Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise

Pierre BRAS and Gilles PAGÈS

Sorbonne Université

December 5, 2022

SORBONNE UNIVERSITÉ

Optimization problem

Let $V : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^1$, coercive (i.e. $V(x) \to +\infty$ as $|x| \to \infty$) and let $\text{argmin}(V) := \{x \in \mathbb{R}^d : \ V(x) = \min_{\mathbb{R}^d} V\}$ and $V^\star := \min V$.

Objective : find argmin(V).

- **Example : Regression as an optimization problem**

– $\{\Phi_x : \ x \in \mathbb{R}^d\}$ family of functions $\Phi_x : \mathbb{R}^{d'} \to \mathbb{R}$ parametrized by $x \in \mathbb{R}^d$ (e.g. Φ_x is a neural function).
– for $1 \leq i \leq N$, $(u_i, v_i) \in \mathbb{R}^{d'} \times \mathbb{R}$: data associated to a regression problem
– We want to find x such that for all i, $\Phi_x(u_i) \approx v_i$

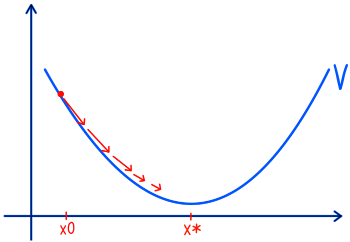
$$\Longrightarrow \quad \text{Find } \min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} (\Phi_x(u_i) - v_i)^2 =: \min_{x \in \mathbb{R}^d} V(x).$$

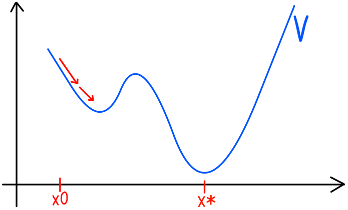
Introduction - Gradient descent

• Gradient descent algorithm : compute the gradient and "go down" the gradient with decreasing step sequence (γ_k):

$$x_0 \in \mathbb{R}^d$$
$$x_{n+1} = x_n - \gamma_{n+1} \nabla V(x_n).$$

• The continuous version is $dX_s = -\nabla V(X_s)ds$.





• **Problem** : x_n can be "trapped" !

- We add a white noise to x_n, hoping to escape traps :

$$x_{n+1} = x_n - \gamma_{n+1}\nabla V(x_n) + \sqrt{\gamma_{n+1}}\sigma\xi_{n+1}, \quad \xi_{n+1} \sim \mathcal{N}(0, I_d).$$

$\Longrightarrow$ called SGLD algorithms (Stochastic Gradient Langevin Dynamics)

- The continuous version becomes:

$$dX_s = -\nabla V(X_s)ds + \sigma dW_s \qquad \textbf{(Langevin Equation)}$$

where (W_s) is a Brownian motion and $\sigma > 0$.

- Assuming that $e^{-2V/\sigma^2} \in L^1(\mathbb{R}^d)$, it is invariant measure is the **Gibbs measure**

$$\nu_\sigma(x)dx = C_\sigma e^{-2(V(x)-V^\star)/\sigma^2}dx$$

$$C_\sigma := \left(\int_{\mathbb{R}^d} e^{-2(V(x)-V^\star)/\sigma^2}dx\right)^{-1}.$$

- Exogenous noise σdW_t added to escape local minima ('traps') and explore the state space.
- For small σ, ν_σ is concentrated around $\operatorname{argmin}(V)$:

Solve the Langevin equation $\Longrightarrow$ approximation of ν_σ $\Longrightarrow$ approximation of $\operatorname{argmin}(V)$.

- We have $\nu_\sigma \underset{\sigma\to 0}{\longrightarrow} \operatorname{argmin}(V)$ in law.
- One possibility : solve the Langevin equation for small σ
- Another possibility : make $\sigma \to 0$ while iterating the algorithm :

$$x_{n+1} = x_n - \gamma_{n+1}\nabla V(x_n) + a(\gamma_1 + \cdots + \gamma_{n+1})\sigma\sqrt{\gamma_{n+1}}\xi_{n+1}, \quad \xi_{n+1} \sim \mathcal{N}(0, I_d),$$

where $a(t)$ is decreasing and $a(t) \underset{t\to 0}{\longrightarrow} 0$.

