Convergence and Efficiency of Adaptive Importance Sampling techniques with partial biasing

Gersende Fort

Institut de Mathématiques de Toulouse
CNRS
France

Joint work with B. Jourdain, T. Lelièvre and G. Stoltz

Talk based on the paper

G.F., B. Jourdain, T. Lelièvre, G. Stoltz *Convergence and Efficiency of Adaptive Importance Sampling techniques with partial biasing*, J. Stat. Phys (2018)

The goal

- Assumption Let $\pi \cdot \mathrm{d}\mu$ be a probability distribution on $\mathsf{X} \subseteq \mathbb{R}^p$ assumed to be highly metastable and (possibly) known up to a normalizing constant.

- Question 1: How to design a MC sampler for an approximation of

$$\int_{\mathsf{X}} f \ \pi \, \mathrm{d}\mu$$

- Question 2: How to compute the free energy

$$-\ln \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu \qquad \mathsf{X}_i \subset \mathsf{X}$$

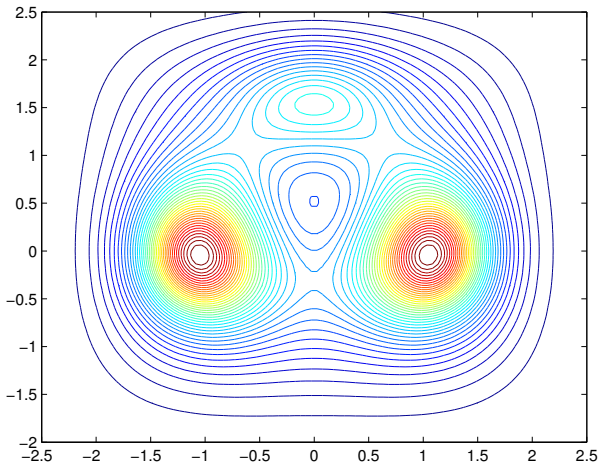
In this talk,

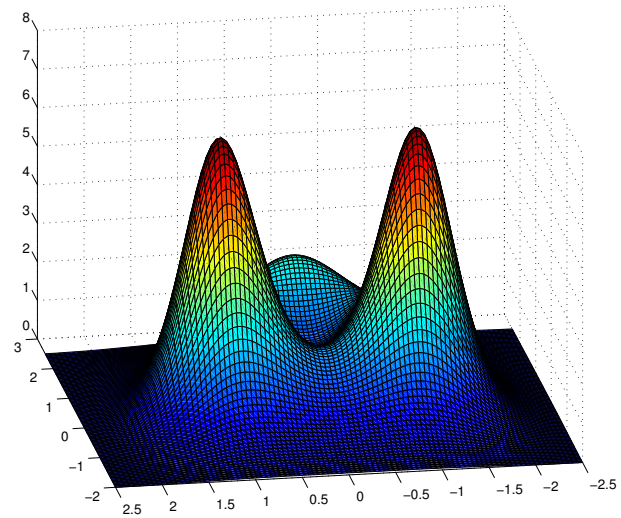
- an approach by Free Energy-based Adaptive Importance Sampling technique
- which is a generalization of Wang Landau, Self Healing Umbrella Sampling, Well tempered metadynamics.

The intuition (1/3) - a family of auxiliary distributions

$$\pi(x) = \frac{1}{Z} \exp(-V(x))$$

- The auxiliary distribution
  - **Choose** a partition $\mathsf{X}_1, \cdots, \mathsf{X}_d$ of X





$$\theta_{*,i} \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu$$

The intuition (1/3) - a family of auxiliary distributions

$$\pi(x) = \frac{1}{Z} \exp(-V(x))$$

- The auxiliary distribution
  - **Choose** a partition $\mathsf{X}_1, \cdots, \mathsf{X}_d$ of X
  - and for positive weights* $\underline{\theta} = (\theta_1, \cdots, \theta_d)$ set

$$\pi_{\underline{\theta}}(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\mathsf{X}_i}(x) \exp\left(-V(x) - \ln \theta_i\right)$$

