Stochastic optimization: when Langevin comes into the game

Gilles Pagès

(including joint works with P. Bras & F. Panloup)

LPSM-Sorbonne-Université (Paris)

Spring School Le Mans

27-31 May 2024

Definitions

Definition (Gibbs measure)

Let $V : \mathbb{R}^d \to \mathbb{R}_+$ be a coercive continuous function such that

$$e^{-\frac{V}{\sigma_0^2}} \in L^1(\lambda_d) \quad \text{for some} \quad \sigma_0 > 0 \tag{1}$$

(λ_d Lebesgue measure on $\mathbb{R}^d$). Then the Gibbs (probability) measures are defined for every $\sigma \in (0, \sigma_0]$ by

$$\pi_\sigma = \pi_\sigma^V := C_\sigma e^{-\frac{V}{\sigma^2}} \cdot \lambda_d$$

where $C_\sigma = \left(\int_{\mathbb{R}^d} e^{-\frac{V(\xi)}{\sigma^2}} d\xi\right)^{-1}$.

- The Gibbs measures are well-defined since, for every $\sigma \leq \sigma_0$

$$0 \leq e^{-\frac{V}{\sigma^2}} \leq e^{-\frac{V}{\sigma_0^2}} \in L^1(\lambda_d) \quad \text{since } V \geq 0.$$

First properties

- As V is coercive and non-negative

$$v_* = \min_{\mathbb{R}^d} V \quad \text{exists and} \quad \mathrm{argmin}_{\mathbb{R}^d} V \quad \text{is compact}$$

- and $\pi_\sigma^V = \pi_\sigma^{V-v_*}$ by homogeneity.
- Hence, we may assume w.l.g. that

$$v_* = 0 \quad \text{and} \quad \mathrm{argmin}_{\mathbb{R}^d} V = \{V = 0\}.$$

- For every $\sigma \in (0, \sigma_0)$, if $\lambda \in \left(0, \frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right)$,

$$\int_{\mathbb{R}^d} e^{\lambda V} d\pi_\sigma < +\infty.$$

  since $e^{\lambda V} e^{-\frac{V(\xi)}{\sigma^2}} \leq e^{-\frac{V(\xi)}{\sigma_0^2}}$.
- By the way, why Gibbs measures ?

Fundamental theorem of Gibbs measures

Theorem

Let $V : \mathbb{R}^d \to \mathbb{R}_+$ *be a coercive continuous function s.t.* $e^{-V/\sigma_0^2} \in L^1(\lambda_d)$ *for some* $\sigma_0 > 0$ *(and* $v_* = 0$*).*

(a) *Then*

$$\forall \varepsilon > 0, \quad \pi_\sigma\big(\{V \geq \varepsilon\}\big) \longrightarrow 0 \quad \textit{as} \quad \sigma \to 0.$$

(b) *Equivalently, if* $X_\sigma \overset{\mathcal{L}}{\sim} \pi_\sigma$ *then*

$$\mathrm{dist}(X_\sigma, \{V = 0\}) \xrightarrow{\mathbb{P}} 0 \quad \textit{as} \quad \sigma \to 0.$$

In particular, if $\{V = 0\} = \{x^\star\}$ *then* $X_\sigma \xrightarrow{\mathbb{P}} x^\star$*.*

- The theorem remains true if continuity and coercivity are replaced by the lighter condition

$$\mathrm{argminV}_{\mathbb{R}^d} = \{V = 0\} \neq \varnothing \quad \text{and} \quad \lambda_d\big(V \in [0, \varepsilon)\big) > 0 \text{ for every } \varepsilon > 0.$$

Proof of (a)

- One has

$$\forall\, x \in \mathbb{R}^d,\; e^{-\frac{V(x)}{\sigma^2}} \to \mathbf{1}_{\{V=0\}}(x) \;\text{ as }\; \to 0$$

since … $V_{|\{V=0\}} = 0$ and $V_{|\{V=0\}^c} > 0$.

- On the other hand $e^{-\frac{V(x)}{\sigma^2}} \le e^{-\frac{V(x)}{\sigma_0^2}} \in L^1(\lambda_d)$ so that, by Lebesgue's dominated convergence theorem,

$$C_\sigma^{-1} = \int_{\mathbb{R}^d} e^{-\frac{V(\xi)}{\sigma^2}} d\xi \searrow \lambda_d(\{V=0\}) < +\infty \;\text{ as }\; \sigma \to 0.$$

- One shows that, for every $\varepsilon > 0$,

$$\begin{aligned}\lambda_d(V \le \varepsilon/3) &= \lambda_d\big(e^{-V/\sigma^2} \ge e^{-\varepsilon/(3\sigma^2)}\big)\\ &\le e^{\varepsilon/(3\sigma^2)} \int_{\mathbb{R}^d} e^{-\frac{V}{\sigma^2}} d\lambda_d = e^{\varepsilon/(3\sigma^2)} C_\sigma^{-1}\end{aligned}$$

so that

$$C_\sigma \le e^{\frac{\varepsilon}{3\sigma^2}} \big(\lambda_d(V \le \varepsilon/3)\big)^{-1}.$$

Proof of (a)

- Note that by continuity of V, $\{V \le \varepsilon/3\}$ contains a ball $B(x^\star, \eta_\varepsilon)$ where $V(x^\star) = 0$ and $\eta_\varepsilon > 0$ so that $\lambda_d(\{V \le \varepsilon/3\}) > 0$.
- Now, we have (keep in mind $C_\sigma \le e^{\frac{\varepsilon}{3\sigma^2}}\big(\lambda_d(V \le \varepsilon/3)\big)^{-1}$)

$$
\begin{aligned}
\pi_\sigma\big(V \ge \varepsilon\big) &= C_\sigma \int_{\{V \ge \varepsilon\}} e^{-\frac{V}{\sigma^2}} d\lambda_d \\
&= C_\sigma \int_{\{V \ge \varepsilon\}} e^{-\frac{2V}{3\sigma^2}} e^{-\frac{V}{3\sigma^2}} d\lambda_d \\
&\le C_\sigma e^{-\frac{\varepsilon}{3\sigma^2}} e^{-\frac{\varepsilon}{3\sigma^2}} \int_{\{V \ge \varepsilon\}} e^{-\frac{V}{3\sigma^2}} d\lambda_d \\
&\le \big(\lambda_d\big(V \le \varepsilon/3\big)\big)^{-1} e^{-\frac{\varepsilon}{3\sigma^2}} \int_{\{V \ge \varepsilon\}} e^{-\frac{V}{3\sigma^2}} d\lambda_d \\
&\le \big(\lambda_d\big(V \le \varepsilon/3\big)\big)^{-1} e^{-\frac{\varepsilon}{3\sigma^2}} \int_{\mathbb{R}^d} e^{-\frac{V}{3\sigma^2}} d\lambda_d \\
&= \big(\lambda_d\big(V \le \varepsilon/3\big)\big)^{-1} e^{-\frac{\varepsilon}{3\sigma^2}} C_{\sqrt{3}\,\sigma} \overset{\sigma \to 0}{\longrightarrow} 0.
\end{aligned}
$$

Proof of (b)

- Let $\varepsilon > 0$, and $\eta_\varepsilon := \inf\big\{V(x) : \mathrm{dist}\big(x, \{V=0\}\big) \geq \varepsilon\big\} > 0$.
- Hence

$$\mathbb{P}\big(\mathrm{dist}(X_\sigma, \{V=0\}) \geq \varepsilon\big) \leq \mathbb{P}\big(V(X_\sigma) \geq \eta_\varepsilon\big) \to 0 \quad \text{as} \quad \sigma \to 0.$$

- Conversely, if $\mathrm{dist}(X_\sigma, \{V=0\}) \xrightarrow{\mathbb{P}} 0$ then $V(X_\sigma) \xrightarrow{\mathbb{P}} 0$. Now

$$\mathcal{L}(V(X_\sigma)) = \pi_\sigma \circ V^{-1}$$

so that

$$\forall\, \varepsilon > 0, \quad \pi_\sigma\big(V \geq \varepsilon\big) \to 0 \quad \text{as} \quad \sigma \to 0.$$

Unique non degenerate (strict) minima $x^\star$

W.l.g. we may assume, up to a change of variable, that $x^\star = 0$.

Theorem (Athreya-Hwang I, 2010)

Let $V : \mathbb{R}^d \to \mathbb{R}_+$ *be a continuous coercive function such that* $\operatorname{argmin}_{\mathbb{R}^d} V = \{0\}$, $V(0) = 0$ *and* $\nabla^2 V(0)$ *exist and is positive definite. Assume furthermore*

1. $e^{-V/\sigma_0^2} \in L^1(\mathbb{R}^d, \lambda_d)$ *for some* $\sigma_0 > 0$.
2. $\forall\, x \in \mathbb{R}^d$, $\dfrac{V(\sigma x)}{\sigma^2} \longrightarrow g(x) := \frac{1}{2} x^\top \nabla^2 V(0) x \in \mathbb{R}$ *as* $\sigma \to 0$.
3. *One has* $\displaystyle\int_{\mathbb{R}^d} \sup_{0<\sigma<\sigma_0} e^{-\frac{V(\sigma x_1, \ldots, \sigma x_d)}{\sigma^2}} dx_1 \ldots dx_d < +\infty$.

Then $e^{-g} \in L^1(\mathbb{R}^d, \lambda_d)$ *and if* $X_\sigma \overset{\mathcal{L}}{\sim} \pi_\sigma$ *for every* $\sigma \in (0, \sigma_0)$, *one has*

$$\frac{X_\sigma}{\sigma} \xrightarrow{\mathcal{L}} C_d e^{-g(x_1,\ldots,x_d)} dx_1, \ldots, dx_d \quad as \quad \sigma \to 0.$$

- If $\nabla^2 V(0)$ has a null eigenvalue, then

$$\int_{\mathbb{R}^d} e^{-g} d\lambda_d = +\infty$$

- Let $V(x_1, x_2) = x_1^2 + x_2^4$. Then $\alpha_1 = 1$, $\alpha_2 = 2$ and $g = V$. One checks that

$$\Big(\frac{(X_\sigma)_1}{\sigma}, \frac{(X_\sigma)_2}{\sigma^2}\Big) \xrightarrow{\mathcal{L}} C_V e^{-V}.$$

- How to handle when this happens ?

Degenerate minima

What happens when, e.g., $\nabla^2 V(0)$ is degenerate ?

Theorem (Athreya-Hwang II, 2010)

Let $V : \mathbb{R}^d \to [0, \infty)$ be a continuous and coercive function such that :

1. $e^{-V/\sigma_0^2} \in L^1(\mathbb{R}^d)$.
2. *There exist $\alpha_1, \ldots, \alpha_d > 0$ such that for all $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,*

$$\frac{1}{\sigma^2} V(\sigma^{\alpha_1} x_1, \ldots, \sigma^{\alpha_d} x_d) \longrightarrow g(x_1, \ldots, x_d) \in \mathbb{R} \quad \text{as} \quad \sigma \to 0.$$

3. $\displaystyle\int_{\mathbb{R}^d} \sup_{0<\sigma<\sigma_0} e^{-\frac{V(\sigma^{\alpha_1} x_1, \ldots, \sigma^{\alpha_d} x_d)}{\sigma^2}} dx_1 \ldots dx_d < +\infty.$

Then $e^{-g} \in L^1(\mathbb{R}^d)$ and if $X_\sigma \stackrel{\mathcal{L}}{\sim} \pi_\sigma$ for every $\sigma \in (0, \sigma_0)$, one has

$$\left(\frac{(X_t)_1}{\sigma^{\alpha_1}}, \ldots, \frac{(X_t)_d}{\sigma^{\alpha_d}}\right) \xrightarrow{\mathcal{L}} C_d e^{-g(x_1, \ldots, x_d)} \quad \text{as } \sigma \to 0.$$

Multiple wells

- Assume now $\operatorname*{argmin}_{\mathbb{R}^d} V = \{x^{1,\star}, \ldots, x^{m,\star}\}$ for some $m \in \mathbb{N}$.
- The limiting measure of π_σ as $\sigma \to 0$ will be supported by a subset $\{x_1^\star, \ldots, x_m^\star\}$, with different weights.

Theorem (Athreya-Hwang III, 2010)

Let $V : \mathbb{R}^d \to [0, \infty)$ be continuous and coercive such that:

1. $e^{-V/\sigma_0^2} \in L^1(\lambda_d, \mathbb{R}^d)$.
2. *For all i, there exist $(\alpha_{ij})_{1 \le j \le d}$ such that $\alpha_{ij} \ge 0$ for all j and*
$$\frac{1}{\sigma^2} V\big(x^{i,\star} + (\sigma^{\alpha_{i1}} x_1, \ldots, \sigma^{\alpha_{id}} x_d)\big) \longrightarrow g_i(x_1, \ldots, x_d) \in [0, \infty) \text{ as } \sigma \to 0.$$
3. $\forall\, i \in \{1, \ldots, m\}$, $\displaystyle\int_{\mathbb{R}^d} \sup_{0<t<1} e^{-\frac{V(x^{i,\star} + (\sigma^{\alpha_{i1}} x_1, \ldots, \sigma^{\alpha_{id}} x_d))}{\sigma^2}} dx_1 \ldots dx_d < +\infty.$

[to be continued ...]

...

Theorem (Athreya-Hwang III, 2010)

Let $\alpha := \min_{1\le i\le m}\Big\{\sum_{j=1}^{d}\alpha_{ij}\Big\}$ *and let*
$J := \Big\{i\in\{1,\ldots,m\} : \ \sum_{j=1}^{d}\alpha_{ij}=\alpha\Big\}$. *Let* $X_\sigma \sim \pi_t$, $0<\sigma<\sigma_0$.
Then:

$$X_\sigma \xrightarrow{\mathcal{L}} \frac{1}{\sum_{j\in J}\int_{\mathbb{R}^d} e^{-g_j(x)}dx}\sum_{i\in J}\int_{\mathbb{R}^d} e^{-g_i(x)}dx\cdot\delta_{x^{i,\star}} \ \ \textit{as} \ \ \sigma\to 0.$$

- The non-empty index set J represent the dominating elements or "less degenerate") of $\{V=0\}$.
- Only the less degenerate minima are asymptotically "visible" by the Gibbs measure.

How to use this theorem ?

- Checking Condition 2 is the core of the problem.
- It has been extensively investigated in a recent paper by P. Bras (*Bernoulli* 2022) when V has $x^\star$ is a "higher order" strict minimum...
- It relies on the analysis of the tensors $\nabla^{2k} V(x^\star)$ which are associated to homogenous polynomials of degree $2k$ on $\mathbb{R}^d$
- A curiosity: it involves the answer to the 17th Hilbert's problem (1900): Can such a polynomial be represented as sum of squares of other polynomials? The answer is "no"
- It can be proved that it boils down to look at homogenous polynomials with even degree. Thus (Motzkin, 1967) exhibited

$$f(x, y, z) = z^6 + x^4y^2 + x^2y^4 - 3x^2y^2z^2.$$

cannot be decomposed

Gibbs measures as invariant distributions of Langevin equations

- To localize $\underset{\mathbb{R}^d}{\operatorname{argmin}} V$ estimate π_σ for small enough $\sigma > 0$ is a natural idea.
- Several ways to estimate a distribution, usually as the invariant distribution of a Markov dynamics
  - Metropolis algorithm ...
  - MCMC
  - Diffusions
  - Combination of the above (ULA)
- We will opt for diffusions due to its compatibility with recursive stochastic approximation, flexibility, etc.

