Convergence and Efficiency of Adaptive Importance Sampling techniques with partial biasing

Benjamin Jourdain

École des Ponts ParisTech and INRIA, France

Stochastic Sampling and Accelerated Time Dynamics on Multidimensional Surfaces, IPAM, october 17 2017

Joint work with

- Gersende Fort (CNRS, IMT Toulouse)
- Tony Lelièvre (École des Ponts ParisTech and INRIA)
- Gabriel Stoltz (École des Ponts ParisTech and INRIA)

Talk based on the paper

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

Goal:

Explore the support of a distribution $\pi\, \mathrm{d}\lambda$ with density π w.r.t. the Lebesgue measure λ on $\mathcal{D} \subseteq \mathbb{R}^d$

and/or compute integrals w.r.t. π

$$\int_{\mathcal{D}} f(x)\ \pi(x) \mathrm{d}\lambda(x)$$

when π is highly metastable, d 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

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

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

- Define a partition of the support $\mathcal{D}$ in I strata

$$\mathcal{D} = \bigcup_{i=1}^{I} \mathcal{D}_i \qquad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \text{ for } i \neq j$$

- A family of auxiliary distribution based on a local biasing
For all probability $\theta = (\theta(1), \cdots, \theta(I))$ on $\{1, 2, \ldots, I\}$ with $\theta(i) > 0, \forall i$, let

$$\pi_\theta(x) \stackrel{\text{def}}{=} \left(\sum_{i=1}^{I} \frac{\theta_\star(i)}{\theta(i)} \right)^{-1} \sum_{i=1}^{I} \frac{\pi(x)}{\theta(i)} \mathbb{1}_{\mathcal{D}_i}(x),$$

where

$$\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathcal{D}_i} \pi \mathrm{d}\lambda$$

If $\mathcal{D}_i = \xi^{-1}([a_i, a_{i+1}))$ with $\xi : \mathbb{R}^d \to \mathbb{R}$ a collective variable (reaction coordinate) and $a_1 < a_2 < \ldots < a_{I+1}$ then $\log \theta_\star(i)$ is the free-energy (up to an additive constant)

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

- Define a partition of the support $\mathcal{D}$ in I strata

$$\mathcal{D} = \bigcup_{i=1}^{I} \mathcal{D}_i \qquad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \text{ for } i \neq j$$

- A family of auxiliary distribution based on a local biasing
  For all probability $\theta = (\theta(1), \cdots, \theta(I))$ on $\{1, 2, \ldots, I\}$ with $\theta(i) > 0, \forall i$, let

$$\pi_\theta(x) \stackrel{\text{def}}{=} \left(\sum_{i=1}^{I} \frac{\theta_\star(i)}{\theta(i)} \right)^{-1} \sum_{i=1}^{I} \frac{\pi(x)}{\theta(i)} \mathbb{1}_{\mathcal{D}_i}(x),$$

  where

$$\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathcal{D}_i} \pi \mathrm{d}\lambda$$

  If $\mathcal{D}_i = \xi^{-1}([a_i, a_{i+1}))$ with $\xi : \mathbb{R}^d \to \mathbb{R}$ a collective variable (reaction coordinate) and $a_1 < a_2 < \ldots < a_{I+1}$ then $\log \theta_\star(i)$ is the free-energy (up to an additive constant)

Key property: $\pi_{\theta_\star}(\mathcal{D}_i) = 1/I$ – 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
$$\int H(\theta, x)\ \pi_\theta(x)\, \mathrm{d}\lambda(x) = 0.$$
- from draws $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where $P_\theta(x, \cdot)$ is a kernel with invariant distribution π_θ (e.g. a Metropolis-Hastings kernel)

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 $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$ where $P_\theta(x, \cdot)$ is a kernel with invariant distribution π_θ (e.g. a Metropolis-Hastings kernel)

If convergence is established, this yields

- an estimator of the free energy: $\lim_n \theta_n = \theta_\star$.
- an approximation of the target distribution π - computed on the fly/online
$$\int f \; \pi \mathsf{d}\lambda = \lim_n \frac{I}{n} \sum_{k=1}^{n} f(X_k) \left(\sum_{i=1}^{I} \theta_k(i) \mathbb{1}_{\mathcal{D}_i}(X_k) \right)$$

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$_\rho^g$

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

$$C_{n+1}(i) = C_n(i) + (\cdots)^2 \; \mathbb{1}_{\mathcal{D}_i}(X_{n+1}) \qquad i = 1, \cdots, I$$

