Convergence and Efficiency 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`

Wang-Landau: a biasing technique (1/3)

- 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.
- There are interactions between the particles $x_1, \cdots, x_N$, described through a *potential/Hamiltonian* $\mathcal{H}(x_1, \cdots, x_N)$.
- A state of the system is characterized by a probability $\pi(\mathbf{x})$: e.g. in the canonical ensemble NVT

$$\pi(\mathbf{x}) \propto \exp(-\beta \mathcal{H}(\mathbf{x})) \qquad \beta \stackrel{\text{def}}{=} \frac{1}{k_B\, T} \text{(inverse temperature)}$$

where $\mathbf{x} = (x_1, \cdots, x_N) \in \mathbb{X}$.

- The goal is to compute derivatives of the *partition function* i.e. expectations under the distribution π when

  the dimension of the support $\mathbb{X}$ is very large,
  π is multimodal (or metastable).

Wang-Landau: a biasing technique (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.

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

Wang-Landau: a biasing technique (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 points remain trapped in local modes for a very long time

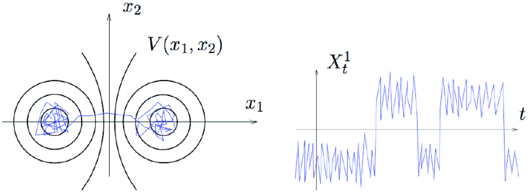


FIG.: [left] level curves of a potential in $\mathbb{R}^2$ which is metastable in the first direction. [right] path of the first component of $(\mathbf{X}_t)_t$

Wang-Landau: a biasing technique (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 points remain trapped in local modes for a very long time

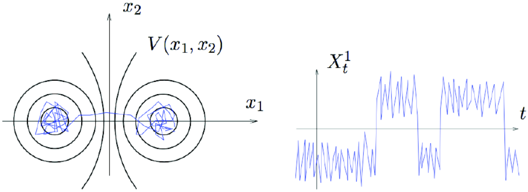


FIG.: [left] level curves of a potential in $\mathbb{R}^2$ which is metastable in the first direction. [right] path of the first component of $(\mathbf{X}_t)_t$

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

Wang-Landau: a biasing technique (3/3)

- It is not possible to answer the metastability problem in full generality (number of modes, size of the barriers between metastable states which increase with the dimension N, $\cdots$).
- Nevertheless, in Molecular Dynamics, it is often possible to identify a reaction coordinate that is, in some sense a "direction of metastability".

Wang-Landau: a biasing technique (3/3)

- It is not possible to answer the metastability problem in full generality (number of modes, size of the barriers between metastable states which increase with the dimension N, $\cdots$).
- 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:

- ▶ sample from a biased distribution $\pi_\star$ such that
  - the image of $\pi_\star$ by the reaction coordinate $\mathcal{O}$ is **uniform**:

$$\mathcal{O}(\mathbf{X}) \quad \text{when } \mathbf{X} \sim \pi_\star \text{ has a uniform distribution}$$

  - the conditional distribution of $\pi_\star$ given $\mathcal{O}(\mathbf{x})$ is equal to the conditional distribution of π given $\mathcal{O}(\mathbf{x})$.
- ▶ approximate integrals w.r.t. π by an importance sampling algorithm with proposal $\pi_\star$

Outline

The original Wang-Landau algorithm (1/3)

Assume

$$\pi(\mathbf{x}) \propto \exp(-\beta\ \mathcal{H}(\mathbf{x}))$$

on a discrete (but large) space $\mathbb{X}$, and the goal is to compute

$$\sum_{\mathbf{x}\in\mathbb{X}} \Phi(\mathcal{H}(\mathbf{x}))\ \pi(\mathbf{x})$$

Then,

$$\sum_{\mathbf{x}} \Phi(\mathcal{H}(\mathbf{x}))\ \pi(\mathbf{x}) = \sum_{e\in\mathcal{H}(\mathbb{X})} \Phi(e)\ \frac{g(e)}{\sum_{e'\in\mathcal{H}(\mathbb{X})} g(e')}$$

where g is the density of state:

$$g(e) \stackrel{\text{def}}{=} \sum_{\mathbf{x}\in\mathbb{X}} \mathbb{1}_{\mathcal{H}(\mathbf{x})=e}$$

The original Wang-Landau algorithm (2/3)

Density of state:

$$g(e) \stackrel{\text{def}}{=} \sum_{\mathbf{x} \in \mathbb{X}} \mathbb{I}_{\mathcal{H}(\mathbf{x})=e}$$

- $g(e)$ can not be calculated exactly for large systems.
- Although the total number of configurations increases exponentially with the size of the system, the total number of possible energy levels increases linearly with the size of system. example: q^{L^2} compared to $2L^2$ for a q-state Potts on a $L \times L$ lattice withe nearest-neighfor interactions

Wang and Landau (2001) proposed to perform a random walk in the energy space in order to estimate $g(e)$ for any e.

