Contents

1 Introduction

1.1 Goal: Approximating configuration-space bias of Langevin integrators

Our goal is to approximate the KL divergence $\mathcal{D}_{\mathrm{KL}}(\rho_{\mathrm{conf}} \| p)$ between the equilibrium configuration distribution p and the one sampled by a discrete-time Langevin integrator ρ_{conf}. We will estimate this quantity by relating the KL divergence to a nonequilibrium free energy difference. This free energy difference, in turn, can be measured near equilibrium by estimating the average work performed by the Langevin integrator, applied to specially prepared ensembles of starting conditions.

1.2 Outline of argument

To justify this procedure, we will:

- Express the desired KL divergence as a nonequilibrium free energy difference.
- Show that this free energy difference can be approximated in terms of work-like quantities.
- Argue that these work-like quantities can be expressed in terms of easily computed quantities, for a wide family of plausible Langevin integrators.
- Provide evidence that these quantities can be estimated reliably from simulation.

1.3 Notation: ensembles

- the **equilibrium distribution** is π (and the configuration distribution and velocity distribution are independent)
$$\pi(\mathbf{x}, \mathbf{v}) \equiv p(\mathbf{x}) q(\mathbf{v})$$

- the **nonequilibrium steady state** sampled by our Langevin integrator is ρ (where there may be arbitrary correlations between $\mathbf{x}$ and $\mathbf{v}$, and we can't compute the density)
$$\rho(\mathbf{x}, \mathbf{v}) = ?$$

- the **configuration-marginal of the nonequilibrium steady state** sampled by the Langevin integrator is ρ_{conf}
$$\rho_{\mathrm{conf}}(\mathbf{x}) \equiv \int d\mathbf{v} \rho(\mathbf{x}, \mathbf{v})$$

- the **nonequilibrium macrostate** we will prepare with re-sampled velocities is ω
$$\omega(\mathbf{x}, \mathbf{v}) \equiv \rho_{\mathrm{conf}}(\mathbf{x}) q(\mathbf{v})$$

2 Expressing configuration space bias in terms of the nonequilibrium free energy difference $F_\omega - F_\pi$

We would like to express the KL divergence between the configuration distribution at equilibrium and the configuration distribution in the nonequilibrium steady state $\mathcal{D}_{\text{KL}}(\rho_{\text{conf}} \| p)$. First, we show that the KL divergence of the configuration marginals is equivalent to the full phase space KL divergence.

2.1 The KL divergence between these ensembles is the KL divergence between their configuration marginals

We've constructed the ensemble ω so that the KL divergence ω and π equals the KL divergence between ρ_{conf} and p:

$$
\begin{aligned}
\mathcal{D}_{\text{KL}}(\omega \| \pi) &\equiv \int \int d\mathbf{x} d\mathbf{v} \left(\omega(\mathbf{x}, \mathbf{v}) \log \frac{\omega(\mathbf{x}, \mathbf{v})}{\pi(\mathbf{x}, \mathbf{v})} \right) \\
&= \int \int d\mathbf{x} d\mathbf{v} \left(\omega(\mathbf{x}, \mathbf{v}) \log \frac{\rho_{\text{conf}}(\mathbf{x})}{p(\mathbf{x})} \overbrace{\frac{q(\mathbf{v})}{q(\mathbf{v})}}^{1} \right) \\
&= \int \int d\mathbf{x} d\mathbf{v} \left(\omega(\mathbf{x}, \mathbf{v}) \log \frac{\rho_{\text{conf}}(\mathbf{x})}{p(\mathbf{x})} \right) \\
&= \int d\mathbf{v} \left(\int d\mathbf{x} \rho_{\text{conf}}(\mathbf{x}) q(\mathbf{v}) \log \frac{\rho_{\text{conf}}(\mathbf{x})}{p(\mathbf{x})} \right) \\
&= \left(\overbrace{\int d\mathbf{v} q(\mathbf{v})}^{1} \right) \left(\int d\mathbf{x} \rho_{\text{conf}}(\mathbf{x}) \log \frac{\rho_{\text{conf}}(\mathbf{x})}{p(\mathbf{x})} \right) \\
&= \mathcal{D}_{\text{KL}}(\rho_{\text{conf}} \| p)
\end{aligned}
$$

Note that $\mathcal{D}_{\text{KL}}(\omega \| \pi) = \beta(F_\omega - F_\pi)$ where $\beta = \frac{1}{k_B T}$ (see proof proceeding eq. 1 of http://threeplusone.com/Sivak2012a.pdf).

2.2 Relation between KL divergence and free energy difference

The free energy F of any distribution over microstates $P(x)$ is:

$$
F \equiv \langle E \rangle - S/\beta
$$

where

1. $\langle E \rangle$ is the mean energy $\langle E \rangle = \sum_x P(x) E(x)$
2. S is the entropy $S = \sum_x P(x) \log P(x)$
3. β is the inverse temperature

Let $P_{\text{eq}}(x) = \exp[-\beta(E(x) - F_{\text{eq}})]$ and let P_{neq} be a nonequilibrium distribution that's close to P_{eq}.

$$D(P_{\text{neq}} \| P_{\text{eq}}) \equiv \sum_x P_{\text{neq}}(x) \log \left[\frac{P_{\text{neq}}(x)}{P_{\text{eq}}(x)} \right] \tag{1}$$

$$= \underbrace{\sum_x P_{\text{neq}}(x) \log P_{\text{neq}}(x)}_{-S_{\text{neq}}} - \sum_x P_{\text{neq}}(x) \log P_{\text{eq}}(x) \tag{2}$$

$$= -S_{\text{neq}} - \sum_x P_{\text{neq}}(x)\beta(F_{\text{eq}} - E(x) \tag{3}$$

$$= -S_{\text{neq}} - \underbrace{\sum_x P_{\text{neq}}(x)\beta F_{\text{eq}}}_{\beta F_{\text{eq}}} + \underbrace{\sum_x P_{\text{neq}}(x)\beta E(x)}_{\beta \langle E \rangle_{\text{neq}}} \tag{4}$$

$$= \beta(F_{\text{eq}} - F_{\text{neq}}) \tag{5}$$

$$= \beta \Delta F_{\text{neq}} \tag{6}$$

Thus, if we can approximate the nonequilibrium free energy difference $F_\omega - F_\pi$, then we can approximate the configuration-space bias.

