Estimating Boltzmann Averages for Protein Structural Quantities Using Sequential Monte Carlo

Zhaoran Hou and Samuel W.K. Wong

University of Waterloo

Supplementary Material

This Supplementary Material contains algorithmic descriptions for SISR (referenced in Section 1 of the main text) and lookahead SISR (referenced in Section 4 of the main text), the proofs of Propositions 1 and 2 in Section 2.2 of the main text, the details of the protein energy function in Section 3 of the main text, the experiment that assesses the Monte Carlo variance of different values of M with overall computational budget fixed (referenced in Section 4 of the main text), and a graphical example of decay in the diversity of the paths of the particles (referenced in Section 4 of the main text).

S1   Pseudocode for SISR algorithm

Require: particle size N;

Initialization: Sample $\{\mathbf{x}_0^{(n)}\}_{n=1}^N$ from $\eta(\mathbf{x}_0)$, and the weight

$w(\mathbf{x}_0^{(n)}) = p_0(\mathbf{x}_0^{(n)})/\eta(\mathbf{x}_0^{(n)})$;

for $t = 1, \ldots, T$ **do**

Step 1: Sample $\widetilde{\mathbf{x}}_t$ from $\eta(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}^{(n)})$ and set $\widetilde{\mathbf{x}}_{0:t}^{(n)} = (\mathbf{x}_{0:t-1}^{(n)}, \widetilde{\mathbf{x}}_t)$ for

each n;

Step 2: Evaluate the weight

$w(\widetilde{\mathbf{x}}_{0:t}^{(n)}) = w(\mathbf{x}_{0:t-1}^{(n)})p_t(\widetilde{\mathbf{x}}_{0:t}^{(n)})/\{p_{t-1}(\mathbf{x}_{0:t-1}^{(n)})\eta(\widetilde{\mathbf{x}}_t \mid \mathbf{x}_{0:t-1}^{(n)})\}$ for each n;

Step 3: Resample N particles $\{\mathbf{x}_{0:t}^{(n)}\}_{n=1}^N$ from $\{\widetilde{\mathbf{x}}_{0:t}^{(n)}\}_{n=1}^N$ based on

$\{w(\widetilde{\mathbf{x}}_{0:t}^{(n)})\}_{n=1}^N$ and update the weights $\{w(\mathbf{x}_{0:t}^{(n)})\}_{n=1}^N$;

end

Sequential Importance Sampling with Resampling (SISR) is a common frame-work to implement propagation with the help of importance distributions $\eta(\mathbf{x}_0), \eta(\mathbf{x}_1 \mid \mathbf{x}_0), \ldots, \eta(\mathbf{x}_T \mid \mathbf{x}_{0:T-1})$ and resampling (Liu and Chen (1995, 1998)). To briefly note some key features of SISR in the algorithm above:

1. **Step 1**: In this step, one descendant is sampled for each particle.

2. **Step 2**: In this step, propagation occurs with the help of importance dis-tributions. The importance weights are usually uneven in practice.

3. **Step 3**: This step involves resampling from the propagated particles. If Step 3 is omitted, the SISR framework reduces to sequential importance sampling (SIS).

The necessity of Step 3 depends on the importance weights: if the importance weights are all constant, resampling only reduces the distinction of the particles and thus increases Monte Carlo variance. However, the importance weights are usually uneven in practice; then without Step 3, some of the importance weights evolving in Step 2 may eventually decay to zero during propagation, which is known as weight degeneracy.

S2 Proper weighting of the proposed sampling scheme

Recall $p_t(\mathbf{x}_{0:t})$ denotes the auxiliary distribution for step t in Section 3.2 in this context. Let $\mathcal{S}_0 = \{\mathbf{x}_0^{(n_0)}; n_0 = 1, \ldots, NM\}$ and $\mathcal{S}_t = \mathcal{S}_{t-1} \cup \{(\mathbf{x}_0^{(n_0)}, \ldots, \mathbf{x}_t^{(n_t)}); n_0 = 1, \ldots, NM; n_1 = 1, \ldots, M; \ldots; n_t = 1, \ldots, M\}$, i.e., $\mathcal{S}_t$ is the collection of up-sampled particles up to time t. Any $\mathbf{x}_{0:t}^* = (\mathbf{x}_0^{(n_0^*)}, \ldots, \mathbf{x}_t^{(n_t^*)}) \in \mathcal{S}_t$ satisfies $\mathbf{x}_{0:s}^* = (\mathbf{x}_0^{(n_0^*)}, \ldots, \mathbf{x}_s^{(n_s^*)}) \in \mathcal{S}_t$ for $s \leq t$. Let $\eta(\mathbf{x}_{0:t})$ denote $\eta(\mathbf{x}_0) \prod_{s=1}^t \eta(\mathbf{x}_s \mid \mathbf{x}_{0:s-1})$ and $\mathcal{Q}_{0:t}$ denote the σ-algebra generated by $Q(\mathbf{x}_{0:t})$ conditional on $\mathcal{S}_t$.

