Convergence and Efficiency of the Wang Landau algorithm

Gersende FORT

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.

Convergence analysis of a Monte Carlo sampler to sample from

$$\pi(x)\ d\lambda(x) \qquad \text{on } \mathbb{X} \subseteq \mathbb{R}^p$$

when π is multimodal

Wang Landau : a biasing potential approach

- Instead of sampling from π, sample from $\pi_\star$

$$\pi_\star(x) \propto \pi(x) \exp(A_\star(x))$$

  where $A_\star$ is a biasing potential chosen such that $\pi_\star$ satisfies some efficiency criterion.

- Such a "perfect" $A_\star$ is unknown: it has to be estimated on the fly, when running the sampler.

- To obtain samples approximating π, use an *importance sampling* strategy.

Wang Landau : definition of $\pi_\star$?

$$\pi_\star(x) \propto \pi(x)\ \exp(-A_\star(x))$$

- Choose a partition $\mathbb{X}_1, \cdots, \mathbb{X}_d$ of $\mathbb{X}$
- and choose $A_\star$ constant on $\mathbb{X}_i$

$$\pi_\star(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\mathbb{X}_i}(x)\ \pi(x)\ \exp(-A_\star(i))$$

- and such that under $\pi_\star$, each subset $\mathbb{X}_i$ has the same weight: $\pi_\star(\mathbb{X}_i) = 1/d$

$$\frac{1}{d} = \pi(\mathbb{X}_i)\ \exp(-A_\star(i))$$

Then,

$$\pi_\star(x) = \frac{1}{d} \sum_{i=1}^{d} \frac{\pi(x)}{\pi(\mathbb{X}_i)} \mathbb{1}_{\mathbb{X}_i}(x)$$

Wang Landau: an adaptive biasing potential algorithm

$\pi(\mathbb{X}_i)$ is unknown and we can not sample under $\pi_\star$.

- Define the family of biased densities, indexed by a weight vector $\theta = (\theta(1), \cdots ,\theta(d))$,

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

- The algorithm produces iteratively a sequence $((\theta_t,X_t))_t$ s.t.

(i) $X_t \sim \pi_{\theta_t}$
or, if not possible, $X_t \sim P_{\theta_t}(X_{t-1},\cdot)$ where $\pi_\theta P_\theta = \pi_\theta$.

(ii) $\lim_t \theta_t = (\pi(\mathbb{X}_1), \cdots ,\pi(\mathbb{X}_d))$

Wang Landau: Update rules for the bias θ_t

By definition, $\pi_\star(\mathbb{X}_i) = 1/d$. The update rules consist in penalizing the subsets $\mathbb{X}_i$ which are visited in order to force the sampler to spend the same time in each subset $\mathbb{X}_i$.

Since $\pi_\theta(\mathbb{X}_i) \propto \pi(\mathbb{X}_i)/\theta(i)$

$$\text{Rules: } \begin{cases} \text{if } X_{t+1} \in \mathbb{X}_i \qquad \theta_{t+1}(i) > \theta_t(i) \qquad \theta_{t+1}(k) < \theta_t(k),\ k \neq i \\ \lim_t \theta_t = (\pi(\mathbb{X}_1), \cdots, \pi(\mathbb{X}_d)) \end{cases}$$

Wang Landau: Update rules for the bias θ_t

By definition, $\pi_\star(\mathbb{X}_i) = 1/d$. The update rules consist in penalizing the subsets $\mathbb{X}_i$ which are visited in order to force the sampler to spend the same time in each subset $\mathbb{X}_i$.

