Convergence and Efficiency of Adaptive Importance Sampling techniques with partial biasing

Gersende Fort

Institut de Mathématiques de Toulouse
CNRS and Univ. de Toulouse
France

Joint work with

- Benjamin Jourdain (ENPC, France)
- Tony Lelièvre (ENPC, France)
- Gabriel Stoltz (ENPC, France)

Talk based on the paper

G. Fort, B. Jourdain, T. Lelièvre, G. Stoltz *Convergence and Efficiency of Adaptive Importance Sampling techniques with partial biasing*, arXiv:1610.0919

Motivation (1/4)

Goal:

Explore the support of a distribution $\pi \, \mathrm{d}\lambda$ on $\mathsf{X} \subseteq \mathbb{R}^p$

and/or compute integrals w.r.t. π

$$\int_{\mathsf{X}} f(x) \; \pi(x) \mathrm{d}\lambda(x)$$

when π is highly metastable, p is large.

Solution: based on Importance Sampling (IS)

Sample $X_1, \cdots, X_n, \cdots \overset{i.i.d.}{\sim} \widetilde{\pi} \, \mathrm{d}\lambda$

Define the IS approximation

$$\int_{\mathsf{X}} f \, \pi \mathrm{d}\lambda \approx \frac{1}{n} \sum_{k=1}^{n} \underbrace{\frac{\pi(X_k)}{\widetilde{\pi}(X_k)}}_{\text{importance ratio}} f(X_k).$$

- Define a partition of the support X (Molecular dynamics: Chipot, Pohorille (2007) and Lelievre, Rousset, Stoltz (2010); Statistics: Chopin, Lelievre, Stoltz (2012))

$$\mathsf{X} = \bigcup_{i=1}^{d} \mathsf{X}_i \qquad d \text{ strata}$$

- A family of auxiliary distribution based on a local biasing
  For all *positive vector* $\tau = (\tau(1), \cdots, \tau(d))$ $\quad \tau(i) > 0, \forall i$

$$\pi_\tau(x) \stackrel{\text{def}}{=} \frac{1}{\sum_{i=1}^{d} \frac{\theta_\star(i)}{\tau(i)}} \sum_{i=1}^{d} \frac{\pi(x)}{\tau(i)} \mathbb{1}_{\mathsf{X}_i}(x),$$

where

$$\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi \mathrm{d}\lambda, \qquad \text{up to a constant, } \log \theta_\star(i) \text{ is the free-energy}$$

Motivation (2/4) - How to choose $\widetilde{\pi}$?

- Define a partition of the support X (Molecular dynamics: Chipot, Pohorille (2007) and Lelievre, Rousset, Stoltz (2010); Statistics: Chopin, Lelievre, Stoltz (2012))

$$\mathsf{X} = \bigcup_{i=1}^{d} \mathsf{X}_i \qquad d \text{ strata}$$

- A family of auxiliary distribution based on a local biasing
For all *positive vector* $\tau = (\tau(1), \cdots, \tau(d))$ $\quad \tau(i) > 0, \forall i$

$$\pi_\tau(x) \stackrel{\text{def}}{=} \frac{1}{\sum_{i=1}^d \frac{\theta_\star(i)}{\tau(i)}} \sum_{i=1}^{d} \frac{\pi(x)}{\tau(i)} \mathbb{1}_{\mathsf{X}_i}(x),$$

where

$$\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi \mathrm{d}\lambda, \qquad \text{up to a constant, } \log \theta_\star(i) \text{ is the free-energy}$$

Key property: $\pi_{\theta_\star}(\mathsf{X}_i) = 1/d$ – all the strata have the same weight: efficient to tackle multimodality ! but $\theta_\star$ is unknown.

Motivation - Adaptive Importance Sampling (3/4)

An *iterative* algorithm which

- Will learn on the fly the weight vector $\theta_\star$ though a Stochastic Approximation algorithm
$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1})$$
where H is chosen so that $\theta_\star$ is the unique solution of
$$\int H(\theta, x)\ \pi_\theta(x)\, \mathsf{d}\lambda(x) = 0.$$
- from draws X_{n+1}
$$X_{n+1} \sim P_{\theta_n}(X_n, \cdot) \qquad \text{kernel with inv. dist. } \pi_{\theta_n}$$

Motivation - Adaptive Importance Sampling (3/4)

An *iterative* algorithm which

- Will learn on the fly the weight vector $\theta_\star$ though a Stochastic Approximation algorithm