When $t = 0$, it is obvious that $E_{\mathcal{Q}_0}\{Q(\mathbf{x}_0^*) \mid \mathcal{S}_0\} = \frac{p_0(\mathbf{x}_0^*)}{\eta(\mathbf{x}_0^*)}$ and

$$E_\eta\left[E_{\mathcal{Q}_0}\{Q(\mathbf{x}_0^*) \mid \mathcal{S}_0\}\right] = E_\eta\left\{\frac{p_0(\mathbf{x}_0^*)}{\eta(\mathbf{x}_0^*)}\right\} = \int \frac{p_0(\mathbf{x}_0^*)}{\eta(\mathbf{x}_0^*)}\eta(\mathbf{x}_0^*)d\mathbf{x}_{0:t}^* = 1$$

and for any square integrable function $f_0(\mathbf{x}_0)$,

$$E_\eta \left[E_{\mathcal{Q}_0} \{ f_0(\mathbf{x}_0^*) Q(\mathbf{x}_0^*) \mid \mathcal{S}_0 \} \right] = E_\eta \left[f_0(\mathbf{x}_0^*) E_{\mathcal{Q}_0} \{ Q(\mathbf{x}_0^*) \mid \mathcal{S}_0 \} \right]$$

$$= E_\eta \left\{ f_0(\mathbf{x}_0^*) \frac{p_0(\mathbf{x}_0^*)}{\eta(\mathbf{x}_0^*)} \right\}$$

$$= \int f_0(\mathbf{x}_0^*) \frac{p_0(\mathbf{x}_0^*)}{\eta(\mathbf{x}_0^*)} \eta(\mathbf{x}_0^*) d\mathbf{x}_0^*$$

$$= E_{p_0} \{ h(\mathbf{x}_0^*) \} ,$$

which justifies the proper weighting condition.

We now proceed by induction: assume $E_{\mathcal{Q}_{0:t-1}} \{ Q(\mathbf{x}_{0:t-1}^*) \mid \mathcal{S}_{t-1} \} = \frac{p_{t-1}(\mathbf{x}_{0:t-1}^*)}{\eta(\mathbf{x}_{0:t-1}^*)}$

holds. Conditional on $\mathcal{S}_t$ and $Q(\mathbf{x}_{0:t-1})$ for all $\mathbf{x}_{0:t-1} \in \mathcal{S}_{t-1}$, and our SMC pro-

duces $Q(\mathbf{x}_{0:t}^*)$ as

$$Q(\mathbf{x}_{0:t}^*) = \begin{cases} \dfrac{Q(\mathbf{x}_{0:t-1}^*) p_t(\mathbf{x}_{0:t}^*)}{p_{t-1}(\mathbf{x}_{0:t-1}^*) \eta(\mathbf{x}_t^* \mid \mathbf{x}_{0:t-1}^*)} \dfrac{1}{q(\mathbf{x}_{0:t}^*)} & \text{with probability } q(\mathbf{x}_{0:t}^*) \\[6mm] 0 & \text{otherwise} \end{cases}$$

where $q(\mathbf{x}_{0:t}^*) = \min \left\{ \frac{c_t Q(\mathbf{x}_{0:t-1}^*) p_t(\mathbf{x}_{0:t}^*)}{p_{t-1}(\mathbf{x}_{0:t-1}^*) \eta(\mathbf{x}_t^* \mid \mathbf{x}_{0:t-1}^*)}, 1 \right\}$ with c_t being the root of

$$\sum_{\mathbf{x}_{0:t} \in \mathcal{S}_t} \min \left\{ \frac{c_t Q(\mathbf{x}_{0:t-1}) p_t(\mathbf{x}_{0:t})}{p_{t-1}(\mathbf{x}_{0:t-1}) \eta(\mathbf{x}_t \mid \mathbf{x}_{0:t-1})}, 1 \right\} = N.$$

We can rewrite $Q(\mathbf{x}_{0:t}^*)$ as

$$Q(\mathbf{x}_{0:t}^*) = \frac{Q(\mathbf{x}_{0:t-1}^*) p_t(\mathbf{x}_{0:t}^*)}{p_{t-1}(\mathbf{x}_{0:t-1}^*) \eta(\mathbf{x}_t^* \mid \mathbf{x}_{0:t-1}^*)} \frac{1}{q(\mathbf{x}_{0:t}^*)} \mathbf{I}(\mathbf{x}_{0:t}^*),$$

where $\mathbf{I}(\mathbf{x}_{0:t}^*)$ denotes an indicator function defined conditional on $\mathcal{S}_t$ and $Q(\mathbf{x}_{0:t-1}^*)$

with $E_\mathbf{I} \{ \mathbf{I}(\mathbf{x}_{0:t}^*) \mid Q(\mathbf{x}_{0:t-1}^*), \mathcal{S}_t \} = q(\mathbf{x}_{0:t}^*)$. Note that $\mathbf{I}(\mathbf{x}_{0:t}^*)$ is independent of fu-

ture descendants, so for $0 < r < s \le t$,

$$E_{\mathbf{I}}\left\{\mathbf{I}(\mathbf{x}_{0:r}^*) \mid Q(\mathbf{x}_{0:r-1}^*), \mathcal{S}_s\right\} = E_{\mathbf{I}}\left\{\mathbf{I}(\mathbf{x}_{0:r}^*) \mid Q(\mathbf{x}_{0:r-1}^*), \mathcal{S}_r\right\}.$$

Now we can see that $Q(\mathbf{x}_{0:t}^*)$ is proportional to the product of the two random variables $Q(\mathbf{x}_{0:t-1}^*)$ and $\mathbf{I}(\mathbf{x}_{0:t}^*)$, with $\mathbf{I}(\mathbf{x}_{0:t}^*)$ conditional on $Q(\mathbf{x}_{0:t-1}^*)$. By the tower rule, we have