3   Approximating $F_\omega - F_\pi$

Next, we need to show that we can approximate the desired free energy difference in terms of computable quantities.

3.1   A nonequilibrium protocol that transforms π into ω

We can write down a nonequilibrium protocol that transforms π into ω:

1. **Simulate Langevin dynamics** for N steps (performs "shadow work" to turn π into ρ)

2. **Randomize velocities** by taking a single O-step with infinite friction parameter γ.

where N is chosen sufficiently large to reach steady state ρ.

Call the procedure [1,2] the "forward protocol" and the procedure [2,1] the "reverse protocol."

Using the near-equilibrium approximation in http://threeplusone.com/Sivak2012a.pdf, we can approximate $\Delta F_{\text{neq}} \equiv F_\omega - F_\pi$ in terms of quantities we can compute from simulation.

Denote by $\langle W_{\text{F}} \rangle_\pi$ the average work performed by the "forward protocol," where the average is over initial conditions at equilibrium. Denote by $\langle W_{\text{R}} \rangle_\omega$ the average work performed by the "reverse protocol," where the average is over initial conditions distributed according to ω.

$$\Delta F_{\text{neq}} \approx \frac{1}{2}[\langle W_{\text{F}}\rangle_\pi - \langle W_{\text{R}}\rangle_\omega]$$

$$= \frac{1}{2}\left[(\langle W\rangle_{\pi\to\rho} + \langle W\rangle_{\rho\to\omega}) - (\langle W\rangle_{\omega\to\omega} + \langle W\rangle_{\omega\to\rho})\right]$$

$$= \frac{1}{2}[\langle W_{\text{shad}}\rangle_{\pi\to\rho} + \cancelto{0}{\langle W\rangle_{\rho\to\omega}} - \cancelto{0}{\langle W\rangle_{\omega\to\omega}} - \langle W_{\text{shad}}\rangle_{\omega\to\rho}]$$

$$= \frac{1}{2}\left[\langle W_{\text{shad}}\rangle_{\pi\to\rho} - \langle W_{\text{shad}}\rangle_{\omega\to\rho}\right]$$

Where we can approximate:

- $\langle W_{\text{shad}}\rangle_{\pi\to\rho}$ by starting in equilibrium, and measuring the shadow work accumulated by simulating for N steps.

- $\langle W_{\text{shad}}\rangle_{\omega\to\rho}$ by starting in nonequilibrium steady state with velocities resampled from equilibrium, and measuring the shadow work accumulated by simulating for N steps.

4 Interpreting Langevin integrators as nonequilibrium protocols

4.1 More notation: Nonequilibrium protocols and work

We'll define a *protocol* as any sequence of Markov kernels:

$$\Lambda = [k_0, k_1, \ldots, k_T]$$

We define the time-reverse of a protocol as the reverse sequence of kernels, where each individual kernel is time-reversed. We denote time-reversal using tildes: $\tilde{\Lambda} = [\tilde{k}_T, \tilde{k}_{T-1}, \ldots, \tilde{k}_0]$.

The probability of a given trajectory $X = (\mathbf{x}_0, \ldots, \mathbf{x}_T)$, given a protocol and initial ensemble, is defined as the product of the initial state's probability and all the transition probabilities:

$$P[X|\Lambda] = p_0(\mathbf{x}_0)\prod_{t=1}^{T}k_t(\mathbf{x}_t|\mathbf{x}_{t-1})$$

We define the "work" in terms of the differential path action:

4.2 A family of Langevin splitting integrators

We'll define a family of concrete integrators of the Langevin equations, indexed by splitting strings over the alphabet { V, R, O }. First, we'll define the explicit update equations, which accept an initial state (including a stream of i.i.d. standard normal random variables), a timestep h, and a friction coefficient γ, and return a new state:

$$\text{R}(\mathbf{x}, \mathbf{v}; h) \equiv (\mathbf{x} + \mathbf{v}h, \mathbf{v})$$
$$\text{V}(\mathbf{x}, \mathbf{v}; h) \equiv (\mathbf{x}, \mathbf{v} + (f(\mathbf{x})/m)h)$$
$$\text{O}(\mathbf{x}, \mathbf{v}; h, \gamma) \equiv (\mathbf{x}, a_{\gamma,h}\mathbf{v} + b_{\gamma,h}\sqrt{k_B T/m}\xi)$$

where $a_{\gamma,h} = e^{-\gamma h}$, $b_{\gamma,h} = \sqrt{1 - e^{-2\gamma h}}$, $\xi \sim \mathcal{N}(0,1)$ per degree of freedom.

Then executing a single step of the integrator indexed a splitting string Λ is given by:

Accept: $(\mathbf{x}_0, \mathbf{v}_0)$, γ, T, m, total timestep h, length-T splitting Λ, n_R is the number of "R"s in Λ, n_V the number of "V"s, etc.