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1})$$

where H is chosen so that $\theta_\star$ is the unique solution of

$$\int H(\theta, x)\; \pi_\theta(x)\, \mathsf{d}\lambda(x) = 0.$$

- from draws X_{n+1}

$$X_{n+1} \sim P_{\theta_n}(X_n, \cdot) \qquad \text{kernel with inv. dist. } \pi_{\theta_n}$$

If convergence is established

- An estimator of the free energy: $\lim_n \theta_n = \theta_\star$.
- An approximatiton of the target distribution π - computed on the fly/online

$$\int f\, \pi \mathsf{d}\lambda = \lim_n \frac{d}{n} \sum_{k=1}^{n} f(X_k) \left(\sum_{i=1}^{d} \theta_k(i) \mathbb{1}_{\mathsf{X}_i}(X_k) \right)$$

Motivation - Choice of the field $H(\theta, x)$ (4/4)

A family of algorithms: Wang Landau, Self Healing Umbrella Sampling (SHUS), Well-Tempered Metadynamics, SHUS_ρ^g

on the form

1. Given a new draw $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ with inv. dist. π_{θ_n}
2. Update a counter of the visits to a stratum

$$C_{n+1}(i) = C_n(i) + (\cdots)^2 \; \mathbb{1}_{\mathsf{X}_i}(X_{n+1}) \qquad i = 1, \cdots, d$$

3. Normalize the counter to obtain a weight vector

$$\theta_{n+1}(i) = \frac{C_{n+1}(i)}{\sum_{j=1}^d C_{n+1}(j)} = \theta_n(i) + \gamma_{n+1} \cdots + O(\gamma_{n+1}^2) \qquad i = 1, \cdots, d$$

Fundamental: if $X_{n+1} \in \mathsf{X}_i$

$$C_{n+1}(i) > C_n(i), \qquad C_{n+1}(j) = C_n(j), j \neq i$$
$$\Longrightarrow \pi_{\theta_{n+1}}(\mathsf{X}_i) < \pi_{\theta_n}(\mathsf{X}_i), \quad \pi_{\theta_{n+1}}(\mathsf{X}_j) = \pi_{\theta_n}(\mathsf{X}_j).$$

A Wang-Landau (WL) based algorithm

a WL based algorithm - algorithm (1/3)

(adapted from) the Wang-Landau algorithm (Wang and Landau, 2001)

Input:

- *initial values: a point* $X_0 \in \mathsf{X}$ *and a counter* $C_0 \in (\mathbb{R}_+^\star)^d$
- *a positive (deterministic) stepsize sequence* $\{\gamma_n, n \geq 0\}$

For $n = 0, 1, \cdots$

- *Normalize the counter*

$$\theta_n(i) = \frac{C_n(i)}{\sum_{j=1}^d C_n(j)}, \qquad \forall i = 1, \cdots, d$$

- *Draw a new point:* $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ *kernel with inv. dist.* π_{θ_n}
- *Update the counter of the visited stratum*

$$C_{n+1}(i) = C_n(i) + \gamma_{n+1}\ C_n(i)\ \mathbb{1}_{\mathsf{X}_i}(X_{n+1}), \qquad \forall i = 1, \cdots, d$$

a WL based algorithm - convergence results (2/3)

On the form

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \left(\theta_n(i) \mathbb{1}_{\mathsf{X}_i}(X_{n+1}) - \sum_{j=1}^{d} \theta_n(j) \mathbb{1}_{\mathsf{X}_j}(X_{n+1}) \right) + \gamma_{n+1}^2 O_{w.p.1.}(1).$$

On the form

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \left(\theta_n(i) \mathbb{1}_{\mathsf{X}_i}(X_{n+1}) - \sum_{j=1}^{d} \theta_n(j) \mathbb{1}_{\mathsf{X}_j}(X_{n+1}) \right) + \gamma_{n+1}^2 O_{w.p.1.}(1).$$

Under conditions on

- the strata and the target: $0 < \inf_{\mathsf{X}} \pi \leq \sup_{\mathsf{X}} \pi < \infty$, $\theta_\star(i) > 0$.
- the ergodicity of the kernels P_θ
- the stepsize sequence γ_n: $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_n^2 < \infty$

it is proved asymptotic results (F., Jourdain, Kuhn, Lelièvre, Stoltz, 2015a)