- Property 1:

$$\forall i \in \{1, \cdots, d\}, \qquad \int_{\mathsf{X}_i} \pi_{\underline{\theta}} \, \mathrm{d}\mu \propto \frac{\theta_{*,i}}{\theta_i} \qquad \theta_{*,i} \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu$$

- Property 2:

$$\forall i \in \{1, \cdots, d\}, \qquad \int_{\mathsf{X}_i} \pi_{\underline{\theta}_*} \, \mathrm{d}\mu = \frac{1}{d}.$$

*$\theta_i \in (0,1), \sum_{i=1}^{d} \theta_i = 1$

Intuition (2/3) - How to choose θ ?

$$\pi_{\underline{\theta}}(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\mathsf{X}_i}(x) \exp\left(-V(x) - \ln \theta_i\right) \qquad \theta_{*,i} \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu$$

- If $\underline{\theta} = \underline{\theta}_*$
  - Efficient exploration under $\pi_{\underline{\theta}_*}$: each subset X_i has the same weight under $\pi_{\underline{\theta}_*}$
  - Poor ESS: The IS approximation gets into

$$\int_{\mathsf{X}} f \, \pi \mathrm{d}\mu \approx \frac{d}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{d} \mathbb{1}_{\mathsf{X}_i}(X_n) \, \theta_{*,i} \right) f(X_n)$$

- **Choose** $\rho \in (0,1)$ and set $\underline{\theta}_*^{\rho} \propto (\theta_{*,1}^{\rho}, \cdots, \theta_{*,d}^{\rho})$:

$$\int_{\mathsf{X}} f \, \pi \mathrm{d}\mu \approx \left(\sum_{i=1}^{d} \theta_{*,i}^{1-\rho} \right) \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{d} \mathbb{1}_{\mathsf{X}_i}(X_n) \, \theta_{*,i}^{\rho} \right) f(X_n)$$

- But $\underline{\theta}_*$ is unknown

Intuition (3/3) -Estimation of the free energy

$$\theta_{*,i} \overset{\text{def}}{=} \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu \approx \theta_{n,i} \overset{\text{def}}{=} \frac{C_{n,i}}{\sum_{j=1}^{d} C_{n,j}} \text{"Normalized count of the visits to } \mathsf{X}_i\text{"}$$

- Exact sampling If $X_{n+1} \sim \pi \, \mathrm{d}\mu$: $\quad C_{n+1,i} = C_{n,i} + \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$

This yields for all $i = 1, \cdots, d$

$$C_{n+1,i} = \sum_{k=1}^{n+1} \mathbb{1}_{\mathsf{X}_i}(X_k) \qquad S_{n+1} \overset{\text{def}}{=} \sum_{i=1}^{d} C_{n+1,i} = (n+1) = O(n)$$

and

$$\theta_{n+1,i} = \frac{1}{n+1} \sum_{k=1}^{n+1} \mathbb{1}_{\mathsf{X}_i}(X_{n+1}) = \theta_{n,i} + \frac{1}{n+1} \left(\mathbb{1}_{\mathsf{X}_i}(X_{n+1}) - \theta_{n,i} \right)$$

i.e. Stochastic Apprimation scheme with learning rate $1/S_{n+1}$, and limiting point $\theta_{*,i}$

Intuition (3/3) -Estimation of the free energy

$$\theta_{*,i} \overset{\text{def}}{=} \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu \approx \theta_{n,i} \overset{\text{def}}{=} \frac{C_{n,i}}{\sum_{j=1}^{d} C_{n,j}} \text{ "Normalized count of the visits to } \mathsf{X}_i\text{"}$$

- Exact sampling If $X_{n+1} \sim \pi \, \mathrm{d}\mu$: $\quad C_{n+1,i} = C_{n,i} + \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$
- IS sampling If $X_{n+1} \sim \pi_{\widehat{\underline{\theta}}} \, \mathrm{d}\mu$: $\quad C_{n+1,i} = C_{n,i} + \gamma \, \widehat{\theta}_i \, \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$