For $i = 1$ to T:

$(\mathbf{x}_t, \mathbf{v}_t) = \Lambda[t](\mathbf{x}_{t-1}, \mathbf{v}_{t-1}; \delta t / n_{\Lambda[t]})$

4.3   These update equations sample exactly from the following transition densities

To interpret a single timestep of a discrete-time integrator of Langevin dynamics as a protocol, we'll explicitly write down the transition densities these update equations sample. For example, the protocol corresponding to a single step of the integrator "OVRVO" is:

$$\Lambda = [k_{O,\delta t/2}, k_{V,\delta t/2}, k_{R,\delta t}, k_{V,\delta t/2}, k_{O,\delta t/2}]$$

where the substep Markov kernels are defined as:

$$k_{R,h}(\mathbf{x}_{t+h}, \mathbf{v}_{t+h} | \mathbf{x}_t, \mathbf{v}_t) \equiv \delta[\mathbf{x}_{t+h} - (\mathbf{x}_t + \mathbf{v}_t h)]\delta[\mathbf{v}_{t+h} - \mathbf{v}_t]$$
$$k_{V,h}(\mathbf{x}_{t+h}, \mathbf{v}_{t+h} | \mathbf{x}_t, \mathbf{v}_t) \equiv \delta[\mathbf{v}_{t+h} - (\mathbf{v}_t + f(\mathbf{x}_t)/m)h]\delta[\mathbf{x}_{t+h} - \mathbf{x}_t]$$
$$k_{O,h}(\mathbf{x}_{t+h}, \mathbf{v}_{t+h} | \mathbf{x}_t, \mathbf{v}_t) \equiv \mathcal{N}(\mathbf{v}_{t+h}; \mu = a_h \mathbf{v}_t, \Sigma = b\sqrt{k_B T/m})\delta[\mathbf{x}_{t+h} - \mathbf{x}_t]$$

4.4   Observations

Note that the R, V, O update equations draw exactly from each of these conditional distributions.

Note that for Langevin integrators defined by palindromic operator splittings (e.g. "OVRVO" but not "OVR"), the forward and reverse protocols are identical: $\tilde{\Lambda} = \Lambda$.

Note that applying $k_{R,h}$ to an arbitrary distribution $\rho(\mathbf{x}, \mathbf{v})$ leaves the marginal distribution over velocities invariant for any h (i.e. $\int d\mathbf{x} k_{R,h} \circ \rho(\mathbf{x}, \mathbf{v}) = \int d\mathbf{x}\rho(\mathbf{x}, \mathbf{v})$), and applying $k_{V,h}$ or $k_{O,h}$ leaves the marginal distribution over configurations invariant (i.e. $\int d\mathbf{v} k_{V,h} \circ \rho(\mathbf{x}, \mathbf{v}) = \int d\mathbf{v} k_{O,h} \circ \rho(\mathbf{x}, \mathbf{v}) = \int d\mathbf{v}\rho(\mathbf{x}, \mathbf{v})$). (Of course, applying a sequence of these kernels doesn't necessarily preserve an arbitrary starting distribution.)

For finite time step, we can interpret compositions of R and V steps as preserving a "shadow-hamiltonian." [citations!]

Note that, multiple consecutive O steps are equivalent to a single large O step: $k_{O,h_1} \circ k_{O,h_2} \circ \cdots \circ k_{O,h_N} = k_{O,\sum_{i=1}^N h_i}$.

Note that, in the absence of constraints, multiple consecutive R or V steps also satisfy this property.

When constraints are active, however, multiple consecutive R steps are not equivalent in general to a single large R step.

5 Deriving shadow-work accumulation scheme

To estimate shadow work from simulation, we need to measure the shadow work performed by an integrator correctly. Our current accounting is that $\Delta E = W_{\text{shad}} + Q$, and that Q is the sum of substep energy changes during O-steps.

Claim:

$$W_{\text{shad}} = \sum_{i=1}^{N} [\Delta E_i \mathbb{I}[\text{substep}_i \in \{\text{R}, \text{V}\}]]$$

$$= \Delta E - \sum_{i=1}^{N} [\Delta E_i \mathbb{I}[\text{substep}_i = \text{O}]]$$

We will first proceed to show this for a single step of OVRVO with timestep δt. Then we will attempt to show this for arbitrary sequences of R, V, O steps.

The game here will be to write the forward and reverse path probabilities in terms of specific substep energy changes.

We'll say $X = ((\mathbf{x}_0, \mathbf{v}_0), (\mathbf{x}_1, \mathbf{v}_1), (\mathbf{x}_2, \mathbf{v}_2), (\mathbf{x}_3, \mathbf{v}_3), (\mathbf{x}_4, \mathbf{v}_4), (\mathbf{x}_5, \mathbf{v}_5))$ is the trajectory during a single timestep. Since each of the updates only touches the $\mathbf{x}$ or $\mathbf{v}$ component in a given substep, we'll be able to make substitutions, e.g. $\mathbf{x}_0 = \mathbf{x}_1$.

Writing down the probability of the forward trajectory:

$$
\begin{aligned}
P[X|\Lambda] =& \pi(\mathbf{x}_0, \mathbf{v}_0) \\
& \times k_{\text{O}, \delta t/2}((\mathbf{x}_1, \mathbf{v}_1)|(\mathbf{x}_0, \mathbf{v}_0)) \\
& \times k_{\text{V}, \delta t/2}((\mathbf{x}_2, \mathbf{v}_2)|(\mathbf{x}_1, \mathbf{v}_1)) \\
& \times k_{\text{R}, \delta t}((\mathbf{x}_3, \mathbf{v}_3)|(\mathbf{x}_2, \mathbf{v}_2)) \\
& \times k_{\text{V}, \delta t/2}((\mathbf{x}_4, \mathbf{v}_4)|(\mathbf{x}_3, \mathbf{v}_3)) \\
& \times k_{\text{O}, \delta t/2}((\mathbf{x}_5, \mathbf{v}_5)|(\mathbf{x}_4, \mathbf{v}_4)) \\
=& \pi(\mathbf{x}_0, \mathbf{v}_0) \\
& \times \mathcal{N}(\mathbf{v}_1; \mu = a_{\delta t/2}\mathbf{v}_0, \Sigma = a_{\delta t/2}\sqrt{k_B T/m})\mathbb{I}[\mathbf{x}_1 = \mathbf{x}_0] \\
& \times \mathbb{I}[v_2 = v_1 + (f(\mathbf{x}_1)/m)(\delta t/2)\mathbb{I}[\mathbf{x}_2 = \mathbf{x}_1] \\
& \times \mathbb{I}[\mathbf{x}_3 = \mathbf{x}_2 + \mathbf{v}_2 \delta t]\mathbb{I}[v_3 = v_2] \\
& \times \mathbb{I}[v_4 = v_3 + (f(\mathbf{x}_3)/m)(\delta t/2)\mathbb{I}[\mathbf{x}_4 = \mathbf{x}_3] \\
& \times \mathcal{N}(\mathbf{v}_5; \mu = a_{\delta t/2}\mathbf{v}_4, \Sigma = a_{\delta t/2}\sqrt{k_B T/m})\mathbb{I}[\mathbf{x}_5 = \mathbf{x}_4]
\end{aligned}
$$