1. The a.s. convergence of the sequence θ_n to $\theta_\star$.
2. The "convergence" of the samples $\{X_1, \cdots, X_n, \cdots\}$

$$\int f \, \pi \mathrm{d}\lambda = \lim_n \frac{d}{n} \sum_{k=1}^{n} f(X_k) \left(\sum_{i=1}^{d} \theta_k(i) \mathbb{1}_{\mathsf{X}_i}(X_k) \right) \qquad a.s.$$

$\hookrightarrow$ very bad Effective Sample Size and role of the stepsize sequence (F., Jourdain, Kuhn, Lelièvre, Stoltz, 2015b) in the transient phase

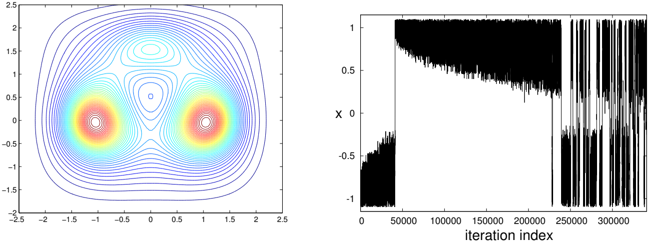


Figure: Left: level curves of the target density. Right: typical trajectory for $\beta = 15$ when $\gamma_n = \gamma_\star / n^{0.6}$ with $\alpha = 0.6$ and $\gamma_\star = 1$.

- The density depends on a parameter β: large values of β increases the metastability phenomenon.
- We choose $\gamma_n = \gamma_\star / n^\alpha \qquad \alpha \in (1/2, 1]$

$$\ln T_{(\alpha<1)} = C(\alpha, \gamma_\star) + \frac{1}{1-\alpha} \ln \beta \qquad \ln T_{(\alpha=1)} = C(\gamma_\star) + \frac{\mu_0}{1+\gamma_\star} \beta$$

$\hookrightarrow$ "self tuned" step size γ_n

An Adaptive Importance Sampling with

- **self-tuned stepsize sequence**
- **partial biasing to improve the IS step**

SHUS$_\rho^g$

Self-tuned and Partially biasing algorithm (F., Jourdain, Leliévre, Stoltz (2016))

Input:

- *initial values: a point* $X_0 \in \mathsf{X}$ *and a counter* $C_0 \in (\mathbb{R}_+^\star)^d$
- *a biasing function* ρ *and a stepsize control function* g

For $n = 0, 1, \cdots$

- *Normalize the counter*

$$\theta_n(i) = \frac{C_n(i)}{\sum_{j=1}^d C_n(j)}, \qquad \forall i = 1, \cdots, d$$

- *Draw a new point:* $X_{n+1} \sim P_{\rho(\theta_n)}(X_n, \cdot)$ *kernel with inv. dist.* $\pi_{\rho(\theta_n)}$
- *Update the counter of the visited stratum* $\forall i = 1, \cdots, d$

$$C_{n+1}(i) = C_n(i) + \frac{\gamma}{g\left(\sum_{j=1}^d C_n(j)\right)} \left(\sum_{j=1}^d C_n(j)\right) \rho\left(\theta_n(i)\right) \mathbb{1}_{\mathsf{X}_i}(X_{n+1}),$$

The intuition for this new update rule of C_n

The samples $X_n \overset{i.i.d.}{\sim} \pi$;

- A counter of the visits to each stratum

$$C_n(i) = C_{n-1}(i) + \gamma \mathbb{1}_{\mathsf{X}_i}(X_n) = C_0(i) + \gamma \sum_{k=1}^{n} \mathbb{1}_{\mathsf{X}_i}(X_k) \;\Rightarrow C_n(i) \sim \gamma n \, \theta_\star(i)$$

$$= C_{n-1}(i) + \underbrace{\frac{\gamma}{\sum_{j=1}^{d} C_{n-1}(j)}}_{\gamma_n = O(1/n)} \left(\sum_{j=1}^{d} C_{n-1}(j) \right) \mathbb{1}_{\mathsf{X}_i}(X_n)$$

- The estimate of $\theta_\star$

$$\theta_n(i) = \theta_{n-1}(i) + \gamma_n \left(\mathbb{1}_{\mathsf{X}_i}(X_n) - \sum_{j=1}^{d} \mathbb{1}_{\mathsf{X}_j}(X_n) \right) + O(\gamma_n^2)$$

