Mathematical aspects of adaptive samplers: application to free energy calculation

Gersende FORT

CNRS
(LTCI)
Paris, France

Free Energy Calculations: A mathematical perspective
Oaxaca, July 2015

Introduction

Goal:

Explore the support of a distribution π

and/or compute integrals w.r.t. π

$$\int_{\{x \in \mathbb{R}^n : \xi(x) \in \mathcal{O}\}} \mathrm{d}\pi(x).$$

when

- π highly metastable
- and π is a distribution on $\mathbb{R}^n$, n large.

Monte Carlo: a stochastic approximation

- propose sample from a (proposal) distribution $\pi_\star$
- correct the samples in order to approximate π

1. Markov chain Monte Carlo: $\{X_1, X_2, \cdots\}$ Markov chain with stationary distribution π

$$\int \phi(x) \mathrm{d}\pi(x) \approx \frac{1}{N} \sum_{k=1}^{N} \phi(X_k)$$

   its construction depends on a proposal distribution $\pi_\star$

2. Importance Sampling: $\{X_1, X_2, \cdots\}$ from a proposal distribution $\pi_\star$

$$\int \phi(x) \mathrm{d}\pi(x) = \int \phi(x) \frac{\mathrm{d}\pi(x)}{\mathrm{d}\pi_\star(x)} \mathrm{d}\pi_\star(x) \approx \frac{1}{N} \sum_{k=1}^{N} \phi(X_k) \frac{\mathrm{d}\pi(X_k)}{\mathrm{d}\pi_\star(X_k)}$$

Monte Carlo: a stochastic approximation

- propose sample from a (proposal) distribution $\pi_\star$
- correct the samples in order to approximate π

1. Markov chain Monte Carlo: $\{X_1, X_2, \cdots\}$ Markov chain with stationary distribution π

$$\int \phi(x) \mathrm{d}\pi(x) \approx \frac{1}{N} \sum_{k=1}^{N} \phi(X_k)$$

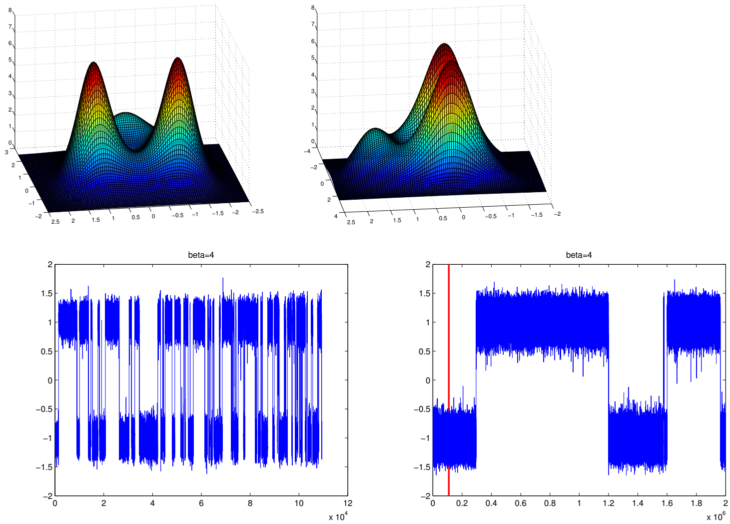
   its construction depends on a proposal distribution $\pi_\star$

2. Importance Sampling: $\{X_1, X_2, \cdots\}$ from a proposal distribution $\pi_\star$

$$\int \phi(x) \mathrm{d}\pi(x) = \int \phi(x) \frac{\mathrm{d}\pi(x)}{\mathrm{d}\pi_\star(x)} \mathrm{d}\pi_\star(x) \approx \frac{1}{N} \sum_{k=1}^{N} \phi(X_k) \frac{\mathrm{d}\pi(X_k)}{\mathrm{d}\pi_\star(X_k)}$$

$\hookrightarrow$ the efficiency of these samplers depends on the choice of $\pi_\star$

Naïve Monte Carlo samplers



Adaptive Monte Carlo

- choose a family of proposal distributions $\{\pi_\theta, \theta \in \Theta\}$
- at iteration $t+1$,

  based on the past behavior of the sampler: $X_1, \cdots, X_t$,

  (a) Sample X_{t+1} by using the proposal distribution π_{θ_t}

  (b) Update the parameter θ: $\theta_t \to \theta_{t+1}$

Example: Wang-Landau based-algorithms (Self Healing Umbrella Sampling, Well Tempered Metadynamics, $\cdots$) are Adaptive Importance Samplers

