Convergence of the Wang-Landau algorithm

Gersende FORT

LTCI
CNRS & Telecom ParisTech
Paris, France

Joint work with

- Benjamin Jourdain, Tony Lelièvre and Gabriel Stoltz - from ENPC, France.
- Estelle Kuhn - from INRA Jouy-en-Josas, France.

Paper `arXiv math.PR 1207.6880`

Introduction (1/3)

- The goal is to compute expectations under the distribution π when
  - the dimension of the support/state space $\mathbb{X}$ is very large,
  - π is multimodal (or metastable).

- **Example: in Molecular dynamics**, the models consist in the description of the state of the system: the location of the N particles x_ℓ (e.g. the set of N points in $\mathbb{R}^3$) and sometimes the speed of the particles.

  A state of the system is characterized by a probability $\pi(\mathbf{x})$:

  $$\pi(\mathbf{x}) \propto \exp(-\mathcal{H}(\mathbf{x})) \qquad \text{where} \quad \mathbf{x} = (x_1, \cdots, x_N) \in \mathbb{X}.$$

  The *potential/Hamiltonian* $\mathcal{H}(\mathbf{x})$ describes interactions between the $x_1, \cdots, x_N$.

  The goal is to compute derivatives of the *partition function*.

Introduction (2/3)

- Exact computations of $\int \phi \, d\pi$ are not possible (π is known up to a normalizing constant, the domain of integration is very large, $\cdots$)
- (Markov chain) Monte Carlo methods allow to sample points $(\mathbf{X}_t)_t$ s.t.

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \phi(\mathbf{X}_t) \xrightarrow{\text{a.s.}} \int \phi \, d\pi.$$

Introduction (2/3)

- Exact computations of $\int \phi \, d\pi$ are not possible (π is known up to a normalizing constant, the domain of integration is very large, $\cdots$)
- (Markov chain) Monte Carlo methods allow to sample points $(\mathbf{X}_t)_t$ s.t.

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \phi(\mathbf{X}_t) \xrightarrow{\text{a.s.}} \int \phi \, d\pi.$$

- Unfortunately, in mestastable systems, the samples remain trapped in local modes for a very long time

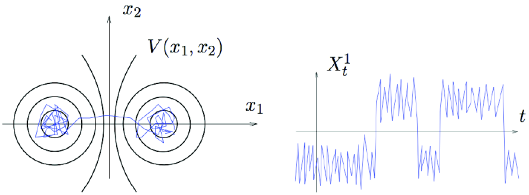


FIG.: [left] level curves of a potential in $\mathbb{R}^2$ which is metastable in the first direction. [right] path of the first component of $(\mathbf{X}_t)_t$

Introduction (2/3)

- Exact computations of $\int \phi \, d\pi$ are not possible (π is known up to a normalizing constant, the domain of integration is very large, $\cdots$)
- (Markov chain) Monte Carlo methods allow to sample points $(\mathbf{X}_t)_t$ s.t.

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \phi(\mathbf{X}_t) \xrightarrow{\text{a.s.}} \int \phi \, d\pi .$$

- Unfortunately, in mestastable systems, the samples remain trapped in local modes for a very long time

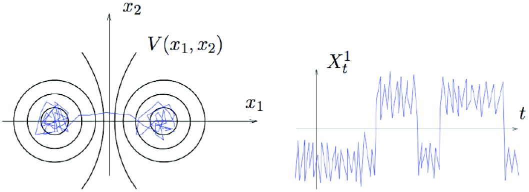


FIG.: [left] level curves of a potential in $\mathbb{R}^2$ which is metastable in the first direction. [right] path of the first component of $(\mathbf{X}_t)_t$

In such situations, the convergence is very long to obtain!

Introduction (3/3)

- Nevertheless, in Molecular Dynamics, it is often possible to identify a reaction coordinate that is, in some sense a "direction of metastability".

Introduction (3/3)

- Nevertheless, in Molecular Dynamics, it is often possible to identify a reaction coordinate that is, in some sense a "direction of metastability".

A new approach to define samplers robust to metastability:

1) sample from a biased distribution $\pi_\star$ such that
   - the image of $\pi_\star$ by the reaction coordinate $\mathcal{O}$ is **uniform**:
   $$\mathcal{O}(\mathbf{X}) \text{ when } \mathbf{X} \sim \pi_\star \text{ has a uniform distribution}$$
   - the conditional distribution of $\mathbf{x}$ given $\mathcal{O}(\mathbf{x})$ under $\pi_\star$ and π are the same.