With the density of states,

- we can calculate most of thermodynamic quantities in all inverse temperature β
- we can access many thermodynamic properties (free energy, internal energy, specific heat i.e. normalizing constant, expectation and variance under π)

The original Wang-Landau algorithm (3/3)

Algorithm:

- Initialisation:

  density of state: $g(e) = 1$ for any e

  modification factor: f_0
- LOOP 1:
  - Repeat

    Run a Markov chain with transition matrix

    $$Q(e,e') = 1 \wedge \frac{g(e)}{g(e')}$$

    Update the histogram in the energy space: if E is the new point,

    $$\ln g(E) \leftarrow \ln g(E) + \ln f_t$$
  - Until the *flat histogram* is reached.
- LOOP 2: Repeat LOOP1 with a new modification factor $f_{t+1} \leftarrow \sqrt{f_t}$ until the modification factor is smaller than a predefined value.

The original Wang-Landau algorithm (3/3)

Algorithm:

- Initialisation:

  density of state: $g(e) = 1$ for any e

  modification factor: f_0
- LOOP 1:
  - Repeat

    Run a Markov chain with transition matrix

    $$Q(e,e') = 1 \wedge \frac{g(e)}{g(e')}$$

    Update the histogram in the energy space: if E is the new point,

    $$\ln g(E) \leftarrow \ln g(E) + \ln f_t$$
  - Until the *flat histogram* is reached.
- LOOP 2: Repeat LOOP1 with a new modification factor $f_{t+1} \leftarrow \sqrt{f_t}$ until the modification factor is smaller than a predefined value.

Why does it work? the intuition:

- The chain Q is reversible w.r.t. $\propto 1/g(e)$
- The distribution of $g(E)$ when $E \sim 1/g(e)$ is the uniform distribution.

General Wang-Landau (1/3)

How to sample a metastable target distribution π on a general state space $\mathbb{X}$?

- Choose a partition $\mathcal{X}_1, \cdots, \mathcal{X}_d$ of $\mathbb{X}$. Then

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

General Wang-Landau (1/3)

How to sample a metastable target distribution π on a general state space $\mathbb{X}$?

- Choose a partition $\mathcal{X}_1, \cdots, \mathcal{X}_d$ of $\mathbb{X}$. Then

$$\pi(\mathbf{x}) = \sum_{i=1}^{d} \mathbb{I}_{\mathcal{X}_i}(\mathbf{x})\pi(\mathbf{x})$$

- Consider a family of biased distributions $(\pi_\theta, \theta \in \mathbb{R}^d)$ on $\mathbb{X}$

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

where $\theta = (\theta(1), \cdots, \theta(d))$ satisfies $\sum_i \theta(i) = 1$ and $\theta(i) \geq 0$.

General Wang-Landau (1/3)

How to sample a metastable target distribution π on a general state space $\mathbb{X}$?

- Choose a partition $\mathcal{X}_1, \cdots, \mathcal{X}_d$ of $\mathbb{X}$. Then

$$\pi(\mathbf{x}) = \sum_{i=1}^{d} \mathbb{I}_{\mathcal{X}_i}(\mathbf{x})\pi(\mathbf{x})$$

- Consider a family of biased distributions $(\pi_\theta, \theta \in \mathbb{R}^d)$ on $\mathbb{X}$

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

  where $\theta = (\theta(1), \cdots, \theta(d))$ satisfies $\sum_i \theta(i) = 1$ and $\theta(i) \geq 0$.