- Let $V : \mathbb{R}^d \to \mathbb{R}_+$ be a coercive, continuously differentiable function with Lipschitz gradient and satisfying our standing assumption

$$e^{-\frac{V}{\sigma_0^2}} \in L^1(\lambda_d) \ \text{ for some } \ \sigma_0 > 0.$$

(When this holds true for any $\sigma_0 > 0$ we will consider by convention that $\sigma_0 = +\infty$.)

- We associate to π_σ, $\sigma \in (0, \sigma_0)$, the Langevin (Brownian) SDE on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$

$$(\mathcal{L}_\sigma) \ \equiv \ dX_t = -\nabla V(X_t)dt + \sigma\sqrt{2}dW_t.$$

- This *SDE* has a unique strong solution starting from any random variable $X_0 \perp\!\!\!\perp W$.
- If $\sigma = 0$, then $\dot{V}(X_t) = -|\nabla V(X_t)|^2 \leq 0$ so that $V(X_t) \searrow$ and $\int_0^{+\infty} |\nabla V(X_s|^2 ds < +\infty$ so that $(\dots)$ $X_t \to \{V = 0\}$.
- When $\sigma \in (0, \sigma_0)$, we will prove that the SDE has π_σ is a unique invariant distribution i.e. if $X_0 \stackrel{d}{=} \pi_\sigma$, then $X_t \sim \pi_\sigma$ for every $t \geq 0$, (and much more...)

Necessary conditions

- The infinitesimal generator of the Langevin equation $(\mathcal{L}_\sigma)$ reads for $f \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$
$$\mathcal{A}f = -(\nabla f \,|\, \nabla V) + \sigma^2 \mathrm{Tr}(\nabla^2 f).$$
- Assume $\nu_\sigma = p_\sigma \cdot \lambda_d$ is an invariant distribution.
$$\forall\, f \in \mathcal{C}^2_K(\mathbb{R}^d, \mathbb{R}), \quad \mathbb{E}\, f(X_t) = \mathbb{E}\, f(X_0) = \int_{\mathbb{R}^d} f(\xi)\nu_\sigma(d\xi),\ t \geq 0.$$
Hence
$$\mathbb{E}\, f(X_t) = \mathbb{E}\, f(X_0) + \mathbb{E}\int_0^t \mathcal{A}f(X_s)ds + \sigma\sqrt{2}\,\mathbb{E}\int_0^t (\nabla f(X_s) \,|\, dW_s),$$
i.e.
$$\forall\, t \geq 0, \quad \mathbb{E}\int_0^t \mathcal{A}f(X_s)ds = 0$$
- $\mathcal{A}f$ is bounded and $X_s \stackrel{d}{=} \nu_\sigma = p_\sigma(\xi)d\xi$, hence by Fubini's Theorem
$$\int_0^t \mathbb{E}\,\mathcal{A}f(X_s)ds = 0 \ \text{ i.e. } \quad \int_0^t \Big[\int \mathcal{A}f(\xi)p_\sigma(\xi)d\xi\Big]ds = 0.$$

Stationary Fokker-Planck equation

- Then $$\forall f \in \mathcal{C}_K^2, \quad \int \mathcal{A}f(\xi)p_\sigma(\xi)d\xi = 0.$$
- Let $\mathcal{A}^*$ denote the adjoint operator of $\mathcal{A}$ on $\mathcal{C}_K^2(\mathbb{R}^d, \mathbb{R})$ defined by

$$\forall\, f,\, g \in \mathcal{C}_K^2(\mathbb{R}^d, \mathbb{R}), \quad \int_{\mathbb{R}^d} (\mathcal{A}^* g)(\xi) f(\xi)\, d\xi = \int_{\mathbb{R}^d} g(\xi)(\mathcal{A}f(\xi))\, d\lambda_d.$$

- Elementary computations show that, if V is $\mathcal{C}^2$, it reads

$$\forall\, g \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R}), \quad \mathcal{A}^* g = \mathrm{div}(g\nabla V) + \sigma^2 \Delta g.$$

(div denotes the divergence operator and Δ the Laplacian operator.)

- As a consequence, if $p_\sigma \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d, \lambda_d)$, then it is a non-negative λ_d-integrable weak solution and in fact a classical solution by approximation arguments to the (elliptic) PDE

$$\sigma^2 \Delta p_\sigma + \mathrm{div}(p_\sigma \nabla V) = 0.$$

The converse is more demanding...

- Conversely if a non-negative function $g_\sigma \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R}) \cap L^1(\mathbb{R}^d, \lambda_d)$ satisfies the elliptic PDE

$$\sigma^2 \Delta g_\sigma + \mathrm{div}(g_\sigma \nabla V) = 0,$$

then it is clear that $p_\sigma = \Big(\int_{\mathbb{R}^d} g_\sigma d\lambda_d\Big)^{-1} g_\sigma$ is a probability density and, by a backward reasoning,

$$\forall f \in \mathcal{C}^2_K(\mathbb{R}^d, \mathbb{R})\lambda_d), \quad \int_{\mathbb{R}^d} \mathcal{A}f(\xi)\, \underbrace{p_\sigma(\xi)d\xi}_{=:\, \nu_\sigma(d\xi)} = 0.$$

- Does it imply stationarity of $\nu_\sigma = p_\sigma \cdot \lambda_d$?
- Can we make ν_σ explicit ?

Echeverria-Weiss Theorem

- The fact that then $\nu_\sigma = p_\sigma \cdot \lambda_d$ is an invariant distribution for the Langevin equation is a consequence of Echeverria-Weiss Theorem (see [3, Theorem 9.17]).

Theorem (Echeverria-WeissTheorem)

Let $\mathcal{A}$ be a linear operator defined on $\mathcal{C}^2_K(\mathbb{R}^d)$ satisfying

- Posit. max. princ. $\forall\, f \in \mathcal{C}^2_K(\mathbb{R}^d)$, $\sup_{\mathbb{R}^d} f(x) = f(x_0) \geq 0 \Rightarrow \mathcal{A}f(x_0) \leq 0$.
- $\exists\, f_n$, $n \geq 1$, *s.t.* $\sup_n(\|f_n\|_\infty + \|\mathcal{A}f_n\|_\infty) < +\infty$, $f_n \to 1$ *and* $\mathcal{A}f_n \to 0$.
- $\forall\, g \in \mathcal{C}^2_K(\mathbb{R}^d, \mathbb{R})$, $\nu_\sigma(g) = 0$.

then there exists a stationary solution for the martingale problem $(\mathcal{A}, \nu)$ i.e. there exists a stationary continuous-time homogeneous Markov process with infinitesimal generator $\mathcal{A}$ and ν as an invariant distribution.

- Switch from $\mathbb{R}^d$ to E locally compact Polish space and $\mathcal{C}^2_K(\mathbb{R}^d)$ to a dense subset in $\mathcal{C}_0(E)$.

Heuristics to understand?

- Let $P_t f(x) = \mathbb{E}\, f(X_t^x)$. By Itô's formula to $f(X_t^x)$, $f \in \mathcal{C}_K^2(\mathbb{R}^d, \mathbb{R})$ and taking expectation implies

$$P_s f(x) = \mathbb{E}\, f(X_s^x) = f(x) + \int_0^s \mathbb{E}\, \mathcal{A}f(X_u^x)du = f(x) + \int_0^s P_u \mathcal{A}f(x)du$$

so that (as $u \mapsto P_u \mathcal{A}f(x)$ is continuous),

$$(\star) \qquad \mathcal{A}f(x) = \lim_{s\to 0} \frac{P_s f(x) - f(x)}{s}.$$

- Assume that for "enough" functions f $(\star)$ is also true for $P_t f$ (needs $P_t f$ to be at least $\mathcal{C}^2$). The (Markovian) semi-group property $P_s \circ P_t = P_{s+t} = P_{s+t} = P_t \circ P_s$ yields (formally)

$$\mathcal{A}P_t f(x) = \lim_{s\to 0} \frac{P_{s+t} f(x) - P_t f(x)}{s} = P_t \left(\lim_{s\to 0} \frac{P_s f(x) - f(x)}{s} \right) = P_t \mathcal{A}f(x)$$

(this interchange of P_t and the limit is the blocking point in fact) i.e.

$$\forall\, t \geq 0, \qquad \mathcal{A}P_t = P_t \mathcal{A}$$

- Assume that $X_0 \stackrel{d}{=} \nu_\sigma$. As $\mathbb{E}\, f(X_t) = \int \nu_\sigma(d\xi) P_t f(\xi)$, we get

$$\begin{aligned}\mathbb{E}\, f(X_t) &= \int f d\nu + \int_0^t \mathbb{E} \mathcal{A} f(X_s) ds \\ &= \int d\nu + \int_0^t \Big[\int P_s \mathcal{A} f(x_0) \nu(dx_0)\Big] ds \\ &= \int f d\nu + \int_0^t \underbrace{\int \mathcal{A} P_s f(x_0) \nu(dx_0)}_{=0} ds = \int f d\nu\end{aligned}$$

- ... provided $P_s f$ lies in the class of functions g such that $\nu(\mathcal{A} g) = 0$.

Proposition

Assume $V : \mathbb{R}^d \to \mathbb{R}_+$ be a coercive $\mathcal{C}^2$ function with a bounded Hessian $\nabla^2 V$ s.t. $e^{-V/\sigma_0^2} \in L^1(\lambda_d)$. Let $\sigma \in (0, \sigma_0)$. The Gibbs measure $\pi_\sigma = C_\sigma e^{-\frac{V}{\sigma^2}} \cdot \lambda_d$ is the unique invariant distribution of the Langevin SDE

$$dX_t = -\nabla V(X_t)dt + \sigma dW_t.$$

Moreover, for every $\lambda \in \left(0, \frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right)$, $\int_{\mathbb{R}^d} e^{\lambda V} d\pi_\sigma < +\infty$.

Proof (*Existence*). One computes for every $x = (x^1, \ldots, x^d) \in \mathbb{R}^d$ and every $i \in \{1, \ldots, d\}$,

$$\frac{\partial}{\partial x^i}\left(e^{-\frac{V}{\sigma^2}} \frac{\partial V}{\partial x^i}\right) = \left(\frac{\partial^2 V}{\partial (x^i)^2} - \frac{1}{\sigma^2}\left(\frac{\partial V}{\partial x^i}\right)^2\right) e^{-\frac{V}{\sigma^2}}$$

and

$$\frac{\partial^2 e^{-\frac{V}{\sigma^2}}}{\partial (x^i)^2} = \left(-\frac{1}{\sigma^2}\frac{\partial^2 V}{\partial (x^i)^2} + \frac{1}{(\sigma^2)^2}\left(\frac{\partial V}{\partial x^i}\right)^2\right) e^{-\frac{V}{\sigma^2}}$$

so that

$$\sigma^2 \Delta\, e^{-\frac{V}{2\sigma^2}} + \mathrm{div}\left(e^{-\frac{V}{2\sigma^2}} \nabla V\right) = 0.$$

- Several approaches are possible to establish uniqueness of the invariant distribution by probabilistic methods.
- We choose the most general one based on ellipticity of $(\mathcal{L}_\sigma)$ (also valid for a wide class of homogenous Markov processes)
- We know by Girsanov theorem $(\dots)$ that for every $x \in \mathbb{R}^d$ and every $t > 0$, the distribution $P_t(x, dy)$ of X_t^x is absolutely continuous

$$P_t(x, dy) = p_t(x, y)\lambda_d(dy) \quad \text{with} \quad p_t(x, y) > 0.$$

- Now let ν be any invariant distribution of $(\mathcal{L}_\sigma)$. For every non-negative Borel function $g : \mathbb{R}^d \to \mathbb{R}_+$, one derives from the identity $\nu P_t = \nu$ and Fubini-Tonelli's Theorem that

$$\int g\, d\nu = \mathbb{E}\, g(X_t^\nu) = \int\int \nu(dx)\mathbb{E}\, g(X_t^x) = \int \nu(dx)P_t g(x)$$
$$= \int_{\mathbb{R}^d} \nu(dx) \int_{\mathbb{R}^d} g(y)p_t(x, y)dy = \int_{\mathbb{R}^d} g(y)\Big[\int_{\mathbb{R}^d} p_t(x, y)\nu(dx)\Big]dy.$$

Hence, as $p_t(x, y) > 0$ for every x, $y > 0$, $\forall\, y,\ \int_{\mathbb{R}^d} p_t(x, y)\nu(dx) > 0$,

$$\nu = \Big[\int_{\mathbb{R}^d} p_t(x, y)\mu(dx)\Big] \cdot \lambda_d \sim \lambda_d.$$

- As a consequence, any two invariant distributions are equivalent on $\mathbb{R}^d$.
- Let μ be another invariant distribution. Then $\mu \sim \pi_\sigma$ so that

$$\mu = h\cdot\pi_\sigma, \quad h:\mathbb{R}^d \to \mathbb{R}_+ \text{ probability density function.}$$

- If $h \le 1$ π_σ-*a.s.* then $\displaystyle\int (1-h)d\pi_\sigma = \pi_\sigma(\mathbb{R}^d) - \mu(\mathbb{R}^d) = 0$ so that $h = 1$ π_σ-*a.s.* *i.e.* $\mu = \pi_\sigma$.
- Otherwise $\pi_\sigma(\{h>1\}) > 0$ and set $\tilde\mu = (h\wedge 1)\cdot\pi_\sigma$. One has $\tilde\mu(\mathbb{R}^d) \le 1$

$$\tilde\mu P_t g = \int_{\mathbb{R}^d} (h(x)\wedge 1) P_t g(x)\pi_\sigma(dx) \le \mu P_t g \wedge \pi_\sigma P_t g \ (= \mu(g)\wedge\pi_\sigma(g)).$$

Then, using the above upper-bound successively in the second line

$$\begin{aligned}\int P_t g d\tilde\mu &= \int P_t(g\mathbf{1}_{\{h\le 1\}})\, d\,\tilde\mu + \int P_t(g\mathbf{1}_{\{h>1\}})\, d\,\tilde\mu \\ &\le \int g\mathbf{1}_{\{h\le 1\}}\, d\,\mu + \int g\mathbf{1}_{\{h>1\}}\, d\,\pi_\sigma \\ &= \int_{\{h\le 1\}} g\, h\, d\pi_\sigma + \int_{\{h>1\}} g\, d\pi_\sigma = \int g(h\wedge 1)d\pi_\sigma = \int g d\tilde\mu.\end{aligned}$$

- Consequently $\qquad \tilde\mu P_t \le \tilde\mu.$

- Consequently $\tilde{\mu} P_t \leq \tilde{\mu} \ldots$ with the same mass.
- Hence $\tilde{\mu} = \tilde{\mu} P_t$ is also an invariant measure.
- As $\tilde{\mu} \leq \pi_\sigma$ by construction it is clear that $\tilde{\pi}_\sigma = \tilde{\mu} = (1-h)_+ \cdot \pi_\sigma$ is also a finite invariant measure.
- If $\tilde{\pi}_\sigma \equiv 0$ then $h \geq 1$ π_σ-a.s. which implies $\int h d\pi_\sigma > 1$ since $\pi_\sigma(\{h > 1\}) > 0$. Impossible.
- Consequently $\tilde{\pi}_\sigma \not\equiv 0$. Then $\frac{\tilde{\pi}_\sigma}{\tilde{\pi}_\sigma(\mathbb{R}^d)} \sim \pi_\sigma$ is an invariant distribution which in turn implies that $(1-h)_+ > 0$ π_σ-*a.s.* or, equivalently, $h < 1$ π_σ-*a.s.*. Then $\mu(\mathbb{R}^d) < \pi_\sigma(\mathbb{R}^d)$ which is also impossible. Hence $\mu = \pi_\sigma$. $\square$

Remarks. • An alternative and more straightforward proof based on a "confluence" argument (*e.g.* when V is α-convex is possible (see the exercise later on).