2) approximate integrals w.r.t. π by an importance sampling algorithm with proposal $\pi_\star$

General Wang-Landau (1/4)

- Choose a partition $\mathcal{X}_1, \cdots, \mathcal{X}_d$ of $\mathbb{X}$.
- Set $\mathcal{O}(\mathbf{x}) = i$ iff $\mathbf{x} \in \mathcal{X}_i$.

Then

$$\pi_\star(\mathbf{x}) \text{ denoted hereafter } \pi_{\theta_\star}(\mathbf{x}) \propto \sum_{i=1}^{d} \frac{\pi(\mathbf{x})}{\theta_\star(i)} \mathbb{1}_{\mathcal{X}_i}(\mathbf{x})$$

where $\theta_\star = (\theta_\star(1), \cdots, \theta_\star(d))$ is the **weight vector** i.e. $\sum_i \theta_\star(i) = 1$ and $\theta_\star(i) \geq 0$

$$\theta_\star(i) = \int_{\mathcal{X}_i} d\pi(\mathbf{x})$$

$\hookrightarrow$ BUT: $\pi_{\theta_\star}$ is unknown since in non-trivial applications $\theta_\star$ is unknown

General Wang-Landau (2/4)

Wang-Landau is an iterative algorithm designed to simultaneously

- learn the weight vector $\theta_\star$
- draw samples $\{X_k, k \geq 0\}$ approximating the distribution $\pi_{\theta_\star}$.

- Note that $\pi_\star \neq \pi$ but

roughly: $$\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k} \approx \pi_{\theta_\star} \Longrightarrow \frac{1}{n}\sum_{k=1}^{n} \theta_\star(i) \mathbb{1}_{X_k \in \mathbb{X}_i} \, \delta_{X_k} \approx \pi$$

an empirical approximation of π is obtained by importance sampling from samples approximating $\pi_{\theta_\star}$.

General Wang-Landau (3/4): the algorithm

Iteratively, define $\{(X_t, \theta_t), t \geq 0\}$

(i) sample $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$ where P_{θ_t} is a Markov kernel with invariant distribution π_{θ_t}

$$\pi_\theta(\mathbf{x}) \propto \sum_{i=1}^{d} \frac{\pi(\mathbf{x})}{\theta(i)} \mathbb{1}_{\mathcal{X}_i}(\mathbf{x}).$$

(ii) Update the weights

$$\theta_{t+1} = \text{function}(t, \theta_t, X_{t+1})$$

For θ_{t+1}, updating strategy based on stochastic approximation

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1})$$

where $(\gamma_t)_t$ is a positive stepsize sequence and, the field H is chosen so that $\theta_\star$ is a zero of

$$\theta \mapsto \int \pi_\theta(d\mathbf{x}) H(\theta, \mathbf{x}).$$

General Wang-Landau (4/4): the algorithm

Many updating strategies for (θ_t, γ_t) such that: the chain is pushed towards strata with weaker frequency of visit thus improving the exploration of the space. Among examples

- (exponential update) for any $i \in \{1, \cdots, d\}$

$$\theta_{t+1}(i) = \frac{\theta_t(i) \exp\left(\gamma_{t+1}\left(\mathbb{1}_{\mathcal{X}_i}(X_{t+1}) - 1/d\right)\right)}{\sum_{\ell=1}^{d} \theta_t(\ell) \exp\left(\gamma_{t+1}\left(\mathbb{1}_{\mathcal{X}_\ell}(X_{t+1}) - 1/d\right)\right)}$$

- (linearized version) if $X_{t+1} \in \mathcal{X}_i$,

$$\begin{cases} \theta_{t+1}(i) = \theta_t(i) + \gamma_{t+1}\, \theta_t(i)(1 - \theta_t(i)) & \\ \theta_{t+1}(k) = \theta_t(k) - \gamma_{t+1}\, \theta_t(k)\theta_t(i) & k \neq i \end{cases}$$

General Wang-Landau (4/4): the algorithm

Many updating strategies for (θ_t,γ_t) such that: the chain is pushed towards strata with weaker frequency of visit thus improving the exploration of the space. Among examples