- For approximation of integrals

$$\int f \pi \mathrm{d}\lambda \approx \frac{1}{n} \sum_{k=1}^{n} f(X_k)$$

The intuition for this new update rule of C_n

The samples $X_n \overset{i.i.d.}{\sim} \pi$; $X_n \overset{i.i.d.}{\sim} \pi_{\rho(\theta_\star)} \propto \sum_{i=1}^d \frac{\pi}{\rho(\theta_\star(i))} \mathbb{1}_{\mathsf{X}_i}$;

- A counter of the visits to each stratum

$$C_n(i) = C_{n-1}(i) + \underbrace{\frac{\gamma}{\sum_{j=1}^d C_{n-1}(j)}}_{\gamma_n = O(1/n)} \left(\sum_{j=1}^d C_{n-1}(j) \right) \rho(\theta_\star(i)) \mathbb{1}_{\mathsf{X}_i}(X_n)$$

- The estimate of $\theta_\star$

$$\theta_n(i) = \theta_{n-1}(i) + \gamma_n \left(\rho(\theta_\star(i)) \mathbb{1}_{\mathsf{X}_i}(X_n) - \sum_{j=1}^d \rho(\theta_\star(j)) \mathbb{1}_{\mathsf{X}_j}(X_n) \right) + O_{w.p.1}(\gamma_n^2)$$

- For approximation of integrals

$$\int f \pi \mathrm{d}\lambda \approx \frac{1}{n} \sum_{k=1}^n f(X_k) \left(\sum_{j=1}^d \rho(\theta_\star(j)) \mathbb{1}_{\mathsf{X}_j}(X_k) \right) \left(\sum_{j=1}^d \frac{\theta_\star(j)}{\rho(\theta_\star(j))} \right)$$

The discrepancy between the weights is modified through ρ. ex. $t^a, 0 < a < 1$

The intuition for this new update rule of C_n

The samples $X_n \overset{i.i.d.}{\sim} \pi$; $X_n \overset{i.i.d.}{\sim} \pi_{\rho(\theta_\star)} \propto \sum_{i=1}^d \frac{\pi}{\rho(\theta_\star(i))} \mathbb{I}_{\mathsf{X}_i}$;

- A counter of the visits to each stratum

$$C_n(i) = C_{n-1}(i) + \underbrace{\frac{\gamma}{g\left(\sum_{j=1}^d C_{n-1}(j)\right)}}_{\gamma_n \to 0} \left(\sum_{j=1}^d C_{n-1}(j)\right) \rho(\theta_\star(i)) \, \mathbb{I}_{\mathsf{X}_i}(X_n)$$

- The estimate of $\theta_\star$

$$\theta_n(i) = \theta_{n-1}(i) + \gamma_n \left(\rho(\theta_\star(i)) \mathbb{I}_{\mathsf{X}_i}(X_n) - \sum_{j=1}^d \rho(\theta_\star(j)) \mathbb{I}_{\mathsf{X}_j}(X_n)\right) + O_{w.p.1}(\gamma_n^2)$$

- For approximation of integrals

$$\int f \pi \mathrm{d}\lambda \approx \frac{1}{n} \sum_{k=1}^n f(X_k) \left(\sum_{j=1}^d \rho(\theta_\star(j)) \mathbb{I}_{\mathsf{X}_j}(X_k)\right) \left(\sum_{j=1}^d \frac{\theta_\star(j)}{\rho(\theta_\star(j))}\right)$$

The discrepancy between the weights is modified through ρ. ex. $t^a, 0 < a < 1$

Control the step size through a function g

The intuition for this new update rule of C_n

The samples $X_n \overset{i.i.d.}{\sim} \pi$; $X_n \overset{i.i.d.}{\sim} \pi_{\rho(\theta_\star)} \propto \sum_{i=1}^{d} \frac{\pi}{\rho(\theta_\star(i))} \mathbb{1}_{\mathsf{X}_i}$; The weight $\theta_\star$ is learnt along iterations

- A counter of the visits to each stratum

$$C_n(i) = C_{n-1}(i) + \underbrace{\frac{\gamma}{g\left(\sum_{j=1}^{d} C_{n-1}(j)\right)}}_{\gamma_n \to 0} \left(\sum_{j=1}^{d} C_{n-1}(j)\right) \rho(\theta_{n-1}(i)) \, \mathbb{1}_{\mathsf{X}_i}(X_n)$$