This yields for all $i = 1, \cdots, d$

$$C_{n+1,i} = \gamma \, \widehat{\theta}_i \sum_{k=1}^{n+1} \mathbb{1}_{\mathsf{X}_i}(X_{n+1}) \qquad S_{n+1} \overset{\text{def}}{=} \sum_{i=1}^{d} C_{n+1,i} = O_{wp1}(n)$$

and

$$\theta_{n+1,i} = \theta_{n,i} + \frac{\gamma}{S_{n+1}} H_i(\underline{\theta}_n, X_{n+1}) + O(\frac{1}{n^2})$$

i.e. Stochastic Apprimation scheme with learning rate $1/S_{n+1}$, and limiting point $\theta_{*,i}$

Intuition (3/3) -Estimation of the free energy

$$\theta_{*,i} \overset{\text{def}}{=} \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu \approx \theta_{n,i} \overset{\text{def}}{=} \frac{C_{n,i}}{\sum_{j=1}^{d} C_{n,j}} \text{ "Normalized count of the visits to } \mathsf{X}_i\text{"}$$

- Exact sampling If $X_{n+1} \sim \pi \, \mathrm{d}\mu$: $\quad C_{n+1,i} = C_{n,i} + \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$
- IS sampling If $X_{n+1} \sim \pi_{\underline{\theta}} \, \mathrm{d}\mu$: $\quad C_{n+1,i} = C_{n,i} + \gamma \, \widehat{\theta}_i \, \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$
- IS sampling with a leverage effect If $X_{n+1} \sim \pi_{\underline{\theta}} \, \mathrm{d}\mu$:

$$C_{n+1,i} = C_{n,i} + \gamma \, \frac{\mathbf{S_n}}{\mathbf{g(S_n)}} \widehat{\theta}_i \, \mathbb{1}_{\mathsf{X}_i}(X_{n+1}) \qquad \lim_{+\infty} g = +\infty, \ \liminf_{s} s/g(s) > 0$$

This yields

$$S_{n+1} \uparrow +\infty \qquad \frac{S_{n+1} - S_n}{S_n} = \frac{\gamma}{g(S_n)} \widehat{\theta}_i \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$$

and

$$\theta_{n+1,i} = \theta_{n,i} + \frac{\gamma}{g(S_n)} H_i(\underline{\theta}_n, X_{n+1}) + O\left(\frac{\gamma^2}{g^2(S_n)}\right)$$

i.e. S.A. scheme with learning rate $\gamma/g(S_n)$, and limiting point $\theta_{*,i}$.

Intuition (3/3) -Estimation of the free energy

$$\theta_{*,i} \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi \, \mathrm{d}\mu \approx \theta_{n,i} \stackrel{\text{def}}{=} \frac{C_{n,i}}{\sum_{j=1}^{d} C_{n,j}} \text{ "Normalized count of the visits to } \mathsf{X}_i\text{"}$$

- Exact sampling If $X_{n+1} \sim \pi \, \mathrm{d}\mu$: $\quad C_{n+1,i} = C_{n,i} + \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$
- IS sampling If $X_{n+1} \sim \pi_{\underline{\widehat{\theta}}} \, \mathrm{d}\mu$: $\quad C_{n+1,i} = C_{n,i} + \gamma \, \widehat{\theta}_i \, \mathbb{1}_{\mathsf{X}_i}(X_{n+1})$
- IS sampling with a leverage effect If $X_{n+1} \sim \pi_{\underline{\widehat{\theta}}} \, \mathrm{d}\mu$:

$$C_{n+1,i} = C_{n,i} + \gamma \, \frac{\mathbf{S_n}}{\mathbf{g(S_n)}} \widehat{\theta}_i \, \mathbb{1}_{\mathsf{X}_i}(X_{n+1}) \qquad \lim_{+\infty} g = +\infty, \ \liminf_{s} s/g(s) > 0$$