- Reaction Coordinate: $\xi : \mathbb{R}^n \to \{1, \cdots, d\}$
- The distributions π_θ:

$$\pi_\theta(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\{x \in \mathbb{R}^n : \xi(x) = i\}} \pi(x) \, \exp(-\ln(\theta(i)))$$

- $\theta = (\theta(1), \cdots, \theta(d)) \in (\mathbb{R}_+)^d$ such that $\sum_{i=1}^{d} \theta(i) = 1$

Example: Wang-Landau based-algorithms (Self Healing Umbrella Sampling, Well Tempered Metadynamics, $\cdots$) are Adaptive Importance Samplers

- Reaction Coordinate: $\xi : \mathbb{R}^n \to \{1, \cdots, d\}$
- The distributions π_θ:

$$\pi_\theta(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\{x \in \mathbb{R}^n : \xi(x) = i\}} \pi(x) \exp(-\ln(\theta(i)))$$

- $\theta = (\theta(1), \cdots, \theta(d)) \in (\mathbb{R}_+)^d$ such that $\sum_{i=1}^d \theta(i) = 1$

Algorithm: repeat

Sample Draw X_{t+1} "from" π_{θ_t}

Update Update the parameter:

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t + \mathsf{C}(\theta_t, X_{t+1}, t)$$

and set

$$\theta_{t+1} = \frac{\tilde{\theta}_{t+1}}{\sum_{\ell=1}^d \tilde{\theta}_{t+1}(\ell)}$$

Example: Wang-Landau based-algorithms (Self Healing Umbrella Sampling, Well Tempered Metadynamics, $\cdots$) are Adaptive Importance Samplers

- Reaction Coordinate: $\xi : \mathbb{R}^n \to \{1, \cdots, d\}$
- The distributions π_θ:

$$\pi_\theta(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\{x \in \mathbb{R}^n : \xi(x) = i\}} \pi(x) \exp(-\ln(\theta(i)))$$

- $\theta = (\theta(1), \cdots, \theta(d)) \in (\mathbb{R}_+)^d$ such that $\sum_{i=1}^d \theta(i) = 1$

Algorithm: repeat

Sample Draw X_{t+1} "from" π_{θ_t}

Update Update the parameter:

$$\tilde{\theta}_{t+1}(i) = \tilde{\theta}_t(i) \left(1 + \gamma_{t+1} \mathbb{1}_{\{x : \xi(x) = i\}}(X_{t+1})\right)$$

and set

$$\theta_{t+1} = \frac{\tilde{\theta}_{t+1}}{\sum_{\ell=1}^d \tilde{\theta}_{t+1}(\ell)}$$

Adaptation for $\cdots$

- Optimal proposal mecanism $\pi_{\theta_\star}$ where

$$\theta_\star \qquad \text{solves} \qquad h(\theta_\star) = 0$$

- No explicit solution: define an iterative mecanism $\theta_1, \theta_2, \cdots$, such that (hopefully) $\lim_t \theta_t = \theta_\star$.

Example: Wang Landau based algorithms

$$\pi_\theta(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\{x\in\mathbb{R}^n:\xi(x)=i\}}\pi(x)\ \exp(-\ln(\theta(i)))$$

Optimality criterion: $\theta_\star$ is such that

$$\forall \ell, \qquad \int_{\{x\in\mathbb{R}^n:\xi(x)=\ell\}} \pi_{\theta_\star}(x)\mathrm{d}x = \frac{1}{d}$$

- we look for the $(-\ln)$ free energy: $\theta_\star = (\theta_\star(1),\cdots,\theta_\star(d))$

$$\theta_\star(\ell) = \int_{\{x\in\mathbb{R}^n:\xi(x)=\ell\}} \pi(x)\mathrm{d}x$$

Example: Wang Landau based algorithms

$$\pi_\theta(x) \propto \sum_{i=1}^{d} \mathbb{1}_{\{x\in\mathbb{R}^n:\xi(x)=i\}}\,\pi(x)\,\exp(-\ln(\theta(i)))$$