- Run an algorithm which combines

  sampling under π_{θ_t} (exact or MCMC)
  update of the biasing factor $\theta_{t+1} \leftarrow \theta_t + \cdots$

  in such a way that $(\theta_t)_t$ and $(\pi_{\theta_t})_t$ converge to

$$\theta_\star = (\pi(\mathcal{X}_1), \cdots, \pi(\mathcal{X}_d)) \qquad \pi_{\theta_\star}(\mathcal{X}_i) = \frac{1}{d}$$

General Wang-Landau (2/3)

When it converges

- $\theta_t(i) \approx \pi(\mathcal{X}_i)$

- Integrals w.r.t. π by Importance Sampling

$$\int \phi \, d\pi \approx \frac{1}{T} \sum_{t=1}^{T} \left(d \sum_{i=1}^{d} \theta_t(i) \mathbb{1}_{\mathbf{X}_t \in \mathcal{X}_i} \right) \phi(\mathbf{X}_t)$$

General Wang-Landau (3/3)

Set $\theta_\star = (\pi(\mathcal{X}_1), \cdots, \pi(\mathcal{X}_d))$.

Algorithm

- Initialisation: X_0 and $\theta_0 = (1/d, \cdots, 1/d)$
- Repeat: given (X_t, θ_t)

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

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

  where the field H is chosen so that $\theta_\star$ is a zero of

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

  and $(\gamma_t)_t$ is a positive stepsize sequence.

Wang-Landau in Statistics

- Multicanonical sampling (Atchadé & Liu, 2010)
- Simulated Tempering (Atchadé & Liu, 2010)

  Target: ρ on $\tilde{\mathbb{X}}$.
  Temperatures: $T_1 > T_2 > \cdots > T_d = 1$.

$$\mathbb{X} = \tilde{\mathbb{X}} \times \{1, \cdots, d\} \qquad \theta_\star(i) = \int \rho^{1/T_i}(d\mathbf{x}) \qquad \pi_\theta(\mathbf{x}, i) \propto \frac{1}{\theta(i)} \rho^{1/T_i}(\mathbf{x})$$

- Trans-dimensional MCMC (Atchadé & Liu, 2010)

  $\tilde{\mathbb{X}} = \bigcup_{k=1}^{K} \mathbb{X}_k$
  Target $\propto \sum_{k=1}^{K} \rho_k(\mathbf{x})\ \mathbb{1}_{\mathbb{X}_k}(\mathbf{x})$ on $\tilde{\mathbb{X}}$.

$$\mathbb{X} = \tilde{\mathbb{X}} \times \{1, \cdots, d\} \qquad \theta_\star(i) = \int_{\mathbb{X}_i} \rho_i(d\mathbf{x}) \qquad \pi_\theta(\mathbf{x}, i) \propto \frac{1}{\theta(i)} \rho_i(\mathbf{x}) \mathbb{1}_{\mathbb{X}_i}(\mathbf{x})$$

- Variable selection (Bornn et al, 2013)

  Target: a posteriori distribution π of a binary vector.
  reaction coordinate: partition of the energy state $-\log \pi(\mathbb{X})$

- Bayesian inference in mixture models (Bornn et al, 2013)

Outline

WL: an example of adaptive MCMC (1/2)

- A family of target distributions $(\pi_\theta)_{\theta \in \Theta}$.
- A family of transition kernels $(P_\theta)_{\theta \in \Theta}$ such that $\pi_\theta P_\theta = \pi_\theta$.

- WL defines a random sequence $((X_t,\theta_t))_t$ such that

$$\mathbb{E}\left[\phi(X_{t+1}) | \theta_0, X_0, \cdots, \theta_t, X_t\right] = \int P_{\theta_t}(X_t, dy)\phi(y).$$

and the parameter θ_t is updated by a Stochastic Approximation algorithm

WL: an example of adaptive MCMC (2/2)

In the literature, different strategies for the update of (θ_t, γ_t) in such a way that $\sum_{i=1}^d \theta_t(i) = 1$ and $\theta_t(i) \geq 0$.

- (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 $X_{t+1} \in \mathcal{X}_i$,

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

$\hookrightarrow$ For the next move, the probability of sampling a point in the current stratum $\mathcal{X}_i$ is reduced. The chain is pushed towards strata which weaker frequency of visit thus improving the exploration of the space.

WL: an example of adaptive MCMC (2/2)

In the literature, different strategies for the update of (θ_t, γ_t) in such a way that $\sum_{i=1}^{d} \theta_t(i) = 1$ and $\theta_t(i) \geq 0$.

