Weighted ensemble sampling and optimization

David Aristoff

Colorado State University

October 2017

Weighted ensemble (WE)

WE (Huber, Kim 1996) is a statistically exact technique for path sampling.

Basic algorithm.

Resample from paths or path endpoints to get "good" spatial sampling.
Then assign weights so that the resulting statistical distribution is exact.

"Good" sampling usually obtained by binning – keeping a user-determined number of replicas per bin.

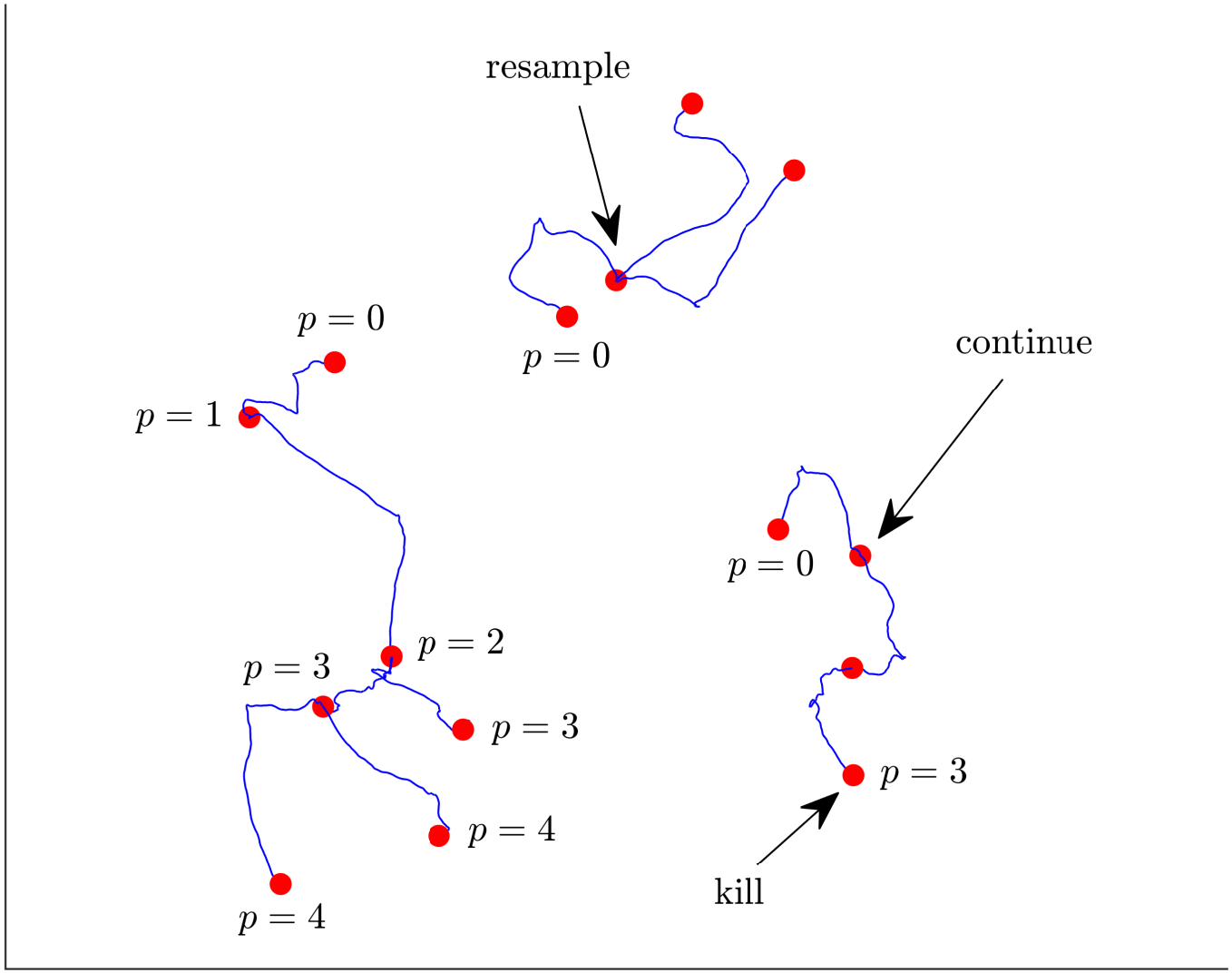


Figure: Visualization of WE. Blue = stochastic trajectory, red = resampling times.

Current usage

- Software: `https://westpa.github.io/westpa/publications.html`
  Zwier, Adelman, Kaus, Pratt, Wong, Rego, Suárez, Lettieri, Wang, Grabe, Zuckerman, Chong
- Package is described in J. Chem. Theory Comput., 11: 800-809 (2015)
- Currently $\sim$ 30 related publications, most after year 2010

1 minute summary

- We derive a new replica allocation strategy for WE (D. Aristoff, ESAIM: M2AN 2017)
- It is a variance-reduction strategy that is optimal in some sense
- The strategy requires a coarse/cheap model to implement in practice

1 minute summary

- We derive a new replica allocation strategy for WE (D. Aristoff, ESAIM: M2AN 2017)
- It is a variance-reduction strategy that is optimal in some sense
- The strategy requires a coarse/cheap model to implement in practice

- It works on toy models. Will it work on real problems?
- Close connection to SMC work of Del Moral, Garnier and others

Setting

Assumption.

$(X_p)_{p=0,1,\ldots}$ is a Markov chain: its future behavior depends only on the present.

Example.

(Y_t) is stochastic MD, and (X_p) is obtained from (Y_t) along a time sequence, e.g.

- $X_p = Y_{p\Delta t}$, where Δt is a fixed time step/resampling time
- $X_p = Y_{\tau_p}$ where τ_p is the pth crossing time of some "milestones"