- (exponential update) for any $i \in \{1, \cdots, d\}$

$$\theta_{t+1}(i) = \frac{\theta_t(i) \exp\left(\gamma_{t+1}\left(\mathbb{I}_{\mathcal{X}_i}(X_{t+1}) - 1/d\right)\right)}{\sum_{\ell=1}^{d} \theta_t(\ell) \exp\left(\gamma_{t+1}\left(\mathbb{I}_{\mathcal{X}_\ell}(X_{t+1}) - 1/d\right)\right)}$$

- (linearized version) if $X_{t+1} \in \mathcal{X}_i$,

$$\begin{cases} \theta_{t+1}(i) = \theta_t(i) + \gamma_{t+1}\,\theta_t(i)(1 - \theta_t(i)) \\ \theta_{t+1}(k) = \theta_t(k) - \gamma_{t+1}\,\theta_t(k)\theta_t(i) \qquad k \neq i \end{cases}$$

- Deterministic or random decreasing stepsize sequence $(\gamma_t)_t$.

In our work: the **linearized update** and a **deterministic** sequence $\{\gamma_t, t \geq 0\}$.

A numerical illustration (1/3)

Target density: $\pi(x_1,x_2) \propto \exp(-\beta\,\mathcal{H}(x_1,x_2))\mathbb{1}_{[-R,R]}(x_1)$

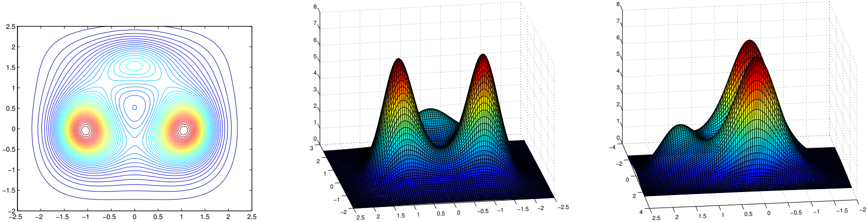
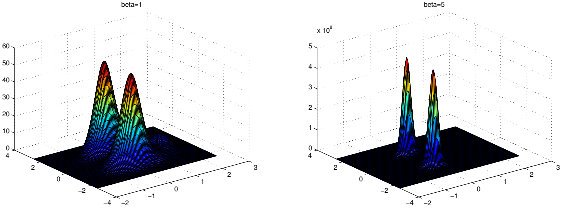
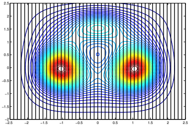


FIG.: [left] Level curves of the potential $\mathcal{H}$. [center, right] Density π up to a normalizing constant.



The larger β is, the larger is the ratio between the weight of the strata located near to the main metastable states and the weight of the transition region (near $x_1 = 0$).

A numerical illustration (2/3)



$R = 2.4$. $d = 48$ strata, partition along the x-axis.

P_θ are Hastings-Metropolis kernels with proposal distribution $\mathcal{N}(0,(2R/d)^2 I)$ and target π_θ.
$X_0 = (-1,0)$.

The stepsize sequence is $\gamma_t \sim c/t^{0.8}$.

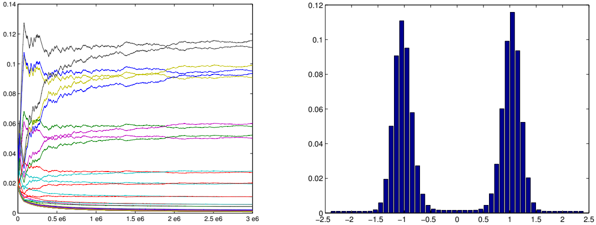


FIG.: [left] The sequences $(\theta_t(i))_t$. [right] The limiting value $\theta_\star(i)$

A numerical illustration (3/3)

Path of the x_1-component of $(X_t)_t$, when X_t is the WL chain (left) and the HM chain (right).

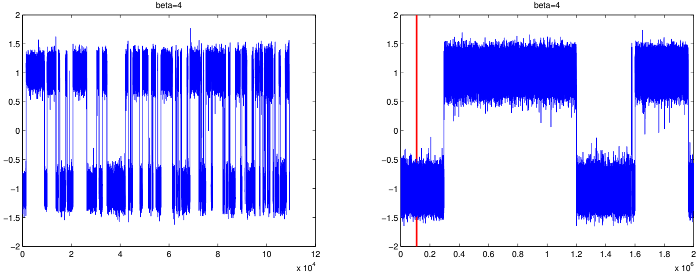


FIG.: [left] Wang Landau, $T = 110\,000$. [right] Hastings Metropolis, $T = 2\ 10^6$; the red line is at $x = 110\,000$

Outline

Limiting behavior: Two points of view

1. Limiting behavior of $\{\theta_t, t \geq 0\}$ i.e. of a stochastic approximation procedure
$$\theta_{t+1} = \theta_t + \gamma_{t+1} \, H(\theta_t, X_{t+1})$$
with controlled Markov chain dynamic $\{X_t, t \geq 0\}$.

