SUPPLEMENTAL MATERIAL: DYNAMICAL COMPUTATION OF THE DENSITY OF STATES AND BAYES FACTORS USING NONEQUILIBRIUM IMPORTANCE SAMPLING

GRANT M. ROTSKOFF AND ERIC VANDEN-EIJNDEN

1. **Derivation of Eqs. (6) and (7).** To derive Eq. (6) for the density $\rho_{\rm ne}(\boldsymbol{x})$, consider first the forward time trajectories alone and let us define the forward density $\rho^+_{\rm ne}(\boldsymbol{x})$ via

$$\begin{aligned}\langle\phi\rangle^+_{\rm ne} &= \frac{1}{\langle\tau^+\rangle}\int_\Omega\int_0^{\tau^+(\boldsymbol{x})}\phi(\boldsymbol{X}(t,\boldsymbol{x}))dt\,\rho(\boldsymbol{x})d\boldsymbol{x}\\ &\equiv \int_\Omega \phi(\boldsymbol{x})\rho^+_{\rm ne}(\boldsymbol{x})d\boldsymbol{x}\end{aligned} \tag{SM.1}$$

The nonequilibrium density $\rho^+_{\rm ne}(\boldsymbol{x})$ satisfies the stationary Liouville equation

$$\langle\tau^+\rangle^{-1}\rho(\boldsymbol{x}) = \nabla\cdot\left(\boldsymbol{b}(\boldsymbol{x})\rho^+_{\rm ne}(\boldsymbol{x})\right), \tag{SM.2}$$

with the boundary condition $\rho^+_{\rm ne}(\boldsymbol{x})=0$ on the regions of $\partial\Omega$ where $\boldsymbol{b}(\boldsymbol{x})$ points inward (since, by construction, no mass can be transported there forward in time). Physically, this equation asserts that, at stationarity, the nonequilibrium probability flux out of a small volume in the vicinity of $\boldsymbol{x}$ is balanced by the rate of reinjection. To solve (SM.2) notice that if we differentiate $\rho^+_{\rm ne}(\boldsymbol{X}(t,\boldsymbol{x}))$ with respect to time, by the chain rule we have

$$\frac{d}{dt}\rho^+_{\rm ne} = \boldsymbol{b}\cdot\nabla\rho^+_{\rm ne} = \langle\tau^+\rangle^{-1}\rho - (\nabla\cdot\boldsymbol{b})\rho^+_{\rm ne} \tag{SM.3}$$

where all functions are evaluated at $\boldsymbol{X}(t,\boldsymbol{x})$ and we used Eq. (1) for $\boldsymbol{X}(t,\boldsymbol{x})$ to derive the first equality and (SM.2) to derive the second. Using the Jacobian (Eq. (5)),

$$J(t,\boldsymbol{x}) = \exp\left(\int_0^t \nabla\cdot\boldsymbol{b}(\boldsymbol{X}(s,\boldsymbol{x}))ds\right), \tag{SM.4}$$

we can write (SM.3) as

$$\frac{d}{dt}\left(\rho^+_{\rm ne}(\boldsymbol{X}(t,\boldsymbol{x}))J(t,\boldsymbol{x})\right) = \langle\tau^+\rangle^{-1}\rho(\boldsymbol{X}(t,\boldsymbol{x}))J(t,\boldsymbol{x}). \tag{SM.5}$$

Integration from $t=\tau^-(\boldsymbol{x})$ to $t=0$ using the boundary condition $\rho^+_{\rm ne}(\boldsymbol{X}(\tau^-(\boldsymbol{x}),\boldsymbol{x}))=0$ gives

$$\rho^+_{\rm ne}(\boldsymbol{x}) = \langle\tau^+\rangle^{-1}\int_{\tau^-(\boldsymbol{x})}^0 J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))dt. \tag{SM.6}$$

Date: February 26, 2019.