The continuous version becomes :

Langevin-Simulated Annealing Equation

$$dX_t = -\nabla V(X_t)dt + a(t)\sigma dW_t,$$

- The 'instantaneous' invariant measure $\nu_{a(t)\sigma}(dx) \propto \exp\left(-2V(x)/(a^2(t)\sigma^2)\right)$ converges itself to $\operatorname{argmin}(V)$
- Schedule $a(t) = A\log^{-1/2}(t)$ then $X_t \underset{t\to\infty}{\longrightarrow} \operatorname{argmin}(V)$ in law [Chiang-Hwang 1987], [Miclo 1992]
- [Gelfand-Mitter 1991] proves the convergence of the algorithm (x_n).

Multiplicative noise

- Noise $\sigma > 0 \implies$ isotropic, homogeneous noise $\implies$ not adapted to V
- Instead : $\sigma(X_t)$ is a matrix depending on the position
- In Machine Learning literature, a good choice is $\sigma(x)\sigma(x)^\top \simeq (\nabla^2 V(x))^{-1}$ as in the Newton algorithm.

$$dY_t = -(\sigma\sigma^\top \nabla V)(Y_t)dt + a(t)\sigma(Y_t)dW_t + \underbrace{\left(a^2(t) \left[\sum_{j=1}^{d} \partial_i(\sigma\sigma^\top)(Y_t)_{ij} \right]_{1\le i\le d} \right)}_{\text{correction term} \quad \Upsilon(Y_t)} dt$$

$$a(t) = \frac{A}{\sqrt{\log(t)}},$$

• Correction term so that $\nu_{a(t)} \propto \exp\left(-2V(x)/a^2(t)\right)$ is still the "instantaneous" invariant measure

Objectives and assumptions

- Prove the convergence in of Y_t and $\bar{Y}_t$ to $\nu^\star$ (supported by argmin(V))
- We use the L^1-Wasserstein distance:

$$\mathcal{W}_1(\pi_1, \pi_2) = \sup \left\{ \int_{\mathbb{R}^d} f(x)\pi_1(dx) - \int_{\mathbb{R}^d} f(x)\pi_2(dx) : \; f : \mathbb{R}^d \to \mathbb{R}, \; [f]_{\text{Lip}} = 1 \right\}.$$

  and we show that $\mathcal{W}_1([Y_t], \nu^\star) \to 0$ and $\mathcal{W}_1([\bar{Y}_t], \nu^\star) \to 0$.

- We have

$$\mathcal{W}_1(Y_t, \nu^\star) \leq \mathcal{W}_1(Y_t, \nu_{a(t)}) + \mathcal{W}_1(\nu_{a(t)}, \nu^\star)$$

  The convergence is limited by the slowness of $a(t)$ as $\mathcal{W}_1(\nu_{a(t)}, \nu^\star) \asymp a(t) \asymp \log^{-1/2}(t)$. In fact we also prove

$$\mathcal{W}_1(Y_t^{x_0}, \nu_{a(t)}) \leq C_\alpha \max(1 + |x_0|, V(X_0))t^{-\alpha}$$
$$\mathcal{W}_1(\bar{Y}_t^{x_0}, \nu_{a(t)}) \leq C_\alpha \max(1 + |x_0|, V^2(X_0))t^{-\alpha}$$

  for every $\alpha < 1$.

- **Assumptions:**
  1. V is strongly convex outside some compact set
  2. σ is bounded and elliptic: $\sigma\sigma^\top \geq \sigma_0 I_d$, $\sigma_0 > 0$.
  3. ∇V is Lipschitz
  4. Decreasing steps (γ_n) for the Euler scheme, with $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 < \infty$, $\Gamma_n := \gamma_1 + \cdots + \gamma_n$.

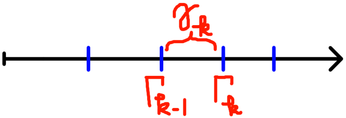
Domino strategy

- [Pages-Panloup 2020] proves the convergence of the Euler scheme of a general SDE $dX_t = b(X_t)dt + \sigma(X_t)dW_t$ to the invariant measure $\pi^\star$ for $\mathcal{W}_1$:

$$\mathcal{W}_1(\bar{X}_t, \pi^\star) \to 0.$$

- *Domino strategy:* for f 1-Lipschitz (P, $\bar{P}$: kernels of X, $\bar{X}$):