Since $\pi_\theta(\mathbb{X}_i) \propto \pi(\mathbb{X}_i)/\theta(i)$

Rules: $\begin{cases} \text{if } X_{t+1} \in \mathbb{X}_i \qquad \theta_{t+1}(i) > \theta_t(i) \qquad \theta_{t+1}(k) < \theta_t(k),\ k \neq i \\ \lim_t \theta_t = (\pi(\mathbb{X}_1), \cdots, \pi(\mathbb{X}_d)) \end{cases}$

Ex. Strategy 1: Non-linear update with deterministic step size $(\gamma_t)_t$

$$\theta_{t+1}(i) = \theta_t(i) \frac{1+\gamma_{t+1}}{1+\gamma_{t+1}\theta_t(i)} \qquad \theta_{t+1}(k) = \theta_t(k) \frac{1}{1+\gamma_{t+1}\theta_t(i)}$$

Wang Landau: Update rules for the bias θ_t

By definition, $\pi_\star(\mathbb{X}_i) = 1/d$. The update rules consist in penalizing the subsets $\mathbb{X}_i$ which are visited in order to force the sampler to spend the same time in each subset $\mathbb{X}_i$.

Since $\pi_\theta(\mathbb{X}_i) \propto \pi(\mathbb{X}_i)/\theta(i)$

$$\text{Rules: } \begin{cases} \text{if } X_{t+1} \in \mathbb{X}_i \qquad \theta_{t+1}(i) > \theta_t(i) \qquad \theta_{t+1}(k) < \theta_t(k),\ k \neq i \\ \lim_t \theta_t = (\pi(\mathbb{X}_1), \cdots, \pi(\mathbb{X}_d)) \end{cases}$$

Ex. Strategy 1: Non-linear update with deterministic step size $(\gamma_t)_t$

$$\theta_{t+1}(i) = \theta_t(i) \frac{1+\gamma_{t+1}}{1+\gamma_{t+1}\theta_t(i)} \qquad \theta_{t+1}(k) = \theta_t(k) \frac{1}{1+\gamma_{t+1}\theta_t(i)}$$

Ex. Strategy 2: Linear update with deterministic step size $(\gamma_t)_t$

$$\theta_{t+1}(i) = \theta_t(i) + \gamma_{t+1}\theta_t(i)\left(1-\theta_t(i)\right)$$
$$\theta_{t+1}(k) = \theta_t(k) - \gamma_{t+1}\theta_t(i)\ \theta_t(k)$$

Herefater, in the talk

WL is an iterative algorithm: each iteration consists in

(i) sampling a point $X_{t+1} \sim P_{\theta_t}(X_t,\cdot)$ where $\pi_\theta P_\theta = \pi_\theta$

(ii) updating the biasing potential: $\theta_{t+1} = \Xi(\theta_t, X_{t+1}, t)$

We now prove that

1. $\lim_t \theta_t = (\pi(\mathbb{X}_1), \cdots, \pi(\mathbb{X}_d))$ a.s.
2. as $t \to \infty$, X_t "approximates" $\pi_\star$: for a large class of functions f

$$\lim_t \mathbb{E}[f(X_t)] = \pi_\star(f)$$

$$\lim_T T^{-1} \sum_{t=1}^{T} f(X_t) = \pi_\star(f) \text{ a.s.}$$

and we propose an adaptive importance sampling estimator of π.

Outline

In this section, the update of θ_t is one of the tow previous strategies

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

where $(\gamma_t)_t$ is a non increasing positive sequence chosen by the user controlling the adaption rate of the weight sequence $(\theta_t)_t$.

We address

1. the convergence
2. the rate of convergence

of the weight sequence $(\theta_t)_t$

WL as a Stochastic Approximation algorithm

WL is a stochastic approximation algorithm with Markov controlled dynamics

- it produces a sequence of weights $(\theta_t)_t$ defined by

$$\theta_{t+1} = \theta_t + \gamma_{t+1} \, H(\theta_t, X_{t+1}) + O\left(\gamma_{t+1}^2\right)$$

where

$$H_i(\theta, x) = \theta(i) \left(\mathbb{1}_{\mathbb{X}_i}(x) - \theta(I(x))\right) \qquad i \in \{1, \cdots, d\}$$