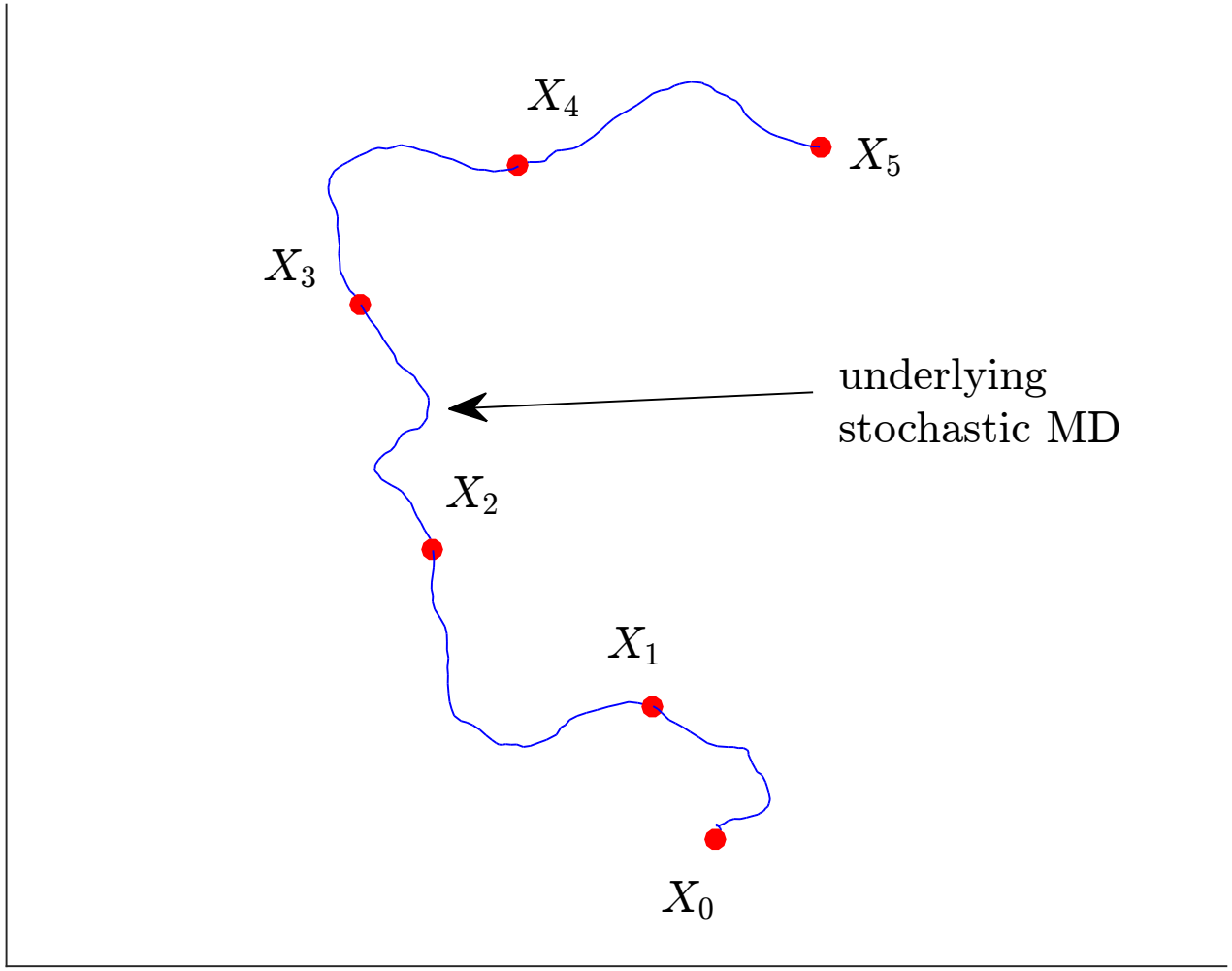


Figure: Blue curve $= (Y_t)$, red dots $= (X_p)$. E.g. $X_p = Y_{p\Delta t}$, or...

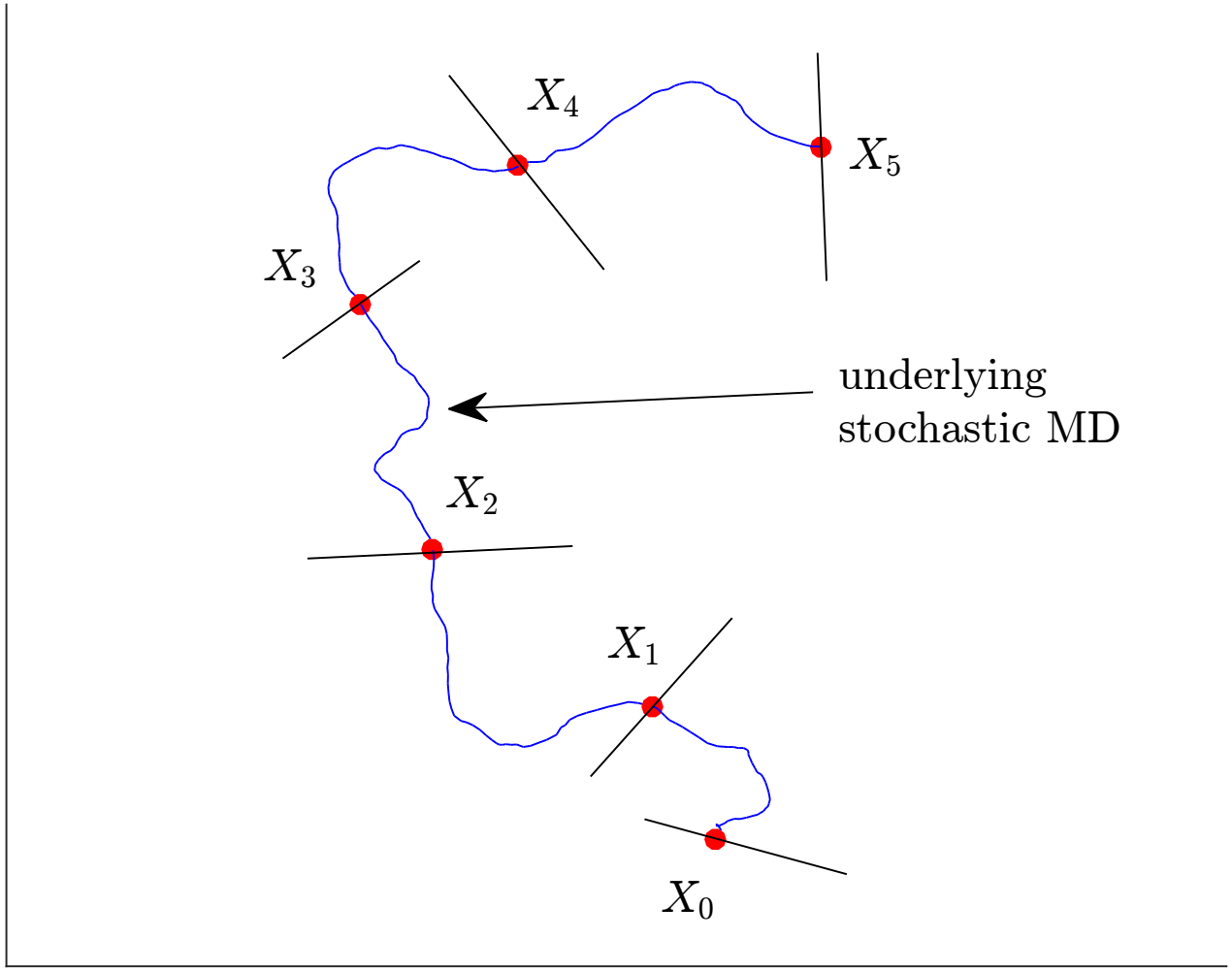


Figure: $X_p = Y_{\tau_p}$, where τ_p is pth crossing of a "milestone."

Sampling rules

Rules.

- Each replica always has a positive probability to survive
- weight of child $=$ (weight of parent)$/\mathbb{E}(\#$ of children)
- Children evolve independently according to the law of (X_p)

Parents have a *random* # of children. Total weight not conserved (but it is on average).

