When Monte Carlo and Optimization met in a Markovian dance

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse, France

INSTITUT
de MATHEMATIQUES
de TOULOUSE

ICTS "Advances in Applied Probability", Bengaluru, August 2019.

Part III.

Stochastic Approximation

with
Markovian dynamics

Stochastic Approximation (SA) methods with Markovian dynamics

- Designed to solve on $\Theta \subseteq \mathbb{R}^p$: $h(\theta) = 0$ when h is not explicit but

$$h(\theta) = \int_{\mathsf{X}} H(\theta, x) \, \mathrm{d}\pi_\theta(x)$$

Robbins-Monro, 1951; Benveniste-Métivier-Priouret, 1990; Kushner-Yin, 2003; Borkar, 2008

- Algorithm:

- Choose: a deterministic positive sequence $\{\gamma_t\}_t$ s.t. $\sum_t \gamma_t = +\infty$

- Initialisation: $\theta_0 = \theta_{\mathrm{init}} \in \Theta, X_0 = x_{\mathrm{init}}$

- Until convergence:

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot) \qquad \theta_{t+1} = \theta_t + \gamma_{t+1} \; H(\theta_t, X_{t+1})$$

where P_θ admits $\mathrm{d}\pi_\theta$ as its unique inv distribution.

- A perturbation of a time-discretized ODE $\dot{\theta}_s = h(\theta_s)$

$$\theta_{t+1} = \theta_t + \gamma_{t+1} \; h(\theta_t) + \gamma_{t+1}\eta_{t+1} \qquad \eta_{t+1} := H(\theta_t, X_{t+1}) - h(\theta_t)$$

A biased perturbation

$$h(\theta) = \int_{\mathsf{X}} H(\theta, x) \, \mathrm{d}\pi_\theta(x)$$

At each iteration

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

- Usually, cond. to the past, $X_{t+1} \sim \mathrm{d}\pi_{\theta_t}$ so that

$$\mathbb{E}\left[H(\theta_t, X_{t+1}) | \mathcal{F}_t\right] = h(\theta_t) \qquad \text{i.e. } \mathbb{E}\left[\eta_{t+1} | \mathcal{F}_t\right] = 0.$$

- In the present case, cond. to the past, $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$ so that

$$\mathbb{E}\left[H(\theta_t, X_{t+1}) | \mathcal{F}_t\right] - h(\theta_t) = \int_{\mathsf{X}} \left(P_{\theta_t}(X_t, \mathrm{d}x) - \mathrm{d}\pi_{\theta_t}(x)\right) H(\theta_t, x)$$

A **biased** approximation !

Is the bias vanishing ?

At each iteration

$$\theta_{t+1} = \theta_t + \gamma_{t+1}\ h(\theta_t) + \gamma_{t+1}\eta_{t+1} \qquad \eta_{t+1} := H(\theta_t, X_{t+1}) - h(\theta_t)$$

Two strategies:

• **The mini-batch approach.** Make the perturbation vanishing by sampling more and more points at each iteration

$$h(\theta_t) \approx \frac{1}{m_{t+1}} \sum_{j=1}^{m_{t+1}} H(\theta_t, X_{j,t+1}) \qquad \mathbb{E}\left[\eta_{t+1}|\mathcal{F}_t\right] = O(m_{t+1}^{-1})$$

Possible: choose $\gamma_t = \gamma$ if $m_t \to +\infty$ at some convenient rate.

• **($\star\star\star$) The SA regime.** Choose a vanishing stepsize $\gamma_t \to 0$ but a single Monte Carlo sample at each iteration.

Convergence analysis for SA: the successive steps

- Required: there exists a non-negative *Lyapunov* function V:

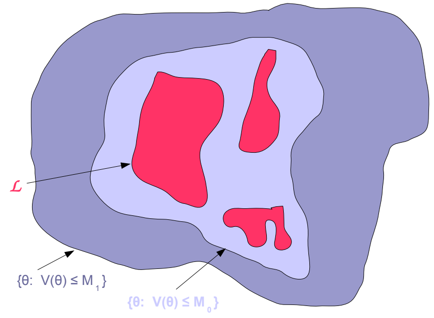
$$V(\theta_{t+1}) \leq V(\theta_t) - \gamma_{t+1}\, \phi^2(\theta_t) + \gamma_{t+1}\, \underbrace{W_{t+1}}_{\text{signed}}.$$

whose level sets are compact subsets of Θ, and ϕ is s.t. that

$$\inf_{\text{compact} \subset \Theta \setminus \mathcal{L}} \phi^2 > 0,$$

with

$$\mathcal{L} := \{\phi^2 = 0\} \subset \{V \leq M_0\}.$$



- Step 1: The sequence $\{\theta_t\}_t$ is stable i.e. (w.p.1) there exists a compact subset $\mathcal{K}$ of Θ such that $\theta_t \in \mathcal{K}$ for any t.