Stochastic Gradient Descent (SGD)

- We start from the standard *SGD* related to a differentiable function V with Markovian representation

$$Y_{n+1} = Y_n - \gamma_{n+1} H(Y_n, Z_{n+1}), \quad Y_0 = \xi_0$$

- where
  - $(Z_n)_{n\geq 1}$ is an i.i.d sequence of "innovations" on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$,
  - ξ_0 is independent of $(Z_n)_{n\geq 1}$ on $(\Omega, \mathcal{A}, \mathbb{P})$,
  - $\nabla V(y) = \mathbb{E}\, H(y, Z_1),\ y \in \mathbb{R}^d$, $H : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^d$ Borel,
  - $(\gamma_n)_{n\geq 1}$ a sequence of (small) constant or decreasing steps.
- Its canonical decomposition

$$Y_{n+1} = Y_n - \gamma_{n+1}\nabla V(Y_n) + \gamma_{n+1}\Delta M_{n+1} \ \text{ with } \ \Delta M_{n+1} = \nabla V(Y_n) - H(Y_n, Z_{n+1})$$

is a sequence of martingale increments (called natural) since

$$\mathbb{E}\, H(Y_n, Z_{n+1})\,|\mathcal{F}_n^{Y_0,Z}) = \big[\mathbb{E}\, H(y, Z)\big]_{|y=Y_n} = \nabla V(Y_n)$$

where $\mathcal{F}_n^{Y_0,Z} = \sigma(Y_0, Z_1, \ldots, Z_n),\ n \geq 0$, denotes the natural filtration of the *SGD*.

Gradient Descent (GD)

- If $H(y, z) = \nabla V(y)$ then $\Delta M_n \equiv 0$ and the recursion reads

$$y_{n+1} = y_n - \gamma_{n+1}\nabla V(y_n), \quad y_0 = \xi_0 \in \mathbb{R}^d$$

- This recursion is called a *Gradient Descent*.

Discussion: Datascience *vs* Numercial Probability

- In Numerical Probability, usually

$$Z \stackrel{\mathcal{L}}{\sim} p(z)\lambda_q(dz), \quad q \text{ large}$$

so if the only access to ∇V is

$$\nabla V(y) = \mathbb{E}\, H(y, Z) = \int_{\mathbb{R}^d} H(y,z)p(z)dz,$$

simulation becomes the only way out.

- We don't know how to bypass this problem.
- In DataScience, one can samples from a (huge) database $(z_k)_{k1:N}$ since

$$Z \stackrel{\mathcal{L}}{\sim} \frac{1}{N}\sum_{k=1}^{N} \delta_{z_k} \sim Z_{I_N}, \quad I_N \sim U\big(\{1:N\}\big)$$

- No ! We can't compute $\nabla V(y) = \frac{1}{N}\sum_{k=1}^{N} H(y, z_k)$ at each timestep.

Mini-batch: the art of "en même temps"

- One defines $(Y_n)_{n\geq 1}$ recursively by

$$Y_{n+1} = Y_n - \gamma_{n+1}\frac{1}{M}\sum_{k=1}^{M} H(Y_n, Z_k^{(n)}), \quad Y_0 = \xi_0.$$

with $(Z_k^{(n)})_{k=1:M,n\geq 1}$ i.i.d., Z-distributed;

- In fact it is a *SGD* since associated to

$$\widetilde{H}(y,\widetilde{z}) = \frac{1}{M}\sum_{k=1}^{M} H(y,\widetilde{z}_k), \quad \widetilde{z} = \begin{pmatrix}\widetilde{z}_1\\ \vdots \\ \widetilde{z}_M\end{pmatrix} \in (\mathbb{R}^q)^M$$

and $\widetilde{Z}^{(n)} = \begin{pmatrix}\widetilde{Z}_1^{(n)}\\ \vdots \\ \widetilde{Z}_M^{(n)}\end{pmatrix}$.

- It is clear that $\mathbb{E}\, H(y,\widetilde{Z}^{(1)}) = \nabla V(y)$.

A.s.convergence theorem

Theorem (Stochastic optimization: Stochastic Gradient Descent)

- *Let* $V : \mathbb{R}^d \to \mathbb{R}_+$ *be a differentiable function* $\lim_{|y|\to+\infty} V(y) = +\infty$, ∇V *Lipschitz,* $|\nabla V|^2 \le C(1+V)$ *and* $\{\nabla V = 0\} = \{y_*\}$.
- *Let* $h(y) = \nabla V(y) = \mathbb{E}\, H(y, Z)$ *with* H *s.t.* $\|H(y,Z)\|_2 \le C\sqrt{1+V(y)}$ *and that* $V(Y_0) \in L^1(\mathbb{P})$ *(and* $Y_0 \perp\!\!\!\perp (Z_n)_{n\ge1}$*).*
- *Assume* $(\gamma_n)_{n\ge1}$ *satisfies (DS).*

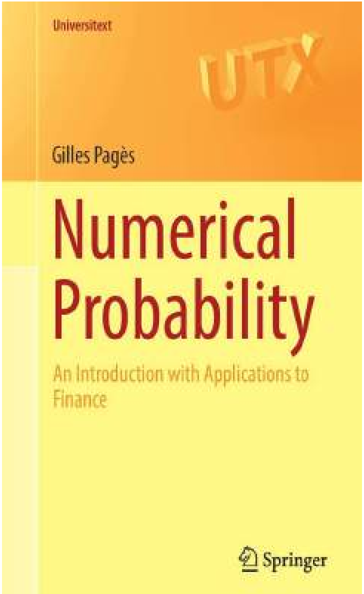
Then

$$V(Y_*) = \min_{\mathbb{R}^d} V \quad \text{and} \quad Y_n \xrightarrow{a.s.} y_* \quad \text{as} \quad n \to +\infty.$$

Moreover, $\nabla V(Y_n)$ *converges to* 0 *in every* L^p, $p \in (0,2)$ *(and* $\big(V(Y_n)\big)_{n\ge0}$ *is* L^1*-bounded so that* $\big(\nabla V(Y_n)\big)_{n\ge0}$ *is* L^2*-bounded).*

- **Remark !** If $H(y,z) = hy) = \nabla V(y)$: Convergence thm for Gradient descent (GD)!!

Aternative: Chapter 6 of . . .



Discussion I

- Practitioners often prefer *SGD* to *GD* up to adding noise to *GD*. Why ?
- Randomness induced by the "natural noise" of the martingale increments" $\rightsquigarrow$ better exploration of the state space.
- Randomness in *SGD* allows avoiding "traps" [Bradière-Duflo, Pemantle, Lazarev, Fort-Pagès, Benaïm in the 1980's] i.e.

$$\nabla V(y) = 0 \quad \text{and } \mathbb{E}\,|H(y,Z)|^2 > 0 \quad (\ldots)$$

- Note that Mini-batch implementation reduces these positive effects by its averaging effects
- Idea: add exogenous noise. How ?
- WARNING ! We will switch from $Y \rightsquigarrow \xi$ or $\bar{X}$ (in discrete time) and X (in continuous time) to make the connection with standard notations in stochastic calculus and SDE theory.

Discussion II:

- The continuous time counterpart of a *GD* $x_{n+1} = x_n - \gamma_{n+1}\nabla V(x_n)$ is the *ODE*

$$\dot{x}(t) = -\nabla V\big(x(t)\big), \quad x(0) \in \mathbb{R}^d.$$

- Thinking of the Gibbs measures

$$\pi_\sigma = C_\sigma e^{-V/\sigma^2}.\lambda_d \xrightarrow{w} \underset{\mathbb{R}^d}{\operatorname{argmin}} V \quad \text{as} \quad s \to 0$$

- and the fact that π_σ is the invariant measure of

$$dX(t) = -\nabla V(X(t))dt + \sigma\sqrt{2}dW(t), \quad X_0 \sim \pi_\sigma$$

- whose Euler scheme with (possibly) decreasing step γ_n reads with $\Gamma_n = \gamma_1 + \cdots + \gamma_n$.

$$\bar{X}_{\Gamma_{n+1}} = \bar{X}_{\Gamma_n} - \gamma_{n+1}\nabla V(\bar{X}_{\Gamma_n}) + \sigma\sqrt{2}\big(\Delta W_{n+1} := W_{\Gamma_{n+1}} - W_{\Gamma_n}\big)$$

- or, with the lighter notations $\bar{X}_n := \bar{X}_{\Gamma_n}$,

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}\nabla V(\bar{X}_n) + \sigma\sqrt{2}\underbrace{\sqrt{\gamma_{n+1}}\,\zeta_{n+1}}_{\text{exogenous noise}}, \ (\zeta_n)_n \ i.i.d. \ \sim \mathcal{N}(0, I_d).$$

Application to *PSLGD*

- If we apply the same treament to the *SLGD* we obtain

$$\xi_{n+1} = \xi_n - \gamma_{n+1} H(\xi_n, Z_{n+1}) + \sigma\sqrt{2}\big(W_{\Gamma_{n+1}} - W_{\Gamma_n}\big)$$

Application to *PSLGD*

- If we apply the same treament to the *SLGD* we obtain

$$\xi_{n+1} = \xi_n - \gamma_{n+1} H(\xi_n, Z_{n+1}) + \sigma\sqrt{2}\big(W_{\Gamma_{n+1}} - W_{\Gamma_n}\big)$$

- After a canonical decomposition

$$\xi_{n+1} = \xi_n - \gamma_{n+1}\nabla V(\xi_n) + \gamma_{n+1}\Delta M_{n+1} + \sigma\sqrt{2\,\gamma_{n+1}}\zeta_{n+1}.$$

where

- $\gamma_{n+1}\Delta M_{n+1}$ is the natural "noise" with variance $\simeq O(\gamma^2_{n+1})$.
- $\sigma\sqrt{2\,\gamma_{n+1}}\zeta_{n+1}$ is the exogenous "noise" with variance $\simeq 2d\sigma^2\gamma_{n+1})$.
- and

$$O(\gamma^2_{n+1}) = 2d\sigma^2 o\big(\gamma_{n+1}\big).$$

- The natural noise of the *SGD* is negligible w.r.t. the exogenous noise of the *PSLGD*.

Roadmap

1. Prove that we can "forget" the natural noise in the recursion i.e. if $(\bar{X}_n)_{n\geq 0}$ denotes the Euler scheme with steps γ_n of $(\mathcal{L})_\sigma$ starting from $\bar{X}_0 = \xi_0$:
$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}\nabla V(\bar{X}_n) + \sigma\sqrt{2}\big(W_{\Gamma_{n+1}} - W_{\Gamma_n}\big), \quad n \geq 0,\ \bar{X}_0 = \xi_0,$$
i.e. if $\xi_0 \in L^2(\mathbb{P})$, then
$$\big\|\xi_n - \bar{X}_n\big\|_{L^2(\mathbb{P})} \longrightarrow 0 \quad \text{as} \quad n \to +\infty \quad \text{(with a rate).}$$
2. As a preliminary step prove that both sequences are $L^2(\mathbb{P})$-bounded.
3. Let $X_0^{(\star,\sigma)} \sim \pi_\sigma$ so that $(X_t^{(\star,\sigma)})_{t\geq 0}$ is a stationary process. Prove
$$\big\|X_t^{(\star,\sigma)} - X_t^{(\xi_0,\sigma)}\big\|_{L^2(\mathbb{P})} \longrightarrow 0 \ \text{ as } \ n \to +\infty \text{ with a rate}$$
We will show that this rate of convergence depends on the regularity of V.
4. Prove that $\big\|\bar{X}_n - X_{\Gamma_n}^{(\xi_0,\sigma)}\big\|_{L^2(\mathbb{P})} \longrightarrow 0$ as $n \to +\infty$ with a rate.
5. Collecting all these results proves
$$\big\|\xi_n - X_{\Gamma_n}^{(\star,\sigma)}\big\|_{L^2(\mathbb{P})} \longrightarrow 0 \Longrightarrow \Big(\xi_n \xrightarrow{\mathcal{W}_2} \pi_\sigma\Big) \quad \text{as} \quad n \to 0 \text{ with a rate.}$$

Standing assumption

- We will assume in the rest of this section that the potential function V is α-convex in the sense that

$$\exists\, \alpha > 0 \text{ such that } V_\alpha(x) = V(x) - \frac{\alpha}{2}|x|^2 \text{ is convex.}$$

- The lemma below sums up he main consequences of this assumption. This assumption can be at least partially relaxed (see e.g. [2] or [7])
- as well as others . . . on σ.

Standing assumption

Lemma

Assume V is α-convex for some the $\alpha > 0$ and differentiable.

(a) *There exists a real constant $C_{\alpha,V} = V(0) - \frac{1}{2\alpha}|\nabla V(0)|^2$ such that*

$$\forall\, x \in \mathbb{R}^d, \quad V(x) \geq \frac{\alpha}{2}|x|^2 + C_{V,\alpha}.$$

In particular, for every $\sigma > 0$,

$$e^{-\frac{V}{\sigma^2}} \in L^1(\mathbb{R}^d, \lambda_d) \quad \text{and} \quad \int_{\mathbb{R}^d} |\xi|^2 \pi_\sigma(d\xi) < +\infty.$$

(b) *The vector field ∇V satisfies*

$$\forall\, x, y \in \mathbb{R}^d, \quad (\nabla V(x) - \nabla V(y)\,|\,x - y) \geq \alpha|x - y|^2$$

(c) *If furthermore ∇V is Lipschitz, then there exists $\alpha' > 0$ and $\beta' \in \mathbb{R}_+$ such that*

$$|\nabla V|^2 \geq (\alpha' V - \beta')^+.$$

Theorem (Forgetting the *SGLD*)

(a) *Assume* $V : \mathbb{R}^d \to \mathbb{R}_+$ *is* $\mathcal{C}^1$ *with a Lipschitz continuous gradient* ∇V *and* α*-convex. Assume that* $H : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^d$ *satisfies*

$$\forall \xi \in \mathbb{R}^d, \quad \mathbb{E}\, H(\xi, Z) = \nabla V(\xi) \quad \text{and} \quad \|H(\xi, Z)\|_2 \leq C(1 + V(\xi))^{\frac{1}{2}}, \tag{2}$$

that $(\gamma_n)_{n\geq 1}$ *is non-increasing and satisfies*

$$\sum_{n\geq 1} \gamma_n = +\infty \quad \text{and} \quad \gamma_n \searrow 0$$

and that $\mathbb{E}\, V(\xi_0) < +\infty$. *Then*

$$\sup_{n\geq 1} \mathbb{E}\big(V(\xi_n) + V(\bar{X}_n)\big) < +\infty \quad \text{and} \quad \big\|\xi_n - \bar{X}_n\big\|_2 \longrightarrow 0 \quad \text{as} \quad n \to +\infty$$

Remark. (2) implies by Jensen's inequality $|\nabla V(\xi)|^2 \leq C(1 + V(\xi))^{\frac{1}{2}}$ so that $V(\xi) = O(|\xi|^2)$. Combined with the Lemma(a)

$$|\nabla V(x)| \asymp |x| \quad \text{and} \quad V(\xi) \asymp |\xi|^2 \quad \text{and} \quad \forall\, p > 0, \int_{\mathbb{R}^d} |\xi|^p \pi_\sigma(d\xi) < +\infty.$$

Theorem (With rates)

(*b*) *If furthermore the sequence* $(\gamma_n)_{n\geq 1}$ *satisfies*

$$\varpi_1 = \limsup_n \frac{\gamma_n - \gamma_{n+1}}{\gamma_{n+1}^2} < 2\alpha$$

then

$$\|\xi_n - \bar{X}_n\|_2 = O(\sqrt{\gamma_n}).$$

In particular, when $\gamma_n = \frac{\gamma_1}{n^r}$, *Then* $\varpi_1 < 2\alpha$ *iff* $(0 < r < 1)$ *or* $(r = 1$ *and* $\gamma_1 > \frac{1}{2\alpha})$.