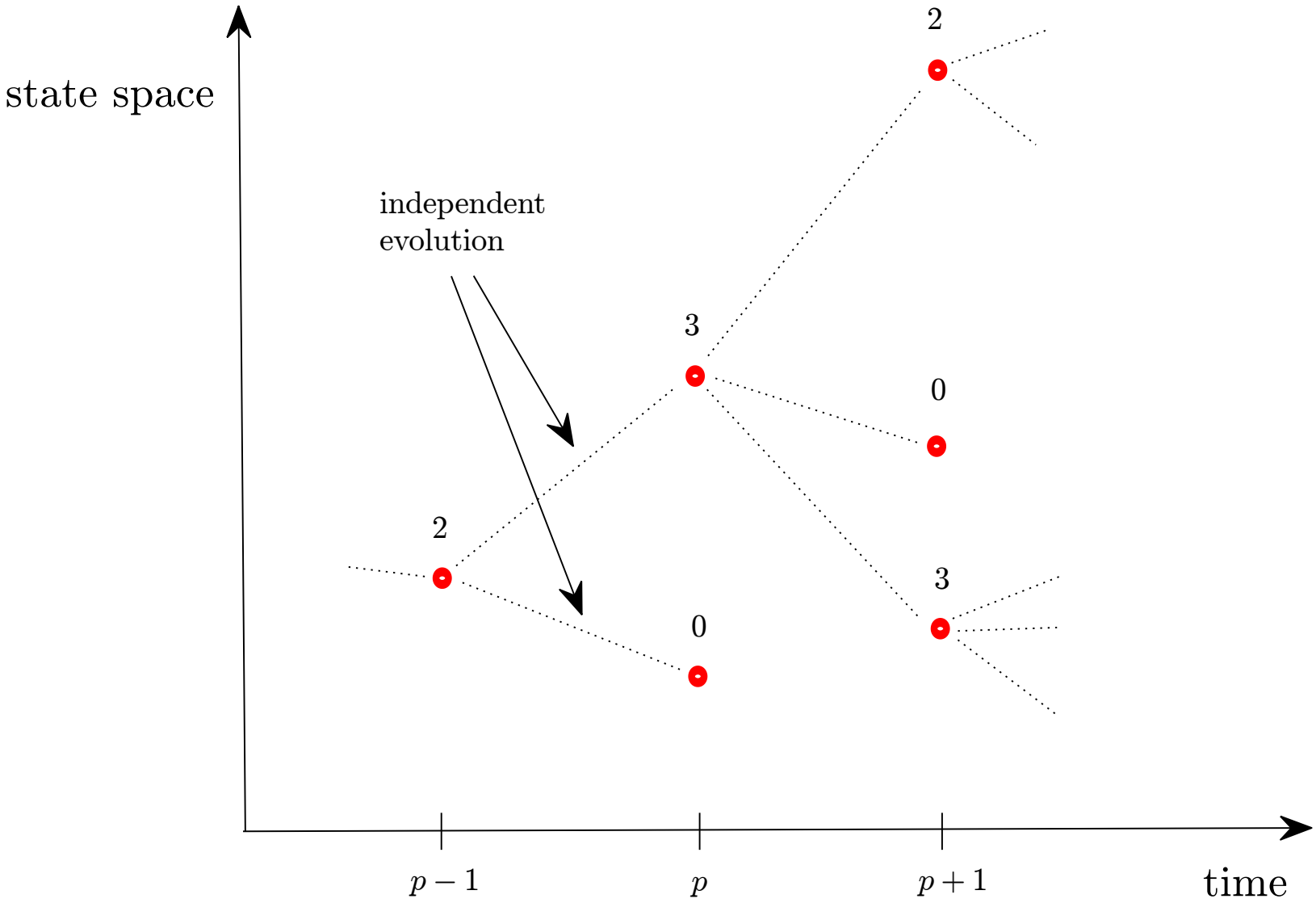


Figure: Red dots: replicas of (X_p). Above dots: # of times replica is resampled.

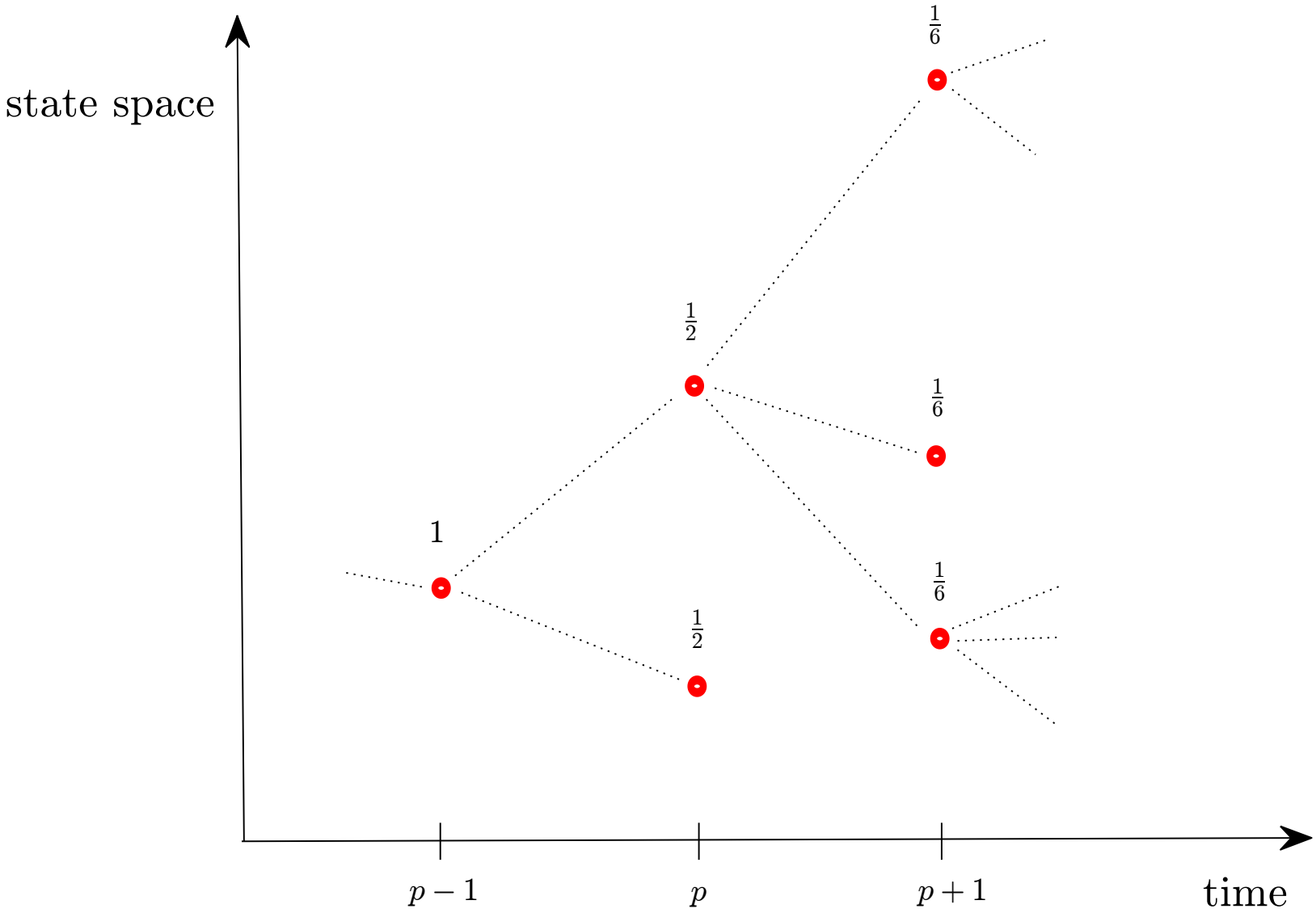


Figure: Red dots: replicas of (X_p). Above dots: (possible) weights of replicas.