- Step 2: Convergence of $\{\theta_t\}_t$ to $\mathcal{L}$ (or to a connected component of $\mathcal{L}$; or to a point $\theta_\star \in \mathcal{L}$).

Stability: a crucial point (1/2)

- Roughly, the control of the noise is

$$\sup_t \left| \sum_{k=1}^{t} \gamma_{k+1} \Big(H(\theta_k, X_{k+1}) - h(\theta_k) \Big) \right| < \infty$$

- In our case,

$$X_{k+1} \sim P_{\theta_k}(X_k, \cdot), \qquad h(\theta) = \int H(\theta, x) \mathrm{d}\pi_\theta(x).$$

(a) How to control "uniformly-in-θ" the difference

$$P_\theta(x, \cdot) - \mathrm{d}\pi_\theta$$

when $\theta \in \{\theta_t, t \geq 0\}$ a random set ? <see Lecture 2> such a "containment condition" is realistic when "θ_t is stable".

(b) note that $X_k \sim P_{\theta_{k-1}}(X_{k-1}, \cdot)$ and the "diminishing adaptation condition" <see Lecture 2> will also play a role.

Stability: a crucial point (2/2) - Different strategies

- Stable by definition:

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

quite unlikely $\cdots$ *hum, really unlikely !*

- Force the stability by a projection on a compact subset $\mathcal{K}$

$$\theta_{t+1} = \Pi_{\mathcal{K}} \left(\theta_t + \gamma_{t+1} H(\theta_t, X_{t+1})\right)$$

Limiting points: in $\mathcal{L} \cap \mathcal{K}$. *How to choose* $\mathcal{K}$ *?*

- Use the Chen's technique: projection on growing compact subsets.

(Chen-Guo-Gao, 1988)

Self-stabilized Stochastic Approximation (the Chen's technique)

Choose compact subsets $\{\mathcal{K}_i\}_{i \geq 0}$ of Θ s.t. $\bigcup_i \mathcal{K}_i = \Theta$ and $\mathcal{K}_i \subset \mathcal{K}_{i+1}$.

- (Start - Block 1):

$\theta_0 = \theta_{\text{init}} \in \mathcal{K}_0$ and $X_0 = x_{\text{init}}$ and repeat for $t \geq 0$

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot) \qquad \theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1})$$

until $\theta_{t+1} \notin \mathcal{K}_0$. Set $T_1 = t + 1$.

- $\cdots$

- (Stop & re-start, Block $q+1$)

$\theta_{T_q} = \theta_{\text{init}}, \quad X_{T_q} = x_{\text{init}}$ and repeat for $t \geq 0$

$$X_{T_q+t+1} \sim P_{\theta_{T_q+t}}(X_{T_q+t}, \cdot) \qquad \theta_{T_q+t+1} = \theta_{T_q+t} + \gamma_{q+t+1} H(\theta_{T_q+t}, X_{T_q+t+1})$$

until $\theta_{T_q+t+1} \notin \mathcal{K}_q$. Set $T_{q+1} = T_q + t + 1$.

- $\cdots$

When is self-stabilized SA successful ? (1/4)

- Ans.: when the number of "stop & re-start" is finite !

then there exists L s.t.

(a) $\{\theta_t\}_t$ is in the compact set $\mathcal{K}_L$

(b) at T_L: $X_{T_L} = x_{\text{init}}$ and $\theta_{T_L} = \theta_{\text{init}}$

(c) for any $t \geq 0$

$$X_{T_L+t+1} \sim P_{\theta_{T_L+t}}(X_{T_L+t}, \cdot) \qquad \theta_{T_L+t+1} = \theta_{T_L+t} + \gamma_{L+t+1} H(\theta_{T_L+t}, X_{T_L+t+1})$$

- If the number is not finite: as if with $\rho_{t+1} \leftarrow \gamma_{L+t+1}$ for arbitrarily large L:

$$\theta_0 = \theta_{\text{init}}, X_0 = x_{\text{init}},$$

$$X_{t+1} \sim P_{\theta_t}(X_t, \cdot), \qquad \theta_{t+1} = \theta_t + \rho_{t+1} H(\theta_t, X_{t+1}),$$

until some **finite** time $t_\star$ such that $\theta_{t_\star+1} \notin \mathcal{K}_L$.