A similar calculation gives the nonequilbrium stationary density resulting from the reverse time propagation of the dynamics, $\rho^-_{\rm ne}(\boldsymbol{x})$, defined via

$$\begin{aligned}\langle\phi\rangle^-_{\rm ne} &= \frac{1}{\langle\tau^-\rangle}\int_\Omega\int_{\tau^-(\boldsymbol{x})}^0 \phi(\boldsymbol{X}(t,\boldsymbol{x}))dt\,\rho(\boldsymbol{x})d\boldsymbol{x}\\ &\equiv \int_\Omega \phi(\boldsymbol{x})\rho^-_{\rm ne}(\boldsymbol{x})d\boldsymbol{x}\end{aligned} \tag{SM.7}$$

We obtain

$$\rho^-_{\rm ne}(\boldsymbol{x}) = \langle\tau^-\rangle^{-1}\int_0^{\tau^+(\boldsymbol{x})} J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))dt. \tag{SM.8}$$

The nonequilibrium density defined via Eq. (4) can then be expressed by superposing $\rho^+_{\rm ne}(\boldsymbol{x})$ and $\rho^-_{\rm ne}(\boldsymbol{x})$:

$$\begin{aligned}\rho_{\rm ne}(\boldsymbol{x}) &= \frac{\langle\tau^+\rangle\rho^+_{\rm ne}(\boldsymbol{x}) - \langle\tau^-\rangle\rho^-_{\rm ne}(\boldsymbol{x})}{\langle\tau\rangle}\\ &= \langle\tau\rangle^{-1}\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))dt\end{aligned} \tag{SM.9}$$

This is Eq. (6).

To derive Eq. (7), start from Eq. (2) and write

$$\begin{aligned}\langle\phi\rangle &= \langle\phi\rho/\rho_{\rm ne}\rangle_{\rm ne}\\ &= \frac{1}{\langle\tau\rangle}\int_{\mathbb{R}^d}\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} \frac{\phi(\boldsymbol{X}(t,\boldsymbol{x}))\rho(\boldsymbol{X}(t,\boldsymbol{x}))}{\rho_{\rm ne}(\boldsymbol{X}(t,\boldsymbol{x}))}dt\,\rho(\boldsymbol{x})d\boldsymbol{x}\\ &= \int_{\mathbb{R}^d}\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} \frac{\phi(\boldsymbol{X}(t,\boldsymbol{x}))\rho(\boldsymbol{X}(t,\boldsymbol{x}))}{\int_{\tau^-(\boldsymbol{X}(t,\boldsymbol{x}))}^{\tau^+(\boldsymbol{X}(t,\boldsymbol{x}))} J(s,\boldsymbol{X}(t,\boldsymbol{x}))\rho(\boldsymbol{X}(s,\boldsymbol{X}(t,\boldsymbol{x})))ds}dt\,\rho(\boldsymbol{x})d\boldsymbol{x}.\end{aligned} \tag{SM.10}$$

Using $\boldsymbol{X}(s,\boldsymbol{X}(t,\boldsymbol{x})) = \boldsymbol{X}(s+t,\boldsymbol{x})$ we have

$$\begin{aligned}J(s,\boldsymbol{X}(t,\boldsymbol{x})) &= \exp\left(\textstyle\int_0^s \nabla\cdot\boldsymbol{b}(\boldsymbol{X}(u,\boldsymbol{X}(t,\boldsymbol{x})))du\right)\\ &= \exp\left(\textstyle\int_0^s \nabla\cdot\boldsymbol{b}(\boldsymbol{X}(u+t,\boldsymbol{x}))du\right)\\ &= \exp\left(\textstyle\int_t^{s+t} \nabla\cdot\boldsymbol{b}(\boldsymbol{X}(u',\boldsymbol{x}))du'\right)\\ &= \frac{J(s+t,\boldsymbol{x})}{J(t,\boldsymbol{x})}\end{aligned} \tag{SM.11}$$