3. Normalize the counter to obtain a probability measure on $\{1, 2, \ldots, I\}$

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

Fundamental: if $X_{n+1} \in \mathcal{D}_i$

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

Wang-Landau (WL) update

a WL based algorithm - algorithm (1/3)

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

Input:

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

For $n = 0, 1, \cdots$

- *Normalize the counter*

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

- *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*

$$C_{n+1}(i) = C_n(i) + \gamma_{n+1}\ C_n(i)\ \mathbb{1}_{\mathcal{D}_i}(X_{n+1}), \qquad \forall i = 1, \cdots, I$$

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \overbrace{\theta_n(i) \left(\mathbb{1}_{\mathcal{D}_i}(X_{n+1}) - \sum_{j=1}^{I} \theta_n(j) \mathbb{1}_{\mathcal{D}_j}(X_{n+1}) \right)}^{H_i(\theta_n, X_{n+1})} + \gamma_{n+1}^2 \mathcal{O}_{w.p.1.}(1).$$

$$\int_{\mathbb{R}^d} H(\theta, x) \pi_\theta(x) \mathrm{d}x = (\sum_{i=1}^{I} \theta_\star(i)/\theta(i))^{-1} (\theta_\star - \theta)$$

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

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \overbrace{\theta_n(i) \left(\mathbb{1}_{\mathcal{D}_i}(X_{n+1}) - \sum_{j=1}^{I} \theta_n(j) \mathbb{1}_{\mathcal{D}_j}(X_{n+1}) \right)}^{H_i(\theta_n, X_{n+1})} + \gamma_{n+1}^2 \mathcal{O}_{w.p.1.}(1).$$

$$\int_{\mathbb{R}^d} H(\theta, x) \pi_\theta(x) \mathrm{d}x = (\sum_{i=1}^{I} \theta_\star(i)/\theta(i))^{-1} (\theta_\star - \theta)$$

Under conditions on

- the strata and the target: $0 < \inf_{\mathcal{D}} \pi \leq \sup_{\mathcal{D}} \pi < \infty$.
- the kernels P_θ : satisfied by Metropolis-Hastings kernels, with proposal $q(x, y)\mathrm{d}\lambda(y)$ such that $q(x, y) = q(y, x)$ and $\inf_{(x,y)\in\mathcal{D}^2} q(x, y) > 0$.
- the stepsize sequence γ_n: $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_n^2 < \infty$

it is proved asymptotic results (Fort, J., Kuhn, Lelièvre, Stoltz, 2015a)

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

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

$\hookrightarrow$ bad Efficiency Factor

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

and role of the stepsize sequence (Fort, J., Kuhn, Lelièvre, Stoltz, 2015b) in the transient phase

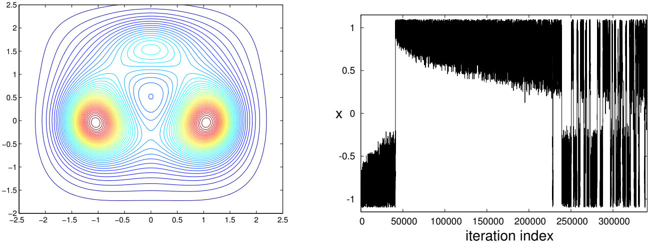

Figure : Left: level curves of the target density. Right: typical trajectory for $\beta = 15$ when $\gamma_n = \gamma_\star / n^{0.6}$ with $\alpha = 0.6$ and $\gamma_\star = 1$.

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

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

$\hookrightarrow$ "self tuned" step size γ_n

An Adaptive Importance Sampling Algorithm 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* $X_0 \in \mathcal{D}$ *and a counter* $C_0 \in (\mathbb{R}_+^\star)^I$
- *a biasing function* $\rho : (0,1) \to \mathbb{R}_+^*$ *and a stepsize function* $g : \mathbb{R}_+^* \to \mathbb{R}_+^*$,

Set $\pi_{\rho(\theta)}(x) \stackrel{\text{def}}{=} \left(\sum_{i=1}^I \frac{\theta_\star(i)}{\rho(\theta(i))}\right)^{-1} \sum_{i=1}^I \frac{\pi(x)}{\rho(\theta(i))} \mathbb{1}_{\mathcal{D}_i}(x)$.