(*c*) *When* $\mathbb{E}\, V(\xi_0)^2 < +\infty$, *then one also has (for the future)*

$$\sup_{n\geq 1} \mathbb{E}\big[(\bar{X}_n)^2 + \mathbb{E}\,|\nabla V(\bar{X}_n)|^4\big] < +\infty.$$

Lemma (Magic Step Lemma)

Let $p \geq 1$ *and let* $(\gamma_n)_{n\geq 1}$ *be a non-increasing positive sequence s.t.*

$$\varpi_p = \limsup_n \frac{\gamma_n^p - \gamma_{n+1}^p}{\gamma_{n+1}^{p+1}} < +\infty.$$

(i) *Let* $\varrho > \varpi_p$ *and let*

$$u_n = e^{-\varrho \Gamma_n} \sum_{k=1}^{n} \gamma_k^{p+1} e^{\varrho \Gamma_k}, \quad n \geq 0.$$

Then, $$u_n = O(\gamma_n^p).$$

(ii) *Moreover, if for any* $a < p\frac{\varrho}{\varpi_p}$,

$$e^{-\varrho \Gamma_n} = o(\gamma_n^a).$$

Proof of the lemma

Set $\widetilde{u}_n = \frac{u_n}{\gamma_n^p}$, $n \geq 1$. We have:

$$\widetilde{u}_{n+1} = \widetilde{u}_n \theta_n + \gamma_{n+1} \quad \text{with} \quad \theta_n = \left(\frac{\gamma_n}{\gamma_{n+1}}\right)^p e^{-\varrho\gamma_{n+1}}.$$

Under the assumption, there exists $c \in (\varpi_p, \varrho)$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^p \leq 1 + c\gamma_{n+1} \leq e^{c\gamma_{n+1}}.$$

Thus, for $n \geq n_0$, $\theta_n \leq e^{-(\varrho - c)}\gamma_{n+1}$ so that plugging this inequality into the above one, we deduce

$$\widetilde{u}_{n+1} \leq \widetilde{u}_n e^{-(\varrho-c)}\gamma_{n+1} + \gamma_{n+1}$$

or, equivalently,

$$e^{(\varrho-c)\Gamma_{n+1}}\widetilde{u}_{n+1} \leq e^{(\varrho-c)\Gamma_n}\widetilde{u}_n + C' e^{(\varrho-c)\Gamma_n}\gamma_{n+1}$$

where $C' = \sup_{k\geq 1} e^{(\varrho-c)\gamma_k}$. Hence, by induction, for every $n \geq n_0$,

$$e^{(\varrho-c)\Gamma_n}\widetilde{u}_n \leq e^{(\varrho-c)\Gamma_{n_0}}\widetilde{u}_{n_0} + C'\int_{\Gamma_{n_0}}^{\Gamma_n} e^{(\varrho-c)u}du \leq e^{(\varrho-c)\Gamma_{n_0}}\widetilde{u}_{n_0} + \frac{e^{(\varrho-c)\Gamma_n} - e^{(\varrho-c)\Gamma_{n_0}}}{\varrho - c}$$

so that $\widetilde{u}_n \leq \widetilde{u}_{n_0} + \frac{1}{\varrho-c}$ for $n \geq n_0$ which clearly implies the announced result.

(ii) Set $v_n = e^{-\varrho\Gamma_n}\gamma_n^{-a}$, $n \geq 1$. Let $\eta \in \left(0, \frac{p\varrho}{a} - \varpi_p\right)$. For large enough n, say $n \geq n_1$, $\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^p \leq 1 + (\varpi + \eta)\gamma_{n+1}$ so that

$$v_{n+1} = \left(\frac{\gamma_{n+1}}{\gamma_n}\right)^a e^{-\varrho\gamma_{n+1}} v_n \leq \left(1 + (\varpi_p + \eta)\gamma_{n+1}\right)^{\frac{a}{p}} e^{-\varrho\gamma_{n+1}} v_n \leq e^{-c\gamma_{n+1}} v_n$$

where $c = \varrho - \frac{a}{p}(\varpi + \eta) > 0$. Consequently $v_n \to 0$ as $n \to +\infty$ since $\sum_n \gamma_n = +\infty$. $\square$

L^2-boundedness of (ξ_n) and $(\bar{X}_n)$

- Denote $V_n = V(\xi_n)$ and $\nabla V_n = \nabla V(\xi_n)$. As ∇V is Lipschitz

$$V_{n+1} \le V_n - \gamma_{n+1}\big(\nabla V_n \,|\, H(\xi_n, Z_{n+1})\big) + \sigma\sqrt{2}\,(\nabla V_n \,|\, \Delta W_{\Gamma_{n+1}})$$
$$+ [\nabla V]_{\rm Lip}|\gamma_{n+1} H(\xi_n, Z_{n+1}) + \sigma\sqrt{2}\Delta W_{\Gamma_{n+1}}|^2.$$

- By induction that V_n is $\mathcal{F}_n$-adapted and integrable since $\mathbb{E}\, V(\xi_0) < +\infty$.
- Taking conditional expectations w.r.t. $\mathcal{F}_n$ yields

$$\mathbb{E}_n(V_{n+1} \,|\, \mathcal{F}_n) \le V_n - \gamma_{n+1}(\nabla V_n \,|\, \nabla V_n) + [\nabla V]_{\rm Lip}(\gamma^2_{n+1}\mathbb{E}_n|H(\xi_n, Z_{n+1})|^2 + 2d\sigma^2\gamma_{n+1}$$

since $\mathbb{E}_n H(\xi_n, Z_{n+1}) = \nabla V(\xi_n) = V_n$ owing to the independence of Z_{n+1} and $\mathcal{F}_n$, $\mathbb{E}_n \Delta W_{\Gamma_{n+1}} = 0$ and $\mathbb{E}_n|\Delta W_{\Gamma_{n+1}}|^2 = d\gamma_{n+1}$

- Note that, still owing to Z_{n+1} the independence of Z_{n+1} and $\mathcal{F}_n$ and $Z_{n+1} \overset{d}{=} Z$, we have

$$\mathbb{E}_n|H(\xi_n, Z_{n+1})|^2 = \Big[\mathbb{E}\,|H(\xi, Z)|^2\Big]_{|\xi=\xi_n} \le C(1 + V_n).$$

L^2-boundedness of (ξ_n) and $(\bar{X}_n)$

- It follows from the Lemma(b) that $-|\nabla V|^2 \leq \beta' - \alpha' V$. Consequently,

$$\begin{aligned}\mathbb{E}_n(V_{n+1}\,|\,\mathcal{F}_n) &\leq V_n + \gamma_{n+1}(\beta' - \alpha' V_n) + C_{d,V}(\gamma_{n+1}^2 V_n + \gamma_{n+1}^2 + \gamma_{n+1})\\ &= V_n(1-\alpha'\gamma_{n+1} + C_{d,V}\gamma_{n+1}^2) + \gamma_{n+1}(C_{d,V} + \beta' + C_{d,V}\gamma_{n+1}.)\end{aligned}$$

- As $\gamma_n \to 0$, for every $n \geq n_0$, $\gamma_{n+1} \leq \frac{\alpha'}{2C_{d,V}}$ so that

$$\mathbb{E}_n(V_{n+1}\,|\,\mathcal{F}_n) \leq V_n(1-\tfrac{\alpha'}{2}\gamma_{n+1}) + C'_{d,V}\gamma_{n+1}.$$

- Taking expectation yields, for every $n \geq n_0$

$$\mathbb{E}\,V_{n+1} \leq \mathbb{E}\,V_n(1-\tfrac{\alpha'}{2}\gamma_{n+1}) + C'_{d,V}\gamma_{n+1}.$$

which in turn implies by induction that by induction on n that

$$\sup_{n\geq 1}\mathbb{E}V_{n+1} \leq \max\Big(\max_{k=1,\ldots,n_0}\mathbb{E}\,V_k, \frac{2C'_{d,V}}{\alpha'}\Big).$$

L^2-boundedness of (ξ_n) and $(\bar{X}_n)$

- Proving the same for the Euler scheme $(\bar{X}_n)_{n\geq 0}$ (starting from ξ_0 as well) is for free : reproduce the above proof when $H(\xi, z) = \nabla V(\xi)$.

$L^2(\mathbb{P})$-convergence

- For every $n \geq 0$,

$$\xi_{n+1} - \bar{X}_{n+1} = \xi_n - \bar{X}_n - \gamma_{n+1}(H(\xi_n, Z_{n+1}) - \nabla V(\bar{X}_n)) + 0 \text{ !!}$$
$$= \xi_n - \bar{X}_n \gamma_{n+1}(\nabla V(\xi_n) - \nabla V(\bar{X}_n)) + \gamma_{n+1}(\nabla V(\xi_n) - H(\xi_n, Z_{n+1})).$$

- Consequently,

$$|\xi_{n+1} - \bar{X}_{n+1}|^2 = |\xi_n - \bar{X}_n|^2 - 2\gamma_{n+1}\big(\xi_n - \bar{X}_n \,|\, H(\xi_n, Z_{n+1}) - \nabla V(\bar{X}_n)\big) + \gamma_{n+1}^2 |H(\xi_n, Z_{n+1}) - \nabla V(\bar{X}_n)|^2$$
$$|\xi_{n+1} - \bar{X}_{n+1}|^2 = |\xi_n - \bar{X}_n|^2 - 2\gamma_{n+1}\big(\xi_n - \bar{X}_n \,|\, \nabla V(\xi_n) - \nabla V(\bar{X}_n)\big) + \gamma_{n+1}^2 |H(\xi_n, Z_{n+1}) - \nabla V(\bar{X}_n)|^2 + 2\gamma_{n+1}\big(\xi_n - \bar{X}_n \,|\, \nabla V(\xi_n) - H(\xi_n, Z_{n+1})\big).$$

- Taking conditional expectation $\mathbb{E}_n$ (given $\mathcal{F}_n$) implies

$$\mathbb{E}_n|\xi_{n+1} - \bar{X}_{n+1}|^2 \leq |\xi_n - \bar{X}_n|^2 - 2\gamma_{n+1}\big(\xi_n - \bar{X}_n \,|\, \nabla V(\xi_n) - \nabla V(\bar{X}_n)\big) + 2\gamma_{n+1}^2\big(\mathbb{E}_n|H(\xi_n, Z_{n+1})|^2 + |\nabla V(\bar{X}_n)|^2\big)$$

since $\mathbb{E}_n H(\xi_n, Z_{n+1}) = \nabla V(\xi_n)$ and $\xi_n - \bar{X}_n$ is $\mathcal{F}_n$-measurable.

- The function V being α-convex, we know that $(\xi_n - \bar{X}_n \,|\, \nabla V(\xi_n) - \nabla V(\bar{X}_n)) \geq \alpha|\xi_n - \bar{X}_n|^2$ so that

$$\begin{aligned}\mathbb{E}_n|\xi_{n+1} - \bar{X}_{n+1}|^2 &\leq |\xi_n - \bar{X}_n|^2(1 - 2\alpha\gamma_{n+1}) + 2\gamma_{n+1}^2\big(\mathbb{E}_n|H(\xi_n, Z_{n+1})|^2 + |\nabla V(\bar{X}_n)|^2\big)\\ &\leq |\xi_n - \bar{X}_n|^2(1 - 2\alpha\gamma_{n+1}) + C_V\gamma_{n+1}^2(1 + V_n)\end{aligned}$$

- Consequently, as (V_n) is $L^2(\mathbb{P})$-bounded by Step 1, we derive that

$$\mathbb{E}\,|\xi_{n+1} - \bar{X}_{n+1}|^2 \leq \mathbb{E}\,|\xi_n - \bar{X}_n|^2(1 - 2\alpha\gamma_{n+1}) + C'_V\gamma_{n+1}^2$$

for some positive constant C'_V. If we set $\Gamma_n = \gamma_1 + \cdots + \gamma_n$, $n \geq 1$ then one shows by induction that, for every $n \geq 0$,

$$e^{2\alpha\Gamma_n}\mathbb{E}|\xi_n - \bar{X}_n|^2 \leq C'_V \sum_{k=1}^{n} e^{2\alpha\Gamma_k}\gamma_k^2.$$

$$\text{i.e.}\quad \mathbb{E}|\xi_n - \bar{X}_n|^2 \leq C'_V e^{-2\alpha\Gamma_n}\sum_{k=1}^{n} e^{2\alpha\Gamma_k}\gamma_k^2$$

$$\simeq C'_V e^{-2\alpha\Gamma_n}\int_0^{\Gamma_n} e^{2\alpha s}\underbrace{\gamma_{N(s)}}_{\to 0}\,ds \xrightarrow{\text{(Césaro)}} 0 \quad \text{as} \quad n \to +\infty \quad (3)$$

where $N(t) = k$ if $\Gamma_k \leq t < \Gamma_{k+1}$.

Proof of (b)

(b) It follows from the Lemma(a) applied with $p = 1$ that, under $(\varpi_1 < 2\alpha)$, (3) implies $\mathbb{E}|\xi_n - \bar{X}_n|^2 = O(\gamma_n)$.