If $g(s) = \ln(1+s)^{\alpha/(1+\alpha)}$, the learning rate is $O(t^{-\alpha})$

- Key property: if $X_{n+1} \in \mathsf{X}_i$, then for any $j \neq i$

$$\pi_{\underline{\theta}_{n+1}}(\mathsf{X}_j) > \pi_{\underline{\theta}_n}(\mathsf{X}_j) \qquad \text{the probability of stratum } \#j \text{ increases}$$

The algorithm: Adaptive IS with partial biasing

- Fix: $\rho \in (0,1)$ and $\alpha \in (1/2, 1)$. Set $g(s) \stackrel{\text{def}}{=} (\ln(1+s))^{\alpha/(1-\alpha)}$.
- Initialisation: $X_0 \in \mathsf{X}$, a positive weight vector $\underline{\theta}_0$,
- Repeat, for $n = 0, \cdots, N-1$
  - sample $X_{n+1} \sim P_{\underline{\theta}_n^\rho}(X_n, \cdot)$, a Markov kernel invariant wrt $\pi_{\underline{\theta}_n^\rho} \, \mathrm{d}\mu$
  - compute

$$C_{n+1,i} = C_{n,i} + \frac{\gamma}{g(S_n)} \; S_n \; \theta_{n,i}^\rho \; \mathbb{I}_{\mathsf{X}_i}(X_{n+1})$$

$$S_{n+1} = \sum_{i=1}^{d} C_{n+1,i} \qquad \theta_{n+1,i} = \frac{C_{n+1,i}}{S_{n+1}}$$

- Return $(\underline{\theta}_n)_n$ sequ. of estimates of $\underline{\theta}_*$; and the IS estimator

$$\int f \, \pi \mathrm{d}\mu \approx \; \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{d} \theta_{n-1,i}^{1-\rho} \right) \; \left(\sum_{i=1}^{d} \mathbb{I}_{\mathsf{X}_i}(X_n) \, \theta_{n-1,i}^{\rho} \right) \; f(X_n)$$

Convergence results

1. The limiting behavior of the estimates $(\underline{\theta}_n)_n$
2. The limiting distribution of X_n
3. The limiting behavior of the IS estimator

1. On the target density and the strata X_i:

$$\sup_{\mathsf{X}} \pi < \infty, \qquad \min_{1 \leq i \leq d} \theta_*(i) > 0$$

2. On the kernels $P_{\underline{\theta}}$: Hastings-Metropolis kernel, with symmetric proposal $q(x,y)\mathrm{d}\mu(y)$ such that $\inf_{\mathsf{X}^2} q > 0$.

   for any compact subset K, there exists C and $\lambda \in (0,1)$ s.t.

$$\sup_{\underline{\theta} \in K} \|P_{\underline{\theta}}^n(x,\cdot) - \pi_{\underline{\theta}}\|_{\mathrm{TV}} \leq C\lambda^n$$

3. $\rho \in (0,1)$

4. $g(s) = (\ln(1+s))^{\alpha/(1-\alpha)}$ with $\alpha \in (1/2, 1)$.

Convergence results: on the sequence $\underline{\theta}_n$

- Recall

$$\underline{\theta}_{n+1} = \underline{\theta}_n + \gamma_{n+1}\ H(X_{n+1}, \underline{\theta}_n) + \gamma_{n+1}^2\ \Lambda_{n+1} \qquad \gamma_{n+1} \stackrel{\text{def}}{=} \gamma / g(S_n)$$

where

γ_n is a positive **random** learning rate
$\sup_n \|\Lambda_{n+1}\|$ is bounded a.s.
$\int H(\cdot, \theta)\, \pi_{\underline{\theta}^\rho}\, \mathrm{d}\mu = 0$ iff $\underline{\theta} = \underline{\theta}_*$.