For $n = 0, 1, \cdots$

- *Normalize the counter* $\theta_n(i) = C_n(i) / \sum_{j=1}^I C_n(j), \qquad \forall i = 1, \cdots, I$
- *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* $\qquad \forall i = 1, \cdots, I$

$$C_{n+1}(i) = C_n(i) + \underbrace{\frac{\gamma}{g\left(\sum_{j=1}^I C_n(j)\right)}}_{\textit{stepsize } \gamma_{n+1}} \underbrace{\left(\sum_{j=1}^I C_n(j)\right) \rho\left(\theta_n(i)\right)}_{= C_n(i) \ \textit{if} \ \rho(t) \equiv t} \mathbb{1}_{\mathcal{D}_i}(X_{n+1}),$$

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

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

$$= C_n(i) + \underbrace{\frac{\gamma}{\sum_{j=1}^{I} C_n(j)}}_{\gamma_{n+1} = \frac{\gamma}{n\gamma + \sum_{j=1}^{I} C_0(j)}} \left(\sum_{j=1}^{I} C_n(j) \right) \mathbb{1}_{\mathcal{D}_i}(X_{n+1})$$

- The estimate of $\theta_\star$

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \left(\mathbb{1}_{\mathcal{D}_i}(X_{n+1}) - \theta_n(i) \sum_{j=1}^{I} \mathbb{1}_{\mathcal{D}_j}(X_{n+1}) \right) + \mathcal{O}(\gamma_{n+1}^2)$$

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

- A counter of the visits to each stratum

$$C_{n+1}(i) = C_n(i) + \underbrace{\frac{\gamma}{\sum_{j=1}^{I} C_n(j)}}_{\gamma_{n+1} = \mathcal{O}(1/n)} \left(\sum_{j=1}^{I} C_n(j) \right) \rho(\theta_\star(i)) \; \mathbb{1}_{\mathcal{D}_i}(X_{n+1})$$

$$C_n(i) \sim \left(\sum_{j=1}^{I} \frac{\theta_\star(j)}{\rho(\theta_\star(j)} \right)^{-1} \gamma n \theta_\star(i)$$

- The estimate of $\theta_\star$

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \left(\rho(\theta_\star(i)) \mathbb{1}_{\mathcal{D}_i}(X_{n+1}) - \theta_n(i) \sum_{j=1}^{I} \rho(\theta_\star(j)) \mathbb{1}_{\mathcal{D}_j}(X_{n+1}) \right) + \mathcal{O}(\gamma_{n+1}^2)$$

- For approximation of integrals

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

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

- A counter of the visits to each stratum

$$C_{n+1}(i) = C_n(i) + \underbrace{\frac{\gamma}{g\left(\sum_{j=1}^{I} C_n(j)\right)}}_{\gamma_{n+1} \to 0} \left(\sum_{j=1}^{I} C_n(j)\right) \rho(\theta_\star(i)) \, \mathbb{1}_{\mathcal{D}_i}(X_{n+1})$$

- The estimate of $\theta_\star$

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \left(\rho(\theta_\star(i)) \mathbb{1}_{\mathcal{D}_i}(X_{n+1}) - \theta_n(i) \sum_{j=1}^{I} \rho(\theta_\star(j)) \mathbb{1}_{\mathcal{D}_j}(X_{n+1}) \right) + \mathcal{O}(\gamma_{n+1}^2)$$

- For approximation of integrals

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

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 $X_{n+1} \sim P_{\rho(\theta_n)}(X_n, .)$ and the weight $\theta_\star$ is learnt along iterations

- A counter of the visits to each stratum

$$C_{n+1}(i) = C_n(i) + \underbrace{\frac{\gamma}{g\left(\sum_{j=1}^{I} C_n(j)\right)}}_{\gamma_{n+1} \to 0} \left(\sum_{j=1}^{I} C_n(j)\right) \rho(\theta_n(i)) \, \mathbb{I}_{\mathcal{D}_i}(X_{n+1})$$

- The estimate of $\theta_\star$

$$\theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1} \left(\rho(\theta_n(i)) \mathbb{I}_{\mathcal{D}_i}(X_{n+1}) - \theta_n(i) \sum_{j=1}^{I} \rho(\theta_n(j)) \mathbb{I}_{\mathcal{D}_j}(X_{n+1}) \right) + \mathcal{O}(\gamma_{n+1}^2)$$