2. Limliting behavior of $\{X_t, t \geq 0\}$ i.e. of an adaptive Markov chain Monte Carlo
$$\mathbb{P}\left(X_{t+1} \in A | \mathcal{F}_t\right) = P_{\theta_t}(X_t, A).$$

Assumptions

The limiting behavior of the Wang-Landau is studied under the assumptions

A) The target distribution is $\pi \, d\lambda$ on $\mathbb{X} \subset \mathbb{R}^p$ and $\sup_{\mathbb{X}} \pi < \infty$.

B) The partition $(\mathcal{X}_i)_i$ such that $\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathcal{X}_i} \pi \; d\lambda > 0$.

C) For any $\theta \in \Theta$, P_θ is a Hastings-Metropolis kernel with proposal q and invariant distribution π_θ. It is assumed: $\inf_{\mathbb{X}^2} q > 0$.

D) The stepsize sequence $(\gamma_t)_t$ is non-increasing and satisfies $\sum_t \gamma_t = +\infty$ and $\sum_t \gamma_t^2 < \infty$.

Sufficient conditions for the convergence of $\{\theta_t, t \geq 0\}$

Benveniste et al. (1987), Andrieu et al. (2005), Fort (2013)

$$\theta_{t+1} = \theta_t + \gamma_{t+1} \, H(\theta_t, X_{t+1}) = \theta_t + \gamma_{t+1} h(\theta_t) + \gamma_{t+1} \left(H(\theta_t, X_{t+1}) - h(\theta_t)\right)$$

where the h is the **mean field** defined by

$$h(\theta) \stackrel{\text{def}}{=} \int H(\theta, \mathbf{x}) \pi_\theta(d\mathbf{x}) = \left(\sum_{i=1}^{d} \frac{\theta_\star(i)}{\theta(i)} \right)^{-1} (\theta_\star - \theta)$$

Sufficient conditions for the convergence of $\{\theta_t, t \geq 0\}$

Benveniste et al. (1987), Andrieu et al. (2005), Fort (2013)

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1}) = \theta_t + \gamma_{t+1} h(\theta_t) + \gamma_{t+1} \left(H(\theta_t, X_{t+1}) - h(\theta_t)\right)$$

where the h is the **mean field** defined by

$$h(\theta) \stackrel{\text{def}}{=} \int H(\theta, \mathbf{x}) \pi_\theta(d\mathbf{x}) = \left(\sum_{i=1}^{d} \frac{\theta_\star(i)}{\theta(i)} \right)^{-1} (\theta_\star - \theta)$$

Convergence to $\theta_\star$ when

- the O.D.E $\dot{\theta} = h(\theta)$ converges to $\theta_\star$ (Lyapunov function, $\cdots$)
- (stability condition) the sequence $(\theta_t)_t$ visits infinitely often a compact subset of $\{\theta : \theta(i) > 0 \text{ and } \sum_{i=1}^{d} \theta(i) = 1\}$
- the noise sequence is small enough

  · $\sum_t \gamma_t = \infty$, $\sum_t \gamma_t^2 < \infty$
  · the transition kernels $(P_\theta, \theta \in \Theta)$ are ergodic (enough) and are smooth enough in θ.

Result: stability of $\{\theta_t, t \geq 0\}$

Theorem: F., Jourdain, Kuhn, Lelièvre, Stoltz (2012)
Under the assumptions A to D and $\inf_{\mathbb{X}} \pi > 0$

$$\mathbb{P}\left(\limsup_t \min_{1 \leq i \leq d} \theta_t(i) > 0\right) = 1.$$

Sketch of the proof:

- $T_k < \infty$ w.p.1. where T_k are the successive times when a sample X_n is drawn in the stratum $i_\star$ such that $\theta_n(i_\star) = \min_k \theta_n(k)$.
- We prove that $\mathbb{P}(\limsup_k \left(\min_i \theta_{T_k - 1}(i)\right) > 0) = 1$, and a key property for this proof is

$$P_\theta(x, \mathcal{X}_j) \mathbb{1}_{\mathcal{X}_i}(x) \leq C\, 1 \wedge \frac{\theta(i)}{\theta(j)}.$$

$\hookrightarrow$ Low probability of moving from a stratum with small weight to a stratum with large weight.