Optimality criterion: $\theta_\star$ is such that

$$\forall \ell, \qquad \int_{\{x\in\mathbb{R}^n:\xi(x)=\ell\}} \pi_{\theta_\star}(x)\mathrm{d}x = \frac{1}{d}$$

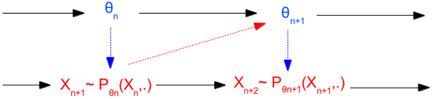
- we look for the $(-\ln)$ free energy: $\theta_\star = (\theta_\star(1),\cdots,\theta_\star(d))$

$$\theta_\star(\ell) = \int_{\{x\in\mathbb{R}^n:\xi(x)=\ell\}} \pi(x)\mathrm{d}x$$

- which is also defined as the root of

$$h(\theta) = \left(\sum_{i=1}^{d} \frac{\theta_\star(i)}{\theta(i)}\right)^{-1} (\theta_\star - \theta) = \int H_\theta(x)\pi_\theta(x)\mathrm{d}x$$

In this talk



Simultaneously

1. How to obtain draws $\{X_n, n \geq 0\}$ approximating $\pi_{\theta_\star}$?

$$\frac{1}{N}\sum_{k=1}^{N}\phi(X_k) \to \int \phi(x)\pi_{\theta_\star}(x)\mathrm{d}x$$

2. How to obtain a converging sequence $\{\theta_n, n \geq 0\}$ with a limit $\theta_\star$ solving

$$h(\theta) = 0$$

when only a Monte Carlo approximation of h is available?

Not in this talk

- We produced $\{X_1, X_2, \cdots\}$ such that

$$\frac{1}{N}\sum_{t=1}^{N}\phi(X_t) \to \int \phi(x)\pi_{\theta_\star}(x)\mathrm{d}x$$

- The samples can be corrected to approximate π

$$\frac{d}{N}\sum_{t=1}^{N}\phi(X_t)\ \theta_t\left(\xi(X_t)\right) \to \int \phi(x)\pi(x)\mathrm{d}x$$

Outline

1 Introduction

2 Adaptive Monte Carlo samplers
- Controlled Markov chains
- Sufficient conditions for the cvg in distribution
- In the literature

3 Stochastic Approximation Algorithms

4 Conclusion

Controlled Markov chains (1/2)

$P_\theta, \theta \in \Theta$: family of Markov kernels.
π_θ invariant distribution of P_θ.

- The draws $(X_t)_t$ are from a controlled Markov chain

$$X_{t+1} | \mathrm{past}_t \sim P_{\theta_t}(X_t, \cdot)$$

- Question: Even in the case $\pi_\theta = \pi$ for all θ: does $(X_t)_t$ converge (say in distribution) to π?

Controlled Markov chains (2/2)

- Answer: No.
- For example:

$$X_{t+1} \sim \begin{cases} P_0(X_t,\cdot) & \text{if } X_t = 0 \\ P_1(X_t,\cdot) & \text{if } X_t = 1 \end{cases}$$

where

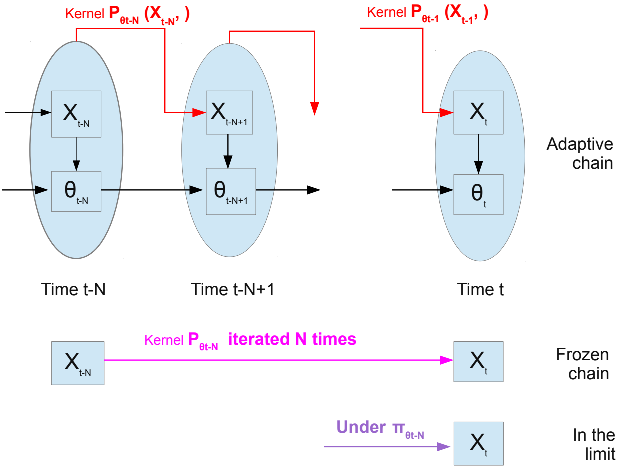
$$P_\ell = \begin{pmatrix} 1-t_\ell & t_\ell \\ t_\ell & 1-t_\ell \end{pmatrix}.$$

We have $\pi P_\ell = \pi$ with $\pi \propto (1,1)$.

