\langle e^{-\beta W_{A\to B}} \right\rangle_A . \tag{3}$$

This form highlights that the Boltzmann factor corresponding to ΔF is the *average* of $e^{-\beta W}$ evaluated over the relevant ensemble.

2 Quantifying phase space overlaps

As a simple analytical example, consider two N-dimensional harmonic systems

$$U_A(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_A x_i^2, \tag{4}$$

$$U_B(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_B (x_i - x_0)^2, \tag{5}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)$, and $\omega_A, \omega_B > 0$ are force constants. The particles (coordinates) are uncorrelated.

Analytically, the free energy difference is independent of the shift x_0 and given by

$$\Delta F_{A\to B} = \frac{1}{2} N k_B T \ln\left(\frac{\omega_B}{\omega_A}\right). \tag{6}$$

Thus ΔF cannot tell us anything about how far the wells are shifted (how much the corresponding phase space regions overlap), only about the relative "stiffness" of the two harmonic basins.

3 Why consider phase space overlaps

Work-based approaches are prone to *bias*, especially when the overlap between the important regions of phase space for A and B is poor. To make this precise, we can relate phase space overlap to dissipated work and relative entropy.

Suppose we perform M nonequilibrium work measurements in the $A \to B$ direction, obtaining $\{W_{i,A\to B}\}_{i=1}^{M}$, and M' measurements in the $B \to A$ direction, obtaining $\{W_{i,B\to A}\}_{i=1}^{M'}$. Define the sample averages

$$\bar{W}_{A\to B} = \frac{1}{M} \sum_{i=1}^{M} W_{i,A\to B}, \tag{7}$$

$$\bar{W}_{B\to A} = \frac{1}{M'} \sum_{i=1}^{M'} W_{i,B\to A}. \tag{8}$$

Using Jarzynski's identity, we estimate the free energy differences

$$\beta \Delta F_{A\to B} = -\ln\left[\frac{1}{M} \sum_{i=1}^{M} \exp\left(-\beta W_{i,A\to B}\right)\right], \tag{9}$$

and similarly

$$\beta \Delta F_{B\to A} = -\ln\left[\frac{1}{M'} \sum_{i=1}^{M'} \exp\left(-\beta W_{i,B\to A}\right)\right]. \tag{10}$$

Note that $\Delta F_{B\to A} = -\Delta F_{A\to B}$ in exact equilibrium.

We can now define (dimensionless) quantities

$$s_A = \beta \bar{W}_{A\to B} - \beta \Delta F_{B\to A}, \tag{11}$$

$$s_B = \beta \bar{W}_{B\to A} - \beta \Delta F_{A\to B}, \tag{12}$$

which will be related to relative entropies between the underlying equilibrium distributions.

Let $p_A(\gamma)$ and $p_B(\gamma)$ be the equilibrium probability densities on phase space Γ for systems A and B, respectively. The (Gibbs) relative entropies

$$s_A = \int_\Gamma d\gamma \, p_A(\gamma) \ln \left[\frac{p_A(\gamma)}{p_B(\gamma)} \right], \tag{13}$$

$$s_B = \int_\Gamma d\gamma \, p_B(\gamma) \ln \left[\frac{p_B(\gamma)}{p_A(\gamma)} \right], \tag{14}$$

quantify how different the two ensembles are. These are difficult to compute directly because they require accurate sampling of *both* distributions over all of phase space.

Using

$$p_A(\gamma) = \frac{e^{-\beta U_A(\gamma)}}{Q_A}, \qquad p_B(\gamma) = \frac{e^{-\beta U_B(\gamma)}}{Q_B}, \tag{15}$$

with Q_A and Q_B the partition functions, one can show that these relative entropies can be written in terms of averages of work and free energies. For example, one finds (schematically)

$$s_A = \beta \left(\langle W_{A\to B} \rangle_A - \Delta F_{A\to B} \right), \tag{16}$$

and similarly

$$s_B = \beta \left(\langle W_{B\to A} \rangle_B - \Delta F_{B\to A} \right). \tag{17}$$

In terms of sample estimates,

$$s_B \approx \beta \bar{W}_{B\to A} - \beta \Delta F_{B\to A}, \tag{18}$$

and so on (up to statistical error).

Large values of s_A and s_B mean that p_A and p_B are very different (little overlap); the important regions of phase space for one ensemble have negligible weight in the other. In such cases,

- the effective free energy "barrier" between the ensembles is large,
- sampling becomes dominated by rare events (entropic barriers),
- nonequilibrium work estimators become highly biased and converge very slowly.

If the two relative entropies are finite and not both extremely large, the ensemble with smaller s has a smaller important phase space region (fewer dominating configurations), and that region will tend to be a subset of the dominant region of the other ensemble. When both s_A and s_B become very large, the dominant regions may become almost disjoint.

Recall Jarzynski's inequality, which follows from convexity:

$$\left\langle e^{-\beta W} \right\rangle = e^{-\beta \Delta F} \quad \Rightarrow \quad \langle W \rangle \geq \Delta F. \tag{19}$$

The equality corresponds to a perfectly reversible process, while the strict inequality reflects **dissipated work**,

$$W_{\text{diss}} = \langle W \rangle - \Delta F \geq 0. \tag{20}$$

The relations above show that the relative entropies s_A and s_B can be expressed in terms of such dissipated work for the forward and reverse processes. Thus, knowledge of the dissipated work directly informs us about the degree of phase space overlap.

4 From phase space overlaps to coordinate overlaps

Phase space overlap is a statement about the similarity of the full distributions $p_A(\gamma)$ and $p_B(\gamma)$, including all coordinates and momenta. In practice, we often want to reason instead in terms of *coordinate overlap*: how similar are the typical configurations $\boldsymbol{x}$ sampled from the two ensembles?

Intuitively:

- If two phase space distributions overlap strongly, then their marginal distributions in configuration space (the distributions of $\boldsymbol{x}$ alone) must also overlap strongly. Typical configurations for A and B look similar.
- Conversely, if the phase space distributions have very small overlap, then almost all configurations typical of A are highly improbable under B, and vice versa. In coordinate space this manifests as little or no overlap between the regions of configuration space that contribute most to each ensemble.

Formally, one obtains the configuration-space distributions by integrating out momenta:

$$\rho_A(\boldsymbol{x}) = \int d\boldsymbol{p}\, p_A(\boldsymbol{x}, \boldsymbol{p}), \qquad \rho_B(\boldsymbol{x}) = \int d\boldsymbol{p}\, p_B(\boldsymbol{x}, \boldsymbol{p}). \tag{21}$$

The same relative entropy constructions can be applied to ρ_A and ρ_B if desired, and in many classical molecular simulations the momentum contributions factor out, so that overlap properties are controlled by the potential energy landscape and the distribution over coordinates.

In practice, we often:

- simulate in full phase space,
- project trajectories down to coordinates (e.g., atomic positions, collective variables),
- and use these projected distributions to reason about overlap and sampling efficiency.

Thus, the "phase space overlap problem" is usually experienced in terms of *coordinate* overlap: whether the simulation explores regions of configuration space that are simultaneously important for both states A and B. Quantities like the relative entropies s_A and s_B, expressible through dissipated work, provide a rigorous way to quantify this overlap and to diagnose when work-based free energy methods will be efficient versus when they are likely to suffer from severe bias.

References

[1] (Placeholder) Authors, *Title about phase space overlap and free energy calculations*, Journal / arXiv (Year).