- Result 1

$$\lim_n \gamma_n\ n^\alpha = (1-\alpha)^\alpha\ \gamma^{1-\alpha}\ \left(\sum_{j=1}^d \theta_{*,j}^{1-\rho} \right) \qquad \text{a.s.}$$

- Result 2:

$$\lim_n \underline{\theta}_n = \underline{\theta}_* \qquad \text{a.s.}$$

Convergence results - on the samples X_n

- Recall

$$X_{n+1} \sim P_{\underline{\theta}_n^\rho}(X_n, \cdot) \qquad \pi_{\underline{\theta}} P_{\underline{\theta}} = \pi_{\underline{\theta}}$$

- Result 1 For any bounded function f

$$\lim_n \mathbb{E}\left[f(X_n)\right] = \int f \ \pi_{\underline{\theta}_*^\rho} \, \mathrm{d}\mu$$

- Result 2 For any bounded function f

$$\lim_N \frac{1}{N} \sum_{n=1}^{N} f(X_n) = \int f \ \pi_{\underline{\theta}_*^\rho} \, \mathrm{d}\mu \qquad a.s.$$

- Result 1 For any bounded function f

$$\lim_N \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(X_n)\left(\sum_{j=1}^{d}\theta_{n-1,j}^{\rho}\mathbb{1}_{\mathsf{X}_j}(X_n)\right)\left(\sum_{j=1}^{d}\theta_{n-1,j}^{1-\rho}\right)\right] = \int f\,\pi\,\mathrm{d}\mu$$

- Result 1 For any bounded function f, a.s.:

$$\lim_N \frac{1}{N}\sum_{n=1}^{N} f(X_n)\left(\sum_{j=1}^{d}\theta_{n-1,j}^{\rho}\mathbb{1}_{\mathsf{X}_j}(X_n)\right)\left(\sum_{j=1}^{d}\theta_{n-1,j}^{1-\rho}\right) = \int f\,\pi\,\mathrm{d}\mu$$

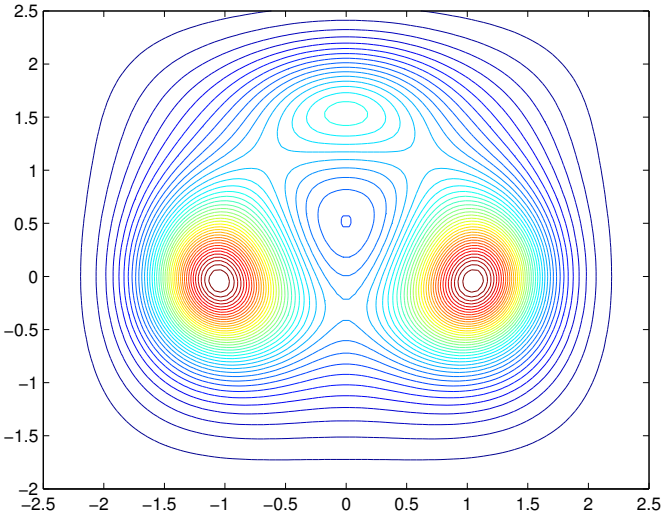
Is it new ?

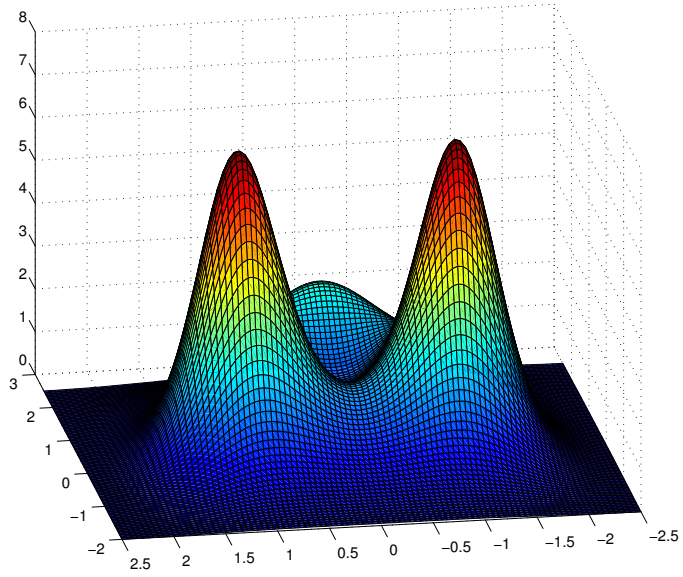
- Theoretical contribution
  - Self Healing Umbrella Sampling
    - $\rho = 1$ (no biasing intensity)
    - $g(s) = s$ (also covered by the theory; not detailed here)
  - Well-tempered metadynamics
    - $\rho \in (0, 1)$ (biasing intensity)
    - $g(s) = s^{1-\rho}$ (also covered by the theory; not detailed here)