- (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 $X_{t+1} \in \mathcal{X}_i$,

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

$\hookrightarrow$ For the next move, the probability of sampling a point in the current stratum $\mathcal{X}_i$ is reduced. The chain is pushed towards strata which weaker frequency of visit thus improving the exploration of the space.

- The stepsize sequence $(\gamma_t)_t$ decreases deterministically OR randomly (based on a flat histogram criterion for example).

In our work, we consider the linearized update and a deterministic stepsize sequence γ_t.

A numerical illustration (1/2)

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

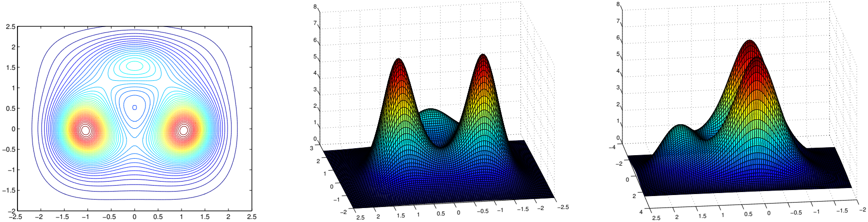
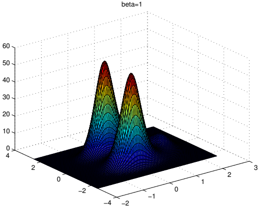
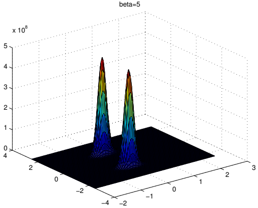


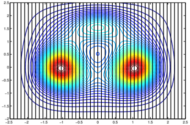
FIG.: [left] Level curves of the potential $\mathcal{H}$. [center, right] Density π up to a normalizing constant.





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 (2/2)



$R = 2.4$. $d = 48$ 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 $\gamma_t \sim c/t^{0.8}$.

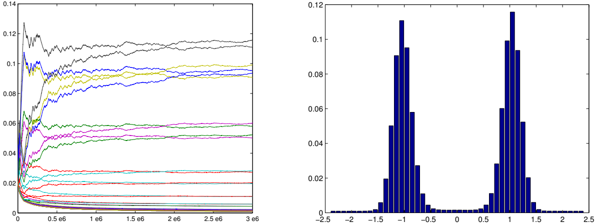


FIG.: [left] The sequences $(\theta_t(i))_t$. [right] The limiting value $\theta_\star(i)$

Sufficient conditions for the convergence of adaptive MCMC (1/2)

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

For the proof of the ergodicity, observe

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

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

$$\|\pi_\theta - \pi_{\theta_\star}\|_{\mathrm{TV}} \leq 2(d-1) \sum_{i=1}^{d} \left| 1 - \frac{\theta(i)}{\theta_\star(i)} \right|$$

Sufficient conditions for the convergence of adaptive MCMC (2/2)

For the convergence of the weight sequence $(\theta_t)_t$, observe

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

where the mean field h is 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 adaptive MCMC (2/2)

For the convergence of the weight sequence $(\theta_t)_t$, observe

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

where the mean field h is 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 $\{\theta : \theta(i) > 0$ 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.

Main results: assumptions (1/5)

1. The target distribution has a density π w.r.t. the measure λ on $\mathbb{X} \subset \mathbb{R}^p$, $\sup_{\mathbb{X}} \pi < \infty$.
2. The partition $(\mathcal{X}_i)_i$ such that $\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathcal{X}_i} \pi \, d\lambda > 0$.
3. For any $\theta \in \Theta$, P_θ is a Hastings-Metropolis kernel with proposal q and invariant distribution π_θ. It is assumed: $\inf_{\mathbb{X}^2} q > 0$.
4. The stepsize sequence $(\gamma_t)_t$ satisfies $\sum_t \gamma_t = +\infty$ and $\sum_t \gamma_t^2 < \infty$.