Assuming that X is a valid trajectory (i.e. that all of the expressions in the indicator functions are satisfied), then we can explicitly write down the path probability as a product of two Gaussians by (1) replacing all the indicator functions with 1, (2) writing $\mathbf{v}_4$ as a deterministic function of $\mathbf{v}_1$ and $\mathbf{v}_0$. (If any of the indicator functions are not satisfied, then the differential path action will be undefined.)

$$P[X|\Lambda] = \pi(\mathbf{x}_0, \mathbf{v}_0)\mathcal{N}(\mathbf{v}_1; \mu = a_{\delta t/2}\mathbf{v}_0, \Sigma = a_{\delta t/2}\sqrt{k_B T/m})\mathcal{N}(\mathbf{v}_5; \mu = a_{\delta t/2}\mathbf{v}_4, \Sigma = a_{\delta t/2}\sqrt{k_B T/m})$$
$$= \pi(\mathbf{x}_0, \mathbf{v}_0)\mathcal{N}(\mathbf{v}_1; \mu = a_{\delta t/2}\mathbf{v}_0, \Sigma = a_{\delta t/2}\sqrt{k_B T/m})$$
$$\times \mathcal{N}(\mathbf{v}_5; \mu = a_{\delta t/2}\mathbf{v}_1 + (f(\mathbf{x}_0)/m)(\delta t/2) + (f(\mathbf{x}_0 + \mathbf{v}_2\delta t)/m)(\delta t/2), \Sigma = a_{\delta t/2}\sqrt{k_B T/m})$$

We can explicitly write the relative path action as:

6 Analyzing the full protocol

Now, we will write down the protocol for performing M steps of discrete Langevin dynamics. Then, we will assume that M is sufficiently large to reach a nonequilibrium steady state (we will later discuss how to determine M from simulation observables). Next, we will add a "velocity-randomization" kernel, completing our description of the protocol. Since this "velocity-randomization" kernel Finally, we will check whether the velocity-randomization kernel performs "work" if applied to an arbitrary distribution ρ.

6.1 Velocity-randomization kernel

Consider the Markov kernel that draws velocities i.i.d. from a fixed distribution q and leaves the positions unchanged:

$$k_{\text{rand}}(\mathbf{x}_{t+h}, \mathbf{v}_{t+h}|\mathbf{x}_t, \mathbf{v}_t) \equiv q(\mathbf{v}_{t+h})\delta[\mathbf{x}_{t+h} - \mathbf{x}_t]$$

Note that applying k_{rand} to an arbitrary distribution ρ is equivalent to marginalizing out the velocities:

$$(k_{\text{rand}} \circ \rho)(\mathbf{x}, \mathbf{v}) = q(\mathbf{v}) \int d\mathbf{v}\rho(\mathbf{x}, \mathbf{v})$$

6.2 Does randomizing velocities to a fixed distribution perform "work?"

How should we treat the energy change performed by applying the "velocity-randomization" kernel to the nonequilibrium steady state $\rho(\mathbf{x}, \mathbf{v})$?

6.2.1 Generic argument looks like it fails: can't randomize to an arbitrary velocity distribution

Using the Crooks Fluctuation Theorem definition of work:

$$W[X] \equiv -\ln\left[\frac{P[\tilde{X}|\tilde{\Lambda}]}{P[X|\Lambda]}\right]$$

Note that since this protocol is only a single step, $\Lambda = \tilde{\Lambda} = [k_{\text{rand}}]$

$X = [(\mathbf{x}_0, \mathbf{v}_0), (\mathbf{x}_1, \mathbf{v}_1)]$

$$\begin{aligned}
P[X|\Lambda] &= p_0(\mathbf{x}_0, \mathbf{v}_0) k_{\text{rand}}(\mathbf{x}_1, \mathbf{v}_1 | \mathbf{x}_0, \mathbf{v}_0) \\
&= p_0(\mathbf{x}_0, \mathbf{v}_0) q(\mathbf{v}_1) \delta(\mathbf{x}_1 - \mathbf{x}_0) \\
P[\tilde{X}|\tilde{\Lambda}] &= p_1(\mathbf{x}_1, \mathbf{v}_1) k_{\text{rand}}(\mathbf{x}_0, \mathbf{v}_0 | \mathbf{x}_0, \mathbf{v}_0) \\
&= p_1(\mathbf{x}_1, \mathbf{v}_1) q(\mathbf{v}_0) \delta(\mathbf{x}_1 - \mathbf{x}_0) \\
\frac{P[\tilde{X}|\tilde{\Lambda}]}{P[X|\Lambda]} &= \frac{p_1(\mathbf{x}_1, \mathbf{v}_1) q(\mathbf{v}_0)}{p_0(\mathbf{x}_0, \mathbf{v}_0) q(\mathbf{v}_1)}
\end{aligned}$$

Where the differential path action is only defined if $\mathbf{x}_1 = \mathbf{x}_0$.

Hmm, I don't think we can compute this in closed form, and I don't think it's zero...