The transition matrix of $(X_t)_t$ is

$$\tilde{P} = \begin{pmatrix} 1-t_0 & t_0 \\ t_1 & 1-t_1 \end{pmatrix} \qquad \text{with invariant distribution } \tilde{\pi} \propto (t_1,t_0)$$

Sufficient conditions for the cvg in distribution



Sufficient conditions for the cvg in distribution

Compare the two mecanisms

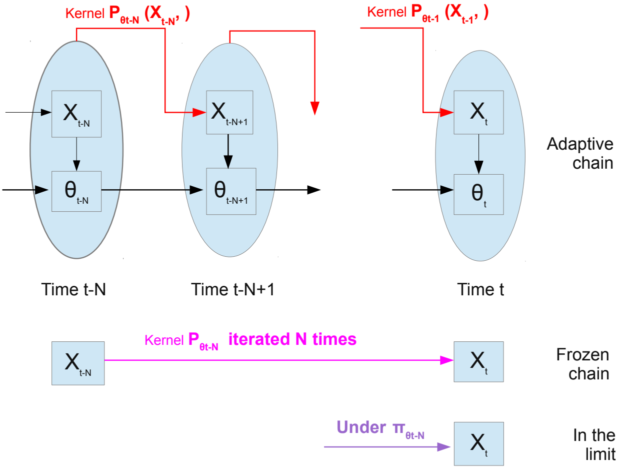
$$P_{\theta_{t-N}} \to P_{\theta_{t-N+1}} \to \cdots \to P_{\theta_{t-1}}$$
$$P_{\theta_{t-N}} \to P_{\theta_{t-N}} \to \cdots \to P_{\theta_{t-N}}$$

which is small as soon as $P_{\theta_{j+1}}$ and P_{θ_j} are close

- Diminishing adaption condition Roughly speaking:

$$\mathrm{dist}(P_\theta,P_{\theta'}) \leq \mathrm{dist}(\theta,\theta') \qquad \text{and} \qquad \lim_t (\theta_{t+1} - \theta_t) = 0$$

Sufficient conditions for the cvg in distribution

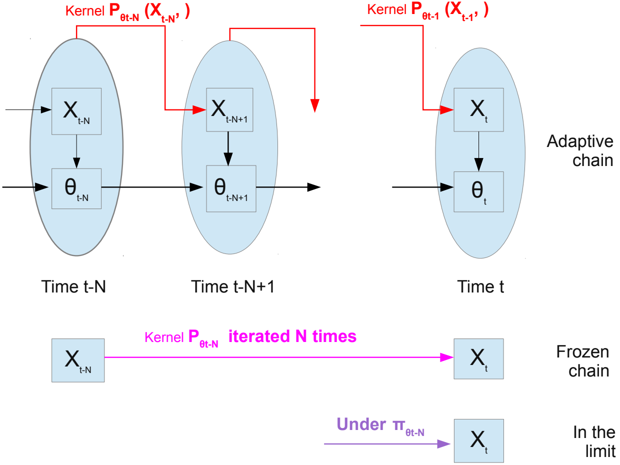


- Containment condition Roughly speaking:

$$\lim_{N \to \infty} \mathrm{dist}(P_\theta^N, \pi_\theta) = 0$$

at some rate depending smoothly on θ.

Sufficient conditions for the cvg in distribution



- Regularity in θ of π_θ so that

$$\lim_t \theta_t = \theta_\star \implies \mathrm{dist}\left(\pi_{\theta_t} - \pi_{\theta_\star}\right) \to 0$$

Sufficient conditions for the cvg in distribution

Theorem (F., Moulines, Priouret (2012))

Assume

A. *(Containment condition)*

- $\exists \pi_\theta$ *s.t.* $\pi_\theta P_\theta = \pi_\theta$
- *for any* $\epsilon > 0$*, there exists a non-decreasing positive sequence* $\{r_\epsilon(n), n \geq 0\}$ *such that* $\limsup_{n\to\infty} r_\epsilon(n)/n = 0$ *and*

$$\limsup_{n\to\infty} \mathbb{E}\left[\| P^{r_\epsilon(n)}_{\theta_{n-r_\epsilon(n)}}(X_{n-r_\epsilon(n)}, \cdot) - \pi_{\theta_{n-r_\epsilon(n)}} \|_{\mathrm{tv}} \right] \leq \epsilon$$