- For approximation of integrals

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

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 : $\sup_{\mathcal{D}} \pi < \infty$ and $\min_{1 \leq i \leq I} \theta_\star(i) > 0$
2. On the kernels P_θ : satisfied by Metropolis-Hastings kernels, with proposal $q(x, y)\mathrm{d}\lambda(y)$ such that $q(x, y) = q(y, x)$ and $\inf_{(x,y)\in\mathcal{D}^2} q(x, y) > 0$
3. On the function $\rho \longrightarrow$ satisfied with $\rho(t) = \max(t_0, t)^a$ with $t_0, a \in [0, 1)$. See (Dama, Hocky, Sun, Voth, 2015) and (McCarty, Valsson, Tiwary, Parrinello, 2015) for motivations to choose $t_0 > 0$.
4. On the function g, chosen of the form

$$g(s) = \begin{cases} (\ln(1+s))^{\alpha/(1-\alpha)} \text{ with } \alpha \in (1/2, 1) \\ s^\mu \text{ with } \mu > 0 \to \text{ corresponds to } \alpha = 1 \end{cases}$$

By using sufficient conditions for convergence of Adaptive MCMC samplers Fort, 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,

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

- On the weight sequence θ_n almost-surely,

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

- On the Importance Sampling step almost-surely,

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

We wrote the results in the case

$\rho(t) = \max(t_0, t)^a$ with $t_0, a \in [0, 1)$

$$g(s) = \begin{cases} (\ln(1+s))^{\alpha/(1-\alpha)} \text{ with } \alpha \in (1/2, 1) \\ s^\mu \text{ with } \mu > 0 \rightarrow \text{ corresponds to } \alpha = 1 \end{cases}$$

Applies to a discrete version of the Well-Tempered metadynamics algorithm (Barducci, Bussi and Parrinello (2008)) where $\rho(t) = t^a$ $\quad g(s) = s^{1-a}$ with $a \in (0, 1)$, $\gamma_n = \mathcal{O}(1/n)$
The "partial biasing" and "self-tuned stepsize" parameters are one to one.

Convergence also holds in the case $\rho(t) = t$ and g as above (Fort, J., Lelièvre, Stoltz, 2016).
Additional assumption $\inf_{\mathcal{D}} \pi > 0$ needed to prove recurrence $\limsup_{n\to\infty} \min_{1\le i\le I} \theta_n(i) > 0$.
Indeed when $\theta_n(i)$ small and $X_{n+1} \in \mathcal{D}_i$, the increase of the counter $C_{n+1}(i+1) - C_n(i) \propto \rho(\theta_n(i))$ is smaller than when $\rho(t) = t^a$ with $a < 1$.
Applies to the Self Healing Umbrella Sampling algorithm (Marsili et al. 2006) where $g(s) = s$ and $\rho(t) = t$ "no partial biasing".

We prove cv of the Generalized Wang-Landau algorithm where for $n \in \mathbb{N}$,

$$C_{n+1}(i) = C_n(i)\left(1 + \gamma_{n+1}\frac{\rho(\theta_n(i))}{\theta_n(i)}\mathbb{1}_{\mathcal{D}_i}(X_{n+1})\right)$$
$$= C_n(i) + \gamma_{n+1}\left(\sum_{j=1}^{I} C_n(j)\right)\rho(\theta_n(i))\mathbb{1}_{\mathcal{D}_i}(X_{n+1}),$$

- γ_{n+1} is a positive random variable only depending on $(C_0, X_0, C_1, X_1, \ldots, C_n, X_n)$ (the past of the algorithm),
- $(\gamma_n)_n$ is non increasing, $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 < \infty$ and $\sup_n \frac{\gamma_n}{\gamma_{n+I-1}} < \infty$,

and then check that these hypotheses are satisfied by $(\gamma_{n+1} = \frac{\gamma}{g(\sum_{j=1}^{I} C_n(j))})_{n\in\mathbb{N}}$.