Proof of (c)

- Set $\bar V_n = V(\bar X_n)$ and and $\nabla \bar V_n = \nabla V(\bar X_n)$ for convenience.
- Revisiting the computations performed for (a) with $(\xi_n)_n$ leads to

$$0 \le \bar V_{n+1} \le \bar V_n(1-\alpha'\gamma_{n+1} + C_V\gamma_{n+1}^2) + C_{V,\beta',\sigma}\gamma_{n+1}^2 + \sigma C'_V\big(\nabla \bar V_n \,|\, \Delta W_{\Gamma_{n+1}}\big).$$

  where $|\nabla V|^2 \ge \alpha' V - \beta'$.
- Consequently

$$0 \le \bar V_{n+1}^2 \le \bar V_n^2(1-\alpha'\gamma_{n+1} + C_V\gamma_{n+1}^2)^2 + \Big(C_{V,\beta',\sigma}\gamma_{n+1}^2 + \sigma C'_V\big(\nabla \bar V_n \,|\, \Delta W_{\Gamma_{n+1}}\big)\Big)^2$$
$$+ 2\bar V_n(1-\alpha'\gamma_{n+1} + C_V\gamma_{n+1}^2)\Big(C_{V,\beta',\sigma}\gamma_{n+1}^2 + \sigma C'_V\big(\nabla \bar V_n \,|\, \Delta W_{\Gamma_{n+1}}\big)\Big).$$

- One easily checks by induction that $\mathbb{E}\,\bar V_n^2 < +\infty$ for every $n \ge 0$ since $\mathbb{E}\,\bar V_0^2 = \mathbb{E}\,V(\xi_0)^2 < +\infty$.
- Taking conditional expectation w.r.t. to $\mathcal{F}_n$, yields

$$0 \le \mathbb{E}_n \bar V_{n+1}^2 \le \bar V_n^2(1-\alpha'\gamma_{n+1} + C_V\gamma_{n+1}^2)^2 + C_{V,\beta',\sigma}^2\gamma_{n+1}^2 + \sigma^2|\nabla \bar V_n|^2 d\gamma_{n+1}$$
$$+ 2C_{V,\beta',\sigma}\gamma_{n+1}^2\bar V_n(1-\alpha'\gamma_{n+1} + C_V\gamma_{n+1}^2).$$

Proof of (c)

- Using that $\sup_{n\geq 0}\mathbb{E}\,|\nabla V(\bar X_n)|^2 < +\infty$ we derive that there exists $n_1 \geq 1$ and a positive constant $\widetilde C = C_{V,\nabla V,\beta'\sigma}\gamma_{n+1}$ such that for every $n\geq n_1$, $1-\alpha'\gamma_{n+1}>0$ and

$$\begin{aligned}\mathbb{E}\,\bar V_{n+1}^2 &\leq \mathbb{E}\,\bar V_n^2\big(1-\tfrac{\alpha'}{2}\gamma_{n+1}\big)^2+\widetilde C\gamma_{n+1}\\ &\leq \mathbb{E}\,\bar V_n^2\big(1-\tfrac{\alpha'}{2}\gamma_{n+1}\big)+\widetilde C\gamma_{n+1}.\end{aligned}$$

- One concludes like in the first step of the proof of Claim (a) that

$$\sup_{n\geq 0}\mathbb{E}\,\bar V_n^2 \leq \max\Big(\max_{k=0,\ldots,n_1}\mathbb{E}\,\bar V_k^2,\frac{2\widetilde C}{\alpha'}\Big).$$

What is left to be done ?

- Compare the solution $X^{\xi_0} = (X_t^{\xi_0})_{t\geq 0}$ of the $(\mathcal{L})_\sigma$ equation starting from $\xi_0 \in L^2(\mathbb{P})$ with the stationary solution $X^{(\star,\sigma)} = (X_t^{(\star,\sigma)})_{t\geq 0}$ starting from $X_0^{(\star,\sigma)} \stackrel{d}{=} \pi_\sigma$ in terms of $L^2(\mathbb{P})$-confluence.
- Compare X^{ξ_0} with its Euler scheme $(\bar{X}_t^{\xi_0})_{t\geq 0}$ in terms of $L^2(\mathbb{P})$-confluence.
- The second task is more demanding, let us start by the first one.

Proposition

Assume V is α-convex and ∇V is Lipschitz continuous. Let $X^x = (X^x_t)_{t\geq 0}$ denote the solution of $(\mathcal{L}_\sigma)$ starting from $X^x_0 = x$.

(a) For every $x,\ y \in R^d$ and every $t \geq 0$,

$$\mathcal{W}_2^2([X^x_t],[X^y_t]) \leq \mathbb{E}\,|X^x_t - X^y_t|^2 \leq e^{-2\alpha t}|x-y|^2.$$

(b) If $\xi_0,\ \xi'_0 \in L^2(\mathbb{P})$, $\perp\!\!\!\perp W$, then (with obvious notations)

$$\mathbb{E}\,|X^{\xi_0}_t - X^{\xi'_0}_t|^2 \leq e^{-2\alpha t}\mathbb{E}\,|\xi_0 - \xi'_0|^2.$$

(c) If $\int_{\mathbb{R}^d} |\xi|^2 \pi_\sigma(d\xi) = \int_{\mathbb{R}^d} |\xi|^2 e^{-\frac{V(\xi)}{\sigma^2}} d\xi < +\infty$ and if $\xi_0^{(\star,\sigma)} \stackrel{d}{=} \pi_\sigma$ then $X^{(\star,\sigma)}$, solution to $(\mathcal{L}_\sigma)$ starting from $\xi_0^{(\star,\sigma)}$, $\perp\!\!\!\perp W$, is a stationary process and, for every $t \geq 0$, $X_t^{(\star,\sigma)} \stackrel{d}{=} \pi_\sigma$ so that

$$\mathcal{W}_2^2([X^{\xi_0}_t],\pi_\sigma) \leq \mathbb{E}\,|X^{\xi_0}_t - X^{(\star,\sigma)}_t|^2 \leq e^{-2\alpha t}\mathbb{E}\,|\xi_0 - \xi_0^{(\star,\sigma)}|^2.$$

Proof (Stochastic processes without stochastic calculus !).

(a) One has

$$X_t^x - X_t^y = x - y - \int_0^t \big(\nabla V(X_s^x) - \nabla V(X_s^y)\big)\,ds + 0 \;!!$$

so that $\langle X^x - X^y\rangle_t \equiv 0$. Itô's formula yields

$$e^{2\alpha t}|X_t^x - X_t^y|^2 = |x-y|^2 + \int_0^t e^{2\alpha s}2\alpha|X_s^x - X_s^y|^2 ds + \int_0^t \big(X_s^x - X_s^y \,|\, d(X_s^x - X_s^y)\big)$$

$$= |x-y|^2 + 2\int_0^t e^{2\alpha s}\big(\underbrace{\alpha|X_s^x - X_s^y|^2 - (X_s^x - X_s^y \,|\, \nabla V(X_s^x) - \nabla V(X_s^y)}_{\le 0 \text{ by } \alpha\text{-convexity of } V}\big)\,ds$$

for every $t \ge 0$, so that, as a non-negative and non-increasing process,

$$0 \le e^{2\alpha t}|X_t^x - X_t^y|^2 \longrightarrow \Xi_\infty^{x,y} \le |x-y|^2 \quad \text{as} \quad t \to +\infty.$$

In particular

$$\forall\, t \ge 0, \quad \mathbb{E}\,|X_t^x - X_t^y|^2 \le e^{-2\alpha t}|x-y|^2.$$

(b) The same reasoning works when replacing x and y by ξ_0 and ξ_0' which yields

$$0 \le e^{2\alpha t}|X_t^{\xi_0} - X_t^{\xi_0'}|^2 \longrightarrow \Xi_\infty^{\xi_0,\xi_0'} \le |\xi_0 - \xi_0'|^2 \in L^1(\mathbb{P}) \quad \text{as} \quad t \to +\infty$$

which in turn implies

$$\forall\, t \ge 0, \quad \mathbb{E}\,|X_t^{\xi_0} - X_t^{\xi_0'}|^2 \le e^{-2\alpha t}\mathbb{E}\,|\xi_0 - \xi_0'|^2.$$

(c) is obvious. $\square$

The final countdown

- Le us introduce the *genuine* continuous time Euler scheme with positive non-increasing step $(\gamma_n)_{n\geq 1}$ of $(\mathcal{L}_\sigma)$ starting from $\xi_0 \in L^2(\mathbb{P})$).

$$\bar{X}_t = \xi_0 - \int_0^t \nabla V(\bar{X}_{\underline{s}})ds + \sigma\sqrt{2}\,W_t$$

where $\underline{t} = \Gamma_n$ if $t \in [\Gamma_n, \Gamma_{n+1})$.

- Then

$$\forall\, t \geq 0, \quad X_t - \bar{X}_t = -\int_0^t \big(\nabla V(X_s) - \nabla V(\bar{X}_{\underline{s}})\big)\,ds.$$

Theorem (Panloup-P. '23 *AAP*, Panloup-Égéa '24, P. '24)

(a) Standard setting. *Assume that $V:\mathbb{R}^d\to\mathbb{R}_+$ is α-convex, $\mathcal{C}^1$, ∇V is Lipschitz. If the step sequence $(\gamma_n)_{n\geq 1}$ is positive, non-increasing and such that*

$$\varpi_1 := \overline{\lim_n}\,\frac{\gamma_n-\gamma_{n+1}}{\gamma_{n+1}^2} < 2\alpha,$$

then,

$$\forall\, n\geq 1,\quad \Big\|\sup_{t\in<[\Gamma_{n-1},\Gamma_n]}\big|X_t^{\xi_0}-\bar X_t^{\xi_0}\big|\Big\|_2 \leq C\sqrt{\gamma_n}.$$

If $\gamma_n=\frac{\gamma_1}{n^r}$, $\varpi_1<2\alpha$ iff $(0<r<1)$ or $(r=1$ and $\gamma_1>\frac{1}{2\alpha})$.

(b) Smoother setting. *Moreover, assume V is $\mathcal{C}^2$ with bounded existing partial derivatives and a Lipschitz continuous Hessian $\nabla^2 V$ and $\xi_0\in L^4(\mathbb{P})$. If*

$$\varpi_2 := \overline{\lim_n}\,\frac{\gamma_n^2-\gamma_{n+1}^2}{\gamma_{n+1}^3} < 2\alpha.$$

then

$$\forall\, n\geq 0,\quad \big\|X_{\Gamma_n}^{\xi_0}-\bar X_{\Gamma_n}^{\xi_0}\big\|_2\leq C\gamma_n.$$

If $\gamma_n=\frac{\gamma_1}{n^r}$, if $\varpi_2<2\alpha$ iff $(0<r<1)$ or $(r=1$ and $\gamma_1>\frac{1}{\alpha})$.

Proof of (a)

- We again start by Itô's formula.
- Let $\tilde{\alpha} = \alpha - \varepsilon$ where ε is small enough so that $\frac{1}{2}\varpi_1 < \bar{\alpha} < \alpha$. Then

$$
\begin{aligned}
e^{2\tilde{\alpha}t}|X_t - \bar{X}_t|^2 &= \int_0^t e^{2\tilde{\alpha}s}\big(2\tilde{\alpha}|X_s - \bar{X}_s|^2 - 2(X_s - \bar{X}_s \,|\, \nabla V(X_s) - \nabla V(\bar{X}_{\underline{s}})\big)ds \\
&= 2\int_0^t e^{2\tilde{\alpha}s}\big(\tilde{\alpha}|X_s - \bar{X}_s|^2 - (X_s - \bar{X}_s \,|\, \nabla V(X_s) - \nabla V(\bar{X}_s)\big)ds \\
&\quad + 2\int_0^t e^{2\tilde{\alpha}s} R(s)ds \\
&\leq 2(\tilde{\alpha} - \alpha)\int_0^t e^{2\tilde{\alpha}s}|X_s - \bar{X}_s|^2 ds + 2\int_0^t e^{2\tilde{\alpha}s} R(s)ds
\end{aligned}
$$

with

$$
R(t) = -\big(X_s - \bar{X}_s \,|\, \nabla V(\bar{X}_s) - \nabla V(\bar{X}_{\underline{s}})\big), \quad t \geq 0.
$$

- Hence, for every $t \geq 0$,

$$
|X_t - \bar{X}_t|^2 \leq 2e^{2-\tilde{\alpha}t}(\tilde{\alpha} - \alpha)\int_0^t e^{2\tilde{\alpha}s}|X_s - \bar{X}_s|^2 ds + 2e^{-2\tilde{\alpha}t}\int_0^t e^{2\tilde{\alpha}s} R(s)ds
$$

Proof of (a)

- By Young's inequality, for every $t \geq 0$

$$|R(t)| \leq \left|X_t - \bar{X}_t\right| \left|\nabla V(X_{\underline{t}}) - \nabla V(\bar{X}_t)\right| \leq \frac{\varepsilon}{2}\left|X_t - \bar{X}_t\right|^2 + \frac{[\nabla V]_{\rm Lip}}{2\varepsilon}\left|\bar{X}_t - \bar{X}_{\underline{t}}\right|^2$$

- so that

$$\begin{aligned}|X_t - \bar{X}_t|^2 &\leq 2(\tilde{\alpha} + \tfrac{\varepsilon}{2} - \alpha)e^{-2\tilde{\alpha}t}\int_0^t e^{2\tilde{\alpha}s}|X_s - \bar{X}_s|^2 ds + \frac{[\nabla V]_{\rm Lip}}{\varepsilon}e^{-2\tilde{\alpha}t}\int_0^t e^{2\tilde{\alpha}s}\left|\bar{X}_s - \bar{X}_{\underline{s}}\right|^2 ds\\ &= \frac{[\nabla V]_{\rm Lip}}{\varepsilon}e^{-2\tilde{\alpha}t}\int_0^t e^{2\tilde{\alpha}s}\left|\bar{X}_s - \bar{X}_{\underline{s}}\right|^2 ds\end{aligned}$$

since $\tilde{\alpha} < \alpha - \frac{\varepsilon}{2}$. Set $\bar{\gamma} = \sup_{k\geq 1}\gamma_k$. One has, for $t \in [\Gamma_{n-1}, \Gamma_n]$, $n \geq 1$,

$$\sup_{t\in[\Gamma_{n-1},\Gamma_n]}|X_t - \bar{X}_t|^2 \leq e^{2\tilde{\alpha}\bar{\gamma}}\frac{[\nabla V]_{\rm Lip}}{\varepsilon}e^{-2\tilde{\alpha}\Gamma_n}\int_0^{\Gamma_n} e^{2\tilde{\alpha}s}\left|\bar{X}_s - \bar{X}_{\underline{s}}\right|^2 ds.$$

Proof of (a)

- By Young's inequality, for every $t \geq 0$

$$|R(t)| \leq \left|X_t - \bar{X}_t\right| \left|\nabla V(X_{\underline{t}}) - \nabla V(\bar{X}_t)\right| \leq \frac{\varepsilon}{2}\left|X_t - \bar{X}_t\right|^2 + \frac{[\nabla V]_{\mathrm{Lip}}}{2\varepsilon}\left|\bar{X}_t - \bar{X}_{\underline{t}}\right|^2$$

- so that

$$\begin{aligned}|X_t - \bar{X}_t|^2 &\leq 2(\tilde{\alpha} + \tfrac{\varepsilon}{2} - \alpha)e^{-2\tilde{\alpha}t}\int_0^t e^{2\tilde{\alpha}s}|X_s - \bar{X}_s|^2 ds + \frac{[\nabla V]_{\mathrm{Lip}}}{\varepsilon}e^{-2\tilde{\alpha}t}\int_0^t e^{2\tilde{\alpha}s}\left|\bar{X}_s - \bar{X}_{\underline{s}}\right|^2 ds\\ &= \frac{[\nabla V]_{\mathrm{Lip}}}{\varepsilon}e^{-2\tilde{\alpha}t}\int_0^t e^{2\tilde{\alpha}s}\left|\bar{X}_s - \bar{X}_{\underline{s}}\right|^2 ds\end{aligned}$$

since $\tilde{\alpha} = \alpha - \frac{\varepsilon}{2}$. Set $\bar{\gamma} = \sup_{k\geq 1}\gamma_k$. One has, for $t = \Gamma_n$, $n \geq 1$,

$$(\star\star) \quad \mathbb{E}\sup_{t\in(\Gamma_{n-1},\Gamma_n]}|X_t - \bar{X}_t|^2 \leq e^{2\tilde{\alpha}\bar{\gamma}}\frac{[\nabla V]_{\mathrm{Lip}}}{\varepsilon}e^{-2\tilde{\alpha}\Gamma_n}\int_0^{\Gamma_n} e^{2\tilde{\alpha}s}\mathbb{E}\left|\bar{X}_s - \bar{X}_{\underline{s}}\right|^2 ds.$$