B. *(Diminishing adaptation)* *For any* $\epsilon > 0$,

$$\lim_{n\to\infty} \sum_{j=0}^{r_\epsilon(n)-1} \mathbb{E}\left[\sup_x \| P_{\theta_{n-r_\epsilon(n)+j}}(x, \cdot) - P_{\theta_{n-r_\epsilon(n)}}(x, \cdot) \|_{\mathrm{tv}} \right] = 0$$

C. *(Convergence of the invariant distributions)* $(\pi_{\theta_t})_t$ *converges weakly to* $\pi_{\theta_\star}$ *almost-surely.*

Then for any bounded and continuous function f

$$\lim_n \mathbb{E}\left[g(X_n) \right] = \int g(x)\, \pi_{\theta_\star}(x) \mathrm{d}x$$

In the literature

Sufficient conditions for

- Convergence in distribution of $(X_t)_t$
- Strong law of large numbers for $(X_t)_t$
- Central Limit Theorem for $(X_t)_t$

▶ Roberts, Rosenthal. Coupling and Ergodicity of Adaptive Markov chain Monte Carlo algorithms. J. Appl. Prob. (2007)
▶ F., Moulines, Priouret. Convergence of adaptive MCMC algorithms: ergodicity and law of large numbers. *Ann. Statist.* (2012)
▶ F., Moulines, Priouret and Vandekerkhove. A Central Limit Theorem for Adaptive and Interacting Markov Chain. *Bernoulli* (2013).

Conditions successfully applied to establish the convergence of Adaptive Hastings-Metropolis, Wang-Landau, SHUS, Well-tempered, $\cdots$

▶ F., Jourdain, Kuhn, Lelièvre and Stoltz. Convergence of the Wang-Landau algorithm. *Mathematics of Computation (2014)*

▶ F., Jourdain, Lelièvre and Stoltz. Self-Healing Umbrella Sampling: convergence and efficiency. *arXiv math.PR 1410.2109* (2014)

Outline

1 Introduction

2 Adaptive Monte Carlo samplers

3 Stochastic Approximation Algorithms
- Introduction
- Stability
- Recurrence of any neighborhood of $\mathcal{L}$
- Convergence of $\{\theta_t, t \geq 0\}$
- Conclusion

4 Conclusion

Stochastic Approximation (SA) Algorithm

Find the roots of: $$h(\theta) = \int H_\theta(x)\ \pi_\theta(\mathrm{d}x)$$

- Natural idea: NOT possible here since h is not explicit

$$\theta_{t+1} = \theta_t + \gamma_{t+1}\ h(\theta_t)$$

- SA algorithm

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

where

- $\{\gamma_t, t \geq 0\}$ sequence of positive stepsizes
- $X_{t+1} \sim \pi_{\theta_t}$ OR $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$ where P_θ is a Markov transition kernel with inv. dist. π_θ.

Example: Wang Landau is a Stochastic Approximation Algorithm

$$
\begin{aligned}
\tilde{\theta}_{t+1}(i) &= \tilde{\theta}_t(i) \ \left\{1 + \gamma_{t+1} \mathbb{1}_{\{\xi(x)=i\}}(X_{t+1})\right\} \\
&= \tilde{\theta}_t(i) + \gamma_{t+1} \tilde{\theta}_t(i) \mathbb{1}_{\{\xi(x)=i\}}(X_{t+1})
\end{aligned}
$$

Example: Wang Landau is a Stochastic Approximation Algorithm

$$\begin{aligned}
\tilde{\theta}_{t+1}(i) &= \tilde{\theta}_t(i) \ \left\{1 + \gamma_{t+1} \mathbb{1}_{\{\xi(x)=i\}}(X_{t+1})\right\} \\
&= \tilde{\theta}_t(i) + \gamma_{t+1} \tilde{\theta}_t(i) \mathbb{1}_{\{\xi(x)=i\}}(X_{t+1}) \\
\sum_{i=1}^{d} \tilde{\theta}_{t+1}(i) &= \sum_{i=1}^{d} \tilde{\theta}_t(i) + \gamma_{t+1} \ \sum_{i=1}^{d} \tilde{\theta}_t(i) \, \mathbb{1}_{\{\xi(x)=i\}}(X_{t+1}) \\
&= \sum_{i=1}^{d} \tilde{\theta}_t(i) + \gamma_{t+1} \ \tilde{\theta}_t \left(\xi(X_{t+1})\right) \\
&= \left(\sum_{i=1}^{d} \tilde{\theta}_t(i)\right) \ \left\{1 + \gamma_{t+1} \ \theta_t \left(\xi(X_{t+1})\right)\right\}
\end{aligned}$$