$$
\begin{aligned}
E_{\mathcal{Q}_{0:t}}\left\{Q(\mathbf{x}_{0:t}^*) \mid \mathcal{S}_t\right\} &= E_{\mathcal{Q}_{0:t-1}}\left[\frac{Q(\mathbf{x}_{0:t-1}^*)p_t(\mathbf{x}_{0:t}^*)}{p_{t-1}(\mathbf{x}_{0:t-1}^*)\eta(\mathbf{x}_t^* \mid \mathbf{x}_{0:t-1}^*)}\frac{1}{q(\mathbf{x}_{0:t}^*)}E_{\mathbf{I}}\left\{\mathbf{I}(\mathbf{x}_{0:t}^*) \mid Q(\mathbf{x}_{0:t-1}^*), \mathcal{S}_t\right\}\,\middle|\,\mathcal{S}_t\right] \\
&= E_{\mathcal{Q}_{0:t-1}}\left\{\frac{Q(\mathbf{x}_{0:t-1}^*)p_t(\mathbf{x}_{0:t}^*)}{p_{t-1}(\mathbf{x}_{0:t-1}^*)\eta(\mathbf{x}_t^* \mid \mathbf{x}_{0:t-1}^*)}\,\middle|\,\mathcal{S}_t\right\} \\
&= E_{\mathcal{Q}_{0:t-1}}\left\{Q(\mathbf{x}_{0:t-1}^*) \mid \mathcal{S}_t\right\}\frac{p_t(\mathbf{x}_{0:t}^*)}{p_{t-1}(\mathbf{x}_{0:t-1}^*)\eta(\mathbf{x}_t^* \mid \mathbf{x}_{0:t-1}^*)}.
\end{aligned}
$$

Note that $Q(\mathbf{x}_{0:t-1}^*)$ is only conditional on $\mathcal{S}_{t-1}$ and thus independent of future descendants, so

$$E_{\mathcal{Q}_{0:t-1}}\left\{Q(\mathbf{x}_{0:t-1}^*) \mid \mathcal{S}_t\right\} = E_{\mathcal{Q}_{0:t-1}}\left\{Q(\mathbf{x}_{0:t-1}^*) \mid \mathcal{S}_{t-1}\right\}$$

and thus

$$E_{\mathcal{Q}_{0:t}}\left\{Q(\mathbf{x}_{0:t}^*) \mid \mathcal{S}_t\right\} = E_{\mathcal{Q}_{0:t-1}}\left\{Q(\mathbf{x}_{0:t-1}^*) \mid \mathcal{S}_{t-1}\right\}\frac{p_t(\mathbf{x}_{0:t}^*)}{p_{t-1}(\mathbf{x}_{0:t-1}^*)\eta(\mathbf{x}_t^* \mid \mathbf{x}_{0:t-1}^*)} = \frac{p_t(\mathbf{x}_{0:t}^*)}{\eta(\mathbf{x}_{0:t}^*)}.$$

Therefore,

$$E_{\eta}\left[E_{\mathcal{Q}_{0:t}}\left\{Q(\mathbf{x}_{0:t}^*) \mid \mathcal{S}_t\right\}\right] = E_{\eta}\left\{\frac{p_t(\mathbf{x}_{0:t}^*)}{\eta(\mathbf{x}_{0:t}^*)}\right\} = \int\frac{p_t(\mathbf{x}_{0:t}^*)}{\eta(\mathbf{x}_{0:t}^*)}\eta(\mathbf{x}_{0:t}^*)d\mathbf{x}_{0:t}^* = 1$$

and for any square integrable function $f_t(\mathbf{x}_{0:t})$,

$$
\begin{aligned}
E_\eta\left[E_{\mathcal{Q}_{0:t}}\{f_t(\mathbf{x}_{0:t}^*)Q(\mathbf{x}_{0:t}^*) \mid \mathcal{S}_t\}\right] &= E_\eta\left[f_t(\mathbf{x}_{0:t}^*)E_{\mathcal{Q}_{0:t}}\{Q(\mathbf{x}_{0:t}^*) \mid \mathcal{S}_t\}\right] \\
&= E_\eta\left\{f_t(\mathbf{x}_{0:t}^*)\frac{p_t(\mathbf{x}_{0:t}^*)}{\eta(\mathbf{x}_{0:t}^*)}\right\} \\
&= \int f_t(\mathbf{x}_{0:t}^*)\frac{p_t(\mathbf{x}_{0:t}^*)}{\eta(\mathbf{x}_{0:t}^*)}\eta(\mathbf{x}_{0:t}^*)d\mathbf{x}_{0:t}^* \\
&= E_{p_t}\{f_t(\mathbf{x}_{0:t})\},
\end{aligned}
$$

which justifies the proper weighting condition.

S3   Minimization of the conditional expected squared error loss

Fearnhead and Clifford (2003) have shown that in the downsampling step, only N of the $Q(\mathbf{x}_{0:t}^{(n,m)})$'s are non-zero so that for some random variable $C_t^{(n,m)}$ with σ-algebra $\mathcal{C}$, we have

$$
Q(\mathbf{x}_{0:t}^{(n,m)}) = \begin{cases} C_t^{(n,m)} & \text{with probability } q(\mathbf{x}_{0:t}^{(n,m)}) \\ \\ 0 & \text{otherwise} \end{cases}
$$

Assume the set of particles $\{\mathbf{x}_{0:t}^{(n,m)}, n = 1, \ldots, N \text{ and } m = 1, \ldots, M\}$ are obtained after upsampling and let $\gamma_t^{(n,m)} = \{p_t(\mathbf{x}_{0:t}^{(n,m)})/\eta(\mathbf{x}_{0:t}^{(n,m)})\}/\sum_{n=1}^{N}\sum_{m=1}^{M}\{p_t(\mathbf{x}_{0:t}^{(n,m)})/\eta(\mathbf{x}_{0:t}^{(n,m)})\}$,

then the expected squared error loss for $Q(\mathbf{x}_{0:t}^{(n,m)})$ can be written as