WL as a Stochastic Approximation algorithm

WL is a stochastic approximation algorithm with Markov controlled dynamics

- it produces a sequence of weights $(\theta_t)_t$ defined by

$$\theta_{t+1} = \theta_t + \gamma_{t+1}\, H(\theta_t, X_{t+1}) + O\left(\gamma_{t+1}^2\right)$$

where

$$H_i(\theta, x) = \theta(i)\left(\mathbb{1}_{\mathbb{X}_i}(x) - \theta(I(x))\right) \qquad i \in \{1, \cdots, d\}$$

- with dynamics $(X_t)_t$: controlled Markov chain

$$\mathbb{P}(X_{t+1} \in A | \text{past}_t) = P_{\theta_t}(X_t, A)$$

Note that the field $H(\theta, X_{t+1})$ is a (random) approximation of the *mean field*

$$h(\theta) = \int H(\theta, x)\, \pi_\theta(x)\, \lambda(dx).$$

Almost-sure convergence of the WL weight sequence

Theorem (F., Jourdain, Kuhn, Lelièvre, Stoltz (2014-a))

Assume

1. *The target distribution* $\pi \, d\lambda$ *satisfies* $0 < \inf_{\mathbb{X}} \pi \leq \sup_{\mathbb{X}} \pi < \infty$ *and* $\inf_i \pi(\mathbb{X}_i) > 0$.
2. *For any* θ, P_θ *is a Hastings-Metropolis kernel with invariant distribution*

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

and proposal distribution $q(x,y)d\lambda(y)$ *such that* $\inf_{\mathbb{X}^2} q > 0$.

3. *The step-size sequence is non-increasing, positive,*

$$\sum_t \gamma_t = \infty \qquad \sum_t \gamma_t^2 < \infty$$

Then

$$\lim_t \theta_t = (\pi(\mathbb{X}_1), \cdots, \pi(\mathbb{X}_d)) \qquad \text{almost-surely}$$

Sketch of the proof (1/2)

$$\theta_{t+1} = \theta_t + \gamma_{t+1}\, H(\theta_t, X_{t+1}) + \gamma_{t+1}^2 O(1)$$

(1.) Rewrite the update rule as a perturbation of a discretized O.D.E. $\dot{u} = h(u)$

$$u_{t+1} = u_t + \gamma_{t+1} h(u_t) + \gamma_{t+1}\xi_{t+1}$$

In our case

$$h(\theta) = \left(\sum_{j=1}^{d} \frac{\theta(j)}{\pi(\mathbb{X}_j)} \right)^{-1} \left(\begin{bmatrix} \pi(\mathbb{X}_1) \\ \cdots \\ \pi(\mathbb{X}_d) \end{bmatrix} - \theta \right)$$

(2.) Show that the ODE $\dot{u} = h(u)$ converges to the set

$$\mathcal{L} = \{\theta : h(\theta) = 0\} = \{(\pi(\mathbb{X}_1), \cdots, \pi(\mathbb{X}_d))\}$$

(3.) Show that the noisy discretization $(u_t)_t$ inherits the same limiting behavior and converges to $\mathcal{L}$.

Sketch of the proof (2/2)

The last step is the most technical
$(3a.)$ The noisy discretization has to visit infinitely often an attractive neighborhood of the limiting set $\mathcal{L}$

$(3b.)$ The noise ξ_t has to be small (at least when t is large)

$$\xi_{t+1} = H(\theta_t, X_{t+1}) - h(\theta_t) + \gamma_{t+1} O(1)$$

and this holds true since we have

– Uniform geometric ergodicity: There exists $\rho \in (0,1)$ s.t.

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

– Regularity-in-θ of π_θ and P_θ: There exists C such that for any $\theta, \theta' \in \Theta$ and any $x \in \mathbb{X}$

$$\|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{\mathrm{TV}} + \|\pi_\theta \, d\lambda - \pi_{\theta'} \, d\lambda\|_{\mathrm{TV}} \leq C \sum_{i=1}^{d} \left| 1 - \frac{\theta'(i)}{\theta(i)} \right|$$

Rate of convergence (1/2)

Theorem (F., Jourdain, Kuhn, Lelièvre, Stoltz (2014-a))

Assume

1. *(the same assumptions as for the convergence result)*
2. *one of the following conditions*
   - (i) $\gamma_t \sim \gamma_0 / t^a$ *for some* $a \in (1/2,1)$
   - (ii) $\gamma_t \sim \gamma_\star / t$ *with* $\gamma_\star > d/2$.