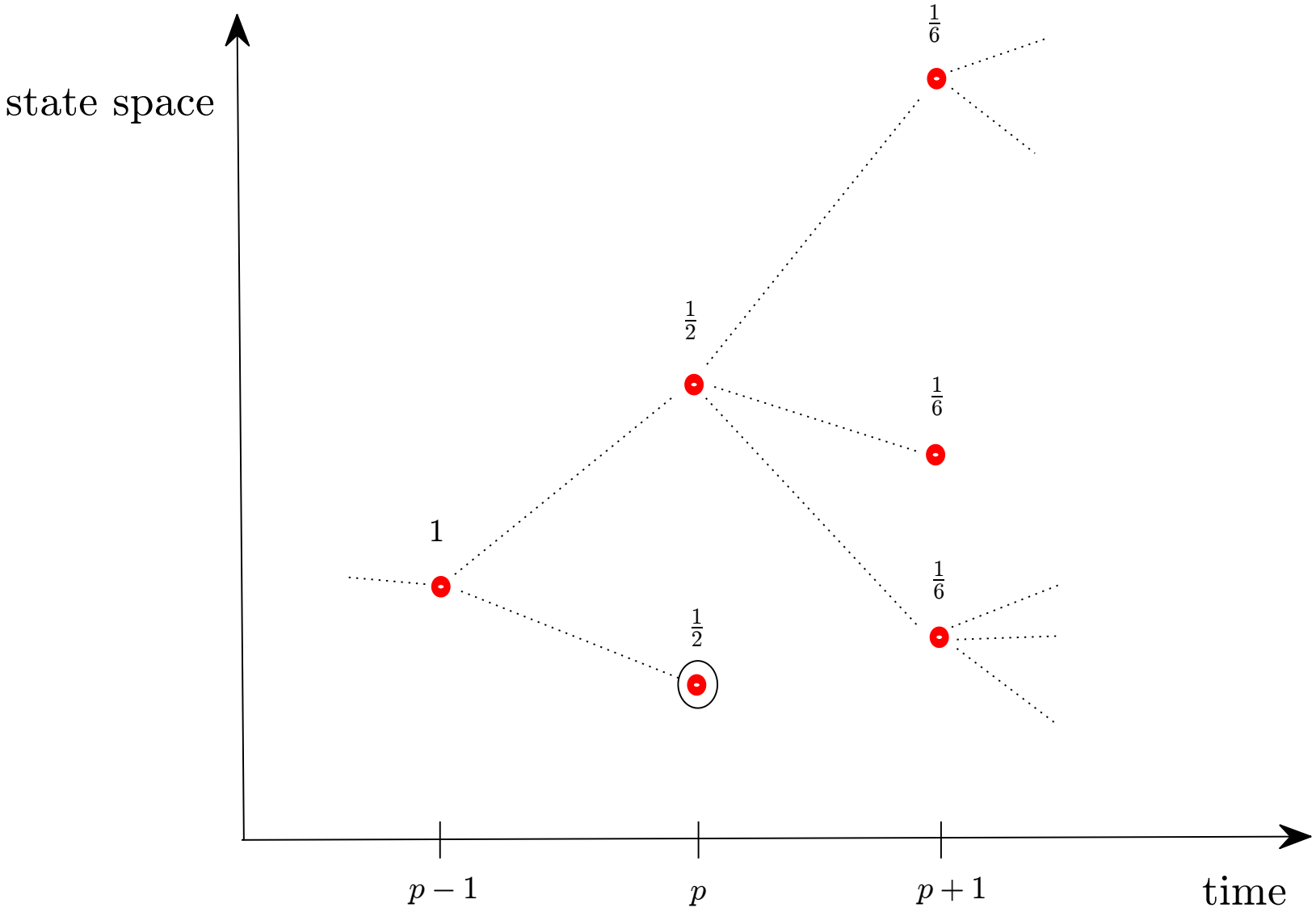


Figure: Suppose the circled replica has 1 child w.p. q, and none otherwise. What is the expected weight of its child (set to 0 if the replica is killed)?

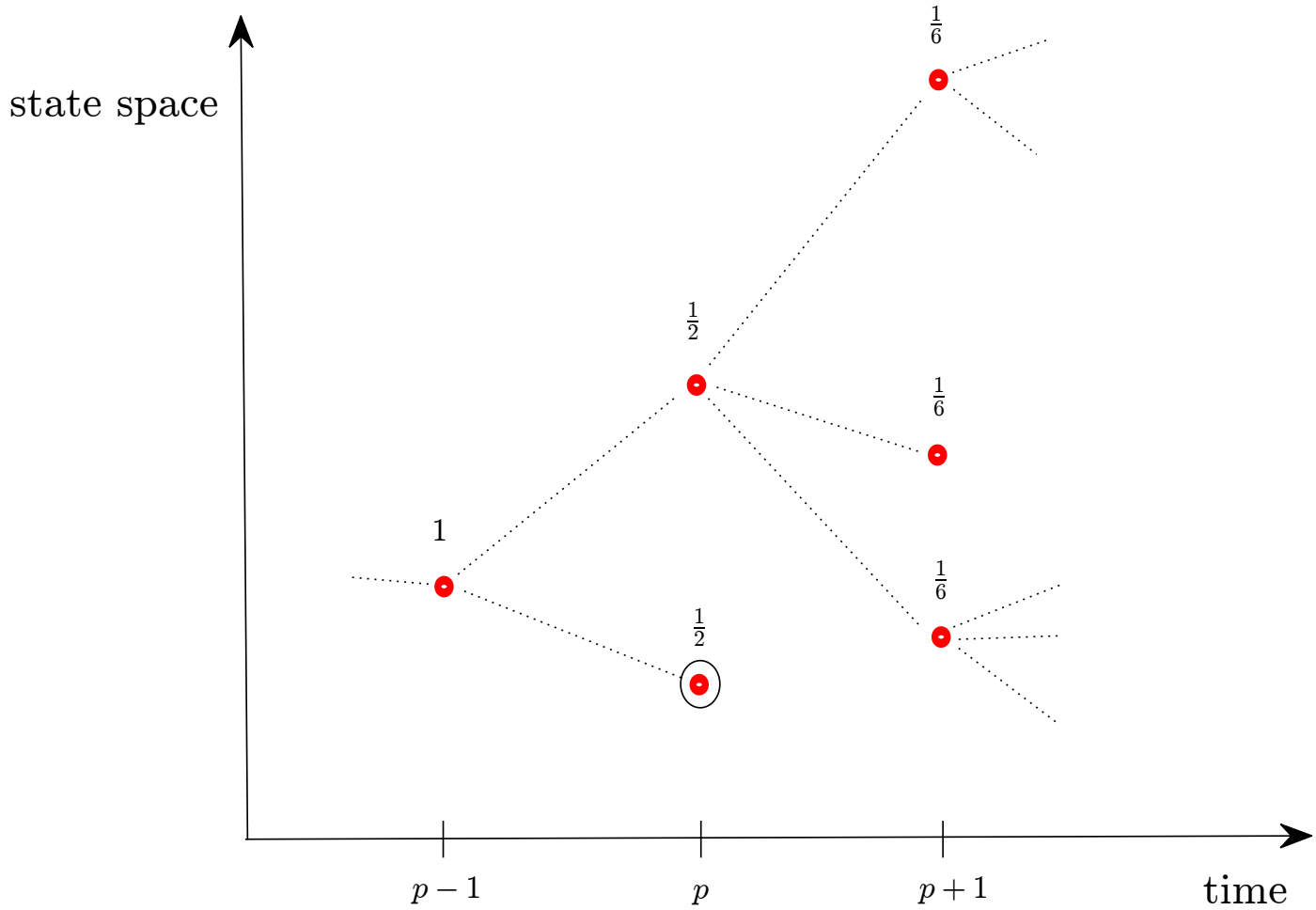


Figure: Suppose the circled replica has 1 child w.p. q, and none otherwise.

Its child's weight $= \begin{cases} \frac{1/2}{q}, & \text{parent survives} \\ 0, & \text{else} \end{cases}$, so $\mathbb{E}(\text{child's weight}) = \frac{1}{2q}q + 0(1-q) = \frac{1}{2}$.

Variance reduction

Problem.

For a function f and final time n, minimize variance in computing $\mathbb{E}(f(X_n))$.

- n can be large (stationary regime) or small (transient regime)
- Important case: f is large in regions of low probability for X_n

Variance reduction

Problem.

For a function f and final time n, minimize variance in computing $\mathbb{E}(f(X_n))$.

- Example: (X_p) is a time discretization of MD, modified so that when it reaches a product set P it immediately goes back to a reactant set R.

  If $f = 1$ on P and 0 elsewhere, and Δt is the time step, then (Hill relation)

$$\text{MFPT from } R \text{ to } P \approx \underbrace{\frac{\Delta t}{\mathbb{E}(f(X_n))}}_{\text{small denominator} \implies \text{variance reduction important!}}, \qquad n \text{ large}$$

Selection value function

Problem.

For a function f and final time n, minimize variance in computing $\mathbb{E}(f(X_n))$.

- Basic idea: we want more replicas in important regions of state space
- Fundamental question: what regions are important, and at which times?

Selection value function.

Let $v_p(x) =$ value of selecting a replica at x (a point in state space) at time $0 \leq p \leq n$.

We will use v_p to decide how to make selections at time p.

The selection value v_p is derived using Doob decomposition:

$$\mathrm{Var}(f(X_n)) = \mathrm{Var}(\text{initial condition}) + \sum_{p=0}^{n-1} \underbrace{\mathrm{Var}(\text{selection and mutation at step } p)}_{\substack{v_p \text{ is obtained by minimizing this, subject} \\ \text{to the constraint: target \#of replicas} = N}}.$$

Explicit formula:

$$v_p(x)^2 = \underbrace{\mathrm{Var}^x(g_{p+1}(X_1))}_{\substack{\text{variance associated to starting} \\ \text{a replica at point } x \text{ at time } p}} \qquad \text{where} \quad g_p(x) = \mathbb{E}^x(f(X_{n-p})).$$

Important simplification: v_p obtained by minimizing only the pth term in variance.