Example: Wang Landau is a Stochastic Approximation Algorithm

$$\begin{aligned}
\tilde\theta_{t+1}(i) &= \tilde\theta_t(i)\ \left\{1+\gamma_{t+1}\mathbb{1}_{\{\xi(x)=i\}}(X_{t+1})\right\} \\
&= \tilde\theta_t(i) + \gamma_{t+1}\tilde\theta_t(i)\mathbb{1}_{\{\xi(x)=i\}}(X_{t+1}) \\
\sum_{i=1}^d \tilde\theta_{t+1}(i) &= \sum_{i=1}^d \tilde\theta_t(i) + \gamma_{t+1}\ \sum_{i=1}^d \tilde\theta_t(i)\ \mathbb{1}_{\{\xi(x)=i\}}(X_{t+1}) \\
&= \sum_{i=1}^d \tilde\theta_t(i) + \gamma_{t+1}\ \tilde\theta_t\left(\xi(X_{t+1})\right) \\
&= \left(\sum_{i=1}^d \tilde\theta_t(i)\right)\ \left\{1+\gamma_{t+1}\ \theta_t\left(\xi(X_{t+1})\right)\right\}
\end{aligned}$$

$$\begin{aligned}
\theta_{t+1}(i) = \frac{\tilde\theta_{t+1}(i)}{\sum_{\ell=1}^d \tilde\theta_{t+1}(\ell)} &= \theta_t(i)\ \frac{1+\gamma_{t+1}\mathbb{1}_{\{\xi(x)=i\}}(X_{t+1})}{1+\gamma_{t+1}\,\theta_t\left(\xi(X_{t+1})\right)} \\
&= \theta_t(i) + \gamma_{t+1}\, H_{\theta_t}(X_{t+1}) + O(\gamma_{t+1}^2)
\end{aligned}$$

Convergence of SA algorithms to the limit set $\mathcal{L}$

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

Prove successively

- Stability: the sequence $\{\theta_t, t \geq 0\}$ is in a compact set of Θ
- Attractive limiting set $\mathcal{L}$:

$$\liminf_t \operatorname{dist}(\theta_t, \mathcal{L}) = 0.$$

- Convergence :

$$\lim_t \operatorname{dist}(\theta_t, \mathcal{L}) = 0.$$

Stability of the SA algorithm: $\theta_{t+1} = \theta_t + \gamma_{t+1} H_{\theta_t}(X_{t+1})$

Theorem

Assume $h : \Theta \to \mathbb{R}^d$ *is continuous and*

(i) **Lyapunov function:** $V : \Theta \to \mathbb{R}^+$ *is* C^1 *and*

- *the level sets* $\{\theta \in \Theta : V(\theta) \leq M\}$ *are compact subsets of* Θ
- $\nabla V(\theta) \cdot h(\theta) \leq 0$
- $\mathcal{L} = \{\theta \in \Theta : \nabla V(\theta) \cdot h(\theta) = 0\}$ *is in a compact level set* $\{V \leq M_0\}$

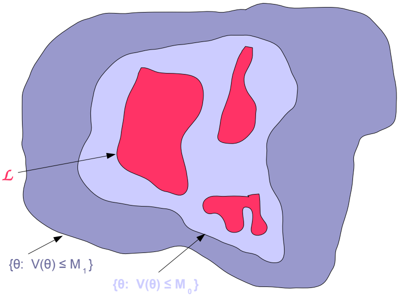
(ii) **Noise:** $H_{\theta_t}(X_{t+1}) = h(\theta_t) + \xi_{t+1}$ *and*

$$\lim_L \sum_{t=1}^{L} \gamma_t \xi_t \text{ exists.}$$

(iii) **Step-size sequence:** $\lim_t \gamma_t = 0$.

(iv) **Recurrence:** $\{\theta_t, t \geq 0\}$ *is i.o. in a compact subset of* $\Theta \subseteq \mathbb{R}^d$

Then: $\{\theta_t, t \geq 0\}$ *remains in a compact subset of* Θ.



Recurrence : θ_t infinitely often in $\{V \leq M\}$.