For arbitrarily small sequence $\{\gamma_{L+t}\}_t$, the algorithm exits from compact subset.

When is self-stabilized SA successful ? (2/4)

When *the number of "stop & restart" is finite*. How to ensure this condition ?

```
Below, a set of sufficient conditions, not the weakest ones;
just to show the mecanism more easily
```

- $\Theta = \mathbb{R}^p$

- Assumption 1. A function $V : \Theta \to (0, +\infty)$, continuously differentiable such that

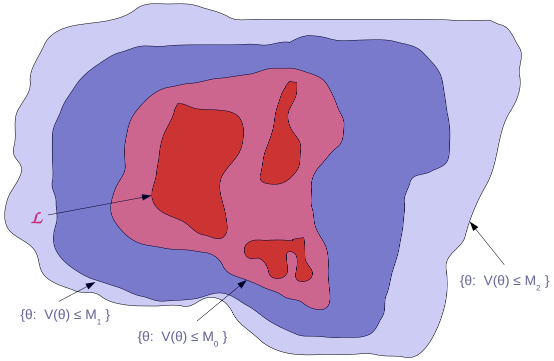
  compact level sets $\{V \leq M\}$, $\langle \nabla V(\tau), h(\tau) \rangle \leq 0$, $\sup_{\text{compact} \subset \mathcal{L}^c} \langle \nabla V(\tau), h(\tau) \rangle < 0$,

  where $\mathcal{L} := \{\tau : \langle \nabla V(\tau), h(\tau) \rangle = 0\}$ is bounded and in $\{V \leq M_0\}$.

- Assumption 2. The function h is Lipschitz on compact subsets.

- Assumption 3. $\lim_t \gamma_t = 0$.

When is self-stabilized SA successful ? (3/4)



Let M_0 be s.t. $\mathcal{L} \cup \mathcal{K}_0 \subset \{V \leq M_0\}$.
Fix $M_0 < M_1 < M_2$.
Given positive step sizes $\{\rho_t\}_t$
and a sequence of "noises" $\{\eta_t\}_t$, define

$$\bar{\tau}_0 = \tau_0 = \tau_{\text{init}} \in \mathcal{K}_0$$
$$\tau_{t+1} = \tau_t + \rho_{t+1}\left(h(\tau_t) + \eta_{t+1}\right), \qquad \bar{\tau}_{t+1} = \bar{\tau}_t + \rho_{t+1} h(\tau_t)$$

- Step 1. For any $t \geq 0$, <proof and expression of Δ_t>

$$\|\tau_{t+1} - \bar{\tau}_{t+1}\|^2 \leq \Delta_{t+1}^2 := xxx$$

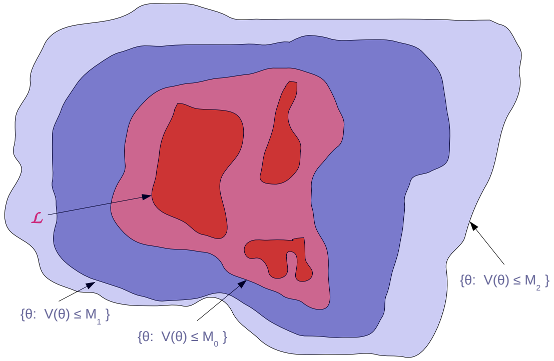
By induction. $\|\tau_0 - \bar{\tau}_0\|^2 = 0 = \Delta_0$.

$$\|\tau_{t+1} - \bar{\tau}_{t+1}\|^2 = \|\tau_t - \bar{\tau}_t + \rho_{t+1}\eta_{t+1}\|^2 = \|\tau_t - \bar{\tau}_t\|^2 + \rho_{t+1}^2\|\eta_{t+1}\|^2 + 2\rho_{t+1}\left\langle \tau_t - \bar{\tau}_t, \eta_{t+1}\right\rangle$$