Result: convergence of $\{\theta_t, t \geq 0\}$

Theorem F., Jourdain, Kuhn, Lelièvre, Stoltz (2012)

Under the assumptions A to D and the stability of the sequence $\{\theta_t, t \geq 0\}$

$$\mathbb{P}\left(\lim_t \theta_t = \theta_\star\right) = 1.$$

Sketch of the proof: Check the conditions of Andrieu, Moulines and Priouret (2005). Main ingredients:

- The Lyapunov function V associated to the mean field h

$$V(\theta) = -\sum_{i=1}^{d} \theta_\star(i) \ \log\left(\frac{\theta(i)}{\theta_\star(i)}\right)$$

- The uniform (in $\mathbf{x}, \theta$) geometric ergodicity of the transition kernels P_θ
- The regularity properties

$$\sup_{x \in \mathbb{X}} \| P_\theta(x, \cdot) - P_{\theta'}(x, \cdot) \|_{\mathrm{TV}} + \| \pi_\theta - \pi_{\theta'} \|_{\mathrm{TV}} C \ \| \theta - \theta' \|$$

Result: Rate of convergence of $\{\theta_t, t \geq 0\}$

Theorem F., Jourdain, Kuhn, Lelièvre, Stoltz (2012)

Under the assumptions A to D, when $\gamma_t \sim \gamma_\star / t^\alpha$ $(1/2 < \alpha \leq 1)$

$$\gamma_t^{-1/2} \left(\theta_t - \theta_\star\right) \xrightarrow{\text{dist.}} \mathcal{N}_d\left(0, \frac{d\gamma_\star}{2\gamma_\star - d} U_\star\right)$$

where

$$U_\star = \frac{d}{2} \int_{\mathbb{X}} \left\{ \hat{H}_\star(\mathbf{x}) \hat{H}_\star^T(\mathbf{x}) - P_{\theta_\star} \hat{H}_\star(\mathbf{x}) \, P_{\theta_\star} \hat{H}_\star^T(\mathbf{x}) \right\} \pi_{\theta_\star}(\mathbf{x}) \lambda(d\mathbf{x})$$

and

$$\hat{H}_\star(\mathbf{x}) = \sum_{\ell \geq 0} P_{\theta_\star}^\ell \left(H(\theta_\star, \cdot) - h(\theta_\star) \right)(\mathbf{x})$$

- The optimal rate is reached with $\gamma_t = d/t$ thus yielding to the optimal covariance matrix $d^2 U_\star$.
- Averaging technique can also be used to reach this optimal rate.

Sufficient conditions for the ergodicity of $\{X_t, t \geq 0\}$

Roberts and Rosenthal (2007); F., Moulines and Priouret (2012)

$$\begin{aligned}\mathbb{E}\left[f(X_t)\right] - \pi_{\theta_\star}(f) &= \mathbb{E}\left[f(X_t) - \mathbb{E}\left[f(X_t)|\mathcal{F}_{t-\ell}\right]\right] \\ &+ \mathbb{E}\left[\mathbb{E}\left[f(X_t)|\mathcal{F}_{t-\ell}\right] - P^\ell_{\theta_{t-\ell}} f(X_{t-\ell})\right] \\ &+ \mathbb{E}\left[P^\ell_{\theta_{t-\ell}} f(X_{t-\ell}) - \pi_{\theta_{t-\ell}}(f)\right] \\ &+ \mathbb{E}\left[\pi_{\theta_{t-\ell}}(f) - \pi_{\theta_\star}(f)\right]\end{aligned}$$

Convergence when

- the first term is null
- the second term is small when adaptation is diminishing
- the third term is small when the transition kernels $(P_\theta, \theta \in \Theta)$ are ergodic (enough), at a rate which is uniform (enough) in θ (containment condition)
- the last term is small provided $(\theta_t, t \geq 0)$ converges to $\theta_\star$ since in our case

$$\|\pi_\theta - \pi_{\theta_\star}\|_{\mathrm{TV}} \leq C \, \|\theta - \theta_\star\|$$