Under these assumptions, there exists $\rho \in (0,1)$ such that for any θ

$$\sup_{x \in \mathbb{X}} \|P_\theta^t(x,\cdot) - \pi_\theta\|_{\mathrm{TV}} \le 2(1-\rho)^t$$

Main result: stability of $(\theta_t)_t$ (2/5)

Theorem

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

Under the stated assumptions and $\inf_{\mathbb{X}} \pi > 0$

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

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

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

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

Main result: convergence of $(\theta_t)_t$ (3/5)

Theorem

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

Under the stated assumptions and the stability of the sequence $(\theta_t)_t$,

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

Main result: convergence of $(\theta_t)_t$ (3/5)

Theorem

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

Under the stated assumptions and the stability of the sequence $(\theta_t)_t$,

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

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 $\mathbf{x}$,θ) geometric ergodicity of the transition kernels P_θ
- The regularity properties

$$\|\pi_\theta - \pi_{\theta'}\|_{\mathrm{TV}} \leq 2(d-1)\sum_{i=1}^{d}\left|1 - \frac{\theta(i)}{\theta'(i)}\right|$$

$$\sup_{x\in\mathbb{X}} \|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{\mathrm{TV}} \leq 4\sup_i\left|1 - \frac{\theta(i)}{\theta'(i)}\right| + 4\sup_i\left|1 - \frac{\theta'(i)}{\theta(i)}\right|$$

Main result: ergodicity and LLN for the samples $(X_t)_t$ (4/5)

Theorem

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

Under the stated assumptions and the stability of the sequence $(\theta_t)_t$*,*

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

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

for any bounded measurable function f*.*

Main result: ergodicity and LLN for the samples $(X_t)_t$ (4/5)

Theorem

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

Under the stated assumptions and the stability of the sequence $(\theta_t)_t$,

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

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

for any bounded measurable function f.

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

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

$$\|\pi_\theta - \pi_{\theta'}\|_{\mathrm{TV}} \leq 2(d-1)\sum_{i=1}^{d}\left|1 - \frac{\theta(i)}{\theta'(i)}\right|$$

$$\sup_{x\in\mathbb{X}} \|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{\mathrm{TV}} \leq 4\sup_i \left|1 - \frac{\theta(i)}{\theta'(i)}\right| + 4\sup_i \left|1 - \frac{\theta'(i)}{\theta(i)}\right|$$

Main result: ergodicity and LLN for the weighted samples $(X_t)_t$ (5/5)

Theorem

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

Under the stated assumptions and the stability of the sequence $(\theta_t)_t$*,*

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

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

for any bounded measurable function f*.*

Outline

Introduction

- 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 convergence results on WL but

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

We now discuss:

- Comparison in terms of *how rapidly does the sampler escape from a metastable state*
- Explicit computation of exit times for a simple model, numerical study for a more complex one.

Central Limit Theorem on the weight sequence

Theorem

F., Jourdain, Kuhn, Lelièvre, Stoltz (2012) *Under the stated assumptions, when* $\gamma_t \sim \gamma_\star / n^\alpha$ *(*$1/2 < \alpha < 1$*)*

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

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

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

Similar result when $\gamma_t \sim \gamma_\star / t$.

Toy example (1/3)

Consider the target distribution on $\mathbb{X} = \{1,2,3\}$

$$\pi(1) = \pi(3) = \frac{1}{2+\epsilon} \qquad \pi(2) = \frac{\epsilon}{2+\epsilon}$$

The proposal distribution in WL (for the kernels P_θ) and in HM is

$$Q = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

Proposal kernel only allowing jumps to the closest strata.

We compute the time $T_{1\to3}$ to reach the state 3 starting from the state 1, for WL and a Hastings-Metropolis (HM) algorithm.

Toy example (2/3)

Here are the transition kernels for HM (top) and WL (bottom)

$$\overline{P} = \begin{bmatrix} 1-\frac{\epsilon}{3} & \frac{\epsilon}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{\epsilon}{3} & 1-\frac{\epsilon}{3} \end{bmatrix}$$

$$P_\theta = \begin{bmatrix} 1-\frac{1}{3}\left(\epsilon\frac{\theta(1)}{\theta(2)} \wedge 1\right) & \frac{1}{3}\left(\epsilon\frac{\theta(1)}{\theta(2)} \wedge 1\right) & 0 \\ \frac{1}{3}\left(\frac{1}{\epsilon}\frac{\theta(2)}{\theta(1)} \wedge 1\right) & 1-\frac{1}{3}\left(\frac{1}{\epsilon}\frac{\theta(2)}{\theta(1)} \wedge 1 + \frac{1}{\epsilon}\frac{\theta(2)}{\theta(3)} \wedge 1\right) & \frac{1}{3}\left(\frac{1}{\epsilon}\frac{\theta(2)}{\theta(3)} \wedge 1\right) \\ 0 & \frac{1}{3}\left(\frac{\epsilon}{1}\frac{\theta(3)}{\theta(2)} \wedge 1\right) & 1-\frac{1}{3}\left(\epsilon\frac{\theta(3)}{\theta(2)} \wedge 1\right) \end{bmatrix}$$

In WL, when the chain gets stuck (say) in state 1, $\theta_n(1)$ increases which penalizes the state 1 and favors moves to state 2.