Hence

$$\Delta_{t+1}^2 = \sum_{k=1}^{t} \rho_{k+1}^2\|\eta_{k+1}\|^2 + 2\rho_{k+1}\left\langle \tau_k - \bar{\tau}_k, \eta_{k+1}\right\rangle$$

When is self-stabilized SA successful ? (3/4)


$\mathcal{L}$
$\{\theta:\ V(\theta) \leq M_2\}$
$\{\theta:\ V(\theta) \leq M_1\}$
$\{\theta:\ V(\theta) \leq M_0\}$

Let M_0 be s.t. $\mathcal{L} \cup \mathcal{K}_0 \subset \{V \leq M_0\}$.
Fix $M_0 < M_1 < M_2$.
Given positive step sizes $\{\rho_t\}_t$
and a sequence of "noises" $\{\eta_t\}_t$, define

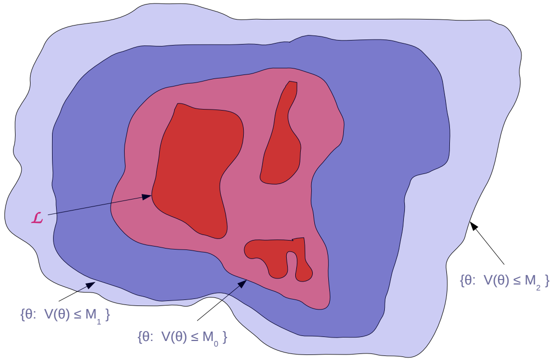
$$\bar{\tau}_0 = \tau_0 = \tau_{\text{init}} \in \mathcal{K}_0$$
$$\tau_{t+1} = \tau_t + \rho_{t+1}\left(h(\tau_t) + \eta_{t+1}\right), \qquad \bar{\tau}_{t+1} = \bar{\tau}_t + \rho_{t+1} h(\tau_t)$$

- Step 2. <proof> There exist $\rho_{\text{max}}, \delta_{\text{max}} > 0$ s.t.

$$\rho_{t+1} \leq \rho_{\text{max}}, \|\tau_t - \bar{\tau}_t\| \leq \delta_{\text{max}}, \bar{\tau}_t \in \{V \leq M_1\}, \tau_t \in \{V \leq M_2\} \Rightarrow \bar{\tau}_{t+1} \in \{V \leq M_1\}.$$

When is self-stabilized SA successful ? (3/4)

Let M_0 be s.t. $\mathcal{L} \cup \mathcal{K}_0 \subset \{V \leq M_0\}$.
Fix $M_0 < M_1 < M_2$.
Given positive step sizes $\{\rho_t\}_t$
and a sequence of "noises" $\{\eta_t\}_t$, define



$$\bar{\tau}_0 = \tau_0 = \tau_{\text{init}} \in \mathcal{K}_0$$

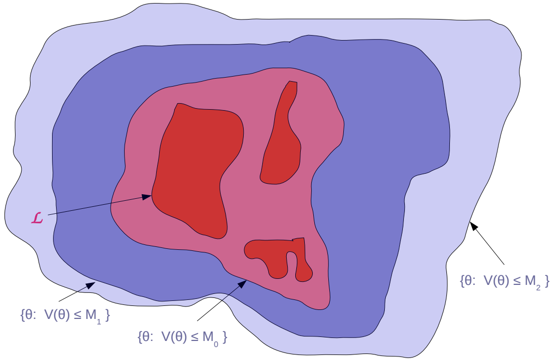
$$\tau_{t+1} = \tau_t + \rho_{t+1} \left(h(\tau_t) + \eta_{t+1} \right), \qquad \bar{\tau}_{t+1} = \bar{\tau}_t + \rho_{t+1} h(\tau_t)$$

- Step 3. <proof> There exist $\delta_{\text{max}} > 0$ s.t.

$$\|\tau_{t+1} - \bar{\tau}_{t+1}\| \leq \delta_{\text{max}}, \bar{\tau}_{t+1} \in \{V \leq M_1\} \Rightarrow \tau_{t+1} \in \{V \leq M_2\}.$$

$$V(\tau_{t+1}) \leq V(\bar{\tau}_{t+1}) + \|\nabla V(\bar{\tau}_{t+1})\| \; \|\tau_{t+1} - \bar{\tau}_{t+1}\| + C \, \|\tau_{t+1} - \bar{\tau}_{t+1}\|^2.$$