Then when $t \to \infty$

$$\frac{1}{\sqrt{\gamma_t}} \left(\theta_t - \begin{bmatrix} \pi(\mathbb{X}_1) \\ \cdots \\ \pi(\mathbb{X}_d) \end{bmatrix} \right) \xrightarrow{w} \mathcal{N}_d \left(0, \sigma^2 \, U_\star \right)$$

where

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

and

$$\sigma^2 = \begin{cases} d/2 & \text{in case (i)} \\ \gamma_\star d / (2\gamma_\star - d) & \text{in case (ii)} \end{cases}$$

Rate of convergence (2/2)

- The limiting variance is the same as in a Stochastic Approximation algorithm with dynamics $(X_t)_t$ sampled from a Markov chain with invariant distribution $\pi_\star$

Rate of convergence (2/2)

- The limiting variance is the same as in a Stochastic Approximation algorithm with dynamics $(X_t)_t$ sampled from a Markov chain with invariant distribution $\pi_\star$
- What is the optimal rate of convergence?

$$\hookrightarrow \text{answer:} \qquad \gamma_t = \frac{\gamma_\star}{t} \quad \text{which yields a rate } O(\sqrt{t})$$

Rate of convergence (2/2)

- The limiting variance is the same as in a Stochastic Approximation algorithm with dynamics $(X_t)_t$ sampled from a Markov chain with invariant distribution $\pi_\star$
- What is the optimal rate of convergence?

$$\hookrightarrow \text{answer:} \qquad \gamma_t = \frac{\gamma_\star}{t} \quad \text{which yields a rate } O(\sqrt{t})$$

- When $\gamma_t = \gamma_\star / t$, the limiting variance is $d\gamma_\star^2(2\gamma_\star - d)\, U_\star$ so: is there an optimal $\gamma_\star$?

$\hookrightarrow$ answer: optimal with $\gamma_\star = d$ and this yields the variance $d^2\, U_\star$

Rate of convergence (2/2)

- The limiting variance is the same as in a Stochastic Approximation algorithm with dynamics $(X_t)_t$ sampled from a Markov chain with invariant distribution $\pi_\star$
- What is the optimal rate of convergence?

$\hookrightarrow$ answer: $\gamma_t = \dfrac{\gamma_\star}{t}$ which yields a rate $O(\sqrt{t})$

- When $\gamma_t = \gamma_\star / t$, the limiting variance is $d\gamma_\star^2 (2\gamma_\star - d)\, U_\star$ so: is there an optimal $\gamma_\star$?

$\hookrightarrow$ answer: optimal with $\gamma_\star = d$ and this yields the variance $d^2\, U_\star$

- In practice: choose $\gamma_t = \gamma_\star / t^\alpha$ with α close to $1/2$ (but larger) and consider an averaging technique:

$$\pi(\mathbb{X}_i) \approx \frac{1}{T} \sum_{t=1}^{T} \theta_t(i)$$

We will have the optimal rate of convergence.

Outline

In this section, the update of θ_t is one of the tow previous strategies

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

where $(\gamma_t)_t$ is a decreasing positive sequence chosen by the user.

We address

1. the convergence of $(X_t)_t$ to $\pi_\star$ in some sense.
2. how to approximate π with the points $(X_t)_t$.

WL as a sampler

WL is an adaptive MCMC sampler

- it produces points $(X_t)_t$:

$$\mathbb{P}\left(X_{t+1} \in A | \text{past}_t\right) = P_{\theta_t}(X_t, A)$$

- and at the same time, updates the adaption parameter

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1}) + O(\gamma_{t+1}^2)$$

Here, each kernel P_θ has its own invariant distribution π_θ
BUT we know that $(\theta_t)_t$ converges and $\pi_{\lim_t \theta_t} = \pi_\star$.