Toy example (3/3)

Yes, the Wang-Landau is less metastable !

- For Hastings-Metropolis, $T_{1\to 3}$ scales like $6/\epsilon$:

$$\frac{\epsilon}{6}\mathbb{E}\left[T_{1\to 3}\right] \sim_{\epsilon\to 0} 1 \qquad\qquad \lim_{\epsilon\to 0}\mathbb{P}(\frac{\epsilon}{6}T_{1\to 3} > c) = \exp(-c)$$

- For Wang-Landau, with a stepsize sequence $\gamma_t = \gamma_\star/t^\alpha$
  - for some $\alpha \in (1/2,1)$

    there exists constants C_1,C_2 such that

$$\lim_{\epsilon\to 0}\mathbb{P}\left(\left|\ln\epsilon\right|^{-1/(1-\alpha)} T_{1\to 3} \in (C_1,C_2)\right) = 1$$

    and $T_{1\to 3}$ scales like $\left|\ln\epsilon\right|^{1/(1-\alpha)}$.

  - for $\alpha = 1$, $T_{1\to 3}$ scales like $\epsilon^{-1/(1+\gamma_\star)}$

A less simple example (1/7)

$$\pi(x_1,x_2) \propto \exp(-\beta\ \mathcal{H}(x_1,x_2))\mathbb{I}_{[-R,R]}(x_1) \qquad \text{on } [-R,R] \times \mathbb{R}^+$$

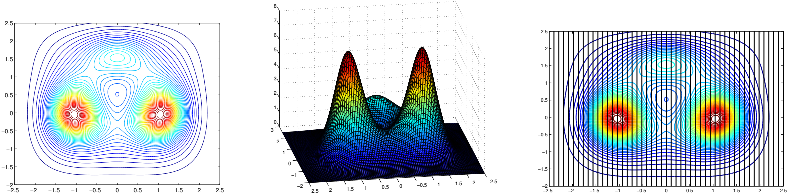


FIG.: [left] level curves of the potential $\mathcal{H}$ [center] Density (up to a normalizing constant) [right] Partition of the state space

In this numerical illustration: $R = 2.4$.
WL is run with $d = 48$; the proposal distribution is $\mathcal{N}(0,v^2 I)$ where $v = 2R/d$.
HM is a symmetric random walk with proposal distribution $\mathcal{N}(0,v^2 I)$ and target π.

A less simple example (2/7)

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

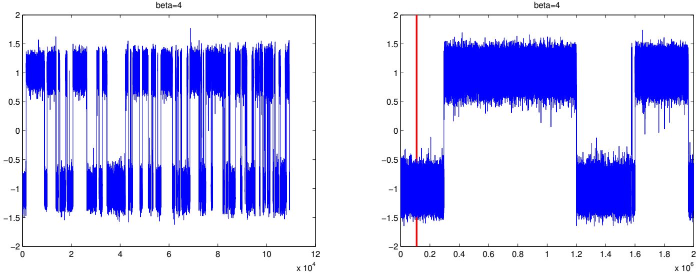


FIG.: [left] Wang Landau, $T = 110\,000$. [right] Hastings Metropolis, $T = 2\ 10^6$; the red line is at $x = 110\,000$

A less simple example (3/7)

- The larger β is, the larger the ratio is between the weight of the strata located near the main metastable states and the weight of the transition region (around $x_1 = 0$).
- The stepsize sequence is $\gamma_t = \gamma_\star / t^\alpha$.

since

- Initialisation of the samplers: $X_0 = (-1,0)$, $\theta_0 = (1/d, \cdots, 1/d)$.
- The algorithm are run until the first time t such that $X_t^1 > 1$.
- We repeat this experiment over M independent runs, and compute the mean value of the exit time ($M \sim 10^2$ to 10^5 depending upon the value of β).

We report the mean value of the exit times

t_β: Wang Landau

$\bar{t}_\beta$: Hastings-Metropolis

as a function of β, for different values of α.