When is self-stabilized SA successful ? (3/4)



Let M_0 be s.t. $\mathcal{L} \cup \mathcal{K}_0 \subset \{V \leq M_0\}$.
Fix $M_0 < M_1 < M_2$.
Given positive step sizes $\{\rho_t\}_t$
and a sequence of "noises" $\{\eta_t\}_t$, define

$$\bar{\tau}_0 = \tau_0 = \tau_{\text{init}} \in \mathcal{K}_0$$
$$\tau_{t+1} = \tau_t + \rho_{t+1} \left(h(\tau_t) + \eta_{t+1} \right), \qquad \bar{\tau}_{t+1} = \bar{\tau}_t + \rho_{t+1} h(\tau_t)$$

- Step 1. For any $t \geq 0$,

$$\|\tau_{t+1} - \bar{\tau}_{t+1}\|^2 \leq \Delta_{t+1}^2 := \sum_{k=1}^{t} \rho_{k+1}^2 \|\eta_{k+1}\|^2 + 2\rho_{k+1} \left\langle \tau_k - \bar{\tau}_k, \eta_{k+1} \right\rangle$$

- Step 2. There exist $\rho_{\text{max}}, \delta_{\text{max}} > 0$ s.t.

$$\rho_{t+1} \leq \rho_{\text{max}}, \|\tau_t - \bar{\tau}_t\| \leq \delta_{\text{max}}, \bar{\tau}_t \in \{V \leq M_1\}, \tau_t \in \{V \leq M_2\} \Rightarrow \bar{\tau}_{t+1} \in \{V \leq M_1\}.$$

- Step 3. There exist $\delta_{\text{max}} > 0$ s.t.

$$\|\tau_{t+1} - \bar{\tau}_{t+1}\| \leq \delta_{\text{max}}, \bar{\tau}_{t+1} \in \{V \leq M_1\} \Rightarrow \tau_{t+1} \in \{V \leq M_2\}.$$

When is self-stabilized SA successful ? (4/4)

- Corollary: there exist $\rho_{\text{max}}, \delta_{\text{max}} > 0$ dependent on M_i but indep of $\{\rho_t\}_t$ and $\{\eta_t\}_t$ s.t.

$$\left.\begin{array}{l} \sup_{t \leq T} \rho_t \leq \rho_{\text{max}}, \\ \sup_{t \leq T} \Delta_t \leq \delta_{\text{max}} \end{array}\right\} \Rightarrow \forall t \leq T : \tau_t \in \{V \leq M_2\}, \bar{\tau}_t \in \{V \leq M_1\}.$$

- Theorem: <comment, proof>

If $\lim_t \gamma_t = 0$ and, for any $\delta > 0$ and any $M_1 < M_2$ (larger than M_0)

$$\lim_{i \to \infty} \mathbb{P}^{\gamma^{(i)}}_{x_{\text{init}}, \theta_{\text{init}}} \left(\sup_t \sum_{k=1}^{t} \{\gamma^2_{i+k+1} \|\eta_{k+1}\|^2 + \gamma_{i+k+1} \left\langle \theta_k - \bar{\theta}_k, \eta_{k+1} \right\rangle\} 1_{\theta_{1:t} \in \{V \leq M_2\}, \bar{\theta}_{1:t} \in \{V \leq M_1\}} \geq \delta \right) = 0$$

then, w.p.1, the number of "stop & restart" is finite.

In case of Markovian dynamics, how to check the condition ? (1/2)

$$\lim_{i\to\infty} \mathbb{P}^{\gamma^{(i)}}_{x_{\text{init}},\theta_{\text{init}}} \left(\sup_t \sum_{k=1}^{t} \{\gamma^2_{i+k+1} \|\eta_{k+1}\|^2 + \gamma_{i+k+1} \left\langle \theta_k - \bar{\theta}_k, \eta_{k+1} \right\rangle\} 1_{\theta_{1:t}\in\{V\leq M_2\}, \bar{\theta}_{1:t}\in\{V\leq M_1\}} \geq \delta \right) = 0$$

$$\eta_{k+1} = H(\theta_k, X_{k+1}) - \int H(\theta_k, x)\ \mathrm{d}\pi_{\theta_k}(x) \qquad X_{k+1} \sim P_{\theta_k}(X_k, \cdot) \qquad \theta_{k+1} = \theta_k + \gamma_{i+k+1} H(\theta_k, X_{k+1})$$