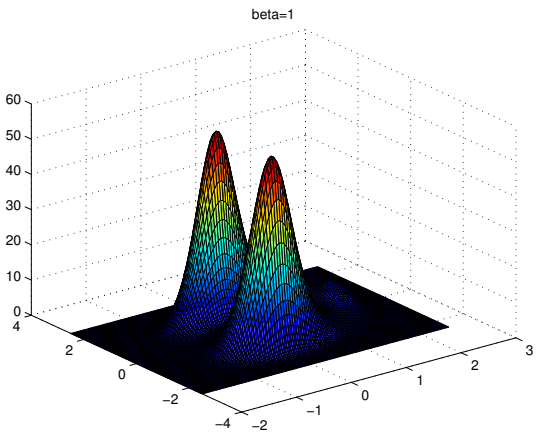
- Methodological contribution: the introduction of a function $g(s)$ in the updating scheme of the estimator θ_n, allowing a random learning rate

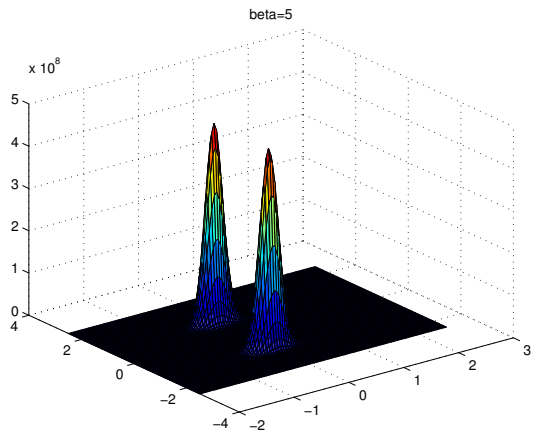
$$\gamma_n \sim O_{wp1}(n^{-\alpha})$$

for $\alpha \in (1/2, 1)$.









Make the metastability larger by increasing β.

Case $\rho \in [0, 1)$ ($\rho = a$ on the plot) and $\alpha \in (1/2, 1)$ $\Rightarrow \gamma_n = O_{wp1}(n^{-\alpha})$