The selection value v_p is derived using Doob decomposition:

$$\mathrm{Var}(f(X_n)) = \mathrm{Var}(\text{initial condition}) + \sum_{p=0}^{n-1} \underbrace{\mathrm{Var}(\text{selection and mutation at step } p)}_{\substack{v_p \text{ is obtained by minimizing this, subject} \\ \text{to the constraint: target \# of replicas} = N}}.$$

Explicit formula:

$$v_p(x)^2 = \underbrace{\mathrm{Var}^x(g_{p+1}(X_1))}_{\substack{\text{variance associated to starting} \\ \text{a replica at point } x \text{ at time } p}} \qquad \text{where} \quad g_p(x) = \mathbb{E}^x(f(X_{n-p})).$$

(Mathematical sketch: Let $M_p = \mathbb{E}^{\eta_p}(f(X_{n-p}))$ where $\eta_p = \sum_{j=1}^{M} \omega^j \delta_{\xi^j}$ with $\xi^1, \ldots, \xi^M$ and $\omega^1, \ldots, \omega^M$ the points/weights at time p. Then we get v_p by minimizing $\mathbb{E}\left[(M_{p+1} - M_p)^2 \middle| \mathcal{F}_p\right]$ subject to the constraint that the expected total # of children $= N$. Here $\mathcal{F}_p =$ information from the WE process up to time p.)

Selection mechanism

Optimal strategy.

Suppose $\xi^1, \ldots, \xi^M$ are the replicas at time p, with weights $\omega^1, \ldots, \omega^M$. Then

$$\text{Target } \# \text{ of children of } \xi^j := \frac{N v_p(\xi^j)\omega^j}{\sum_{j=1}^M v_p(\xi^j)\omega^j} \tag{1}$$

where $N =$ target total $\#$ of replicas.

If the RHS of (1) is noninteger, the $\#$ of children of ξ^j is *random* with mean the target number, and minimal variance. (So if the target $\# = t$, then $\#$ of children is $\lfloor t \rfloor$ or $\lceil t \rceil$).

Understanding the optimal strategy

Let X and Y be random variables and $\alpha, \beta > 0$. Suppose we want to estimate $\mathbb{E}[\alpha X + \beta Y]$ with N total samples of X and Y and minimal variance.

Let $X_1, \ldots, X_R$ and $Y_1, \ldots, Y_S$ be the (independent) samples of X and Y, with $R + S = N$. To minimize variance, we want to choose R and S that minimize

$$\mathrm{Var}\left(\frac{\alpha}{R}\sum_{k=1}^{R} X_k + \frac{\beta}{S}\sum_{k=1}^{S} Y_k\right) = \frac{\mathrm{Var}(X)\alpha^2}{R} + \frac{\mathrm{Var}(Y)\beta^2}{S}.$$

A Lagrange multiplier calculation shows

$$R = \frac{N\,\mathrm{Std}(X)\alpha}{\mathrm{Std}(X)\alpha + \mathrm{Std}(Y)\beta}, \qquad S = \frac{N\,\mathrm{Std}(Y)\beta}{\mathrm{Std}(X)\alpha + \mathrm{Std}(Y)\beta}.$$

Comparison with standard nonlinear filtering/SMC

Toy model: $X_{p+1} = X_p + \xi_p$, where the ξ_p are iid standard Gaussians
Problem: Minimize variance in computing $\mathbb{P}(X_n > a)$ when $a = n = 20$

At time p, let $\xi^1, \ldots, \xi^M$ be the replicas, w/ weights $\omega^1, \ldots, \omega^M$ and parents $\hat{\xi}^1, \ldots, \hat{\xi}^M$.

WE:

- Target # of children of ξ^j

$$= p_j := \frac{N v_p(\xi^j)\omega^j}{\sum_{j=1}^M v_p(\xi^j)\omega^j}$$

  # children of ξ^j is Bernoulli in $\{\lfloor p_j \rfloor, \lceil p_j \rceil\}$

- v_p minimizes pth term of Doob decomposition of finite N variance

SMC:

- Target # of children of ξ^j

$$= q_j := \frac{N G(\xi^j, \hat{\xi}^j)}{\sum_{j=1}^M G(\xi^j, \hat{\xi}^j)}$$

  # of children is $Multinomial(N, q_1, \ldots, q_M)$

- $G(x, y) = e^{\alpha(x-y)}$ is ansatz function, α optimized using $N \to \infty$ variance

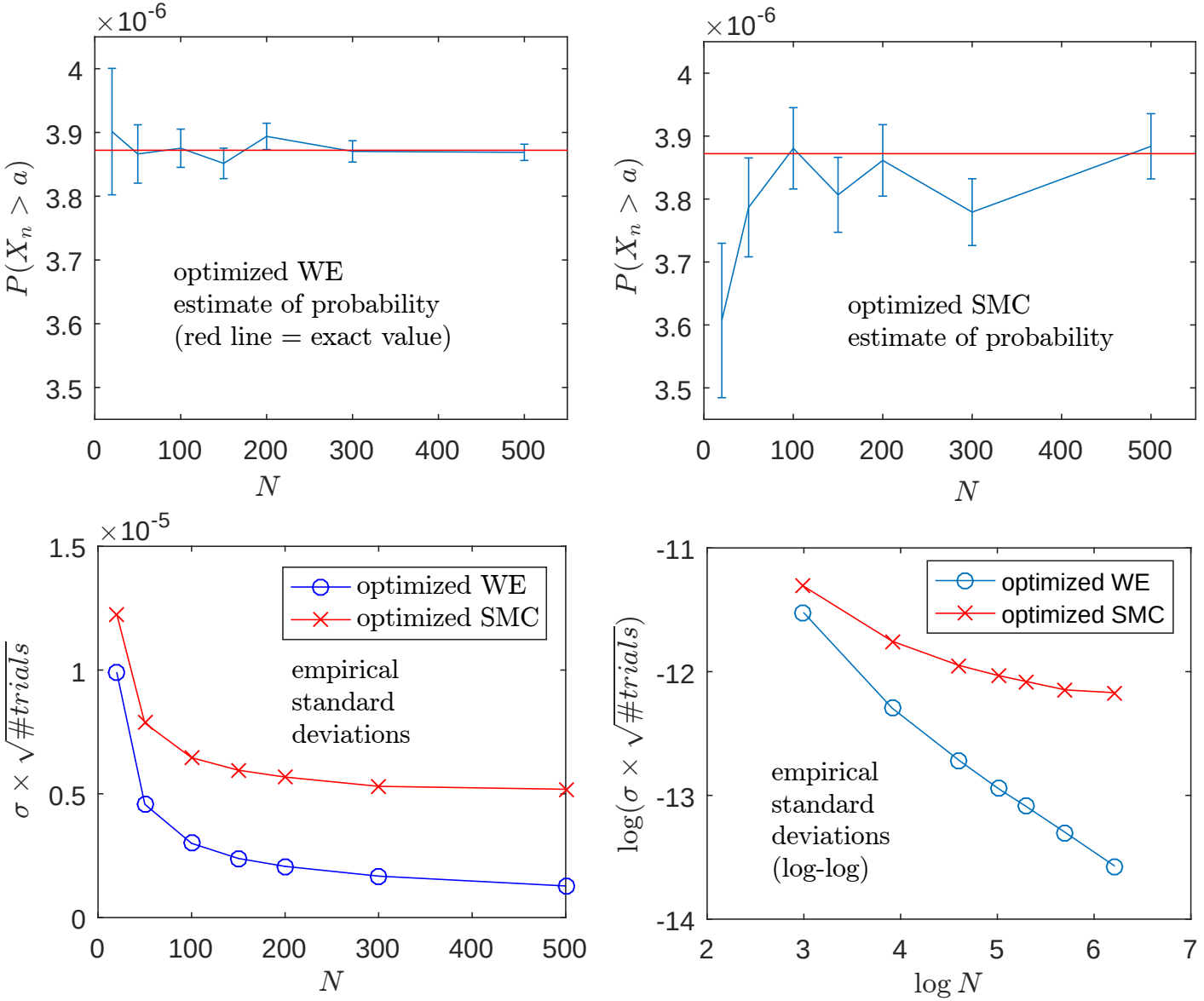


Figure: Estimate of $\mathbb{P}(X_n > a)$ using WE and SMC. (10^4 trials for each value of N.)