which implies that we can transform the denominator in (SM.10) into

$$\begin{aligned}&\int_{\tau^-(\boldsymbol{X}(t,\boldsymbol{x}))}^{\tau^+(\boldsymbol{X}(t,\boldsymbol{x}))} J(s,\boldsymbol{X}(t,\boldsymbol{x}))\rho(\boldsymbol{X}(s,\boldsymbol{X}(t,\boldsymbol{x})))ds\\ =&\frac{1}{J(t,\boldsymbol{x})}\int_{\tau^-(\boldsymbol{X}(t,\boldsymbol{x}))}^{\tau^+(\boldsymbol{X}(t,\boldsymbol{x}))} J(s+t,\boldsymbol{x})\rho(\boldsymbol{X}(s+t,\boldsymbol{x})))ds\\ =&\frac{1}{J(t,\boldsymbol{x})}\int_{\tau^-(\boldsymbol{X}(t,\boldsymbol{x}))+t}^{\tau^+(\boldsymbol{X}(t,\boldsymbol{x}))+t} J(s',\boldsymbol{x})\rho(\boldsymbol{X}(s',\boldsymbol{x})))ds'\\ =&\frac{1}{J(t,\boldsymbol{x})}\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(s',\boldsymbol{x})\rho(\boldsymbol{X}(s',\boldsymbol{x})))ds'\end{aligned} \tag{SM.12}$$

Inserting this expression in (SM.10) gives

$$\begin{aligned}\langle\phi\rangle &= \int_{\mathbb{R}^d}\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} \frac{\phi(\boldsymbol{X}(t,\boldsymbol{x}))J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))}{\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(s',\boldsymbol{x})\rho(\boldsymbol{X}(s',\boldsymbol{x}))ds'}dt\,\rho(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} \frac{\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} \phi(\boldsymbol{X}(t,\boldsymbol{x}))J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))dt}{\int_{\tau^-(\boldsymbol{x})}^{\tau^+(\boldsymbol{x})} J(t,\boldsymbol{x})\rho(\boldsymbol{X}(t,\boldsymbol{x}))dt}\,\rho(\boldsymbol{x})d\boldsymbol{x}\end{aligned} \tag{SM.13}$$

which is Eq. (7).

2. **Comparison with Neal's Annealed Importance Sampling (AIS) method.** While our method is conceptually similar to AIS, the estimator (13) has important philosophical and practical differences from AIS, which we highlight here.

In AIS, one defines a sequence of distributions in order to estimate an expectation with respect to a target density ρ_1, usually known only up to a normalization factor, i.e. we know some $\hat{\rho}_1(\boldsymbol{x})$ can be evaluated pointwise, but the factor $Z_1 = \int_\Omega \hat{\rho}_1(\boldsymbol{x})d\boldsymbol{x}$ needed to get the normalized density $\rho_1(\boldsymbol{x}) = Z_1^{-1}\hat{\rho}_1(\boldsymbol{x})$ is not known. In practice, the sampling is done by transporting from an initial density ρ using a nonequilibrium dynamics: ρ can be sampled e.g. using Metropolis-Hastings Monte-Carlo, and is also only known up to a normalization factor in general. This transport defines a nonequilibrium density $\rho_{\mathrm{ne}}(\boldsymbol{x})$, also known only up to a normalization factor, so that the annealed importance sampling scheme for the expectation of an observable ϕ is

$$\langle\phi\rangle_{\rho_1} = \frac{\langle\phi\hat{\rho}_1/\hat{\rho}_{\mathrm{ne}}\rangle_{\mathrm{ne}}}{\langle\hat{\rho}_1/\hat{\rho}_{\mathrm{ne}}\rangle_{\mathrm{ne}}}, \tag{SM.14}$$

where the ratio $\hat{\rho}_1/\hat{\rho}_{\mathrm{ne}}$ plays the role of the weights in AIS.