The Lyapunov property:

for any $\mathcal{K}$ compact s.t. $\mathcal{K} \cap \mathcal{L} = \emptyset$,

there exist δ,$\gamma_\star$,$\beta_\star > 0$ s.t.

$$[\gamma \leq \gamma_\star,\ |\xi| \leq \beta_\star,\ u \in \mathcal{K}] \Longrightarrow V\left(u + \gamma h(u) + \gamma \xi\right) \leq V(u) - \gamma\delta$$

Any neighboorhood of $\mathcal{L}$ is recurrent

Theorem

Assume $h : \Theta \to \mathbb{R}^d$ *is continuous and*

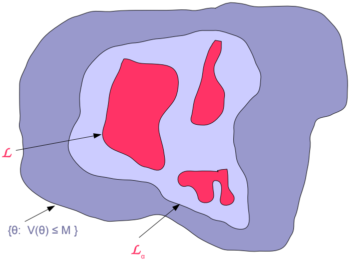
(i) **Lyapunov function:** *(idem)*

(ii) **Noise:** *(idem)*

(iii) **Step-size sequence:** $\sum_t \gamma_t = +\infty$

(iv) **Stability:** $\{\theta_t, t \geq 0\}$ *is in a compact subset of* Θ

Then: $\liminf_t \mathrm{d}\left(\theta_t, \mathcal{L}\right) = 0$.



Stability: $\theta_t \in \{V \leq M\}$ for any t

The Lyapunov property:

for any $\mathcal{K}$ compact s.t. $\mathcal{K} \cap \mathcal{L} = \emptyset$,

there exist δ,$\gamma_\star$,$\beta_\star > 0$ s.t.

$$[\gamma \leq \gamma_\star,\ |\xi| \leq \beta_\star,\ u \in \mathcal{K}] \Longrightarrow V\left(u + \gamma h(u) + \gamma\xi\right) \leq V(u) - \gamma\delta$$

Stepsize: $\sum_t \gamma_t = +\infty$ and $\lim_t \gamma_t = 0$

Convergence of $\{\theta_t, t \geq 0\}$: $\theta_{t+1} = \theta_t + \gamma_{t+1} H_{\theta_t}(X_{t+1})$

Theorem

Assume $h : \Theta \to \mathbb{R}^d$ *is continuous and*

(i) **Lyapunov function:** *(idem)*

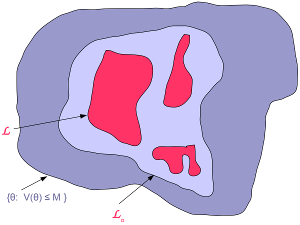
(ii) **Noise:** *(idem)*

(iii) **Step-size sequence:** *(idem)*

(iv) **Stability:** *(idem)*

(v) **Excursions outside $\mathcal{L}$:** $\lim_t V(\theta_t)$ *exists*

Then: $\lim_t \mathrm{dist}(\theta_t, \mathcal{L}) = 0$ *and also: convergence to a connected component of* $\mathcal{L}$



Stability:

The Lyapunov property:

Stepsize:

Shorter length of the excursions outside $\mathcal{L}_\alpha$: "$\lim_t V(\theta_t)$ exists" implies that the time

$$\sum_{j=k}^{\tau_\alpha(k)} \gamma_j$$

arbitrary small when k large enough.

Conclusion: stability and convergence

Theorem

Assume $h : \Theta \to \mathbb{R}^d$ *is continuous and*

(i) **Lyapunov function:** $V : \Theta \to \mathbb{R}^+$ *is* C^1 *and*

- *the level sets* $\{\theta \in \Theta : V(\theta) \leq M\}$ *are compact subsets of* Θ
- $\nabla V(\theta) \cdot h(\theta) \leq 0$
- $\mathcal{L} = \{\theta \in \Theta : \nabla V(\theta) \cdot h(\theta) = 0\}$ *is in a compact level set* $\{V \leq M_0\}$

(ii) **Noise:**

$$\lim_{L \to \infty} \sum_{t=1}^{L} \gamma_t \xi_t \qquad \text{exists}$$

(iii) **Step-size sequence:** $\sum_t \gamma_t = \infty \qquad \lim_t \gamma_t = 0$

(iv) **Recurrence:** $\{\theta_t, t \geq 0\}$ *is i.o. in a compact subset of* $\Theta \subseteq \mathbb{R}^d$

(v) **Excursions outside** $\mathcal{L}$**:** $\lim_t V(\theta_t)$ *exists*

Then: $\{\theta_t, t \geq 0\}$ *remains in a compact subset of* Θ *and* $\lim_t \mathrm{dist}(\theta_t, \mathcal{L}) = 0$.