Proof of (a) (end)

- Now

$$\bar{X}_s - \bar{X}_{\underline{s}} = -(s-\underline{s})\nabla V(\bar{X}_{\underline{s}}) + \sigma\sqrt{2}(W_s - W_{\underline{s}}) \quad \text{for every } s \geq 0$$

- so that by the first step

$$\mathbb{E}\,|\bar{X}_s - \bar{X}_{\underline{s}}|^2 \leq (s-\underline{s})^2 \sup_{u\geq 0}\mathbb{E}\,|\nabla V(\bar{X}_u)|^2 + d\sigma\sqrt{2}(s-\underline{s})$$

$$\leq (s-\underline{s})\Big((s-\underline{s})C_V\big(1+\sup_{u\geq 0}\mathbb{E}\,V(\bar{X}_u)\big) + d\sigma\sqrt{2}\Big) \leq C_{V,\bar{\gamma},d}(s-\underline{s}).$$

- Inserting this in $(\star\star)$ yields

$$\mathbb{E}\sup_{t\in(\Gamma_{n-1},\Gamma_n]}|X_t - \bar{X}_t|^2 \leq \widetilde{C}_{V,\bar{\gamma},\varepsilon}\, e^{-2\tilde{\alpha}\Gamma_n}\int_0^{\Gamma_n} e^{2\tilde{\alpha}s}(s-\underline{s})ds$$

$$\leq \widetilde{C}_{V,\bar{\gamma},\varepsilon}\, e^{-2\tilde{\alpha}\Gamma_n}\sum_{k=1}^{n}\int_{\Gamma_{k-1}}^{\Gamma_k} e^{2\tilde{\alpha}s}(s-\underline{s})ds$$

$$\leq \widetilde{C}_{V,\bar{\gamma},\varepsilon}\, e^{-2\tilde{\alpha}\Gamma_n}\sum_{k=1}^{n} e^{2\tilde{\alpha}\Gamma_k}\gamma_k^2 = O(\gamma_n).$$

owing to the Magic step Lemma applied with $p=1$. □

Proof of (b) = revisiting $R(t)$

- Let $\nabla^2 V$ the Hessian of V. First order Taylor formula to ∇V between $\bar{X}_{\underline{s}}$ and $\bar{X}_s$ yields

$$-R(t) = \big(X_s - \bar{X}_s \,|\, \nabla V(\bar{X}_s) - \nabla V(\bar{X}_{\underline{s}})\big) = \big(X_s - \bar{X}_s \,|\, \nabla^2 V(\bar{X}_{\underline{s}})(\bar{X}_s - \bar{X}_{\underline{s}})\big)$$
$$+ \underbrace{\int_0^1 (X_s - \bar{X}_s)^* \big(\nabla^2 V(\bar{X}_{\underline{s}} + u(\bar{X}_s - \bar{X}_{\underline{s}})) - \nabla^2 V(\bar{X}_{\underline{s}})\big)(\bar{X}_s - \bar{X}_{\underline{s}})du}_{=:(3)_s}$$

- We replace now $\bar{X}_s - \bar{X}_{\underline{s}}$ by its value in the first term

$$\big(X_s - \bar{X}_s \,|\, \nabla^2 V(\bar{X}_{\underline{s}})(\bar{X}_s - \bar{X}_{\underline{s}})\big) = -\underbrace{\big(X_s - \bar{X}_s \,|\, \nabla^2 V(\bar{X}_{\underline{s}})\nabla V(\bar{X}_{\underline{s}})\big)(s - \underline{s})}_{=:(1)_s}$$
$$+ \sigma \underbrace{\big(X_s - \bar{X}_s \,|\, \nabla^2 V(\bar{X}_{\underline{s}})(W_s - W_{\underline{s}})\big)}_{=:(2)_s}.$$

- Now, let us inspect these three terms to bound their expectations.

Proof of (b) : Term $(1)_s$ (easy)

- The first term $(1)_s$ can be upper-bounded by Young's inequality

$$|(1)_s| \leq \frac{\varepsilon}{4}|X_s - \bar{X}_s|^2 + \frac{\|\nabla^2 V\|_{F,\sup}}{\varepsilon}|\nabla V(\bar{X}_{\underline{s}})|^2(s-\underline{s})^2$$

so that

$$|\mathbb{E}\,(1)_s| \leq \mathbb{E}\,|(1)_s| \leq \frac{\varepsilon}{4}\mathbb{E}\,|X_s - \bar{X}_s|^2 + \frac{1}{\varepsilon}\|\nabla^2 V\|_{F,\sup}\sup_{n\geq 0}\mathbb{E}\,|\nabla V(\bar{X}_{\Gamma_n})|^2(s-\underline{s})^2$$

- where $\|\nabla^2 V\|_{F,\sup} = \sup_{x!\in\mathbb{R}^d}\|\nabla^2 V(x)\|$ (Fröbenius norm)
- and $\sup_{n\geq 0}\mathbb{E}\,|\nabla V(\bar{X}_{\Gamma_n})|^2 < +\infty$ since $|\nabla V|^2 \leq C(1+V)$.

Proof of (b) : Term $(3)_s$ (easy but needs 4th moment !)

- Again by Young's inequality, one shows for $(3)_s$ that

$$
\begin{aligned}
|\mathbb{E}\,(3)_s| &\leq [\nabla^2 V]_{\mathrm{Lip}}\mathbb{E}\,|X_s - \bar{X}_s|\,|\bar{X}_s - \bar{X}_{\underline{s}}|^2 \\
&\leq \frac{\varepsilon}{4}\mathbb{E}\,|X_s - \bar{X}_s|^2 + \frac{[\nabla^2 V]^2_{\mathrm{Lip}}}{\varepsilon}\,\mathbb{E}\,|\bar{X}_s - \bar{X}_{\underline{s}}|^4.
\end{aligned}
$$

- It straightforwardly follows that

$$
\mathbb{E}\,|\bar{X}_s - \bar{X}_{\underline{s}}|^4 \leq (s-\underline{s})^2\mathbb{E}\,|\nabla V(\bar{X}_{\underline{s}})|^4 + 2d(d+2)\sigma^2(s-\underline{s})^2.
$$

- As $\xi_0 \in L^4(\mathbb{P})$

$$
\sup_{n\geq 0}\mathbb{E}\,V(\bar{X}_{\Gamma_n})^2 < +\infty
$$

  so that $\sup_{u\geq 0}\mathbb{E}\,|\nabla V(\bar{X}_{\underline{u}})|^4 \leq C\big(1+\sup_{n\geq 0}\mathbb{E}\,V(\bar{X}_{\Gamma_n})^2\big) < +\infty$ owing to the former Theorem(b).

- Finally his in turn implies

$$
\mathbb{E}\,|\bar{X}_s - \bar{X}_{\underline{s}}|^4 \leq C(s-\underline{s})^2.
$$

Proof of (b) : Term $(2)_s$ (the key!)

- Using the expression of $x_s - \bar{x}_s = -\int_0^s (\nabla V(X_s) - \nabla V(\bar{X}_{\underline{s}}))ds$, one gets

$$(2)_s = -\sigma\Big(\int_0^s (\nabla V(X_s) - \nabla V(\bar{X}_{\underline{s}}))\,ds \,|\, \nabla^2 V(\bar{X}_{\underline{s}})(W_s - W_{\underline{s}})\Big).$$

- It is clear that both $\int_0^{\underline{s}} (\nabla V(X_s) - \nabla V(\bar{X}_{\underline{s}}))\,ds$ and $\nabla^2 V(\bar{X}_{\underline{s}})$ are $\mathcal{F}^{\xi_0,W}_{\underline{s}}$-measurable hence independent of $W_s - W_{\underline{s}}$ so that

$$\mathbb{E}((2)_s \,|\, \mathcal{F}_{\underline{s}}) = -\sigma\,\mathbb{E}\Big(\int_{\underline{s}}^s (\nabla V(X_s) - \nabla V(\bar{X}_{us}))\,ds \,|\, \nabla^2 V(\bar{X}_{\underline{s}})(W_s - W_{\underline{s}})\,\Big|\, \mathcal{F}_{\underline{s}}\Big).$$

- Consequently,

$$\mathbb{E}\,(2)_s = -\sigma\mathbb{E}\Big(\int_{\underline{s}}^s (\nabla V(X_s) - \nabla V(\bar{X}_{\underline{s}}))\,ds \,|\, \nabla^2 V(\bar{X}_{\underline{s}})(W_s - W_{\underline{s}})\Big).$$

Proof of (b) : Term $(2)_s$

- This in turn implies, using successively Cauchy-Schwartz and generalized Minkowski's inequalities

$$\left|\mathbb{E}\,(2)_s\right| \le \sigma[\nabla V]_{\mathrm{Lip}}\Big\|\int_{\underline{s}}^{s}|X_s-\bar{X}_{\underline{s}}|ds\Big\|_2\|\nabla^2 V\|_{F,\sup}\big\|W_s-W_{\underline{s}}\big\|_2$$
$$\le \sigma[\nabla V]_{\mathrm{Lip}}\int_{\underline{s}}^{s}\|X_s-\bar{X}_{\underline{s}}\|_2 ds\,\|\nabla^2 V\|_{F,\sup}\sqrt{d}(s-\underline{s})^{1/2}.$$

- Now $\|X_s-\bar{X}_{\underline{s}}\|_2 \le \underbrace{\|X_s-\bar{X}_s\|_2}_{\text{cf. regular rate by } (a)\,!} + \underbrace{\|\bar{X}_s-\bar{X}_{\underline{s}}\|_2}_{C(s-\underline{s})^{1/2}}$. Invoking claim (a) and the former upper bound for the second term yields

$$\sup_{s\in[\Gamma_n,\Gamma_{n+1}]}\|X_s-\bar{X}_{\Gamma_n}\|_2 \le C_{V,\bar{\gamma},d}\gamma_{n+1}^{1/2} \quad \text{and} \quad \sup_{s\in[\Gamma_n,\Gamma_{n+1}]}\|\bar{X}_s-\bar{X}_{\underline{s}}\|_2 \le C\gamma_{n+1}^{1/2}.$$

- Inserting this into the above bound yields

$$\sup_{s\in[\Gamma_n,\Gamma_{n+1}]}\left|\mathbb{E}\,(2)_s\right| \le C'_{V,\bar{\gamma},d}\sigma\;\gamma_{n+1}\,\gamma_{n+1}^{1/2}\,\gamma_{n+1}^{1/2} \le \gamma_{n+1}^2.$$

Proof of (b) : Term $(2)_s$ (end)

- Inserting the resulting bound into $R(s)$,
- re-assigning the two terms $\frac{\varepsilon}{4}\mathbb{E}\,|X_s-\bar{X}_s|^2$ to the other integral of the r.h.s. of the same equation
- and noting that $\tilde{\alpha}+\frac{\varepsilon}{2}+2\frac{\varepsilon}{4}=\alpha$ yields

$$e^{2\tilde{\alpha}t}\mathbb{E}|X_t-\bar{X}_t|^2 \le \widetilde{C}'_{V,\bar{\gamma},d,\varepsilon}\int_0^t e^{2\tilde{\alpha}s}(s-\underline{s})^2ds$$

i.e.

$$\sup_{t\in[G_{n-1},\Gamma_n]}\mathbb{E}|X_t-\bar{X}_t|^2 \le e^{2\tilde{\alpha}\bar{\gamma}}e^{-2\tilde{\alpha}\Gamma_n}\sum_{k=1}^{n}e^{2\tilde{\alpha}\Gamma_k}\gamma_k^3 = O(\gamma_n^2)$$

owing to Lemma (a), applied with $p=2$ since $2\tilde{\alpha}>\varpi_2$ or equivalently

$$\sup_{t\in(\Gamma_{n-1},\Gamma_n]}\big\|X_t-\bar{X}_t\big\|_2 = O(\gamma_n). \qquad \square$$

Synthesis I

- The sequence $(\bar{X}_n)_{n\geq 0}$ is the Euler scheme of $(\mathcal{L})_\sigma$), is also the Langevin "excited" version of the deterministic *gradient descent* (*GD*) induced by V.

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}\nabla(\bar{X}_n) + \sqrt{2\gamma_{n+1}}\sigma\,\zeta_{n+1},\ \bar{X}_0 = \xi_0.$$

- whereas $(\xi_n)_{n\geq 0}$ as mentioned from the beginning is the Langevin excited version of the Stochastic Gradient Descent (*SGD*) induced by V associated to $H(y,Z)$.

$$\xi_{n+1} = \xi_n - \gamma_{n+1}H(\xi_n, Z_{n+1}) + \sqrt{2\gamma_{n+1}}\sigma\,\zeta_{n+1},$$

- In the theorem below, keep in mind that $X^{\star,\sigma}$ the stationary solution of $(\mathcal{L})_\sigma$ starting from $X_0^{\star,\sigma} \stackrel{d}{=} \pi_\sigma$, $n\geq 0$.

Synthesis II: main theorem

Theorem (...Durmus-Moulines '18, ...Panloup-P. '23, Égéa-Panloup '24, P.'24)

Assume V is $\mathcal{C}^1$ and α-convex, $\alpha > 0$, with Lipschitz gradient. Let $\xi_0 \in L^2(\mathbb{P})$. Let $(\xi_n)_{n\geq 0}$ and and let $(\bar{X}_n)_{n\geq 0}$ be the Langevin SGD and GD respectively.

(a) If $(\gamma_n)_{n\geq 1}$ satisfies $\varpi_1 < 2\alpha$,

$$\mathcal{W}_2([\xi_n], \pi_\sigma) \leq \|\xi_n - X^{\star,\sigma}_{\Gamma_n}\|_2 \leq C_{H,X}\sqrt{\gamma_n} + \|\xi_0 - \xi_0^{(\star,\sigma)}\|_2 e^{-\alpha\Gamma_n} = O(\sqrt{\gamma_n}).$$

and

$$\mathcal{W}_2([\bar{X}_n], \pi_\sigma) \leq \left\|\bar{X}_n - X^{\star,\sigma}_{\Gamma_n}\right\|_2 \leq C_X\sqrt{\gamma_n} + \|\xi_0 - \xi_0^{(\star,\sigma)}\|_2 e^{-\alpha\Gamma_n} = O(\sqrt{\gamma_n}).$$

(b) If furthermore V is $\mathcal{C}^2$ with Lipschitz Hessian $\nabla^2 V$, $\xi_0 \in L^4(\mathbb{P})$ and $(\gamma_n)_{n\geq 1}$ satisfies $\varpi_2 < 2\alpha$, then

$$\|\bar{X}_n - X^{\star,\sigma}_{\Gamma_n}\|_2 \leq C_X\gamma_n + \|\xi_0 - \xi_0^{(\star,\sigma)}\|_2 e^{-\alpha\Gamma_n} = O(\gamma_n).$$

Pre-conditioners for ℵ-practitioners (by Panloup-P.'23 & Bras-P.'24)

- To still improve the convergence and in particular to help even more the *SGLD* procedure escape from local minima, practitioners introduced so-called *pre-conditioners* (see [6]) by making σ depend on X_t in $(\mathcal{L})_\sigma$, namely

$$\sigma \rightsquigarrow \sigma\,\vartheta(X_t), \text{ or } \sigma\,\vartheta(\nabla V(X_t)), \text{ or } \sigma\,\vartheta(V(X_t)).$$

- A theoretical background has been provided in [7] to justify en highlight this heuristics.