$$\theta_{n+1}(i) = \frac{C_n(i)}{\sum_{j=1}^{I} C_n(j)} \times \frac{1 + \gamma_{n+1}\frac{\rho(\theta_n(i))}{\theta_n(i)}\mathbb{1}_{\mathcal{D}_i}(X_{n+1})}{1 + \gamma_{n+1}\sum_{j=1}^{I}\rho(\theta_n(j))\mathbb{1}_{\mathcal{D}_j}(X_{n+1})}$$
$$= \theta_n(i) + \gamma_{n+1}\underbrace{\left(\rho(\theta_n(i))\mathbb{1}_{\mathcal{D}_i}(X_{n+1}) - \theta_n(i)\sum_{j=1}^{I}\rho(\theta_n(j))\mathbb{1}_{\mathcal{D}_j}(X_{n+1})\right)}_{H_i(\theta_n, X_{n+1})} + \mathcal{O}(\gamma_{n+1}^2).$$

$$h(\theta) := \int_{\mathbb{R}^d} H(\theta, x)\pi_{\rho(\theta)}(x)\mathrm{d}\lambda(x) = \left(\sum_{j=1}^{I} \frac{\theta_\star(j)}{\rho(\theta(j))}\right)^{-1} (\theta_\star - \theta).$$

- By considering a subsequence of $(\min_{1\leq i\leq I} \theta_n(i))_n$ along well-chosen stopping times $(T_k)_{k\geq 1}$ such that X_{T_k} is in the stratum with smallest weight $\theta_{T_{k-1}}(.)$, we check the recurrence of the algorithm : there is a compact subset $\mathcal{K}$ of the open subset $\Theta = \{\theta \in (\mathbb{R}_+^*)^I : \sum_{i=1}^I \theta(i) = 1\}$ of $\mathbb{R}^I$ such that $(\theta_n)_n$ is infinitely often in $\mathcal{K}$ $\Leftrightarrow$ $\limsup_{n\to\infty} \min_{1\leq i\leq I} \theta_n(i) > 0$.
- Introduce the Lyapunov function $U(\theta) = \sum_{i=1}^T \theta_\star(i) \ln(\theta_\star(i)/\theta(i))$ given by the relative entropy (Kullback-Leibler divergence) of the probability measure θ on $\{1, \ldots, I\}$ w.r.t. $\theta_\star$. Since $\partial_{\theta(i)} U(\theta) = -\frac{\theta_\star(i)}{\theta(i)}$,

$$\left(\sum_{j=1}^{I} \frac{\theta_\star(j)}{\rho(\theta(j))}\right) \nabla U.h(\theta) = -\sum_{i=1}^{I} \frac{\theta_\star^2(i)}{\theta(i)} + \overbrace{\sum_{i=1}^{I} \theta_\star(i)}^{=1=\sum_{i=1}^I (2\theta_\star(i)-\theta(i))}$$

$$= -\sum_{i=1}^{I} \theta(i) \left(\frac{\theta_\star(i)}{\theta(i)} - 1\right)^2 \leq 0.$$

Convergence of the Generalized Wang-Landau algorithm

- Rewrite

$$\theta_{n+1} = \theta_n + \gamma_{n+1} h(\theta_n) + \gamma_{n+1} R_{n+1}$$

  and check using results by Fort, Moulines, Priouret (2012) on the dependence on θ of π_θ and the solution F_θ to the Poisson equation $F_\theta - P_{\rho(\theta)} F_\theta = H(.,\theta) - h(\theta)$ that $\lim_{n\to\infty} \sup_{k \geq n} \left| \sum_{j=n}^{k} \gamma_j R_j \right| = 0$.

- With $\nabla U.h \leq 0$, $\mathcal{L} := \{\theta \in \Theta : \nabla U.h(\theta) = 0\} = \{\theta_\star\}$ and using Andrieu, Moulines, Priouret (2005), deduce stability : $\liminf_{n\to\infty} \min_{1 \leq i \leq I} \theta_n(i) > 0$ and a.s. convergence of $(U(\theta_n))_n$ to the image $\{0\}$ of $\mathcal{L}$ by U.
  By the Pinsker-Csiszar-Kullback inequality,

$$\sum_{i=1}^{I} |\theta_n(i) - \theta_\star(i)| \leq \sqrt{2U(\theta_n)} \longrightarrow_{n\to\infty} 0.$$

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

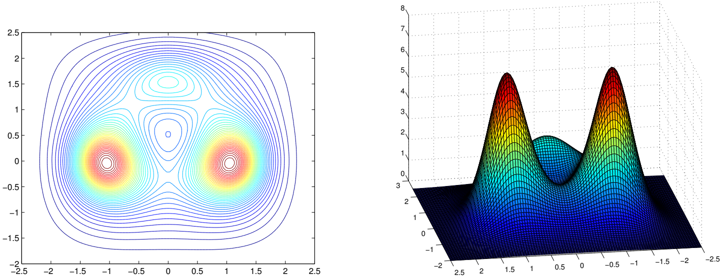

Figure : Left: level curves of the potential. Right: target density.