Ergodicity and Law of large numbers

Theorem (F., Jourdain, Kuhn, Lelièvre, Stoltz (2014-a))

Assume

1. *(the same assumptions as those for the convergence of* $(\theta_t)_t$*)*

Then for any bounded measurable function f

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

$$\lim_T \frac{1}{T}\sum_{t=1}^{T} f(X_t) = \int f(x)\ \pi_\star(x)\ d\lambda(x) \text{ \textit{almost-surely}}$$

Sketch of proof

(1.) The containment condition:
There exist $\rho \in (0,1)$ and C such that

$$\sup_x \sup_\theta \|P_\theta^t(x,\cdot) - \pi_\theta\|_{\mathrm{TV}} \leq C\,\rho^t$$

(2.) The diminishing adaption condition:
There exists C such that for any θ,θ'

$$\sup_x \|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{\mathrm{TV}} \leq C \sum_{i=1}^{d} \left|1 - \frac{\theta(i)}{\theta'(i)}\right|$$

The update of the parameter satisfies: there exists C' such that $\forall t$

$$\|\theta_{t+1} - \theta_t\| \leq C'\,\gamma_{t+1}$$

Approximation of π (1/2)

By definition of $\pi_\star$, on the set $\mathbb{X}_i$: $\pi_\star(x) = \frac{1}{d} \frac{\pi(x)}{\pi(\mathbb{X}_i)}$

Then

$$\int f \; \pi d\lambda = \sum_{i=1}^{d} \int_{\mathbb{X}_i} f \; \pi d\lambda$$
$$= d \sum_{i=1}^{d} \pi(\mathbb{X}_i) \underbrace{\int_{\mathbb{X}_i} f \; \pi_\star d\lambda}_{\substack{\text{approximated by a Monte Carlo sum} \\ \frac{1}{T}\sum_{t=1}^{T} f(X_t) \mathbb{1}_{\mathbb{X}_i}(X_t)}}$$
$$\approx \frac{d}{T} \sum_{t=1}^{T} f(X_t) \sum_{i=1}^{d} \underbrace{\pi(\mathbb{X}_i)}_{\text{approximated by } \theta_t(i)} \mathbb{1}_{\mathbb{X}_i}(X_t)$$

so that

$$\int f \; \pi \, d\lambda \approx \frac{d}{T} \sum_{t=1}^{T} f(X_t) \sum_{i=1}^{d} \theta_t(i) \mathbb{1}_{\mathbb{X}_i}(X_t)$$

Approximation of π (2/2)

Theorem (F., Jourdain, Kuhn, Lelièvre, Stoltz (2014-a))

Assume

1. *(the same assumptions as those for the convergence of $(\theta_t)_t$)*

Then, for any bounded measurable function f

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

$$\lim_T \frac{d}{T}\sum_{t=1}^{T} f(X_t)\left(\sum_{i=1}^{d}\theta_t(i)\mathbb{1}_{\mathbb{X}_i}(X_t)\right) = \int f\ \pi\,d\lambda \qquad \textit{almost-surely}$$

Outline

In this section :

runs are with the *non-linearized Wang-Landau algorithm with deterministic step sizes*

Algorithm: Given (θ_t, X_t)

1. Draw a new sample : $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$
2. Update the weights : if $X_{t+1} \in \mathbb{X}_i$,

$$\theta_{t+1}(i) = \theta_t(i) \; \frac{1 + \gamma_{t+1}}{1 + \gamma_{t+1}\theta_t(i)}$$

$$\theta_{t+1}(k) = \theta_t(k) \; \frac{1}{1 + \gamma_{t+1}\theta_t(i)} \qquad k \neq i$$

A toy example (1/2)

- State space: $\mathbb{X} = \{1,2,3\}$
- Target distribution: $\pi(1) \propto 1 \qquad \pi(2) \propto \epsilon \qquad \pi(3) \propto 1$