A less simple example (4/7)

Plot of $\beta \mapsto \bar{t}_\beta$, the mean exit-time for HM (left) and $\beta \mapsto t_\beta$, the mean exit-time for WL (right).

When $\gamma_t = \gamma_\star / t$ ($\alpha = 1$).

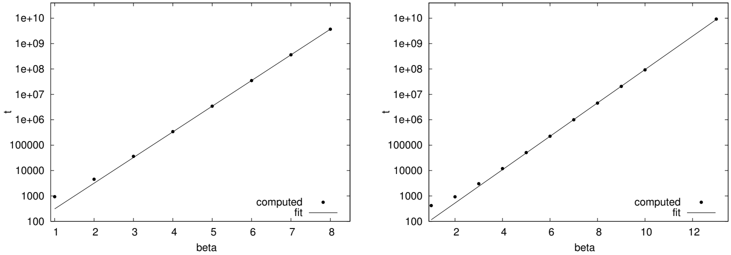


FIG.: When $\gamma_\star = 2$. [left] Hastings-Metropolis. [right] Wang-Landau. Note the logarithmic scale on the y-axis

We also observe (plots not displayed) that the shape depends on $\gamma_\star$.

A less simple example (5/7)

- We observe that

$$\bar{t}_\beta \sim C \exp(\beta \mu_0) \qquad\qquad t_\beta \sim C(\gamma_\star)\ \exp(\beta \mu_{\gamma_\star})$$

- Based on the results for the toy example, it is expected

$$t_\beta \sim C(\gamma_\star)\ \exp(\beta \frac{\mu_0}{1+\gamma_\star})$$

$\gamma_\star$	$\mu_{\gamma_\star}$	$\mu_{\gamma_\star}/\mu_0$	$1/(1+\gamma_\star)$
0	2.32	1	1
1	1.74	0.75	0.5
2	1.51	0.65	0.33
4	1.25	0.54	0.20
8	0.92	0.40	0.11

Comparison of the observed shape $\mu_{\gamma_\star}$ and the expected shape $\mu_0/(1+\gamma_\star)$ for different values of $\gamma_\star$.
Quite bad prediction !

A less simple example (6/7)

Plot of $\beta \mapsto t_\beta$, the mean exit-time for WL.

When $\gamma_t = 1/t^\alpha$ when $\alpha = 0.125$ (left) and $\alpha = 0.75$ (right).

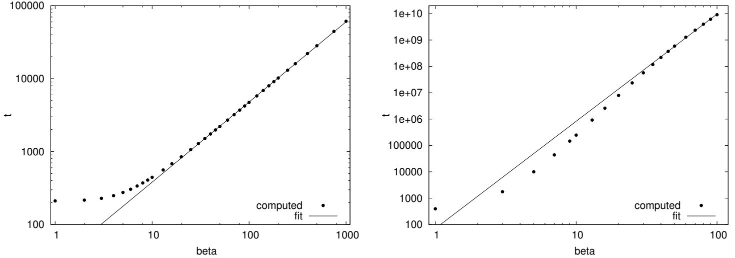


FIG.: [left] $\alpha = 0.125$. [right] $\alpha = 0.75$. Note the logarithmic scale on the y-axis

A less simple example (7/7)

- We observe that

$$t_\beta \sim C(\alpha) t^{\mu_\alpha}$$

- Based on the results for the toy example, it is expected

$$t_\beta \sim C(\alpha) t^{1/(1-\alpha)}$$

α	μ_α	$1/(1-\alpha)$
0.125	1.11	1.14
0.25	1.30	1.33
0.375	1.55	1.60
0.5	2.02	2.00
0.625	2.72	2.67
0.75	4.06	4.00

Comparison of the observed shape μ_α and the expected shape $1/(1-\alpha)$ for different values of α.
Far better prediction !

Outline

Conclusion

- Wang Landau: new methodologies ▸ Next
- Adaptive MCMC - Stochastic Approximation with controlled Markov chains. ▸ Next
- Multimodality, metastability - Molecular Dynamics, Statistical Physics. ▸ Next

Outline

Bibliography

The Wang Landau method in Statistical Physics previous

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 ▸ previous

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.

Convergence of adaptive MCMC previous

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 algorithms. Submittted (2013).

(free energy) Biasing techniques in Molecular Dynamics

previous

[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).