$$E_{\mathcal{Q}_{0:t}}\left\{\left(Q(\mathbf{x}_{0:t}^{(n,m)}) - \gamma_t^{(n,m)}\right)^2 \mid \mathcal{S}_t\right\} = E_{\mathcal{C}}\left[E_{\mathcal{Q}_{0:t}}\left\{\left(Q(\mathbf{x}_{0:t}^{(n,m)}) - \gamma_t^{(n,m)}\right)^2 \mid \mathcal{S}_t, C^{(n,m)}\right\}\middle| \mathcal{S}_t\right]$$

$$= q(\mathbf{x}_{0:t}^{(n,m)})\left\{E_{\mathcal{C}}\left(C^{(n,m)} - \gamma_t^{(n,m)} \mid \mathcal{S}_t\right)\right\}^2$$

$$+ q(\mathbf{x}_{0:t}^{(n,m)})\mathrm{Var}(C^{(n,m)} \mid \mathcal{S}_t) + \left\{1 - q(\mathbf{x}_{0:t}^{(n,m)})\right\}\gamma_t^{(n,m)2},$$

which is minimized when $C^{(n,m)}$ is a constant for each n and m. To ensure proper weighting as shown in Section S2 of the supplement, we set $C^{(n,m)} = w(\mathbf{x}_{0:t}^{(n,m)})/q(\mathbf{x}_{0:t}^{(n,m)})$ where $w(\mathbf{x}_{0:t}^{(n,m)})$ denotes the upsampling weight of $\mathbf{x}_{0:t}^{(n,m)}$, and then the conditional expected squared error loss is minimized subject to $\sum_{n=1}^{N}\sum_{m=1}^{M} q(\mathbf{x}_{0:t}^{(n,m)}) \leq N$, since

$$\min\left[E_{\mathcal{Q}_{0:t}}\left\{\sum_{n=1}^{N}\sum_{m=1}^{M}\left(Q(\mathbf{x}_{0:t}^{(n,m)}) - \gamma_t^{(n,m)}\right)^2\middle| \mathcal{S}_t\right\}\right] = \sum_{n=1}^{N}\sum_{m=1}^{M}\left\{w(\mathbf{x}_{0:t}^{(n,m)})/q(\mathbf{x}_{0:t}^{(n,m)}) - \gamma_t^{(n,m)}\right\}^2.$$

S4 Details of the energy function

Given a target segment represented by the dihedral angles $\mathbf{x}_{0:T}$ and for $t \in \{1,\ldots,T\}$, the details of the two components of the energy function H are as follows:

- *Atomic interactions* − the energy of atomic interactions (denoted by H_a) is derived from the pairwise distances between atoms. Notably, we exclude atom pairs that either share bonds or belong to the same amino acid. To calculate this energy, we employ the distance-scaled, finite ideal-gas reference

(DFIRE) state potential, as proposed by Zhou and Zhou (2002). DFIRE assigns a score to the pairwise distance for each atom type. Additionally, we calculate the pairwise distance between the current amino acid $\mathbf{x}_t$ and the C_α atom after the end of the segment (i.e., C_α^{T+2}) to determine whether the segment will satisfy the constraint of forming a continuous backbone with the rest of the protein (Wong, Liu, and Kou (2017)). In cases where there are *steric clashes* (i.e., overlapping atoms in close proximity defined by a value of 8 in the DFIRE table), or when the conformation cannot form a continuous backbone with the rest of the protein, we assign $H_a = +\infty$.

Let $\mathbf{A}_0$ denote the collection of all the atoms in the protein, excluding those in the segment $\mathbf{x}_{0:T}$. For a given partially sampled conformation $\mathbf{x}_{0:t-1}$, the subsequent propagation step requires evaluating the incremental energy of $\mathbf{x}_t$, denoted as $H(\mathbf{x}_t \mid \mathbf{x}_{0:t-1})$. To calculate this, we compute and accumulate the energy scores of the four backbone atoms C^t, O^t, N^{t+1}, and C_α^{t+1}. For instance, the energy of N^{t+1}, denoted by $En(N^{t+1} \mid \mathbf{x}_{0:t-1})$, is the cumulative energy score between N^{t+1} and (a) all atoms in $\mathbf{A}_0$, (b) all atoms in $\mathbf{x}_{0:t-1}$. The DFIRE potential provides a lookup table for energy scores based on (i) the types of two atoms and (ii) the Euclidean distance between the two atoms, i.e., $DFIRE(atom_1, atom_2)$ can be defined as a function of the atom type of $atom_1$, the atom type of $atom_2$, and the Euclidean distance

between atom$_1$ and atom$_2$. Therefore,

$$En(N^{t+1} \mid \mathbf{x}_{0:t-1}) = \sum_{s=0}^{t-1} \{ \text{DFIRE}(N^{t+1}, C^s) + \text{DFIRE}(N^{t+1}, O^s) +$$

$$\text{DFIRE}(N^{t+1}, N^{s+1}) + \text{DFIRE}(N^{t+1}, C_\alpha^{s+1}) \} \qquad \text{(S4.1)}$$

$$+ \sum_{A \in \mathbf{A}_0} \text{DFIRE}(N^{t+1}, A).$$

Likewise, we can compute $En(C^t \mid \mathbf{x}_{0:t-1})$, $En(O^t \mid \mathbf{x}_{0:t-1})$ and $En(C_\alpha^{t+1} \mid \mathbf{x}_{0:t-1})$ by substituting N^{t+1} with the corresponding atom in (S4.1). $H_a(\mathbf{x}_t \mid \mathbf{x}_{0:t-1})$ is then assigned by

$$H_a(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}) = En(C^t \mid \mathbf{x}_{0:t-1}) + En(O^t \mid \mathbf{x}_{0:t-1})$$

$$+ En(N^{t+1} \mid \mathbf{x}_{0:t-1}) + En(C_\alpha^{t+1} \mid \mathbf{x}_{0:t-1}).$$