Proposition

The diffusion $(\mathcal{L}_{\sigma(x)})$ $\quad dX_t = b(X_t)dt + \sqrt{2}\sigma\,\vartheta(X_t)dW_t, \quad X_0 = \xi_0$

where the drift b is defined by

$$b := -\Big((\vartheta\vartheta^\top)\nabla V\sigma^2 - \Big[\sum_{j=1}^{d}\partial_{x^j}(\vartheta\vartheta^\top)_{ij}\Big]_{i=1:d}\Big)$$

also has $\pi^{(\sigma)}$ as a unique invariant distribution (under an ellipticity assumption on the *preconditioner* ϑ).

Pre-conditioners (by Panloup-P.'23 [7] & Bras-P.'24 [2])

- The implementable version of $(\mathcal{L}_{\sigma(x)})$ is simply its Euler scheme with (constant or decreasing) step $\gamma_n > 0$ and b as above
- It is known as *PGLD* for *Preconditioned Gradient Langevin Dynamics*,

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1} b(\bar{X}_n) + \sqrt{2\gamma_{n+1}}\,\sigma\vartheta(\bar{X}_n)\zeta_{n+1},\ n \geq 0, \quad \bar{X}_0 = \xi_0,$$

were $(\zeta_n)_{n\geq 1}$ is i.i.d. and $\mathcal{N}(0, I_d)$-distributed and b as above.

- This improved version is investigated in [7] in its decreasing step mode w.r.t. $\mathcal{W}_1$-distance.

Theorem (P.-Panloup, *AAP '23* [7])

Under (higher than above) regularity assumptions on ∇V and σ and uniform ellipticity assumptions but only α-confluence outside a compact set of $(\mathbb{R}^d)^2$

$$\mathcal{W}_1\big([\bar{X}_n], \pi_\sigma\big) \leq C_{X,\gamma}\gamma_n \quad \big\|[\bar{X}_n] - \pi_\sigma\big\|_{VT} = o(\gamma_n^{1-\eta}),\ \forall \eta > 0.$$

- It is implementation by ℵ practitioners in order to "improve" a gradient descent is usually carried out with a small enough constant step $\gamma > 0$.

- When adapted to a *SGD*, regular or mini-batch it is called *PSGLD* for *Preconditioned Stochastic Gradient Langevin Dynamics* and reads

$$\xi_{n+1} = \xi_n - \gamma H(\xi_n, Z_{n+1}) + \sqrt{2\gamma_{n+1}}\,\sigma\,\vartheta(\xi_n)\zeta_{n+1}, \quad n \geq 0.$$

- ℵ Practitioners... usually consider diagonal pre-conditioners of the form

$$\forall\,\xi = (\xi^1,\ldots,\xi^d) \in \mathbb{R}^d,\ \vartheta\vartheta^\top(\xi) = \mathrm{Diag}\Big(\big(\varphi(\partial_{\xi^1}V(\xi))\big)^2,\cdots,\big(\varphi(\partial_{\xi^d}V(\xi))\big)^2\Big).$$

- Numerical experiments carried out by practitioners suggest that the resulting additive correcting term in the drift

$$b := -\Big((\vartheta\vartheta^\top)\nabla V\sigma^2 - \Big[\sum_{j=1}^{d}\partial_{x^j}(\vartheta\vartheta^\top)_{ij}\Big]_{i=1:d}\Big)$$

in the drift, which is too computationally demanding in terms of complexity, can be neglected without damage for practical implementation (see [6]).

Simulated annealing regime

- In what precedes, practitioners' strategy, is to set either
  - σ small enough, but not too small
  - or to make σ decrease by "plateaux" toward $\sigma_\infty > 0$ (see [3])

  to get a good compromise between exploration and convergence.
- A simulated annealing version of the above procedures. has been introduced and analyzed in [2, 3], in which, $\sigma = \sigma_n$ is no longer constant but slowly decreasing to 0 to capture the true $\mathrm{agmin}_{\mathbb{R}^d} V$.
- The appropriate tuning turns out to be

$$\sigma_n = \frac{c}{\sqrt{\log n}} \downarrow 0.$$

Simulated annealing regime

- This implementation makes the procedure enter the *simulated annealing regime* and one can show *mutatis mutandis* under the assumptions of the above theorems on the Gibbs measures and the former convergence theorems that
$$\xi_n \xrightarrow{\mathbb{P}} \mathrm{argmin}_{\mathbb{R}^d} V$$
(convergence in probability).
- This simulated annealing regime for (more general) stochastic approximation procedures goes back to the seminal paper [4] by Gelfand & Mitter en 1991.
- However, in practice the tuning of such a variant of the algorithms is very sensitive to the parameters (especially c) and it is not implemented for high dimensional optimization problems like those commonly encountered nowadays in Machine Learning.

Adam algorithm (*Ada*ptive *m*oment estimation)

- The Adam algorithm reads as follows

$$
\begin{aligned}
g_{n+1} &= H(\theta_{n-1}, Z_{n+1}) \quad \text{with} \quad \mathbb{E}\, H(\theta, Z) = \nabla_\theta V(\theta)\\
m_{n+1} &= \beta_1 m_n + (1-\beta_1) g_{n+1} \quad v_{n+1} = \beta_2 m_n + (1-\beta_2) g_{n+1}^2\\
\widehat{m}_{n+1} &= \frac{m_{n+1}}{1-\beta_1^{n+1}}, \quad \widehat{v}_{n+1} = \frac{v_{n+1}}{1-\beta_2^{n+1}}\\
\theta_{n+1} &= \theta_n - \gamma_{n+1} \frac{\widehat{m}_{n+1}}{\sqrt{\widehat{v}_n} + \varepsilon}
\end{aligned}
$$

- with $\gamma_n = \alpha \simeq 10^{-3}$, $\beta_1 \simeq 0.9 \in [0,1]$, $\beta_2 \simeq 0.999 \in [0,1]$, $\varepsilon \simeq 10^{-8}$.
- Initialize m_0, v_0 and θ_0. Then for $n \geq 0$

$$
(\theta_{n+1}, m_{n+1}, v_{n+1}) = \theta_n - \gamma_{n+1}.H_{adam}(\theta_n, m_n, v_n).
$$

- Compromise between
  - AdaGrad (Duchi et al., 2011) (for sparse gradients),
  - RMSProp (Tieleman & Hinton, 2012) (on line algo. for non stationary data).

Langevin Adam algorithm

- Set, for $n \geq 0$

$$\xi_n = \begin{pmatrix} \theta_n \\ m_n \\ v_n \end{pmatrix}.$$

- Langevin Adam algorithm: let $(\zeta_n)_{n\geq 1}$ i.i.d., $\sim \mathcal{N}(0, I_d)$.

$$\xi_{n+1} = \xi_n - \gamma_{n+1}.H_{adam}(\theta_n, m_n, v_n) + \sigma_{n+1}\sqrt{\gamma_{n+1}}\zeta_{n+1}$$

with $\sigma_n = \sigma$ small or $\sigma_n = \sigma/\sqrt{\log n}$ (simulated annealing version).

- PreconditionedLangevin Adam algorithm: cf. [Bras-P. IJCNN2023]

$$\xi_{n+1} = \xi_n - \gamma_{n+1}P_{n+1}.H_{adam}(\xi_n) + \sigma_{n+1}\sqrt{\gamma_{n+1}}\,T_{n+1}\zeta_{n+1}$$

with $T_{n+1}T_{n+1}^{top} = P_{n+1}\ldots$

Bibliography

- B. Athreya, C.-H. Hwang (2010). Gibbs measures asymptotics. *Sankhya A*, **72**(1):191–207.
- P. Bras, G. Pagès (2024). Convergence of Langevin-simulated annealing algorithms with multiplicative noise, *Math. Comp.*, **93**(348):1761–1803.
- P. Bras, G. Pagès (2023). Convergence of Langevin-simulated annealing algorithms with multiplicative noise II: Total variation, *Monte Carlo Methods Appl.* **29**(3):203–219.
- P. Bras, Pierre (2022). Convergence rates of Gibbs measures with degenerate minimum, *Bernoulli*, **28**(4):243–2458.
- P. Bras, Pierre (2022). Langevin algorithms for very deep Neural Networks with applications to image classification. In *International Neural Network Society Workshop on Deep Learning Innovations and Applications*, part of the *IJCNN'23 conference*.
- A. Durmus, É. Moulines (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm, *Bernoulli*, **25**(4A):2854–2882.

M. Egéa and F. Panloup (2024). Multilevel-Langevin pathwise average for Gibbs approximation, to appear in *Mathematics of Operation Research*, [arXiv]

M. Egéa (2024). (Non)-penalized multilevel methods for non-uniformly log-concave distributions, *Electron. J. Probab.*, **29**:1– 43. https://doi.org/10.1214/24-EJP1099.

S.N. Ethier, T.G. Kurtz (1986). *Markov processes. Characterization and convergence*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., John Wiley & Sons, Inc., New York, x+534 pp.ISBN:0-471-08186-8

S.B. Gelfand, S.K. Mitter (1991). Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$, *SIAM J. Control Optimization*, **29**(5):999–1018.

D. P. Kingma, J. Lei Ba (2015). Adam: a method for stochastic optimization. Conference paper at *ICLR 2015*, arXiv:1412.6980;

C. Li , C. Chen, D. E. Carlson, L. Carin (2015). Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks, *AAAI Conference on Artificial Intelligence*.

G. Pagès, F. Panloup (2023). Unadjusted Langevin algorithm with multiplicative noise: total variation and Wasserstein bounds., *Ann. Appl. Probab.*, **33**(1):726–779.

Langevin Algorithms for Markovian Neural Networks and Deep Stochastic Control

Pierre BRAS and Gilles PAGÈS, presented by Pierre BRAS

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université, Paris, France

Presented at the International Joint Conference on Neural Networks 2023
Gold Coast Convention and Exhibition Centre, Queensland, Australia

Acknowledgements

We thank the IEEE Computational Intelligence Society (CIS) for supporting the participation to the conference IJCNN 2023 with the IEEE CIS Conference Travel Grant for Students.

IEEE Computational Intelligence Society

Table of Contents

We consider the following **Stochastic Optimal Control** (SOC) problem associated with a **Stochastic Differential Equation** (SDE):

$$\min_u J(u) := \mathbb{E}\left[\int_0^T G(X_t)dt + F(X_T)\right], \tag{1}$$

$$dX_t = b(X_t, u_t)dt + \sigma(X_t, u_t)dW_t, \;\; t \in [0, T] \tag{2}$$

- X_t: trajectory vector
- u_t: control vector
- $b(X_t, u_t)$: controlled drift vector
- $\sigma(X_t, u_t)$: controlled diffusion matrix
- W_t: Brownian motion (white noise process)

$\implies$ Optimize a functional of a trajectory of a SDE X_t through the control u_t, including a random noise that affects the evolution of the system.

Example: Resource Management

An oil drilling company has to balance the costs of extraction and of storage of oil in a volatile energy market:

- **Trajectory:** Volatile global oil price and quantity of stored (unsold) oil for the company
- **Control:** Quantities of instantaneously extracted, stored and sold oil

Figure: Offshore oil rig - Source: Unsplash

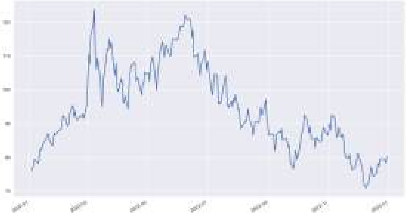


Figure: Crude oil price during the year 2022

Discretization and numerical scheme

Euler-Maruyama scheme

$$\min_{\theta} \bar{J}(\bar{u}_\theta) := \mathbb{E}\Big[\sum_{k=0}^{N-1}(t_{k+1}-t_k)G(\bar{X}^{\theta}_{t_{k+1}}) + F(\bar{X}^{\theta}_{t_N})\Big], \tag{3}$$

$$\begin{aligned}\bar{X}^{\theta}_{t_{k+1}} &= \bar{X}^{\theta}_{t_k} + (t_{k+1}-t_k)b(\bar{X}^{\theta}_{t_k}, \bar{u}_{k,\theta}(\bar{X}^{\theta}_{t_k})) \\ &\quad + \sqrt{t_{k+1}-t_k}\,\sigma(\bar{X}^{\theta}_{t_k}, \bar{u}_{k,\theta}(\bar{X}^{\theta}_{t_k}))\xi_{k+1},\end{aligned} \tag{4}$$

$\xi_k \sim \mathcal{N}(0, I_{d_2})$ i.i.d.

- **Time discretization** of $[0, T]$:

$$t_k := kT/N,\ k \in \{0, \dots, N\}, \quad h := T/N$$

- **Control** u with **parameter** θ using either one time-dependant neural network either N distinct neural networks: $u_{t_k} = \bar{u}_\theta(t_k, X_{t_k})$ or $u_{t_k} = \bar{u}_{\theta^k}(X_{t_k})$
- Since the process is **Markovian**, we assume the control depends only on the running position X_t (instead of the whole previous trajectory $(X_s)_{s\in[0,t]}$).

The parameter θ is optimized by **gradient descent**:

- Simulate batches of trajectories $\bar{X}$ depending on the Brownian motion.
- Compute $\nabla_\theta \bar{J} = \nabla_\theta \bar{J}(\bar{u}_{\theta_n}, (\xi_k^{i,n+1})_{1\le k\le N})$; the gradient is computed by automatic differentiation as the gradient w.r.t. to θ is tracked all along the trajectory of the numerical scheme Giles and Glasserman (2005); Giles (2007)

In the literature:
SOCs are solved using specific techniques: Forward-Backward SDEs, Hamilton-Jacobi-Bellman (HJB) optimality conditions, stochastic dynamic programming. The resolution of SOCs by neural networks scales to the high dimension, contrary to dynamic programming Gobet and Munos (2005); Han and Weinan (2016); Bachouch et al. (2022); Laurière et al. (2023).

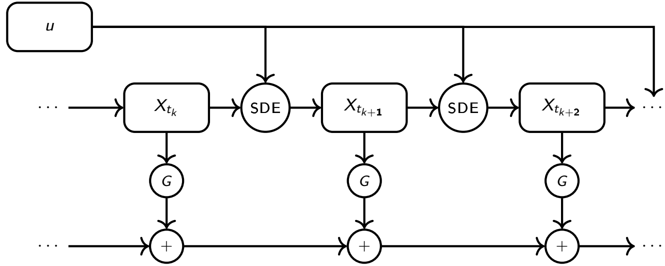


Figure: Markovian Neural Network with one control.