6.2.2 Interpreting velocity-randomization kernel as limits of O-substeps

Observations: The independent velocity randomization step is equivalent to other steps that we know contribute heat not work:

1. It is equivalent to taking a single O step with infinite friction coefficient ($\gamma \to +\infty$).

2. It is equivalent to taking a single O step with the same friction coefficient but an infinitely large timestep $dt \to infty$.

3. It is equivalent to taking an infinite number of O steps, with the current friction coefficient and timestep.

7 One-dimensional example

To demonstrate this approach, we will first check that this produces accurate estimates of the configuration-marginal KL divergence in a low-dimensional system for which precise numerical estimates are available.
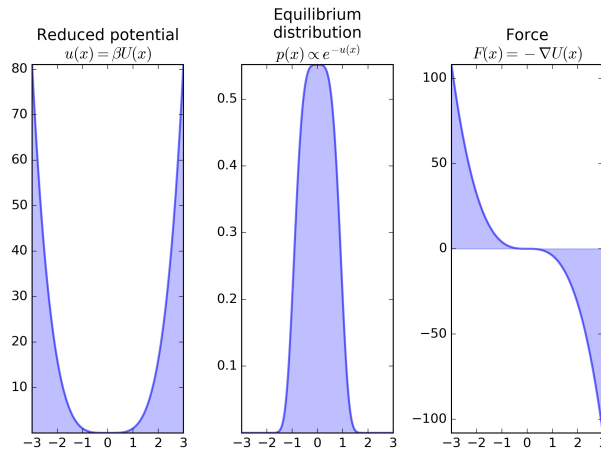
7.1 Methods

Code to reproduce these experiments is available on Github: https://github.com/choderalab/integrator-benchmark/tree/master/code/toy.

1. **System definition:** Quartic potential $U(x) = x^4$, mass $m = 10$, inverse temperature $\beta = 1$, friction $\gamma = 100$, tested timestep range $0.25 - 1.1$ (stability limit of VVVR for this system appears to be about 1.2).

2. **Generating equilibrium and steady state samples** For each condition, I run the integrator for 100 million steps, and save every 100th (x, v) snapshot. To generate equilibrium samples, I run 100 million steps of random-walk Metropolis-Hastings with standard Normal proposals.

3. **Estimating "ground-truth" KL divergence.** I used two methods to measure the KL divergence between the Langevin steady state and equilibrium:

   (a) *Estimate as a free energy difference,* $\mathcal{D}_{\mathrm{KL}}(\rho_{\mathrm{conf}}(x)\|p(x)) = (\langle U(x)\rangle_{\mathrm{neq}} - S_{\mathrm{neq}} - (\langle U(x)\rangle_{\mathrm{eq}} - S_{\mathrm{eq}})$. Where $\langle U(x)\rangle_{\mathrm{neq}} \approx \frac{1}{N}\sum_{i=1}^{N} U(x^{(i)})$ is estimated as the sample mean of the potential energy over the steady-state samples $x^{(i)}$, S_{neq} is the entropy of a 200-bin histogram of the steady-state x samples, and $\langle U(x)\rangle_{\mathrm{eq}}$ and S_{eq} are computed using numerical quadrature.

   (b) *Estimate as the KL divergence between histograms* of the sampled nonequilibrium steady state vs. the exact histogram of the equilibrium x distribution over those same bins.

   These estimates agree to about 4 decimal places, so I just plot the second estimate in all subsequent figures. Perhaps cleverer baseline estimators of the KL divergence would be helpful.

4. **Nonequilibrium protocol** Draw an equilibrium sample $(x, v) \sim \pi$, apply the Langevin integrator for $M = 100$ steps to produce $(x', v') \sim \rho$. Record shadow work accumulated during that trajectory as $W_{\pi \to \rho}$. Randomize $v'' \sim q$, so that $(x', v'') \sim \omega$. Apply the integrator for M steps, record shadow work performed by the second trajectory as $W_{\omega \to \rho}$. Perform 1 million such protocols, and report $\hat{\mathcal{D}}_{\mathrm{KL}}(\rho_{\mathrm{conf}}\|p) \approx \frac{\beta}{2}[\langle W_{\pi \to \rho}\rangle - \langle W_{\omega \to \rho}\rangle]$.

7.2 Results

7.2.1 Validating $\mathcal{D}_{\mathbf{KL}}$ estimator

First, we compare the proposed nonequilibrium estimator of the KL divergence to the precise numerical computation of the KL divergence.

The estimator was within error of the correct answer for 14 of 16 points on this plot. The variance of the naive estimator, even with a million protocol samples, is very high.
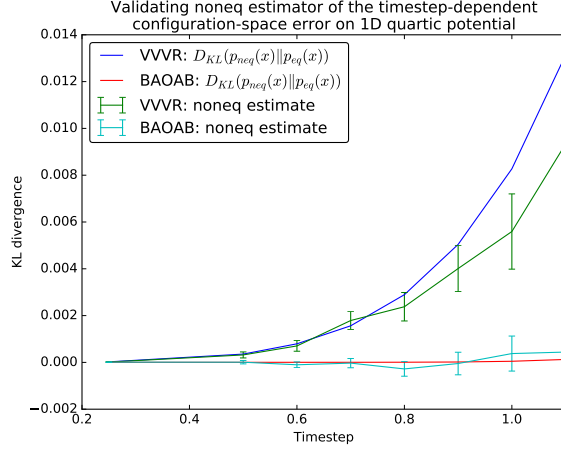
Figure 1: Estimator comparison

7.2.2 Examining the sampled distributions

For this system, BAOAB has extremely low configuration-space bias even up to the stability limit (but doesn't preserve the velocity distribution as well), while VVVR has extremely low velocity bias (but doesn't preserve the configuration distribution as well). At the maximum timestep tested, the configuration-space bias of VVVR was about 100 times greater than that of BAOAB ($\mathcal{D}_{KL}(\rho_{conf}^{BAOAB}\|p) \approx 0.00013$, and $\mathcal{D}_{KL}(\rho_{conf}^{VVVR}\|p) \approx 0.01309$).