- *Dihedral angles* − the energy of backbone dihedral angles (denoted by H_θ) is calculated from the empirical distributions $\tilde{p}(\phi_t, \psi_t, \omega_t)$ derived from historical protein data, independently for each amino acid with

$$\tilde{p}(\phi_t, \psi_t, \omega_t) = \tilde{p}(\phi_t, \psi_t)\tilde{p}(\omega_t),$$

where $\tilde{p}(\phi_t, \psi_t)$ is based on the amino acid type of a_t and consists of discrete bins of every 5° for each angle (which we assume to be uniformly distributed within each bin), i.e., each empirical distribution is stored in a 72 by 72 matrix; $\tilde{p}(\omega_t)$ is a Gaussian distribution with mean 180° and standard deviation 3° independent of (ϕ_t, ψ_t). Following that work of Wong, Liu, and Kou (2017), we then set $H_\theta(\mathbf{x}_t) = -\log\{\tilde{p}(\phi_t, \psi_t, \omega_t)\}$.

Then the incremental energy of $\mathbf{x}_t$ is defined as

$$H(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}) = H_a(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}) + H_\theta(\mathbf{x}_t). \tag{S4.2}$$

The auxiliary distribution at t is

$$p_t(\mathbf{x}_{0:t}) \propto \exp\{-H_a(\mathbf{x}_0)\}\tilde{p}(\mathbf{x}_0) \prod_{s=1}^{t} \exp\{-H_a(\mathbf{x}_s \mid \mathbf{x}_{0:s-1})\}\tilde{p}(\mathbf{x}_s).$$

It is convenient to set $\eta(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}) = \tilde{p}(\mathbf{x}_t)$ as the importance distribution: to draw $\mathbf{x}_t$ from $\tilde{p}(\mathbf{x}_t)$, we sample a bin for (ϕ, ψ) from its empirical distribution then draw a (ϕ_t, ψ_t) value uniformly from the selected bin, and ω_t is drawn from its normal distribution. The equation for upsampled weights thus simplifies to

$$w(\mathbf{x}_{0:t}) = w(\mathbf{x}_{0:t-1}) \frac{p_t(\mathbf{x}_{0:t})}{p_{t-1}(\mathbf{x}_{0:t-1})\eta(\mathbf{x}_t \mid \mathbf{x}_{0:t-1})}$$

$$= w(\mathbf{x}_{0:t-1}) \exp\{-H_a(\mathbf{x}_t \mid \mathbf{x}_{0:t-1})\}. \tag{S4.3}$$

S5   UDSMC experiment for different values of M

The performance of UDSMC will be influenced by the choice of the upsample size M. For a finite space $\mathcal{X}$, a natural choice for M can be to take $M = |\mathcal{X}|$; however, there may not be an intuitive choice for M for a continuous space $\mathcal{X}$. Thus to choose a reasonable value of M for our application, in this section we run a preliminary experiment using different values of M (with a fixed computational budget, i.e., holding MN constant).

For the experiment, we focus on sampling a length 10 segment at amino acid positions 282–291 of the protein with PDB ID *1ds1A*, and thus the dimension of the Boltzmann distribution is 30 in this case. The protein structural quantities of interest are the Boltzmann averages for the number of atomic contacts of C_α at each of the positions 283–292, denoted as $n(C_\alpha{}^{283}), \ldots, n(C_\alpha{}^{292})$. We vary the values of N and M in the experiment, using $NM \in \{10^5, 5 \times 10^5, 10^6\}$ and $M \in \{1, 5, 10, 20, 50, 100, 200, 500, 1000\}$. For each combination of N and M, we conduct 100 independent repetitions of UDSMC. Notably, when $M = 1$, UDSMC is equivalent to SISR.

Table 1 presents the Monte Carlo variance of the estimates for different values of M. While RMSEs to the ground truth would be more informative, the ground truth is very expensive to compute, and Monte Carlo variance can still provide useful guidance for a reasonable choice of M. When M is too large or too small (e.g., $M \in \{1, 200, 500, 1000\}$), some of the 100 repetitions end prematurely due to all propagated particles having zero weights (i.e., $H_a = \infty$). A small M can lead to many duplicate particles among the N particles after each propagation step, resulting in insufficient exploration of the target distribution, which may eventually lead to all dead ends after successive iterations of propagation and resampling. On the other hand, a large M implies a small particle size N (for the same computational cost), which may be inadequate to avoid dead ends.

For values of M in the range of $\{5, 10, 20, 50, 100\}$, all UDSMC repetitions completed successfully, and we report the variance of the 100 estimates for each structural quantity. With M fixed, the Monte Carlo variances of the estimates decreased as NM increased. Furthermore, with NM fixed, the variances tended to decrease as M increased from 5 to 20, and then increased again as M increased from 20 to 100. These results highlight the need for a balance between N and M given a fixed computational budget. In this specific experiment, $M = 20$ produced the lowest variances and can be a reasonable choice for the protein sampling problem, and we adopt this value in the simulation study and the real data example.

Table 1: Variance of the UDSMC estimates over 100 repetitions when estimating the Boltzmann average of the atomic contacts of C_α at each position from 283 to 292, for different combinations of NM and M. "NA" indicates that the algorithm failed to complete for that combination of N and M.