Result: ergodicity and LLN for the samples $\{X_t, t \geq 0\}$ (1/2)

Theorem F., Jourdain, Kuhn, Lelièvre, Stoltz (2012)
Under the assumptions A to D and the stability of the sequence $\{\theta_t, t \geq 0\}$,

$$\lim_t \mathbb{E}\left[f(X_t)\right] = \int f(\mathbf{x})\ \pi_{\theta_\star}(\mathbf{x})\lambda(d\mathbf{x})$$

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow{a.s.} \int f(\mathbf{x})\ \pi_{\theta_\star}(\mathbf{x})\lambda(d\mathbf{x})$$

for any bounded measurable function f.

Proof: Check the conditions of F., Moulines and Priouret (2012). Main ingredients:

- The uniform (in $\mathbf{x}, \theta$) geometric ergodicity of the transition kernels P_θ
- The regularity properties

$$\|\pi_\theta - \pi_{\theta'}\|_{\mathrm{TV}} + \sup_{x \in \mathbb{X}} \|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{\mathrm{TV}} \leq C\|\theta - \theta'\|$$

Result: ergodicity and LLN for the weighted samples $\{X_t, t \geq 0\}$ (2/2)

Theorem F., Jourdain, Kuhn, Lelièvre, Stoltz (2012)
Under the assumptions A to D and the stability of the sequence $\{\theta_t, t \geq 0\}$,

$$\lim_t \mathbb{E}\left[d \sum_{i=1}^{d} \theta_t(i) \ f(X_t) \mathbb{1}_{\mathcal{X}_i}(X_t) \right] = \int f(\mathbf{x}) \ \pi(\mathbf{x}) \lambda(d\mathbf{x})$$

$$\frac{1}{T} \sum_{t=1}^{T} \left(d \sum_{i=1}^{d} \theta_t(i) \mathbb{1}_{\mathcal{X}_i}(X_t) \right) \ f(X_t) \xrightarrow{a.s.} \int f(\mathbf{x}) \ \pi(\mathbf{x}) \lambda(d\mathbf{x})$$

for any bounded measurable function f.

Conclusion

- Wang-Landau algorithms are designed to be able to switch as fast as possible from a metastable state to another metastable state in order to **efficiently** explore the whole configuration space.
- We obtained results on the asymptotic behavior of WL but

  how to study the **efficiency** of the WL and how to compare WL to a non-adaptive MCMC sampler?

$\hookrightarrow$ in a companion paper,

- Comparison in terms of *how rapidly does the sampler escape from a metastable state*
- Computation of exit times for some models

and these discussions evidence the interest of WL over classical MCMC algorithms.

Bibliography

(free energy) Biasing techniques in Molecular Dynamics

[a] on the choice of a "good" reaction coordinate: is it easier to sample from $\pi_\star$ than from π? [b] approximation of $\pi_\star$ on the fly converging to $\pi_\star$ in the long-time limit: either approximation of $\pi_\star$ (adaptive biasing potential) or when the reaction coordinate is a continuous parameter, approximation of (adaptive biasing force).

N. Chopin, T. Lelièvre and G. Stoltz. Free energy methods for efficient exploration of mixture posterior densities. Stat. Comput. 22-897-916 (2012). with a discussion

E. Darve and A. Pohorille. Calculating free energies using average force. J. Chem. Phys. 115:9169-9183 (2001).

Dickson, B. and Legoll, F. and Lelièvre, T. and Stoltz, G. and Fleurat-Lessard, P. Free energy calculations: An efficient adaptive biasing potential method. J. Phys. Chem. B. 114:5823-5830 (2010)

B. Jourdain, T. Lelièvre and R. Roux. Existence, uniqueness and convergence of particle approximation for the adaptive biasing force process. M2AN Math. Model. Numer. Anal. 44:831-865 (2010).

T. Lelièvre and K. Minoukadeh. Longtime convergence of an adaptive biasing force method: the bi-channel case. Arch. Ration. Mech. Anal, 202:1-34 (2011).

T. Lelièvre, M. Rousset and G. Stoltz. Computation of free energy profiles with adaptive parallel dynamics. J. Chem. Phys. 126: (2007).

T. Lelièvre, M. Rousset and G. Stoltz. Long-time convergence of an Adaptive Biasing Force method. Nonlinearity, 21:1155-1181 (2008).