Let us compare

1. **Hastings-Metropolis** P with proposal kernel Q and target π

$$Q = \begin{bmatrix} 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{bmatrix} \qquad P = \begin{bmatrix} 1-\epsilon/3 & \epsilon/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & \epsilon/3 & 1-\epsilon/3 \end{bmatrix}$$

2. **Wang-Landau** P_θ with proposal kernel Q and target π_θ

$$\pi_\theta(i) \propto \frac{\pi(i)}{\theta(i)} \qquad P_\theta = \begin{bmatrix} 1 - \frac{1}{3}\left(\epsilon\frac{\theta(1)}{\theta(2)} \wedge 1\right) & \cdots & 0 \\ \frac{1}{3}\left(\frac{1}{\epsilon}\frac{\theta(2)}{\theta(1)} \wedge 1\right) & \cdots & \frac{1}{3}\left(\frac{1}{\epsilon}\frac{\theta(2)}{\theta(3)} \wedge 1\right) \\ 0 & \cdots & 1 - \frac{1}{3}\left(\epsilon\frac{\theta(3)}{\theta(2)} \wedge 1\right) \end{bmatrix}$$

A toy example (2/2)

Comparison based on the hitting time

$T_{1\to 3}$: hitting-time of state 3, given the chain started from state 1

when $\epsilon \to 0$.

Proposition (F., Jourdain, Kuhn, Lelièvre, Stoltz (2014-b))

When $\epsilon \to 0$

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

$$\lim_{\epsilon \to 0} \frac{\epsilon}{6} \mathbb{E}\left[T_{1\to 3}\right] = 1$$
$$\frac{\epsilon}{6} T_{1\to 3} \to \mathcal{E}(1) \text{ in distribution}$$

- *For Wang-Landau applied with* $\gamma_t = \gamma_\star / t^a$*:* $T_{1\to 3}$ scales like

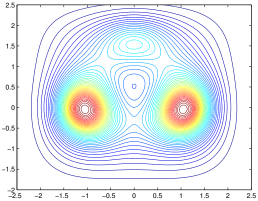
$$\begin{array}{ll} C(a,\gamma_\star)\ |\ln \epsilon|^{1/(1-a)} & \textit{when } 1/2 < a < 1 \\ \epsilon^{-1/(1+\gamma_\star)} & \textit{when } a = 1 \end{array}$$

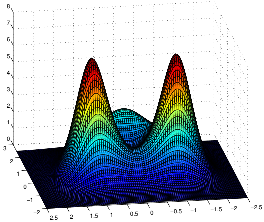
Second example on $\mathbb{R}^2$ (1/5)

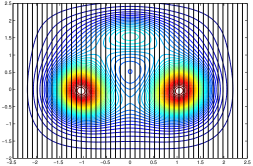
- $\mathbb{X} = [-R,R] \times \mathbb{R}$
- The target density: $\pi \propto \exp(-\beta\ V(x_1,x_2))$ with

$$V(x_1,x_2) = 3\exp\left(-x_1^2 - \left(x_2 - \frac{1}{3}\right)^2\right) - 3\exp\left(-x_1^2 - \left(x_2 - \frac{5}{3}\right)^2\right)$$
$$- 5\exp\left(-(x_1-1)^2 - x_2^2\right) - 5\exp\left(-(x_1+1)^2 - x_2^2\right) + 0.2x_1^4 + 0.2\left(x_2 - \frac{1}{3}\right)^4 .$$

- d strata: obtained by binning the x-axis







Two metastable points $x_- = (-1,0)$, $x+ = (1,0)$

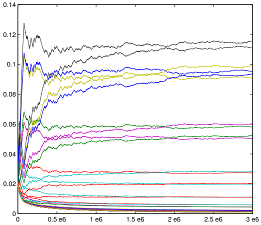
Second example on $\mathbb{R}^2$ (2/5)

$d = 48$ strata, binning along the x-axis.

P_θ are Hastings-Metropolis kernels with proposal distribution $\mathcal{N}(0,(2R/d)^2 I)$ and target π_θ. $R = 2.4$.