	M	$n(C_\alpha^{283})$	$n(C_\alpha^{284})$	$n(C_\alpha^{285})$	$n(C_\alpha^{286})$	$n(C_\alpha^{287})$	$n(C_\alpha^{288})$	$n(C_\alpha^{289})$	$n(C_\alpha^{290})$	$n(C_\alpha^{291})$	$n(C_\alpha^{292})$
$NM = 100000$	1,2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	5	0.277	1.557	3.547	10.004	3.525	3.467	31.481	8.299	4.119	0.599
	10	0.160	0.959	3.378	10.335	2.525	4.194	25.199	10.376	3.843	0.680
	20	0.181	0.814	2.695	7.421	2.633	2.269	12.973	9.321	3.119	0.453
	50	0.313	1.908	4.237	10.802	3.890	3.157	19.476	10.196	4.070	0.576
	100	0.428	3.117	7.379	20.195	6.050	6.695	26.506	14.656	5.471	0.792
	200,500,1000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$NM = 500000$	1,2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	5	0.083	0.671	1.198	5.014	1.179	1.610	13.895	3.041	1.417	0.346
	10	0.063	0.560	1.183	3.489	1.167	1.421	9.391	3.400	1.443	0.258
	20	0.034	0.202	0.578	2.042	0.696	0.749	3.629	2.828	0.856	0.113
	50	0.061	0.345	1.379	2.542	0.945	0.799	4.270	2.824	1.105	0.136
	100	0.139	0.697	1.812	5.946	1.696	2.192	8.269	5.148	1.355	0.228
	200,500,1000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
$NM = 1000000$	1,2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	5	0.051	0.374	0.637	2.504	0.523	1.058	6.682	1.914	0.590	0.096
	10	0.035	0.234	0.499	1.741	0.505	0.860	5.571	1.258	0.527	0.147
	20	0.024	0.132	0.556	0.762	0.315	0.226	1.222	1.327	0.388	0.070
	50	0.036	0.179	0.535	1.421	0.521	0.482	2.498	1.504	0.512	0.065
	100	0.058	0.369	1.070	2.529	0.744	1.088	3.364	2.319	0.857	0.129
	200,500,1000	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

S6　Pseudocode for SISR with one-step lookahead

---

sample $\{X_0^{(n)}, n = 1, \ldots, N\}$ from the proposal distribution $\eta(X_0)$;

for *n in 1:N* **do**

　　sample L descendants of $X_0^{(n)}$ and form $\{X_{0:1}^{(n,l)}, l = 1, \ldots, L\}$ from the proposal distribution

　　$\eta(X_1 \mid X_0^{(n)})$;

　　set $\widetilde{w}_{0:1}(X_{0:1}^{(n,l)}) = \dfrac{p_{0:1}(X_{0:1}^{(n,l)})}{\eta(X_0^{(n,m)})\eta(X_1^{(l)} \mid X_0^{(n)})}$ and $w_{0:1}(X_{0:1}^{(n,l)}) = \widetilde{w}_{0:1}(X_{0:1}^{(n,l)})/\sum_{l=1}^{L}\widetilde{w}_{0:1}(X_{0:1}^{(n,l)})$

end

resample particles $\{\tilde{X}_0^{(n)}, n = 1, \ldots, N\}$ proportional to $\{w_{0:1}(X_{0:1}^{(n,l)}), n = 1, \ldots, N, l = 1, \ldots, L\}$;

set $\{X_0^{(n)}, n = 1, \ldots, N\} = \{\tilde{X}_0^{(n)}, n = 1, \ldots, N\}$ with weights $\{W_0^{(n)}, n = 1, \ldots, N\}$;

for *t in 1:T-2* **do**

　　for *n in 1:N* **do**

　　　　sample a descendants X_t^* of $X_{t-1}^{(n)}$ and form $X_{0:t}^{(n)} = (X_{t-1}^{(n)}, X_t^*)$ from the proposal

　　　　distribution $\eta(X_t \mid X_{t-1}^{(n)})$;

　　　　sample L descendants of $X_{0:t}^{(n)}$ and form $\{X_{0:t+1}^{(n,l)}, l = 1, \ldots, L\}$ from the proposal

　　　　distribution $\eta(X_{t+1} \mid X_{0:t}^{(n)})$;

　　　　set $\widetilde{w}_{0:t+1}(X_{0:t+1}^{(n,l)}) = W_{0:t-1}^{(n)} \dfrac{p_{0:t+1}(X_{0:t+1}^{(n,l)})}{p_{0:t-1}(X_{0:t-1}^{(n)})\eta(X_t^* \mid X_{0:t-1}^{(n)})\eta(X_{t+1}^{(l)} \mid X_{0:t}^{(n)})}$ and