Understanding the selection value function

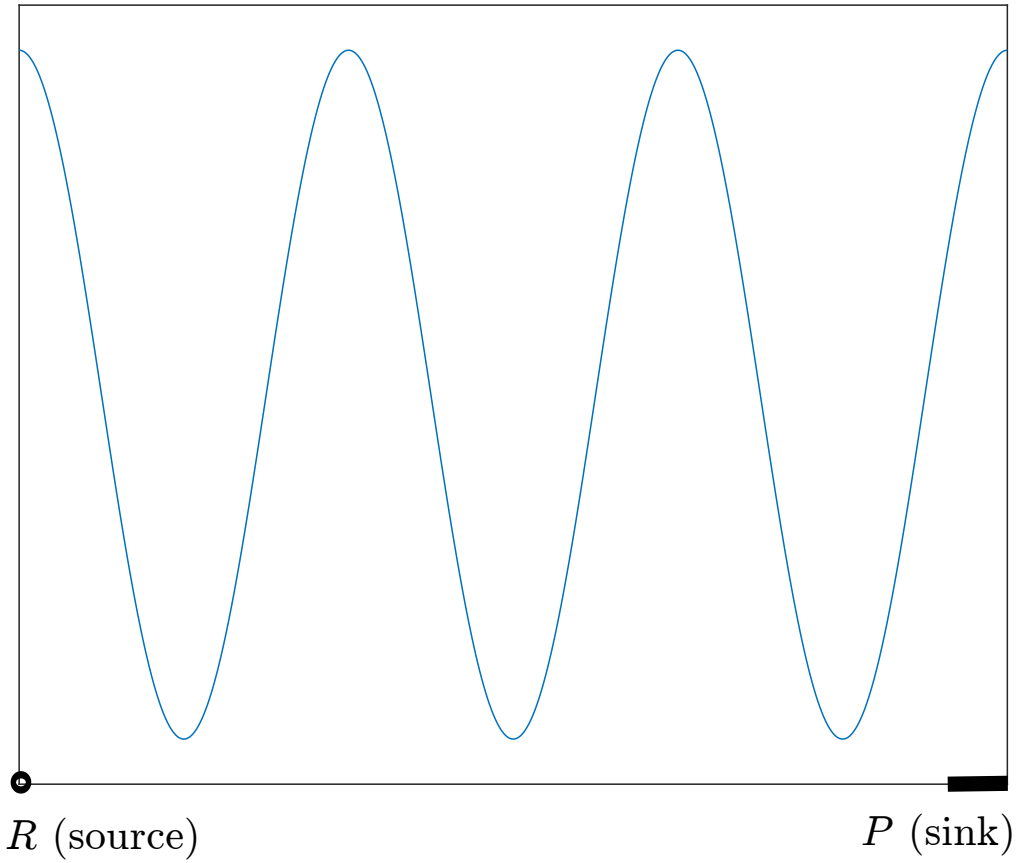
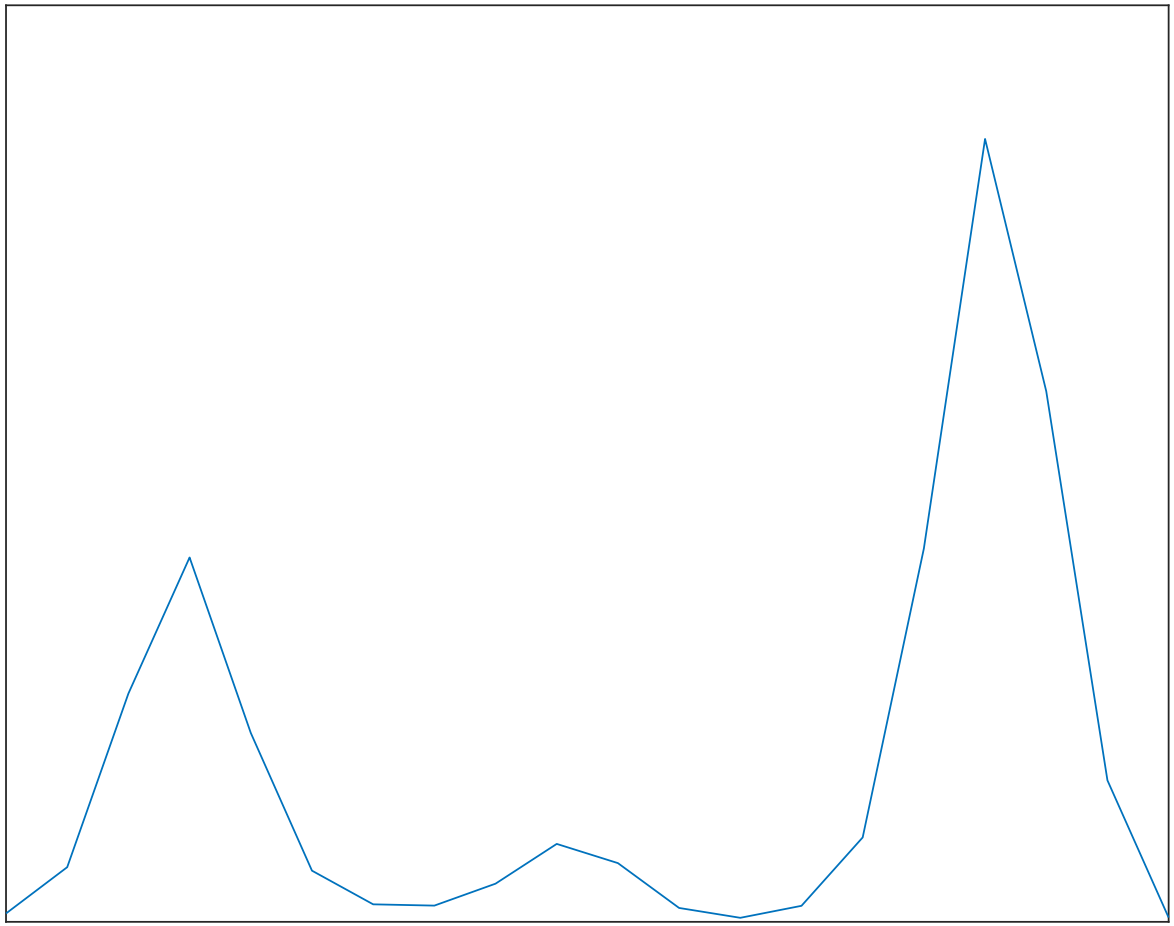
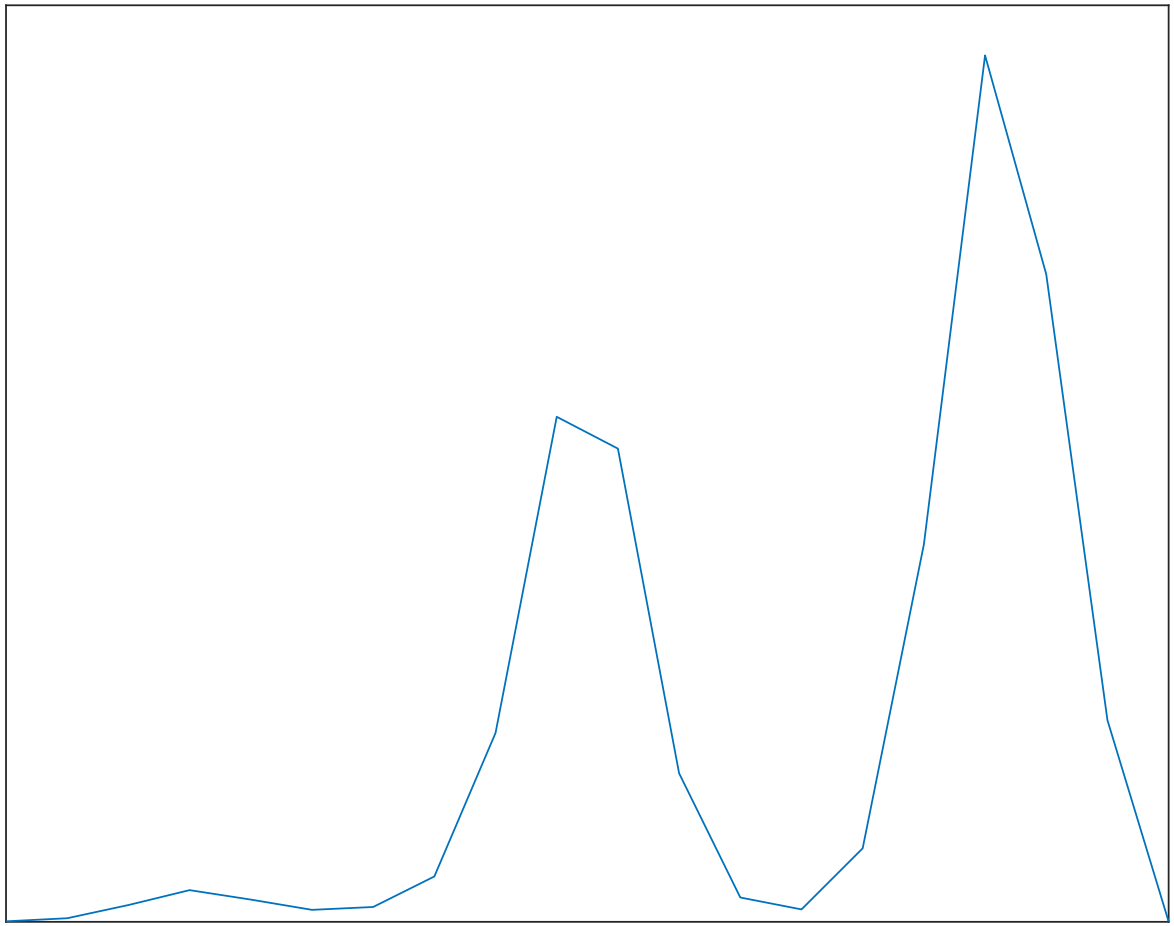