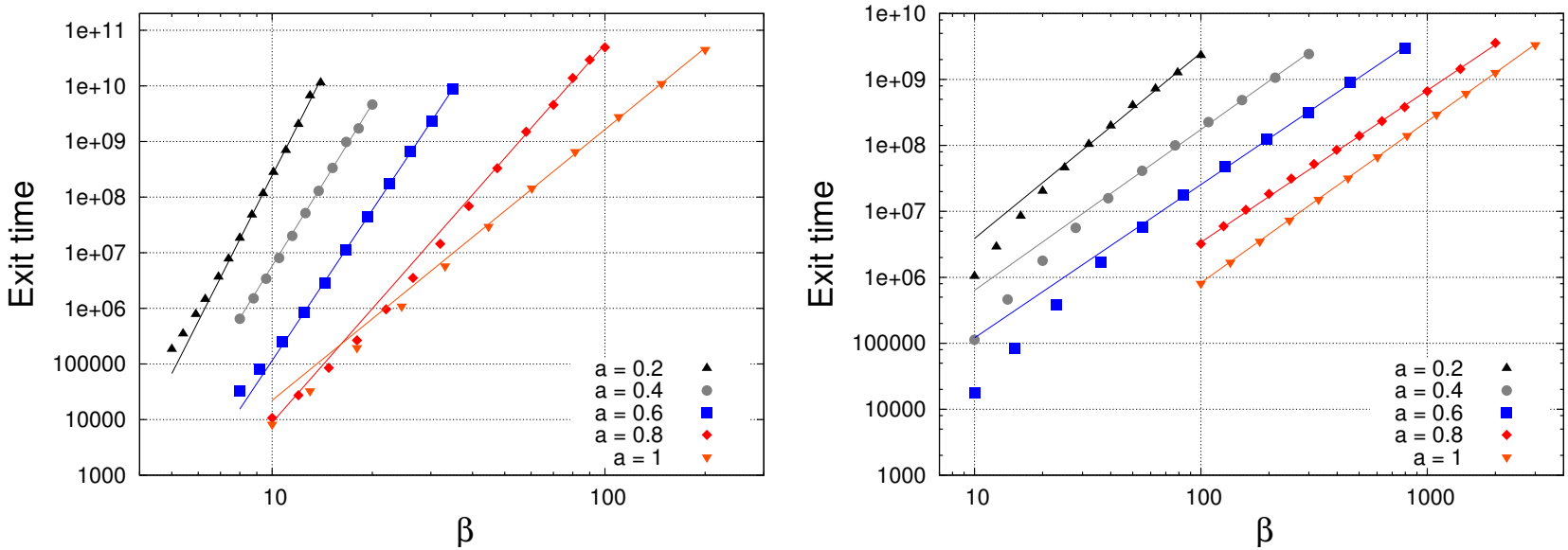


Figure: Left: Exit times for $\alpha = 0.8$. Right: Exit times for $\alpha = 0.6$.

Start from the left mode, compute the exit time T i.e. time to reach $X_{n,1} > 1$

- $T \uparrow$ when $\beta \uparrow$
- fixed β and ρ: $T \downarrow$ when $\alpha \downarrow$ - a slowly $\downarrow$ learning rate is better
- fixed β and α: $T \downarrow$ when $\rho \uparrow$ - a small (or no) bias is better
- Linear fit with a slope indep of ρ: $\ln T = c_\rho + (1 - \alpha)^{-1} \ln \beta$

Case $\rho \in (0,1)$ ($\rho = a$ on the plot) and $\alpha = 1$ $\Rightarrow \gamma_n = O_{wp1}(n^{-1})$ (case well tempered metadynamics)