　　　　$w_{0:t+1}(X_{0:t+1}^{(n,l)}) = \widetilde{w}_{0:t+1}(X_{0:t+1}^{(n,l)})/\sum_{l=1}^{L}\widetilde{w}_{0:t+1}(X_{0:t+1}^{(n,l)})$;
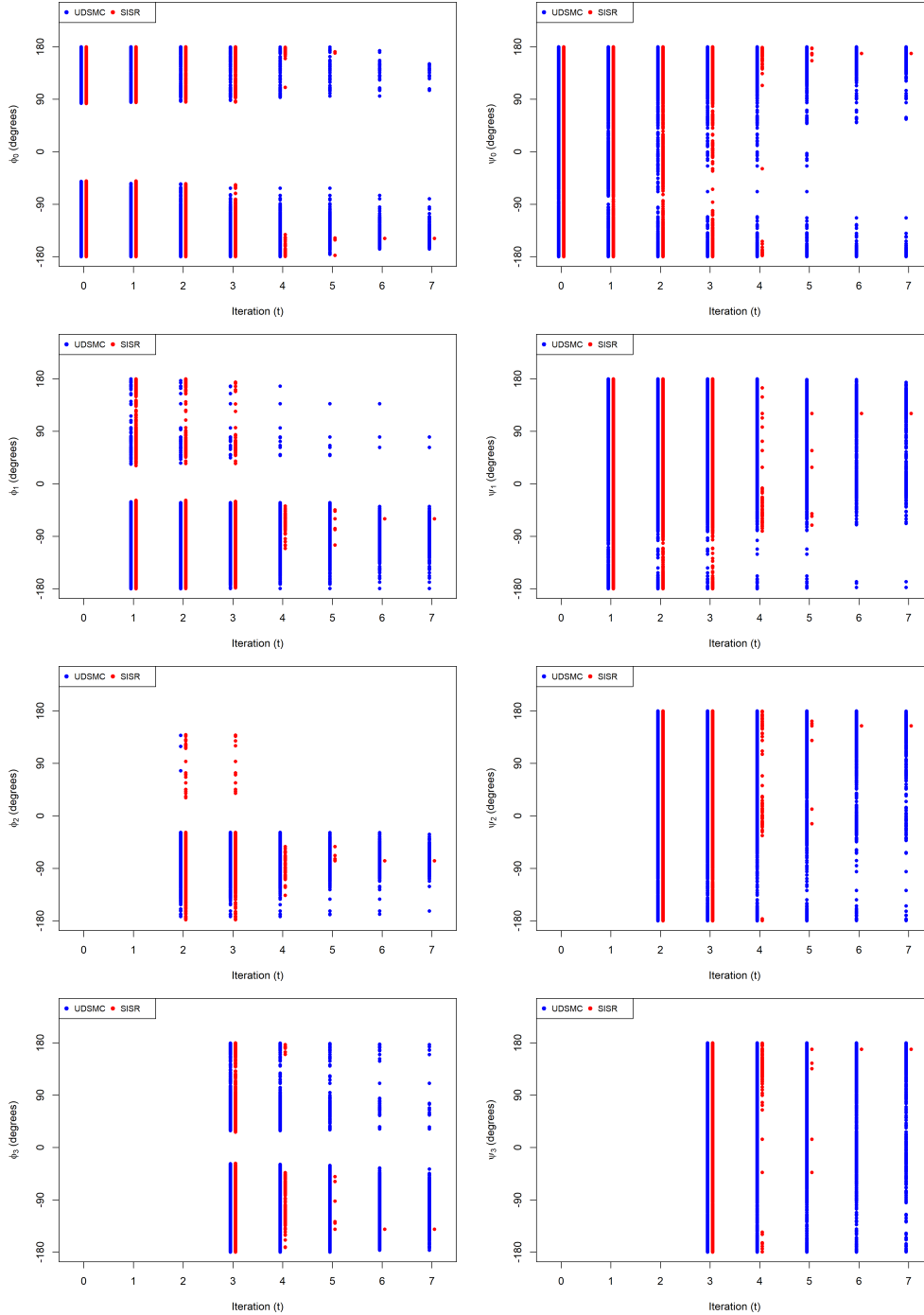
　　end

　　resample particles $\{\tilde{X}_{0:t}^{(n)}, n = 1, \ldots, N\}$ proportional to

　　$\{w_{0:t+1}(X_{0:t+1}^{(n,l)}), n = 1, \ldots, N, l = 1, \ldots, L\}$;

　　set $\{X_{0:t}^{(n)}, n = 1, \ldots, N\} = \{\tilde{X}_{0:t}^{(n)}, n = 1, \ldots, N\}$ with weights $\{W_{0:t}^{(n)}, n = 1, \ldots, N\}$;

end

for *n in 1:N* **do**

　　sample a descendants X_{T-1}^* of $X_{0:T-2}^{(n)}$ and form $X_{0:T-1}^{(n)} = (X_{T-2}^{(n)}, X_{0:T-1}^*)$ from the

　　proposal distribution $\eta(X_{T-1} \mid X_{T-2}^{(n)})$;

　　sample L descendants of $X_{0:T-1}^{(n)}$ and form $\{X_{0:T}^{(n,l)}, l = 1, \ldots, L\}$ from the proposal

　　distribution $\eta(X_T \mid X_{0:T-1}^{(n)})$;

　　set $\widetilde{w}_{0:T}(X_{0:T}^{(n,l)}) = W_{0:T-2}^{(n)} \dfrac{p_{0:T}(X_{0:T}^{(n,l)})}{p_{0:T-2}(X_{0:T-2}^{(n)})\eta(X_{T-1}^* \mid X_{0:T-2}^{(n)})\eta(X_T^{(l)} \mid X_{0:T-1}^{(n)})}$ and

　　$w_{0:T}(X_{0:T}^{(n,l)}) = \widetilde{w}_{0:T}(X_{0:T}^{(n,l)})/\sum_{l=1}^{L}\widetilde{w}_{0:T}(X_{0:T}^{(n,l)})$;

end

resample particles $\{\tilde{X}_{0:T}^{(n)}, n = 1, \ldots, N\}$ proportional to $\{w_{0:T}(X_{0:T}^{(n,l)}), n = 1, \ldots, N, l = 1 \ldots, L\}$;

set $\{X_{0:T}^{(n)}, n = 1, \ldots, N\} = \{\tilde{X}_{0:T}^{(n)}, n = 1, \ldots, N\}$ with weights $\{W_{0:T}^{(n)}, n = 1, \ldots, N\}$;

# S7 Particle diversity of UDSMC and SISR

The resampling step of the t-th SMC iteration ($t = 0, \ldots, 7$) produces N particles for $\mathbf{x}_{0:t}$. Here, we focus on the values of $(\phi_0, \psi_0, \ldots, \phi_t, \psi_t)$ since the ω dihedral angle is usually close to $180°$. Thus, for example, we can examine how the values of (ϕ_0, ψ_0) among resampled particles evolve for $t = 0, \ldots, 7$. Similarly, for any $s = 1, \ldots, 7$, we can examine how the values of (ϕ_s, ψ_s) among resampled particles evolve for $t = s, \ldots, 7$. Due to the possibility of weight degeneracy, the diversity of these values (and hence the diversity of the paths of the particles) may decay as t increases. We compare UDSMC ($N = 20000, M = 20$) and SISR ($N = 400000$) by plotting these values in Figures 1 and 2 for one repetition of the simulation study in Section 4 of the main text. The plots show that the diversity of these values decays much faster for SISR. Taking ϕ_0 as an example, both UDSMC and SISR have values that are similarly diverse up to $t = 3$, and subsequently for $t = 4, \ldots, 7$, only UDSMC is able to maintain diversity in ϕ_0 (while SISR decays to just a few unique values). A similar pattern can be seen for other ϕ_s and ψ_s. In this sense, UDSMC is much better at mitigating decay in the diversity of the paths of the particles compared to SISR, which provides an intuitive explanation for the better performance of UDSMC in this application.