- Step 0. Markov inequality

- Step 1.

$$\lim_{i\to\infty} \mathbb{E}^{\gamma^{(i)}}_{x_{\text{init}},\theta_{\text{init}}} \left[\sum_{k\geq 1} \gamma^2_{i+k+1} \|\eta_{k+1}\|^2 1_{\theta_{1:k}\in\{V\leq M_2\}, \bar{\theta}_{1:k}\in\{V\leq M_1\}} \right] = 0$$

- Step 2.

$$\lim_{i\to\infty} \mathbb{E}^{\gamma^{(i)}}_{x_{\text{init}},\theta_{\text{init}}} \left[\sup_{t\geq 1} \left| \sum_{k=1}^{t} \gamma_{i+k+1} \left\langle \theta_k - \bar{\theta}_k, \eta_{k+1} \right\rangle \right| 1_{\theta_{1:k}\in\{V\leq M_2\}, \bar{\theta}_{1:k}\in\{V\leq M_1\}} \right] = 0$$

In case of Markovian dynamics, how to check the condition ? (2/2)

The tools:

- Uniform control: $\sup_{\theta \in \{V \leq M_2\}} \sup_x \|H(\theta, x)\|/W(x) < \infty$ for $W \geq 1$.
- Uniform-in-θ geometric ergodicity conditions, in W^2-norm. <containment condition>
- The Poisson equation: $\hat{H}_\theta$ s.t. $\hat{H}_\theta(x) - P_\theta \hat{H}_\theta(x) = H(\theta, x) - h(\theta)$.
- Regularity in θ of the solution $\hat{H}_\theta$ to the Poisson equation. <diminishing adapdation>
- The "chaining" rule: $X_{t+1} \sim P_{\theta_t}(X_t, \cdot)$.
- On the step sizes: $\sum_t \gamma_t^2 < \infty$.

Note that, by the self-stabilization: "uniform-in-θ" properties have to be verified *along compact subsets* $\{V \leq M\}$.

Refs: Andrieu-Moulines-Priouret, 2006; Fort-Moulines-Schreck-Vihola, 2016; Fort-Moulines, 2020.

And now, for the convergence of self-stabilized SA

- Similar conditions and techniques:

- after some random time T_L, the sequence is given by

$$\theta_{T_L} = \theta_{\text{init}}, \qquad \theta_{T_L+t+1} = \theta_{T_L+t} - \gamma_{L+t+1}\, H(\theta_{T_L+t}, X_{T_L+t+1})$$

- and remains in a compact subset depending upon the path

In the literature, SA with Markovian dynamics

(Andrieu-Moulines-Priouret, 2005; F,2015; F.-Jourdain-Lelièvre-Stoltz, 2017-2018; F.-Moulines-Schreck-Vihola,2016; Morral-Bianchi-F.,2017; Crepey-F.-Gobet-Stazhinski,2018)

● In the case $\theta \in \mathbb{R}^p$,

- Sufficient conditions for stability and convergence under weaker conditions than those here

- Central Limit Theorems (along a converging path) for both the sequence $\{\theta_t\}_t$ and the averaged sequence

$$\bar{\theta}_t = \frac{1}{t} \sum_{k=1}^{t} \theta_k$$

- Random stepsize sequence $\{\gamma_t\}_t$ (a self-tuned mecanism)

- Distributed SA

In the literature, SA in infinite dimension

- Most of the results on SA are in the case $\theta \in \mathbb{R}^p$. What about the case $\theta : v \mapsto \theta(v)$ is a function in $L^2(\nu)$?

see (Crepey-F.-Gobet-Stazhinski,2018)

- Motivation: uncertainty quantification in SA

$$\int H_v(\theta, x)\, \pi_v(\mathrm{d}x) = 0,$$

with an a priori $\mathrm{d}\nu$ on v.

- Goal: derive an algorithm that at the same time
  - for ν-a.a. v, finds $\theta(v)$ such that $\int H_v(\theta(v), x)\, \pi_v(\mathrm{d}x) = 0$.
  - provides the distribution of $\theta(V)$ when $V \sim \mathrm{d}\nu$.

- Use a chaos expansion, and a SA in **finite but growing dimension**

- Proof of convergence in the case: $\{X_t\}_t$ are i.i.d. in the SA algorithm.