(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$
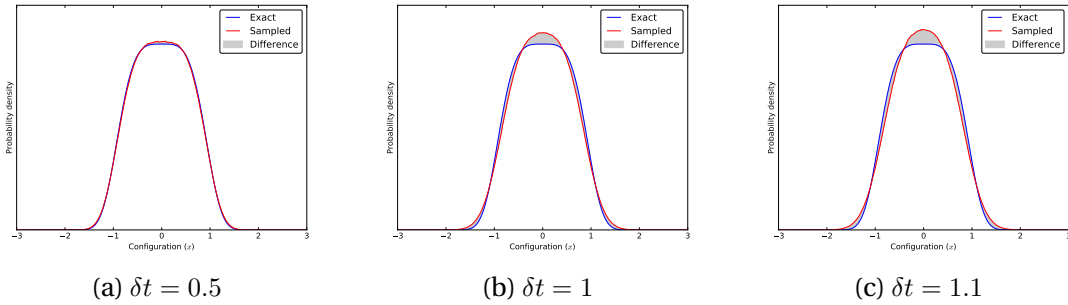
Figure 2: VVVR sampled configuration distribution

(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$
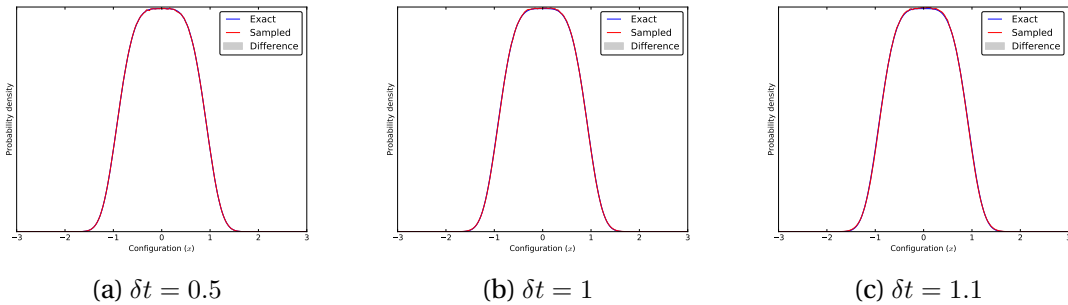
Figure 3: BAOAB sampled configuration distribution

(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$
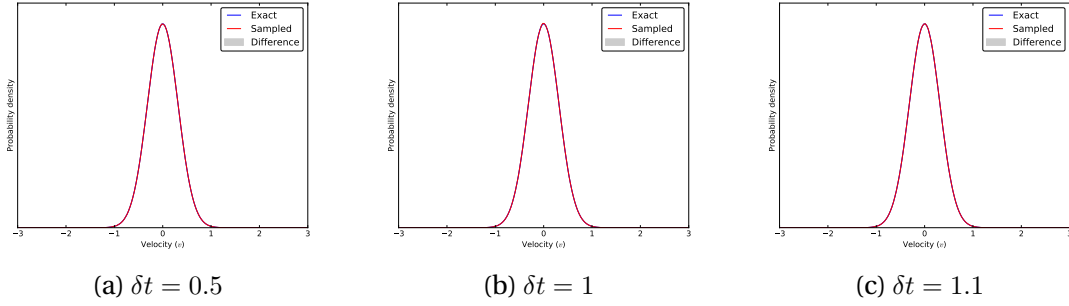
Figure 4: VVVR sampled velocity distribution

(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$
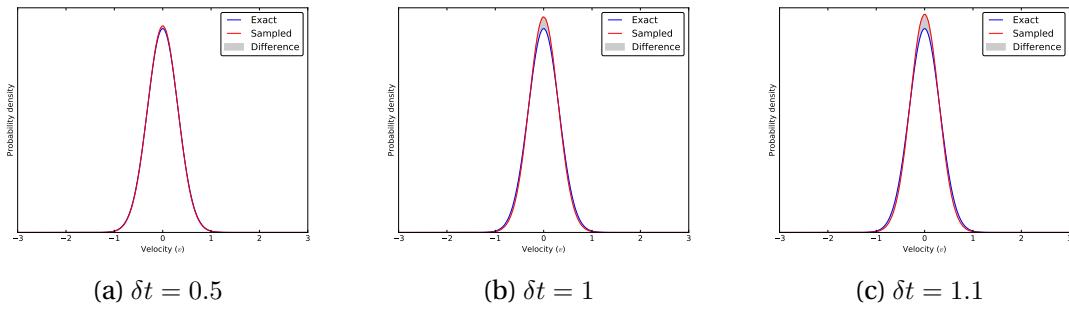
Figure 5: BAOAB sampled velocity distribution

Distortions in the velocity distribution sampled by BAOAB and the configuration distribution sampled by VVVR are visible in the plots above.

7.2.3 Examining work trajectories

To visually confirm that we've reached the nonequilibrium steady state within $M = 100$ integrator steps, we can inspect plots of the accumulated shadow work over time (note that the y-axes have different scales at each of the 3 timesteps.
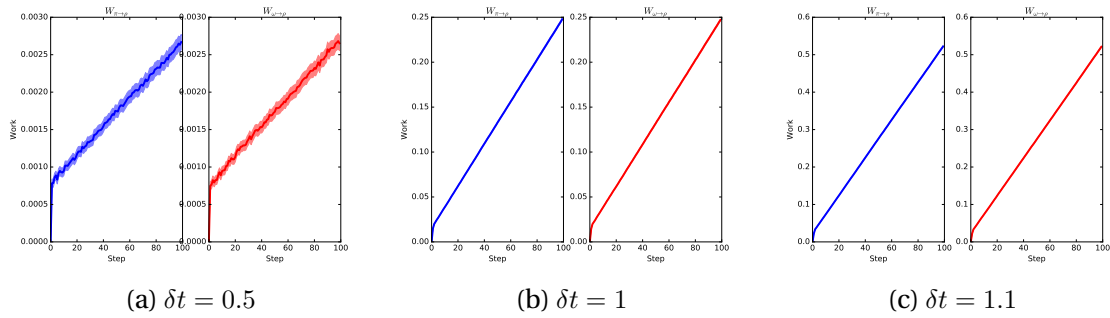