T. Lelièvre, M. Rousset and G. Stoltz. Free energy computations: a mathematical perspective. Imperial Collegee Press (2010).

Minoukadeh, K. and Chipot, C. and Lelièvre, T. Potential of mean force calculations: a multiple-walker adaptive biasing force approach. J. Chem. Th. Comput. 6:1008-1017 (2010).

The Wang Landau method in Statistical Physics

F. Wang and D. P. Landau, Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States external link: Phys. Rev. Lett., 86, 2050, (2001) *Description of the method and its application to the 2D Potts model and Ising model*

F. Wang and D. P. Landau, Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram external link: Phys. Rev. E, 64, 056101, (2001) *Detailed description of the previous study*

D. P. Landau, S.-H. Tsai and M. Exler, A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling external link: Am. J. Phys., 72, 1294, (2004) *Detailed description of the first paper*

M. S. Shell, P. G. Debenedetti and A. Z. Panagiotopoulos, Generalization of the Wang-Landau method for off-lattice simulations external link: Phys. Rev. E, 66, 056703, (2002)

B. J. Schulz, K. Binder, M. Muller, and D. P. Landau, Avoiding boundary effects in Wang-Landau sampling external link. Phys. Rev. E, 67, 067102, (2003) *A slight modification was introduced to avoid the systematic underestimation of the density of states at the higher energy border*

A. G. Cunha-Netto, A. A. Caparica S.-H. Tsai, R. Dickman and D. P. Landau. Improving Wang-Landau sampling with adaptive windows external link. Phys. Rev. E, 78, 055701, (2008) *Using adaptive windows in energy to avoid border effects between energy ranges.*

Methodology and Convergence analysis of Wang-Landau

F. Liang. A general Wang-Landau algorithm for Monte Carlo computation. J. Am. Stat. Assoc. 100:1311-1327 (2005).

F. Liang, C. Liu and R.J. Carroll. Stochastic approximation in Monte Carlo computation. J. Am. Stat. Assoc. 102:305-320 (2007).

Y. Atchadé and J.S. Liu. The Wang-Landau algorithm for Monte Carlo computation in general state space. Stat. Sinica, 20(1):209-233 (2010).

Application of Wang-Landau to Statistics. Convergence results (on the samples $(X_t)_t$) under the assumption that the algorithm is "stable"

L. Bornn, P. Jacob, P. Del Moral and A. Doucet. An Adaptive Wang-Landau Algorithm for Automatic Density Exploration. To appear in Journal of Computational and Graphical Statistics (2013).

New methods for (i) adaptive binning strategy to automate the difficult task of partitioning the state space, (ii) the use ofinteracting parallel chains to improve the convergence speed and use of computational resources, and (iii) the use of adaptive proposal distributions.

P. Jacob and R. Ryder. The Wang-Landau algorithm reaches the flat histogram criterion in finite time. To appear in Ann. Appl. Probab. (2013).

The linearized version of the update of the weight vector θ_t satisfies in finite time the uniformity criterion required in the original Wang-Landau algorithm. This is not guaranteed for some non-linear update.

G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. Convergence of the Wang-Landau algorithm. In revision (2013)

Stability, convergence and rate of convergence of a linearized version of WL.

G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. Efficiency of the Wang-Landau algorithm. Submitted (2013)

Discussion on the efficiency of Wang-Landau when measured in terms of the exit-time from a metastable state.

Convergence of adaptive MCMC
Roberts, G.O. and Rosenthal, J.S. Coupling and ergodicity of adaptive MCMC. J. Appl. Probab. 44:458-475 (2007).

Fort, G., Moulines, E. and Priouret, P. Convergence of interacting MCMC: ergodicity and law of large numbers. Ann. Statist. 39:3262-3289 (2012)

Fort, G., Moulines, E., Priouret, P. and Vandekerkhove, P. Convergence of interacting MCMC: Central Limit Theorem. To appear in Bernoulli (2013).

Convergence of stochastic approximation scheme
A. Benveniste, M. Metivier and P. Priouret. Adaptive algorithms for Stochastic Approximations. Springer-Verlag (1987).

C. Andrieu, E. Moulines and P. Priouret. Stability of stochastic approximation under verifiable conditions. SIAM Journal on Control and Optimisation 44:283-312 (2005).

G. Fort. Central Limit Theorems for Stochastic Approximation with controlled Markov chain dynamics. Submittted (2013).