The situation with our estimator is quite different. First, we use $\rho_1 = \rho$. Secondly, we generate the trajectories in a way such that, even if we do not know the normalization of $\rho_1 = \rho$, we have the property that

$$\langle\hat{\rho}_1/\hat{\rho}_{\mathrm{ne}}\rangle_{\mathrm{ne}} = \langle\hat{\rho}/\hat{\rho}_{\mathrm{ne}}\rangle_{\mathrm{ne}} = 1 \tag{SM.15}$$

by construction. Therefore our reweighting scheme does not require that we estimate a ratio of expectations; therefore, it provides us with an unbaised estimator, unlike (SM.14). Finally, this estimator has lower variance than the direct sample mean estimator, as shown by Eq. (9) in the main text.

3. **The mean-field Ising model.** We next consider a continuous version of the Curie-Weiss magnet, i.e. the mean-field Ising model with d spins and potential (this is Eq. (15))

$$U(\boldsymbol{q}) = -\frac{1}{2d}\sum_{i,j=1}^{d}\cos q_i\cos q_j = -\frac{1}{2d}\left(\sum_{i=1}^{d}\cos q_i\right)^2 \tag{SM.16}$$

The Gibbs (canonical) density for this model is

$$\rho_c(\boldsymbol{q}) = Z^{-1}(\beta)e^{-\beta U(\boldsymbol{q})} \quad \text{where} \quad Z(\beta) = \int_{[-\pi,\pi]^d} e^{-\beta U(\boldsymbol{q})}d\boldsymbol{q}. \tag{SM.17}$$

This system has similar thermodynamics properties as the standard Curie-Weiss magnet with discrete spins, but it is amenable to simulation by Langevin dynamics since the angles q_i vary continuously.

In particular, like the standard Curie-Weiss magnet, the system with potential (SM.16) displays phase-transitions when β is varied. To see why, and also introduce the scaled free energy that we monitor in our numerical experiments, let us marginalize the Gibbs density (SM.17) on the average magnetization m defined as

$$m = \frac{1}{d}\sum_{i=1}^{d} \cos q_i. \tag{SM.18}$$

This marginalized density is given by

$$\bar{\rho}_c(m) = \int_{[-\pi,\pi]^K} \rho_c(\boldsymbol{q})\delta\left(m - \frac{1}{d}\sum_{i=1}^{d}\cos q_i\right) d\boldsymbol{q}. \tag{SM.19}$$

A simple calculation shows that

$$\bar{\rho}_c(m) = \bar{Z}^{-1}(\beta)e^{-\beta d F_d(m,\beta)} \quad \text{where} \quad \bar{Z}(\beta) = \int_{-1}^{1} e^{-\beta d F_d(m,\beta)} dm \tag{SM.20}$$

Here we introduced the (scaled) free energy $F_d(m,\beta)$ defined as

$$F_d(m,\beta) = V(m) - \beta^{-1} S_d(m) \tag{SM.21}$$

with potential term

$$V(m) = -\tfrac{1}{2}m^2 \tag{SM.22}$$

and entropic term

$$S_d(m) = d^{-1}\log \int_{[-\pi,\pi]^d} \delta\left(m - \frac{1}{d}\sum_{i=1}^{d}\cos q_i\right) d\boldsymbol{q}. \tag{SM.23}$$

The marginalized density (SM.20) and the free energy (SM.21) can be used to analyze the properties of the system in thermodynamic limit when $d \to \infty$ and map out its phase transition diagram in this limit. In particular, standard results from large deviation theory recalled below can be used to show that $F_d(m,\beta)$ has a limit as $d \to \infty$ that has a single minimum at high temperature, but two minima at low temperature. Since $F_d(m,\beta)$ is scaled by d in (SM.20), this implies that density can become bimodal at low temperature, indicative of the presence of two strongly metastable states separated by a free energy barrier whose height is proportional to d.