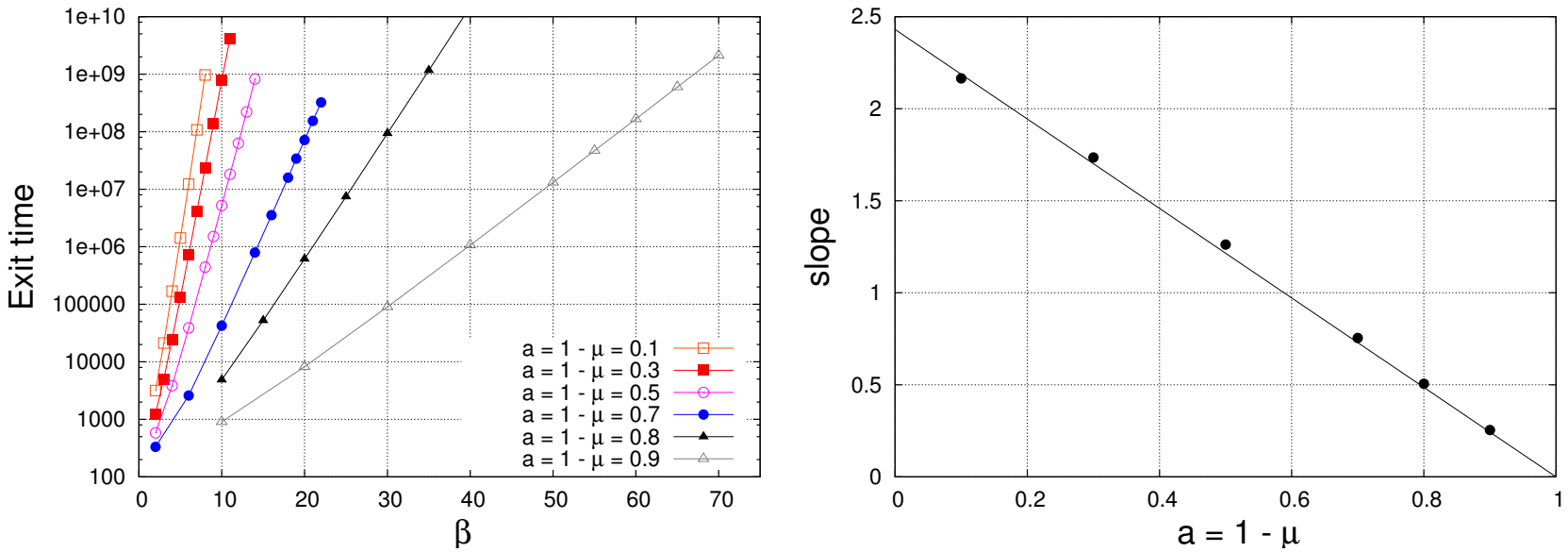


Figure: Left: Exit times for many values of ρ. Right: Associated slopes, fitted by $x \mapsto 2.43(1-x)$.

Exit time T

- For fixed β: $T \downarrow$ when $\rho \uparrow$ - a small bias is better
- Linear fit: $\ln T = c + 2.43(1-\rho)\beta$

Normalized Effective Sample Size (EF)

Case $\gamma_n = O(1/n^{\alpha})$ for $\alpha \in (1/2, 1)$, $\rho \in [0, 1)$

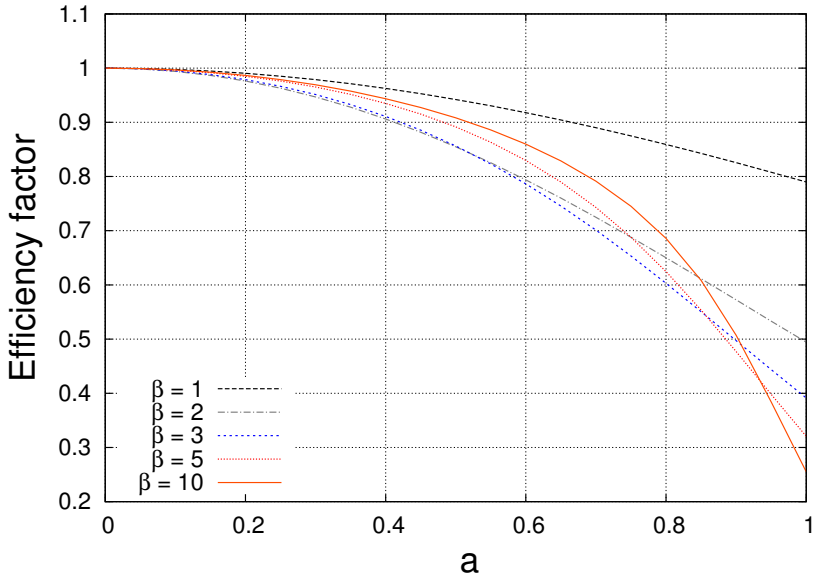


Figure: Efficiency factors $\rho \mapsto \mathrm{EF}(\rho)$ for various values of β.

$$EF = \frac{\left(N^{-1}\sum_{n=1}^{N} w(X_n)\right)^2}{\left(N^{-1}\sum_{n=1}^{N} w^2(X_n)\right)} \in [0, 1]$$

- By definition, when constant weights, $EF = 1$.
- For fixed β, $EF \uparrow$ when $\rho \downarrow$ - a strong bias is better

Conclusion

A new algorithm

- which estimates the free energy of π by a Stochastic Approximation algorithm, where the stepsize sequence $\{\gamma_n, n \geq 0\}$ is tuned on the fly
- which provides an approximation of π by a set of weighted points with a controlled discrepancy of the weights.

- which requires two design parameters (α, ρ) to be fixed by the user
  - Transient phase: ρ close to 1 and α close to $1/2$.
  - At convergence: ρ close to 0 and α close to 1.

- In the transient phase: far more efficient than Well-Tempered Metadynamics, SHUS and WL.