(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$

Figure 6: BAOAB work trajectories

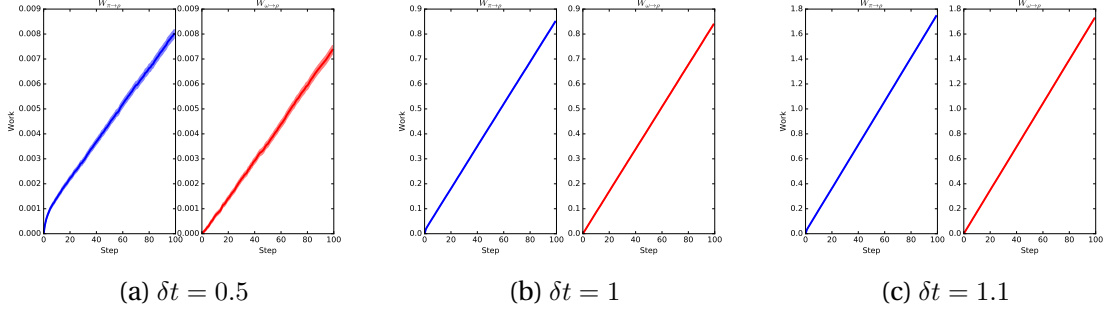(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$

Figure 7: VVVR work trajectories

For this system, it looks like the nonequilibrium steady state is reached quite rapidly, and we could perhaps use the "redundant" information from, say, steps 10 through 200 of each trajectory to reduce the variance of our estimator. I haven't yet explored this, but currently we use just a single number from each trajectory (the shadow work at the endpoint of the trajectory).

7.2.4 Examining work distributions

Next, we can inspect the distributions of shadow work at the end-points of each trajectory (note that the y-axis is log-scale).
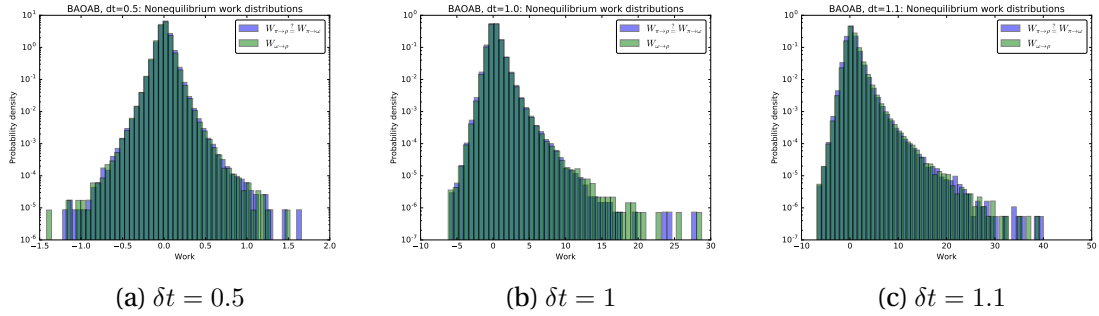

(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$

Figure 8: BAOAB work distributions

(a) $\delta t = 0.5$ (b) $\delta t = 1$ (c) $\delta t = 1.1$
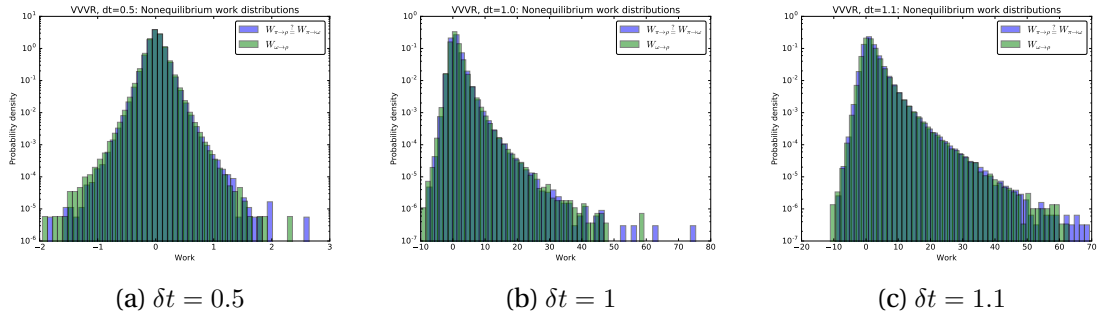
Figure 9: VVVR work distributions

13

For this system, it looks like the differences between the "forward" and "reverse" work distributions are very slight, and these distributions also appear to have fatter tails with increasing step size.

8 Molecular Mechanics Example

To demonstrate that this approach is reasonable for molecular mechanics systems, we will use this procedure to estimate and compare the configuration-space bias of several Langevin integrators on a box of a few hundred TIP3P waters.

Details: Timestep range: 4 evenly spaced timesteps between 1.5 and 2.5 femtoseconds

Protocol length $M = 50$

Number of protocol samples: 100

Temperature: 298 kelvin

Collision rate: 91 / picosecond

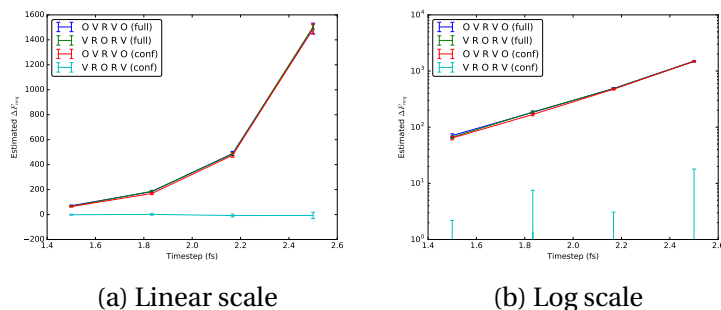PME tolerance: 10^{-5}

Cutoff: 10 angstroms

(a) Linear scale   (b) Log scale

Figure 10: Water box: Full vs. configuration-marginal ΔF_{neq}

9 Appendix

9.1 Validating shadow work accumulation scheme for a specific integrator

Here we work through the CFT definition of work for a single step of a specific integrator: "OVRVO."