Figure: Consider (X_p) a time discretization of overdamped Langevin dynamics (with the pictured potential and added source/sink). Let $f = 1$ on P and 0 otherwise.

Evolution of $v_p(x)$, $p = 300, \ldots, 500$, for (X_p) with $n = 500$ (click for animation):



Evolution of $v_p(x)$, $p = 460, \ldots, 500$, for (X_p) with $n = 500$ (click for animation):



Features of v_p

- When $p \ll n$, v_p is nearly constant.
- When $p \approx n$, v_p is large in regions important for computing f.
- When $0 \ll p < n$, v_p has large values in regions of "high variance" (these regions depend on p and are usually regions around energy/entropy barriers).
- As $p \to n$, the selection "pushes" sampling towards relevant regions.

Estimating v_p with a coarse model

No free lunch.

In problems of interest v_p is not exactly computable. We propose estimating v_p with a MSM, and then tailoring the resampling strategy to the same MSM...

- The basic idea is to use a MSM to guide sampling in an (almost) optimal way.
- Note: sampling is always unbiased, no matter what the choice of MSM is!
- However, useful variance reduction may not be obtained with a bad MSM.

Estimating v_p with a coarse model

No free lunch.

In problems of interest v_p is not exactly computable. We propose estimating v_p with a MSM, and then tailoring the resampling strategy to the same MSM.

- A MSM isn't essential: any sufficiently good cheap/coarse model will do.
- We only assume we have a MSM for definiteness/simplicity.

Estimating v_p with a MSM

Approxiate selection value function.

Use a MSM on states $r = 1, \ldots, R$ to approximate v_p. More precisely set

$$\hat{v}_p(r) = \sqrt{\text{rth entry of } P(P^{n-p-1}u)^2 - (P^{n-p}u)^2}$$

where $P_{rs} \approx \mathbb{P}(X_1 \text{ in state } s | X_0 \text{ in state } r)$, and $u(r) \approx f|_{\text{state } r}$ (and entrywise squaring).

If the MSM is good, then we can expect $\hat{v}_p(r) \approx v_p(x)$ for x in state r.

Adapting resampling strategy to the MSM

Sampling assumption: all children in the same state have equal weight:

Sampling tailored to MSM.

Suppose that at time p, a replica ξ^j with weight ω^j is in state r. Then we define

$$\text{Target } \# \text{ of children of } \xi^j = N_p(r)\frac{\omega^j}{\omega(r)},$$

with $N_p(r) =$ target # of replicas in state r, $\omega(r) =$ total weight in state r. Thus

$$\text{weight of children of } \xi^j = \frac{\omega(r)}{N_p(r)}.$$

The last formula above is obtained using assumption (weight of child) = (weight of parent)/$\mathbb{E}$(# of children)

MSM-guided replica allocation

Replica allocation function.

If $N =$ target total # of replicas, and $0 < \tilde{N} < N/R$, set

$$N_p(r) = \underbrace{(N - \tilde{N}R)\frac{\hat{v}_p(r)\omega(r)}{\sum_{r=1}^{R}\hat{v}_p(r)\omega(r)} + \tilde{N}}_{\text{target number of replicas in state } r \text{ at time } p} .$$

The $\tilde{N}$ ensures positive survival probability. (Compare with equation (1))

This is now a *practical* strategy for allocating replicas in state space.

Application

Example.

Computing $\mathbb{E}(f(X_n))$ when n is large (stationary regime):

- Using the MSM, we may start at $p = 0$ "close" to the stationary regime
- Then pick a time n for relaxation to stationarity (hard to do in general)
- The MSM may also be used to optimize replica allocation in the relaxation.

Toy problem: discrete energy landscape mimicking a potential $V(x) = \sin\left(\frac{6\pi x}{90}\right)$.

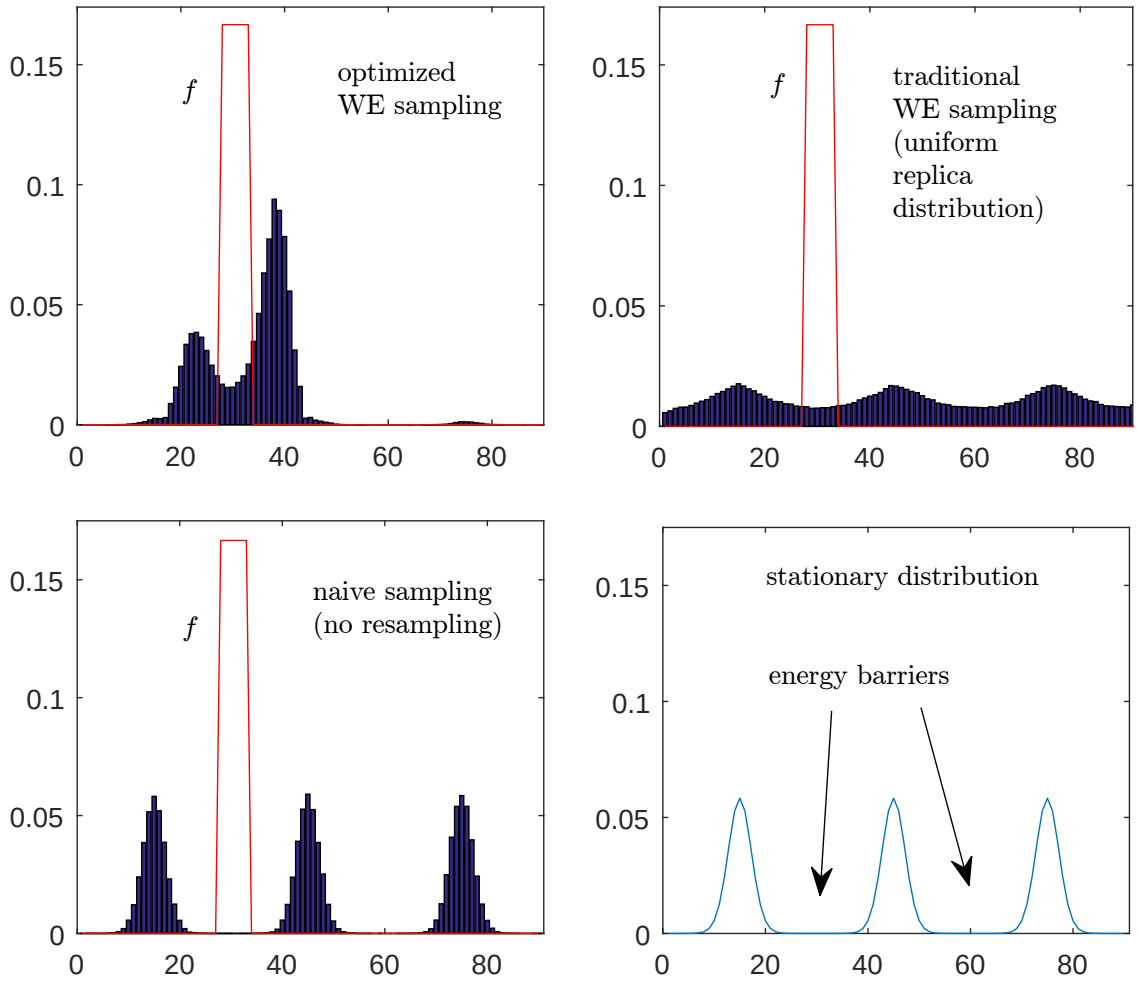


Figure: Distribution of replicas at final time $n = 30$ for different strategies.