Conclusion: on the noise

We write

$$H_{\theta_t}(X_{t+1}) = h(\theta_t) + \xi_{t+1}$$

In practice,

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot) \qquad \text{Markov transition with inv. dist. } \pi_{\theta_t}$$

- Biased approximation: $\mathbb{E}\left[\xi_{t+1}\right] \neq 0$
- Cumulated noise

$$\sum_{t=1}^{L} \gamma_{t+1} \left\{ H_{\theta_t}(X_{t+1}) - h(\theta_t) \right\}$$

  converges under assumptions
  - $\sum_t \gamma_t^2 < \infty$
  - ergodicity of the transition kernels P_θ
  - "smoothness-in-θ" of the transition kernels P_θ

In the literature

- On Stochastic Approximation

  - Andrieu, Moulines, Priouret. Stability of Stochastic Approximation under Verifiable conditions. *SIAM J. Control and Optimization (2005)*
  - Benaïm. Dynamics of stochastic approximation algorithms. *Séminaire de Probabilités de Strasbourg* (1999)
  - Borkar. Stochastic Approximation: a dynamical systems viewpoint. *Cambridge Univ. Press* (2008).
  - F., Moulines, Schreck, Vihola. Convergence of Markovian Stochastic Approximation with discontinuous dynamics *arXiv 1403.6803* (2015).
  - Kushner, Yin. Stochastic Approximation and Recursive Algorithms and Applications *Springer Book* (2003).

- Applied to the convergence of ABP samplers

  - Dama, Parrinello, Voth. Well tempered Metadynamics converges asymptotically. *Physical Review Letters (2014)*
  - F., Jourdain, Kuhn, Lelièvre and Stoltz. Convergence of the Wang-Landau algorithm. *Mathematics of Computation (2014)*
  - F., Jourdain, Lelièvre and Stoltz. Self-Healing Umbrella Sampling: convergence and efficiency. *arXiv math.PR 1410.2109* (2014)

Outline

1 Introduction

2 Adaptive Monte Carlo samplers

3 Stochastic Approximation Algorithms

4 Conclusion

As a global conclusion - a theorem solving the two questions

Theorem (F., Jourdain, Lelièvre, Stoltz (2014))

In the case $\theta_{t+1} = \theta_t + \gamma_{t+1} H_{\theta_t}(X_{t+1})$ *and* $\sup_{\theta,x} |H_\theta(x)| < \infty$

(i) *Lyapunov function with limit set* $\mathcal{L} = \{\theta_\star\}$*;* $\{\theta_t, t \geq 0\}$ *visits i.o. a compact set of* Θ*;* $\sum_t \gamma_t = +\infty$, $\sum_t \gamma_t^2 < \infty$.

(ii) *There exists* $\rho \in (0,1)$ *such that for any* $\theta \in \Theta$

$$\sup_x \|P_\theta^n(x,\cdot) - \pi_\theta\|_{TV} \leq 2\rho^n$$

(iii) *There exists* C *s.t. for all* $\theta, \theta' \in \Theta$

$$\sup_x \|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_{TV} \leq C|\theta - \theta'|$$

(iv) *There exists* C *s.t. for all* $\theta, \theta' \in \Theta$

$$\sup_x |H_\theta(x) - H_{\theta'}(x)| \leq C|\theta - \theta'|$$

Then: $\lim_t \theta_t = \theta_\star$ *a.s.* *and for any bounded function* g

$$\lim_N \frac{1}{N} \sum_{t=1}^{N} g(X_t) = \int g(x)\ \pi_{\theta_\star}(\mathrm{d}x) \qquad \textit{a.s.}$$

Collaborations

Talk based on joint works with

- Eric Moulines (Telecom ParisTech, France)
- Benjamin Jourdain (ENPC, France)
- Tony Lelièvre (ENPC, France)
- Gabriel Stöltz (ENPC, France)
- Pierre Priouret (Univ. Paris VI, France)
- Matti Vihola (Univ. Jyvaskyla, Finland)
- Pierre Vandekerkhove (Univ. Marne-la-Vallée, France)