Figure 1: Values of ϕ_s (left panels) and ψ_s (right panels) among resampled particles, for $s = 0, 1, 2, 3$ after each iteration (t) of UDSMC (blue) and SISR (red), for the protein segment in the simulation study.

Figure 2: Values of ϕ_s (left panels) and ψ_s (right panels) among resampled particles, for $s = 4, 5, 6, 7$ after each iteration (t) of UDSMC (blue) and SISR (red), for the protein segment in the simulation study.
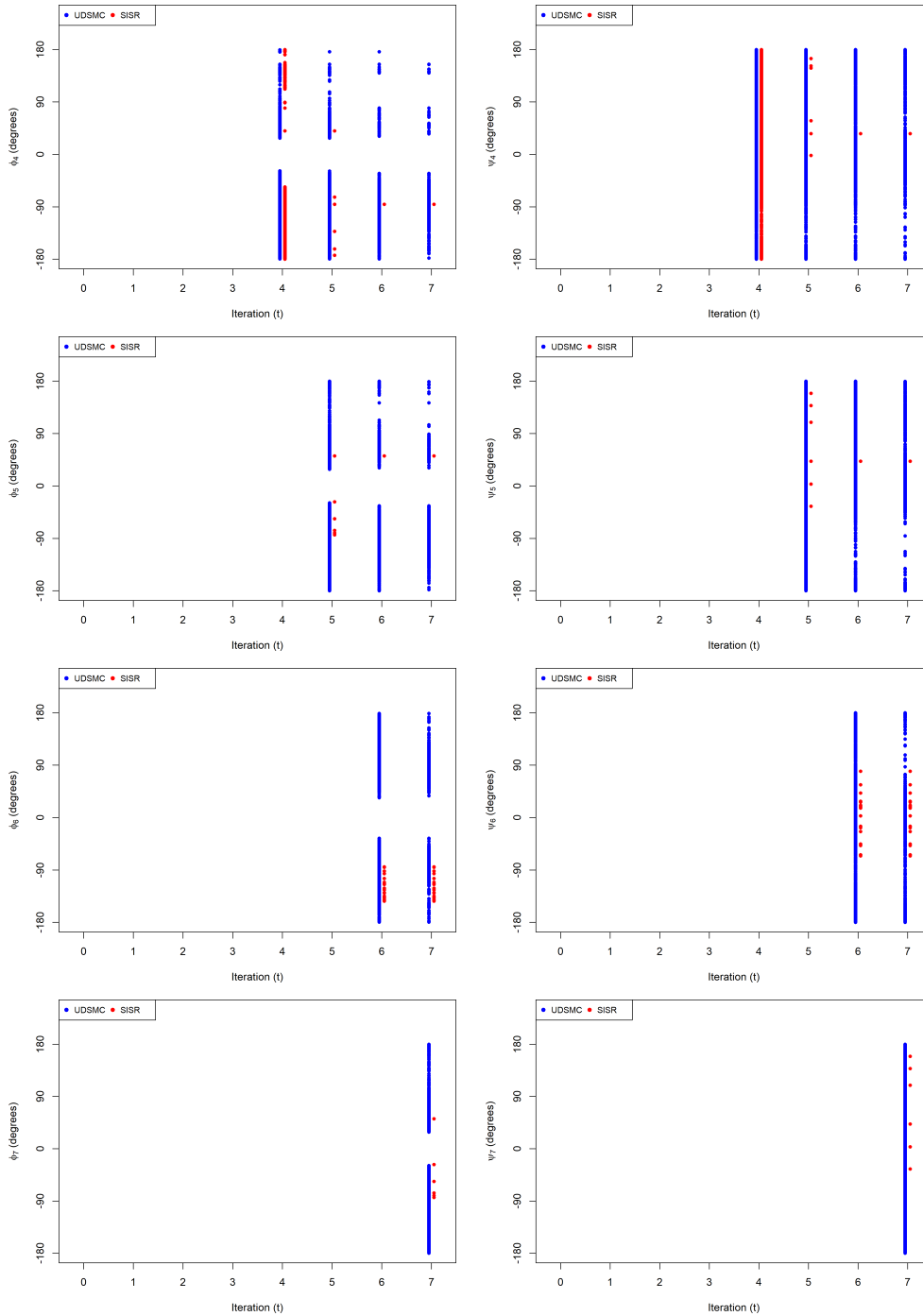
Bibliography

Liu, J. S. and R. Chen (1995). Blind deconvolution via sequential imputations. *Journal of the American Statistical Association 90*(430), 567–576.

Liu, J. S. and R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association 93*(443), 1032–1044.

Wong, S. W., J. S. Liu, and S. Kou (2017). Fast de novo discovery of low-energy protein loop conformations. *Proteins: Structure, Function, and Bioinformatics 85*(8), 1402–1412.

Zhou, H. and Y. Zhou (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science 11*(11), 2714–2726.