Toy problem: discrete energy landscape mimicking a potential $V(x) = \sin\left(\frac{6\pi x}{90}\right)$.

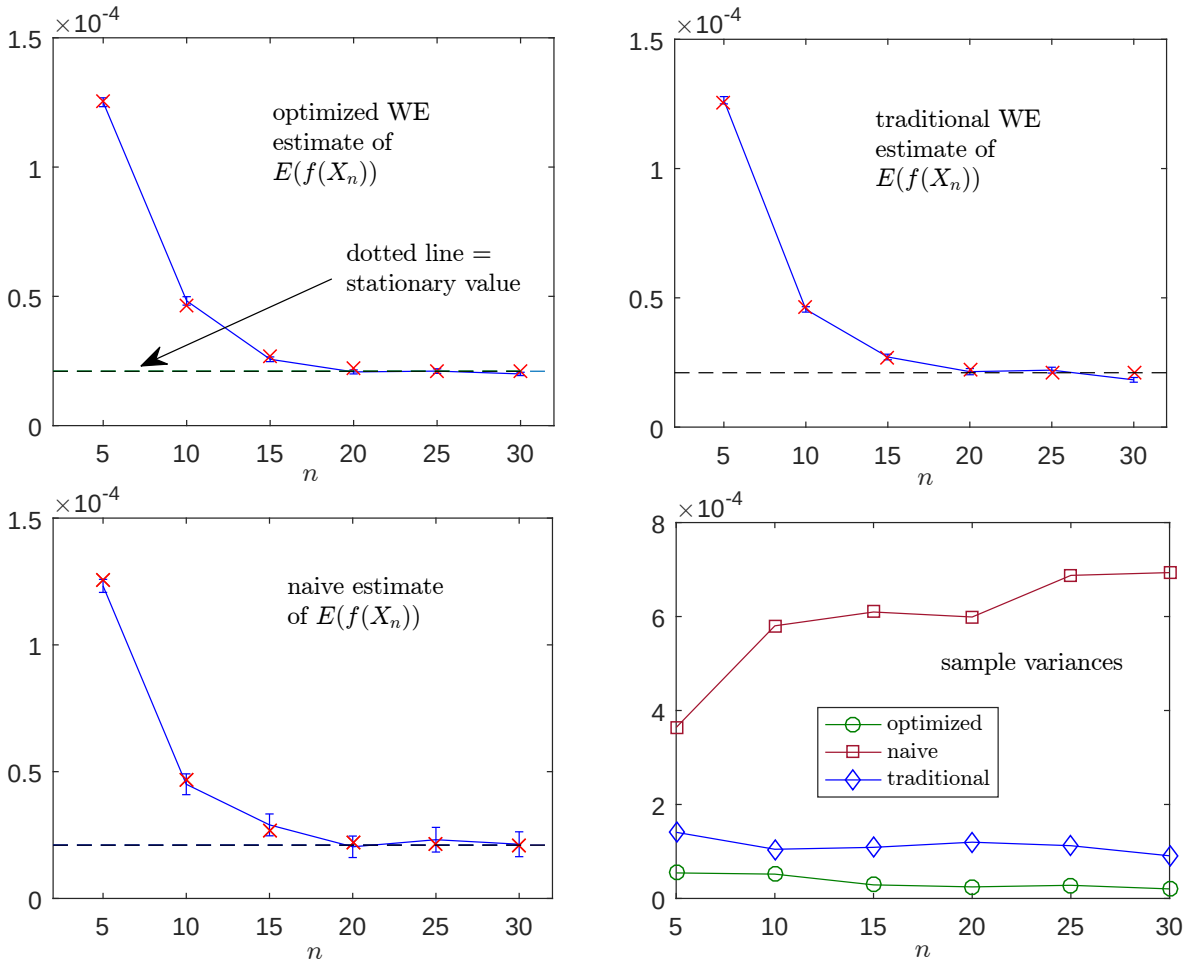


Figure: Estimation of $\mathbb{E}(f(X_n))$ for various relaxation times, and sample variance.

Toy problem: discrete energy landscape mimicking a potential $V(x) = \sin\left(\frac{6\pi x}{90}\right)$.

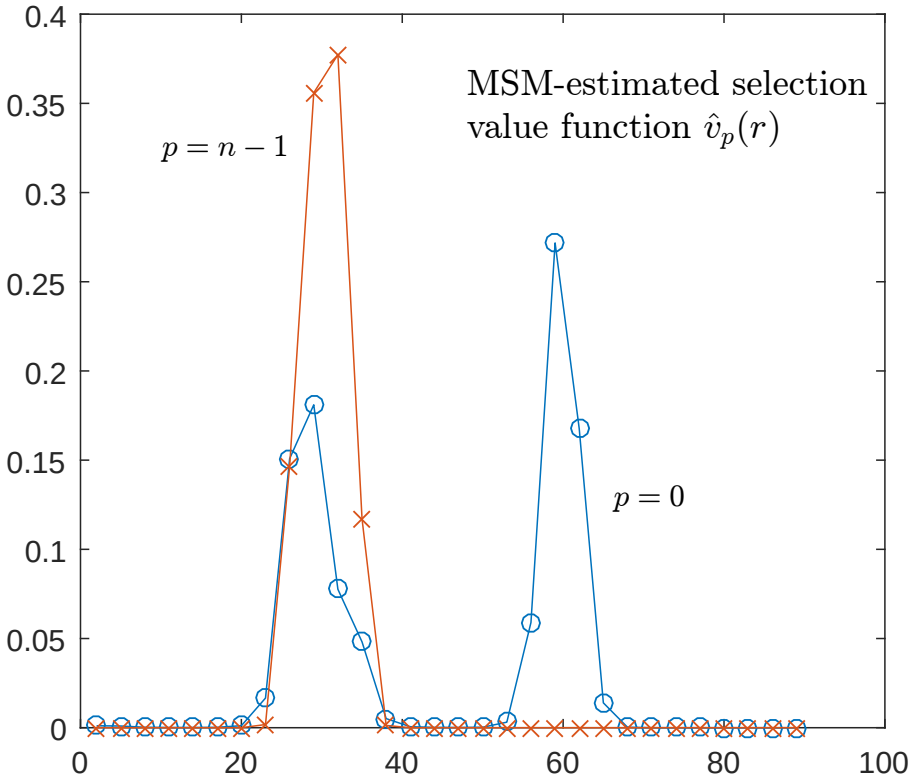


Figure: $\hat{v}_p(r)$ at the initial and penultimate times. (30 equally sized states define the MSM.)

Future work

Practical questions for future work:

- How good does the MSM have to be for a useful variance reduction?
- How much gain can we get in "real" problems (e.g. estimating MFPTs)?

Theoretical questions for future work:

- How does the sequential optimal strategy compare to a globally optimal one?
- What is the asymptotic variance for the sequentially optimal strategy?
- How does this compare to SMC/Gibbs-Boltzmann based selection?

Thanks to M. Rousset, D. Zuckerman, T. Lelièvre, P. Plecháč, , G. Simpson, and T. Wang for enlightening discussions.