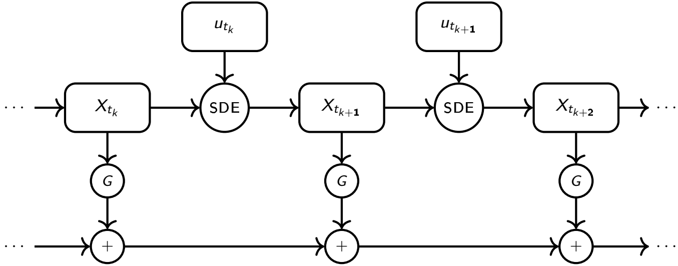


Figure: Markovian neural network with one control for every time step.

- If the control is applied at many discretization times, then the **Markovian Neural Network** becomes a **very deep** neural network, difficult to train directly.
- Adding noise during training is known to improve the learning procedure Neelakantan et al. (2015); Anirudh Bhardwaj (2019):

Gradient Langevin Algorithm

For some choice of **Preconditioner** rule P (Adam, RMSprop...), step size γ_{n+1} and and computed gradient g_{n+1}:

$$\theta_{n+1} = \theta_n - \gamma_{n+1} P_{n+1} \cdot g_{n+1} + \sigma_{n+1}\sqrt{\gamma_{n+1}}\mathcal{N}(0, P_{n+1}) \tag{5}$$

$\implies$ per-dimension adaptive noise rate.

- Bras (2022): the deeper the network is, the greater are the gains provided by Langevin algorithms; introduces the **Layer Langevin** algorithm, consisting in adding Langevin noise only to the deepest layers.

$\implies$ Analysis was conducted especially for deep architectures in **image classification**.

Our objectives

- Side-by-side comparison of non-Langevin/Langevin optimizers on different SOC problems: fishing quotas, financial hedging, energy management.
- If using multiple controls (second case), explore the benefits of Layer-Langevin.

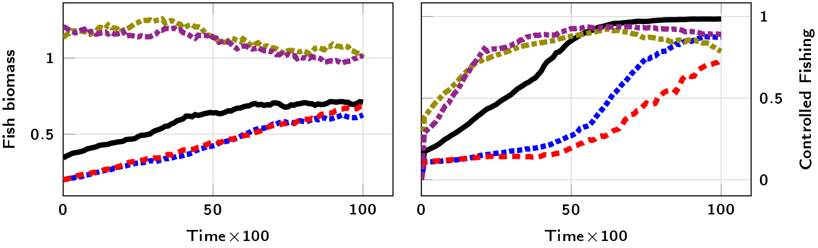
Fish biomass $X_t \in \mathbb{R}^{d_1}$ with:

- Inter-species interaction κX_t
- Fishing following imposed quotas u_t
- Objective: keep X_t close to an ideal state $\mathcal{X}_t$.

Figure: Source: Unsplash

$$dX_t = X_t * ((r - u_t - \kappa X_t)dt + \eta dW_t)$$

$$J(u) = \mathbb{E}\left[\int_0^T (|X_t - \mathcal{X}_t|^2 - \langle \alpha, u_t \rangle)dt + \beta[u]^{0,T}\right]$$



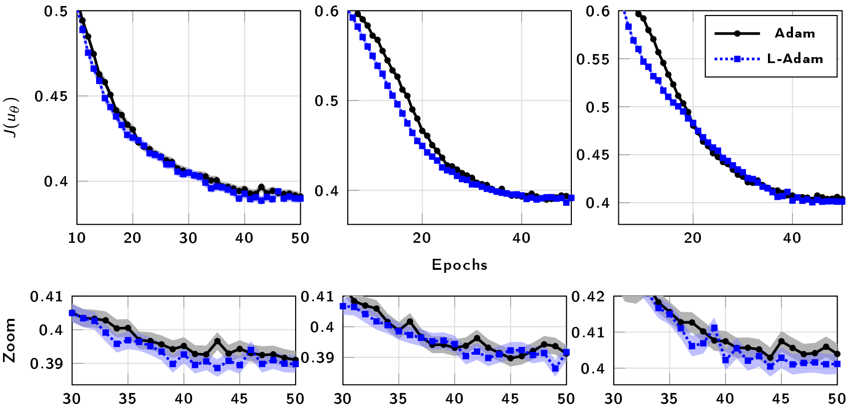


Figure: Comparison of Adam et L-Adam algorithms during the training for the fishing control problem with $N = 20, 50, 100$ respectively. J is estimated over 50×512 trajectories. A zoom on the last epochs is given.

Table: Best performance

	$N = 20$	$N = 50$	$N = 100$
Adam	0.3910	0.3912	0.4029
L-Adam	0.3886	0.3864	0.4011

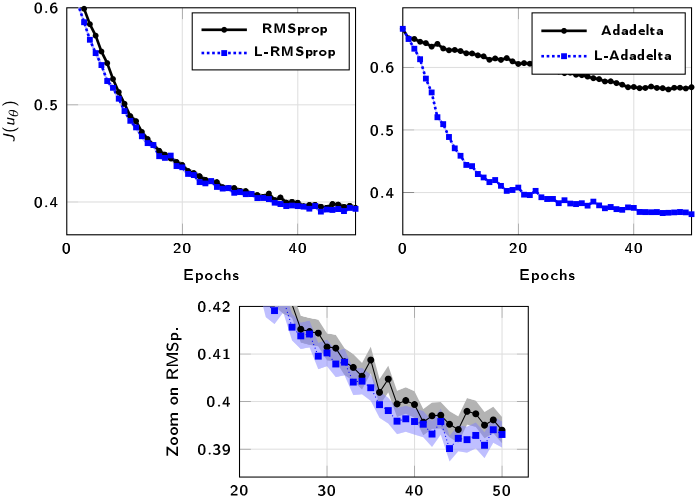


Figure: Comparison of Langevin algorithms with their non-Langevin counterparts during the training for the fishing control problem with $N = 50$.

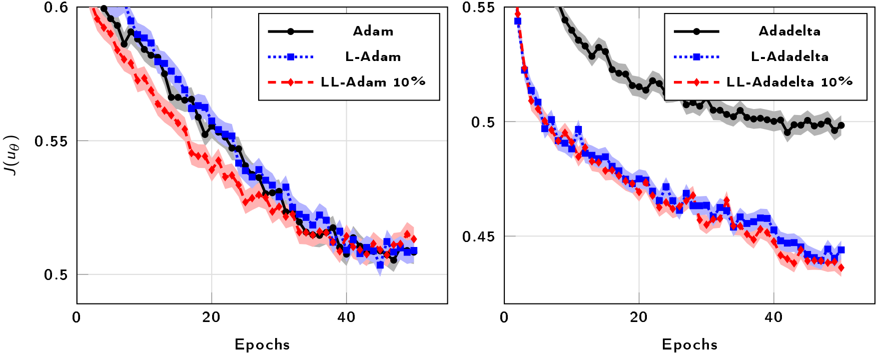


Figure: Training of the fishing problem with multiple controls with $N = 10$

We aim to replicate some payoff Z defined on some portfolio S_t by trading some of the assets with transaction costs; the control u_t is the amount of held assets. The objective is

Figure: Source: Unsplash

$$J(u) = \nu\left(-Z + \sum_{k=0}^{N-1}\langle u_{t_k}, S_{t_{k+1}} - S_{t_k}\rangle - \sum_{k=0}^{N}\langle c_{tr}, S_{t_k} * |u_{t_k} - u_{t_{k-1}}|\rangle\right) \tag{6}$$

where ν is a convex risk measure. We consider the assets S_t to be follow a Heston model and are tradable along with variance swap options.

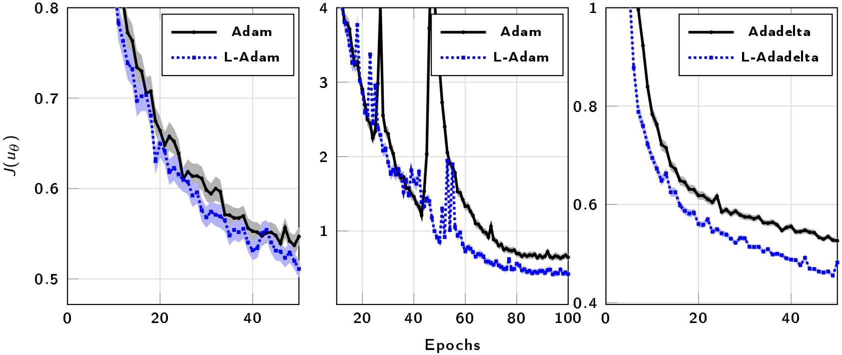


Figure: Comparison of algorithms during the training for the deep hedging control problem with $N = 30, 50, 50$ respectively

Table: Best performance

	Adam, $N = 30$	Adam, $N = 50$	Adadelta, $N = 50$
Vanilla	0.4448	0.6355	0.4671
Langevin	0.4306	0.4182	0.3773

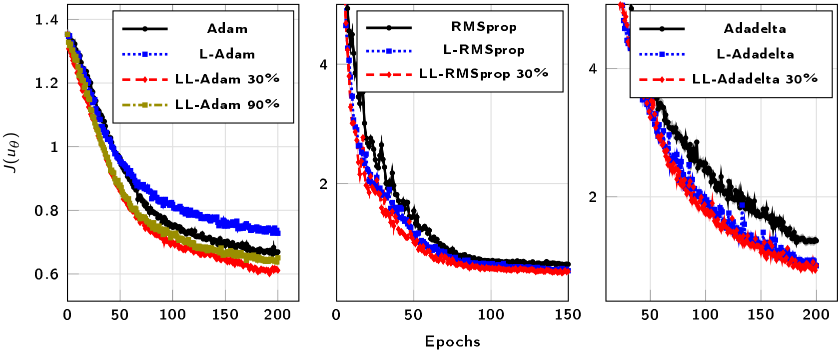


Figure: Training of the deep hedging problem with multiple controls with $N = 10$

Table: Best performance

	Adam	RMSprop	Adadelta
Vanilla	0.6626	0.5618	1.2900
Langevin	0.7278	0.4441	0.9250
Layer Langevin 30%	0.6004	0.4102	0.8554
Layer Langevin 90%	0.6377	–	–

Resource Management and Oil Drilling, Goutte et al. (2018); Gaïgi et al. (2021)

An oil driller has to balance the costs of extraction E_t, storage S_t in a volatile energy market with oil price P_t:

$$
\begin{aligned}
&dP_t = \mu P_t dt + \eta P_t dW_t \\
&J(q) = -\mathbb{E}\left[\int_0^T e^{-\rho r} U\Big(q_r^v P_r + q_r^{v,s}(1-\varepsilon)P_r - (q_r^v + q_r^s)c_e(E_r) - c_s(S_r)\Big)dr\right], \\
&E_t = \int_0^t (q_r^v + q_r^s)dr, \quad S_t = \int_0^t (q_r^s - q_r^{v,s})dr
\end{aligned}
$$

where U is the utility function and $q_t = (q_t^v, q_t^s, q_t^{v,s})$ is the control (extracted, stored, sold from storage).

Results for Oil Drilling

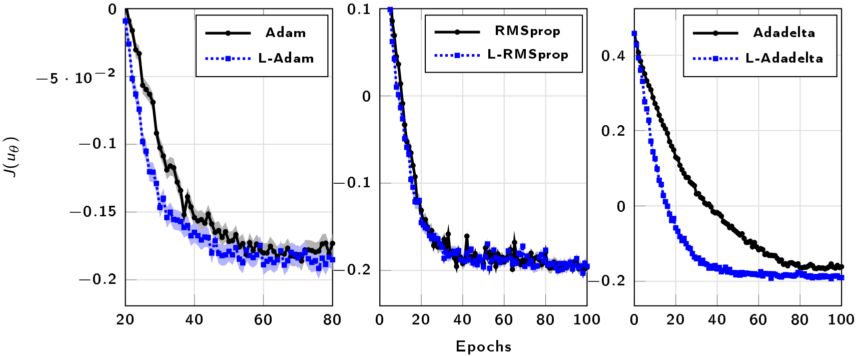


Figure: Comparison of algorithms during the training for the oil drilling control problem with $N = 50$

Table: Best performance

	Adam	RMSprop	Adadelta
Vanilla	-0.1729	-0.1985	-0.1649
Langevin	-0.1915	-0.2032	-0.1929

Conclusion

- In various problems, Langevin and Layer Langevin algorithms show improvements in comparison with their respective non-Langevin counterparts.
- Gains depend on the setting and optimizer; we observe that gains are limited or null for the RMSprop algorithm.
- For SOC with multiple controls, we proved the gains of Layer Langevin algorithms with a small number of layers ($\sim$10%-30%).

Thank you for your attention !

C. Anirudh Bhardwaj. Adaptively Preconditioned Stochastic Gradient Langevin Dynamics. *arXiv e-prints*, art. arXiv:1906.04324, June 2019.

A. Bachouch, C. Huré, N. Langrené, and H. Pham. Deep neural networks algorithms for stochastic control problems on finite horizon: numerical applications. *Methodol. Comput. Appl. Probab.*, 24(1): 143–178, 2022. ISSN 1387-5841. doi: 10.1007/s11009-019-09767-9. URL https://doi.org/10.1007/s11009-019-09767-9.

P. Bras. Langevin algorithms for very deep Neural Networks with application to image classification. *arXiv e-prints*, art. arXiv:2212.14718, Dec. 2022.

H. Buehler, L. Gonon, J. Teichmann, and B. Wood. Deep hedging. *Quant. Finance*, 19(8):1271–1291, 2019. ISSN 1469-7688. doi: 10.1080/14697688.2019.1571683. URL https://doi.org/10.1080/14697688.2019.1571683.

M. Gaïgi, S. Goutte, I. Kharroubi, and T. Lim. Optimal risk management problem of natural resources: application to oil drilling. *Ann. Oper. Res.*, 297(1-2):147–166, 2021. ISSN 0254-5330. doi: 10.1007/s10479-019-03303-1. URL https://doi.org/10.1007/s10479-019-03303-1.

M. B. Giles. Monte Carlo evaluation of sensitivities in computational finance. Technical Report NA07/12, Oxford University Computing Laboratory, 2007.

M. B. Giles and P. Glasserman. Smoking adjoints: fast evaluation of Greeks in Monte Carlo calculations. Technical Report NA05/15, Oxford University Computing Laboratory, 2005.

E. Gobet and R. Munos. Sensitivity analysis using Itô-Malliavin calculus and martingales, and application to stochastic optimal control. *SIAM J. Control Optim.*, 43(5):1676–1713, 2005. ISSN 0363-0129. doi: 10.1137/S0363012902419059. URL https://doi.org/10.1137/S0363012902419059.

S. Goutte, I. Kharroubi, and T. Lim. Optimal management of an oil exploitation. *International Journal of Global Energy Issues*, 41(1/2/3/4):69–85, 2018.

J. Han and W. Weinan. Deep Learning Approximation for Stochastic Control Problems. *Deep Reinforcement Learning Workshop, NIPS (2016)*, Nov. 2016.

M. Laurière, G. Pagès, and O. Pironneau. Performance of a Markovian Neural Network versus dynamic programming on a fishing control problem. *Probability, Uncertainty and Quantitative Risk*, pages –, 2023. ISSN 2095-9672. doi: 10.3934/puqr.2023006. URL /article/id/63c741a4b5351f4889aff727.

A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens. Adding Gradient Noise Improves Learning for Very Deep Networks. *arXiv e-prints*, art. arXiv:1511.06807, Nov. 2015.