The limiting free energy $F(m,\beta)$ is defined as

$$F(m,\beta) = \lim_{d\to\infty} F_d(m) = -\tfrac{1}{2}m^2 - \beta^{-1}\lim_{d\to\infty} S_d(m). \tag{SM.24}$$

To calculate the limit of the entropic term, let us define $H(\lambda)$ via the Laplace transform of (SM.23) through

$$\begin{aligned} e^{-dH(\lambda)} &= \int_{-1}^{1} e^{-d\lambda m + dS_d(m)} dm \\ &= \int_{[-\pi,\pi]^d} e^{-\lambda\sum_{i=1}^{d}\cos(q_i)} d\boldsymbol{q} \\ &= \prod_{i=1}^{d}\int_{-\pi}^{\pi} e^{-\lambda\cos(q_i)} dq_i \\ &= \left(2\pi I_0(\lambda)\right)^d, \end{aligned} \tag{SM.25}$$

where $I_0(\lambda)$ is a modified Bessel function. In the large d limit, $S(m)$ can be calculated from $H(\lambda)$ by Legendre transform

$$\begin{aligned} S(m) = \lim_{d\to\infty} S_d(m) &= \min_{\lambda} \{\lambda m - H(\lambda)\} \\ &= \min_{\lambda} \{\lambda m + \log I_0(\lambda)\} + \log(2\pi). \end{aligned} \tag{SM.26}$$

The minimizer $\lambda(m)$ of (SM.26) satisfies

$$m = -\frac{I_1(\lambda(m))}{I_0(\lambda(m))} \tag{SM.27}$$

which, upon inversion, offers a way to parametrically represent $S(m)$ using

$$S(m(\lambda)) = \lambda m(\lambda) + \log I_0(\lambda) + \log(2\pi), \qquad m(\lambda) = -\frac{I_1(\lambda)}{I_0(\lambda)}, \qquad \lambda \in \mathbb{R}. \tag{SM.28}$$

Similarly we can represent $F(m, \beta)$ as

$$F(m(\lambda), \beta) = -\tfrac{1}{2}m^2(\lambda) - bm(\lambda) - \beta^{-1} S(m(\lambda)), \qquad m(\lambda) = -\frac{I_1(\lambda)}{I_0(\lambda)} \qquad \lambda \in \mathbb{R}. \tag{SM.29}$$

These are the formulae we use to plot the free energy shown in the inset of Fig. 1. We compare this free energy with the one obtained by estimating the following expectation using our estimator with a single ascent / descent trajectory

$$-(d\beta)^{-1} \log \left\langle e^{-\beta U(\boldsymbol{q})} \delta\left(m - \frac{1}{d}\sum_{i=1}^{d} \cos q_j\right)\right\rangle \tag{SM.30}$$

A similar calculation can be used to estimate the factor $Z(\beta)$ in (SM.17) using the identity

$$Z(\beta) = \int_{-1}^{1} e^{\frac{1}{2}\beta d m^2 + d S_d(m)} dm \tag{SM.31}$$

In the $d \to \infty$ limit this implies that

$$G(\beta) \equiv \lim_{d\to\infty} d^{-1} \log Z(\beta) = \max_{m\in[-1,1]} \left(\tfrac{1}{2}\beta d m^2 + d S(m)\right) \tag{SM.32}$$

Since the density of states

$$D(E) = \int_{[-\pi,\pi]^d \times \mathbb{R}^d} \delta\left(E - \frac{1}{2}|\boldsymbol{p}|^2 - U(\boldsymbol{q})\right) d\boldsymbol{q} d\boldsymbol{p} \tag{SM.33}$$

is related to $Z(\beta)$ as (accounting for the extra factor coming from the momenta)

$$(2\pi\beta^{-1})^{d/2} Z(\beta) = \int_{\mathbb{R}} D(E) e^{-\beta E} dE \tag{SM.34}$$

we have

$$\lim_{d\to\infty} d^{-1} \log D(d\mathcal{E}) = Q(\mathcal{E}) \tag{SM.35}$$

with

$$Q(\mathcal{E}) = -\min_{\beta} \left(\beta\mathcal{E} + G(\beta) - \tfrac{1}{2}\log\beta\right) + \tfrac{1}{2}\log\pi. \tag{SM.36}$$

This is the formula we used to plot the red curve in Fig. 1 used for comparison with the estimate of $V(E)$ we obtained directly using our estimator.