1. Write out the state of the system at times $t = 1/4, 1/2, 3/4, 1$, as a function of the initial state (x_0, v_0) and random bath variables $\xi_{1/4}, \xi_{3/4}$ ("forward trajectory")

2. Write out the state of the system after applying the time-reversed integrator to the state $(x_1, -v_1)$

14

3. Write the forward and reverse path probabilities in terms of the probabilities that certain bath variables were drawn

4. Simplify / re-write in terms of measurable substep energy changes

Definitions:

$$\alpha \equiv \exp(\gamma \Delta t)$$
$$\beta \equiv \sqrt{1 - \alpha^2}$$
$$\xi_i \sim \mathcal{N}(0, \sigma^2)$$
$$\sigma^2 \equiv k_B T / m$$
$$a(x) = -\nabla U(x) / m$$

9.1.1 Forward trajectory, starting in (x_0, v_0)

Generate a sequence of states $[(x_0, v_0), (x_{1/4}, v_{1/4}), (x_{1/2}, v_{1/2}), (x_{3/4}, v_{3/4}), (x_1, v_1),]$.

We'll only write down the variables that are changing (e.g. since $x_{1/2} = x_{3/4} = x_1$, we'll just write x_1 at time $t = 1/2$ and not bother with the rest of the explicit updates).

Note that the sequence is a deterministic function of the initial conditions and two bath variables: $x_0, v_0, \xi_{1/4}, \xi_{3/4}$, where ξ_t are drawn i.i.d. from the equilibrium distribution of velocities.

1. O: $v_{1/4} = \beta v_0 + \alpha \xi_{1/4}$

2. V: $v_{1/2} = v_{1/4} + \frac{\Delta t}{2} a(x_0)$

3. R: $x_1 = x_0 + v_{1/2} \Delta t$

4. V: $v_{3/4} = v_{1/2} + \frac{\Delta t}{2} a(x_1)$

5. O: $v_1 = \beta v_{3/4} + \alpha \xi_{3/4}$

9.1.2 Reverse trajectory, starting in $(x_1, -v_1)$

In other words, we generate a sequence of states as a deterministic function of end conditions and two new bath variables: $x_1, v_1, \tilde{\xi}_{3/4}, \tilde{\xi}_{1/2}$.

1. O: $-v_{3/4} = -\beta v_1 + \alpha \tilde{\xi}_{3/4}$

2. V: $-v_{1/2} = -v_{3/4} + \frac{\Delta t}{2} a(x_1)$

3. R: $x_0 = x_1 - v_{1/2} \Delta t$

4. V: $-v_{1/4} = -v_{1/2} + \frac{\Delta t}{2} a(x_0)$

5. O: $-v_0 = -\beta v_{1/4} + \alpha \tilde{\xi}_{1/2}$

9.1.3 Writing the relative path action in terms of substep energy changes

The work is defined in terms of the probability of the forward trajectory, given the forward protocol, and the probability of the time-reversed trajectory, given the reverse protocol

$$W[X] = -\log\left[\frac{P(\tilde{X}|\tilde{\Lambda})}{P(X|\Lambda)}\right]$$

$$= -\log\left[\frac{\pi(x_1, -v_1)J(\tilde{\xi}_{3/4})\phi(\tilde{\xi}_{3/4})J(\tilde{\xi}_{1/2})\phi(\tilde{\xi}_{1/2})}{\pi(x_0, v_0)J(\xi_{1/4})\phi(\xi_{1/4})J(\xi_{3/4})\phi(\xi_{3/4})}\right]$$

where $\pi(x, v)$ is the equilibrium probability as defined above, $J(\xi)$ is the determinant of the Jacobian of the change of variables $|\partial(\alpha\xi)/\partial\xi|$, and $\phi(\xi) = \frac{1}{\sqrt{2\pi}\sigma}$

Substituting the bath variables written in terms of state variables:

$$\xi_{1/4} = (v_{1/4} - \beta v_0)/\alpha$$
$$\xi_{3/4} = (v_1 - \beta v_{3/4})/\alpha$$
$$\tilde{\xi}_{3/4} = (-v_{3/4} + \beta v_1)/\alpha$$
$$\tilde{\xi}_{1/2} = (-v_0 + \beta v_{1/4})/\alpha$$

and noting that the each of these substitutions are equivalent and cancel out $(J(\tilde{\xi}_{3/4}) = J(\tilde{\xi}_{1/2}) = J(\xi_{1/4}) = J(\xi_{3/4}) = \alpha^{-1})$ gives us:

$$W[X] = -\log\left[\frac{\pi(x_1, -v_1)\phi((-v_{3/4} + \beta v_1)/\alpha)\phi((-v_0 + \beta v_{1/4})/\alpha)}{\pi(x_0, v_0)\phi((v_{1/4} - \beta v_0)/\alpha)\phi((v_1 - \beta v_{3/4})/\alpha)}\right]$$
$$= \log\left[\pi(x_0, v_0)\phi((v_{1/4} - \beta v_0)/\alpha)\phi((v_1 - \beta v_{3/4})/\alpha)\right]$$
$$- \log\left[\pi(x_1, -v_1)\phi((-v_{3/4} + \beta v_1)/\alpha)\phi((-v_0 + \beta v_{1/4})/\alpha)\right]$$