- The estimate of $\theta_\star$

$$\theta_n(i) = \theta_{n-1}(i) + \gamma_n \left(\rho(\theta_{n-1}(i)) \mathbb{1}_{\mathsf{X}_i}(X_n) - \sum_{j=1}^{d} \rho(\theta_{n-1}(j)) \mathbb{1}_{\mathsf{X}_j}(X_n) \right) + O_{w.p.1}(\gamma_n^2$$

- For approximation of integrals

$$\int f \pi \mathrm{d}\lambda \approx \frac{1}{n} \sum_{k=1}^{n} f(X_k) \left(\sum_{j=1}^{d} \rho(\theta_{k-1}(j)) \mathbb{1}_{\mathsf{X}_j}(X_k) \right) \left(\sum_{j=1}^{d} \frac{\theta_{k-1}(j)}{\rho(\theta_{k-1}(j))} \right)$$

The discrepancy between the weights is modified through ρ. ex. t^a, $0 < a < 1$

Control the step size through a function g

Assumptions

1. On the target density $0 < \inf_{\mathsf{X}} \pi \leq \sup_{\mathsf{X}} \pi < \infty$ and $\theta_\star(i) > 0$
2. On the ergodic behavior of the kernels Hastings-Metropolis kernel, with proposal $q(x,y)\mathrm{d}\lambda(y)$ such that $\inf_{\mathsf{X}^2} q > 0$
3. On the function $\rho \longrightarrow$ satisfied with $\rho(t) = t^a$ with $a \in [0,1)$
4. On the function g, chosen of the form $g(s) = (\ln(1+s))^{\alpha/(1-\alpha)}$ with $\alpha \in (1/2, 1)$

Convergence results (1/2)

By using sufficient conditions for convergence of Adaptive MCMC samplers F., Moulines, Priouret (2012) and convergence of Stochastic Approximation algo with controlled Markovian dynamics Andrieu, Moulines, Priouret (2005)

- On the random sequence γ_n almost-surely,

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

- On the weight sequence θ_n almost-surely,

$$\lim_n \theta_n = \theta_\star$$

- On the Importance Sampling step almost-surely,

$$\lim_n \frac{1}{n} \sum_{k=1}^n f(X_k) \left(\sum_{j=1}^d \rho(\theta_{k-1}(j)) \mathbb{1}_{\mathsf{X}_j}(X_k) \right) \left(\sum_{j=1}^d \frac{\theta_{k-1}(j)}{\rho(\theta_{k-1}(j))} \right) = \int f \ \pi \mathrm{d}\lambda$$

We wrote the results in the case

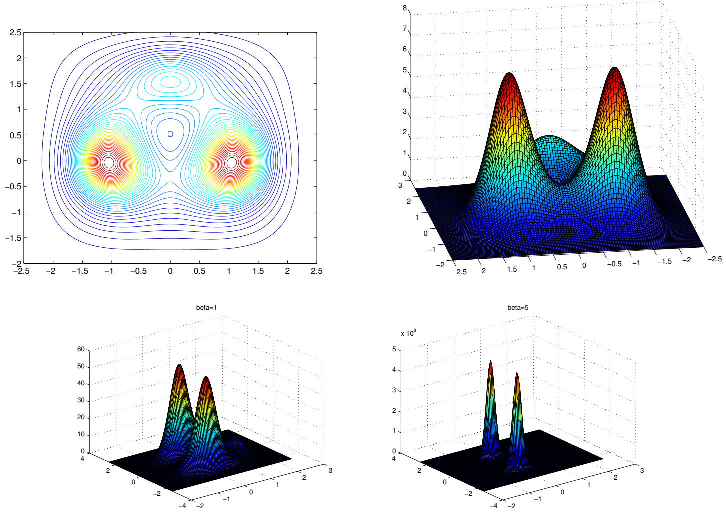
$$\rho(t) = t^a \text{ with } a \in [0, 1)$$
$$g(s) = (\ln(1+s))^{\alpha/(1-\alpha)} \text{ with } \alpha \in (1/2, 1)$$

but our convergence analysis also includes the case

- $\rho(t) = t$ and $g(s) = s$ (F., Jourdain, Lelièvre, Stoltz, 2016)
  In that case, our algorithm is the Self Healing Umbrella Sampling algorithm (Marsili et al. 2006)
  "no partial biasing" and "self-tuned stepsize"

- $\rho(t) = t^a, a \in [0, 1) \qquad g(s) = s^{1-a}$
  In that case, our algorithm is a discrete setting of the Well-Tempered metadynamics algorithm (Barducci, Bussi and Parrinello (2008))
  "partial biasing" and "self-tuned stepsize" with a correlated parameter a.