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 $\alpha \in (1/2, 1)$ $\Rightarrow \gamma_n = \mathcal{O}_{wp1}(1/n^\alpha)$

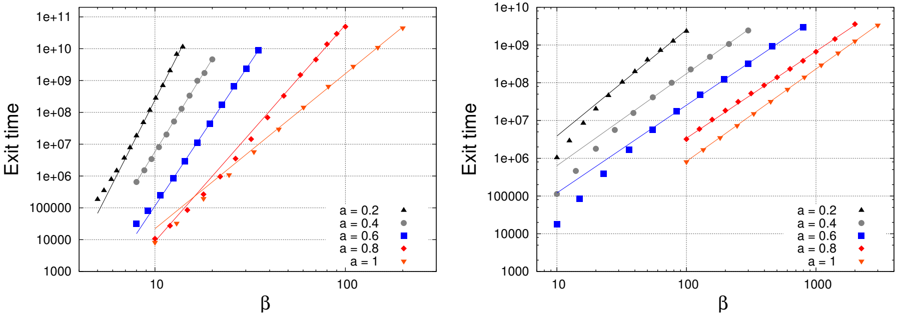

Figure : Left: Exit times for $\alpha = 0.8$. Right: Exit times for $\alpha = 0.6$.

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 $g(s) = s^{1-a}$ $(\Rightarrow \gamma_n = \mathcal{O}(1/n))$ and $\rho(t) = t^a$ for $a \in (0,1)$

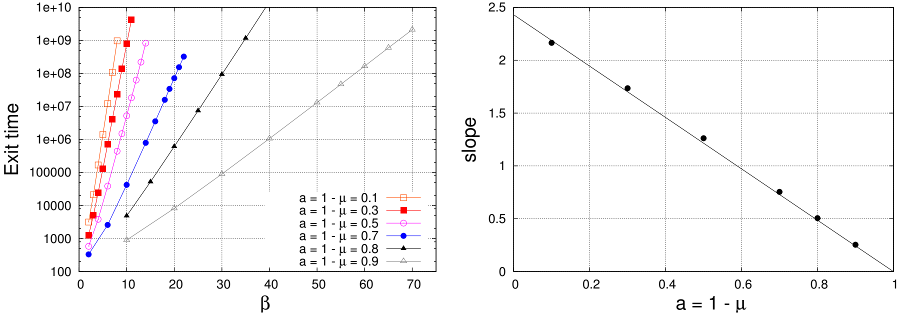

Figure : Left: Exit times for various values of a. Right: Associated slopes, fitted by $2.43(1-a)$.

Exit time T

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

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

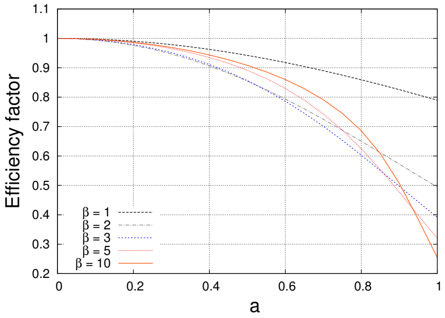

Figure : Efficiency factors EF(a) for various values of β.

$$EF(n) = \frac{\left(n^{-1}\sum_{k=1}^{n}\sum_{i=1}^{I}\theta_\star^a(i)\mathbb{I}_{\mathcal{D}_i}(X_k)\right)^2}{n^{-1}\sum_{k=1}^{n}\left(\sum_{i=1}^{I}\theta_\star^a(i)\mathbb{I}_{\mathcal{D}_i}(X_k)\right)^2} \in [0,1],\ (X_k)_k \text{ i.i.d.} \sim \pi_{\theta_\star^a}$$

$$\lim_{n\to\infty} EF(n) = \left(\sum_{i=1}^{I}\theta_\star^{1-a}(i)\right)^{-1}\left(\sum_{i=1}^{I}\theta_\star^{1+a}(i)\right)^{-1} \uparrow \text{ when } a \downarrow \text{ for fixed } \beta.$$

A convergent 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
  - a stepsize parameter $\alpha \in (1/2, 1]$, $\gamma_n = \mathcal{O}(n^{-\alpha})$ as $n \to \infty$,
  - a biasing parameter $a \in [0, 1]$.