$X_0 = (-1,0)$.

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



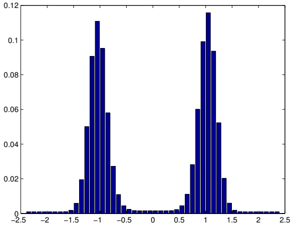


FIG.: [left] The sequences $(\theta_t(i))_t$. [right] The limiting value $\lim_t \theta_t(i)$

Second example on $\mathbb{R}^2$ (3/5)

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

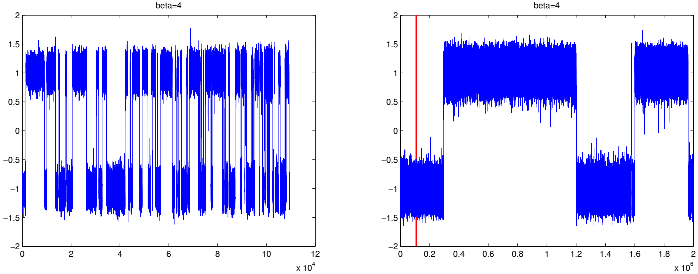


FIG.: [left] Wang Landau, $T = 110\,000$. [right] Hastings Metropolis, $T = 2\ 10^6$; the red line is at $x = 110\,000$

Second example on $\mathbb{R}^2$ (4/5)

For the Wang-Landau algorithm with kernel HM kernels with proposal Q_d

$$Q_d(x,dy) \equiv \mathcal{N}_2(x,v_d I)(y)$$

and target π_θ.

Compute T_β: the hitting-time of the statum containing $\{(x_1,x_2),x_1 > 1\}$, when the chain starts from $x_- = (-1,0)$.

- different (large) values of β are considered.
- the plots show the mean value of this hitting-time over M_β independent runs. M_β chosen such that the variability of T_β is less than few percents

Second example on $\mathbb{R}^2$ (5/5)

- It is expected based on Laplace methods for comparing the weights of strata that $\exp(-\beta\,\mu)$ plays the same role as ϵ in the previous example.
- Therefore, it is expected - and we observe - that T_β *scales as*

$$\begin{array}{ll} C(a,\gamma_\star)'\ \beta^{1/(1-a)} & \text{when } 1/2 < a < 1 \\ C\ \exp(\beta\,\mu/(1+\gamma_\star)) & \text{when } a = 1 \end{array}$$

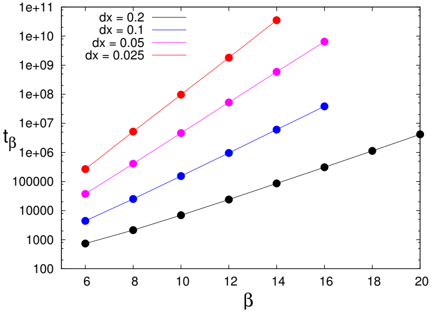


FIG.: $\log T_\beta$ when $\gamma_t = 8/t$. dx is the width of each stratum.

References

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

G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. *Convergence of the Wang-Landau algorithm* Accepted for publication in Mathematics of Computation, 2014. arXiv math.PR 1207.6880.

G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz. *Efficiency of the Wang-Landau algorithm* Accepted for publication in Applied Mathematics Research Express, 2014. arXiv math.NA 1310.6550

Convergence of Stochastic Approximation algorithms
C. Andrieu, E. Moulines and P. Priouret. *Stability of Stochastic Approximation under verifiable conditions*. SIAM J. Control Optim. 44(1):283–312, 2005.

CLT for Stochastic Approximation algorithms
G. Fort. *Central Limit Theorems for Stochastic Approximation algorithms*. Accepted for publication in ESAIM PS, 2014. arXiv math.PR 1309.3116

Ergodicity and Law of Large Numbers for Controlled Markov chains
G. Fort, E. Moulines and P. Priouret. *Convergence of adaptive and interacting Markov chains Monte Carlo algorithms*. Ann. Stat., 39(6):3262–3289, 2012.