Is there a gain in such a self-tuned and partially biasing algorithm ?



Make the metastability larger by increasing β.

Case $\rho(t) = t^a$ for $a \in [0, 1)$

$g(s) = (\ln(1+s))^{\alpha/(1-\alpha)}$ for $\alpha \in (1/2, 1)$ $\Rightarrow \gamma_n = O_{wp1}(1/n^\alpha)$

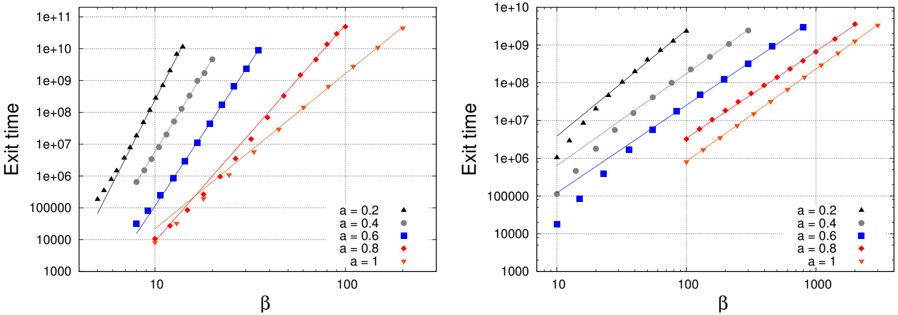


Figure: Left: Exit times for $\alpha = 0.8$. Right: Exit times for $\alpha = 0.6$.

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

- $T \uparrow$ when $\beta \uparrow$
- for fixed β and a: $T \downarrow$ when $\alpha \downarrow$.
- for fixed β and α: $T \downarrow$ when $a \uparrow$.
- Linear fit with a slope indep of a: $\ln T = c + (1-\alpha)^{-1} \ln \beta$

Comparison to the Well-Tempered Metadynamics $g(s) = s^{1-a}$ $(\Rightarrow \gamma_n = O(1/n))$ and $\rho(t) = t^a$ for $a \in (0,1)$

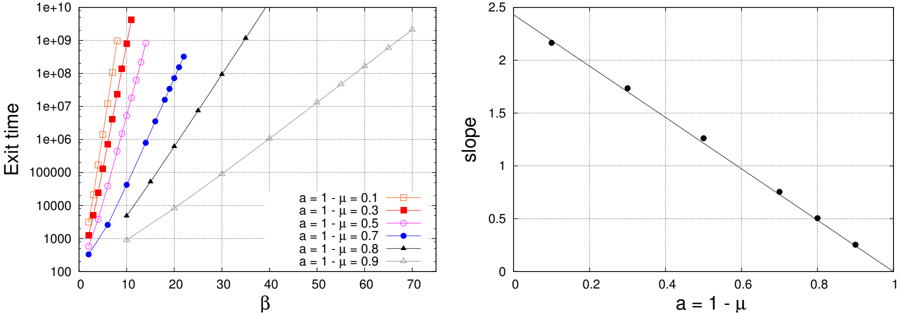


Figure: Left: Exit times for various values of a. Right: Associated slopes, fitted by $2.43(1-a)$.

Exit time T

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

Normalized Effective Sample Size (EF)

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

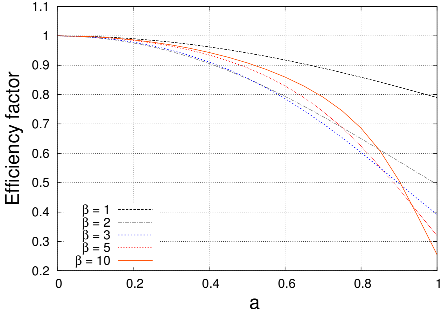


Figure: Efficiency factors EF(a) for various values of β.

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

- By definition, when uniform weights, $EF = 1$.
- For fixed β, $EF \uparrow$ when $a \downarrow$

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, a) to be fixed by the user
  - $\cdot$ a close to 1 in the transient phase, and a close to 0 at convergence.
  - $\cdot$ α close to $1/2$ in the transient phase.
- far more efficient in the transient phase than Well-Tempered Metadynamics or SHUS or WL.