$$\begin{aligned}
\mathcal{W}_1(\bar{X}^x_{\Gamma_n}, X^x_{\Gamma_n}) &\leq |\mathbb{E}f(\bar{X}^x_{\Gamma_n}) - \mathbb{E}f(X^x_{\Gamma_n})| \\
&= |\bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_n} f(x) - P_{\Gamma_n} f(x)| \\
&= \left| \sum_{k=1}^{n} \bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_{k-1}} \circ (\bar{P}_{\gamma_k} - P_{\gamma_k}) \circ P_{\Gamma_n - \Gamma_k} f(x) \right| \\
&\leq \sum_{k=1}^{n} \left| \bar{P}_{\gamma_1} \circ \cdots \circ \bar{P}_{\gamma_{k-1}} \circ (\bar{P}_{\gamma_k} - P_{\gamma_k}) \circ P_{\Gamma_n - \Gamma_k} f(x) \right|,
\end{aligned}$$



1. For large $k \implies$ Error in small time $\implies$ use bounds for $\|X^x_t - \bar{X}^x_t\|_p$
2. For small $k \implies$ Ergodicity contraction properties using the convexity of V outside a compact set and the ellipticity of σ [Wang 2020]:

$$\forall t \geq t_0,\ \mathcal{W}_1(X^x_t, X^y_t) \leq Ce^{-\rho t}|x - y|$$

$$\implies \mathcal{W}_1(X^x_t, \pi^\star) \leq Ce^{-\rho t}(1 + |x|).$$

• Problems before applying the domino strategy: non-homogeneous Markov chain + the ellipticity parameter fades away in $a(t)$.
$\Longrightarrow$ What is the dependency of the constants C and ρ in the ellipticity ?

Consider $dX_t = b(X_t)dt + a\sigma(X_t)dW_t$, $a > 0$ with invariant measure ν_a and with

$$\forall x, y \in \mathcal{B}(0,R)^c,\ \langle b(x) - b(y), x - y\rangle + \frac{a^2}{2}\|\sigma(x) - \sigma(y)\|^2 \le -\alpha|x-y|^2.$$

Then

$$\mathcal{W}_1(X_t^x, X_t^y) \le Ce^{C_1/a^2}|x-y|e^{-\rho_a t}, \quad \rho_a := e^{-C_2/a^2}$$
$$\mathcal{W}_1(X_t^x, \nu_a) \le Ce^{C_1/a^2}e^{-\rho_a t}\mathbb{E}|\nu_a - x|.$$

We first consider the plateau SDE:

$$dX_t = -\sigma\sigma^\top \nabla V(X_t)dt + a_{n+1}\sigma(X_t)dW_t + a_{n+1}^2 \Upsilon(X_t)dt, \quad t \in [T_n, T_{n+1}),$$
$$a_n = A \log^{-1/2}(T_n)$$

We apply the contraction property on every plateau:

$$\mathcal{W}_1(X_{T_{n+1}}, \nu_{a_{n+1}} \, | X_{T_n}) \leq Ce^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \mathbb{E}\left[|\nu_{a_{n+1}} - X_{T_n}| \; |X_{T_n} \right]$$

We integrate over the law of X_{T_n}, giving

$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu_{a_{n+1}}) \leq Ce^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \mathcal{W}_1([X^{x_0}_{T_n}], \nu_{a_{n+1}})$$
$$\leq Ce^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \left(\mathcal{W}_1([X^{x_0}_{T_n}], \nu_{a_n}) + \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) \right).$$

And we iterate:

$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu_{a_{n+1}}) \leq \mu_{n+1} \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) + \mu_{n+1}\mu_n \mathcal{W}_1(\nu_{a_{n-1}}, \nu_{a_n}) + \cdots$$
$$+ \mu_{n+1} \cdots \mu_1 \mathcal{W}_1(\nu_{a_0}, \nu_{a_1}) + \mu_{n+1} \cdots \mu_1 \mathcal{W}_1(\delta_{x_0}, \nu_{a_0}),$$
$$\mu_n := Ce^{C_1/a_n^2} e^{-\rho_{a_n}(T_n - T_{n-1})}.$$

- On the other side, we give bounds for the Gibbs measures:

$$\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) \quad \text{and} \quad \mathcal{W}_1(\nu_{a_n}, \nu^\star).$$

Lemma: Acceptance-rejection Wasserstein bounds

Let μ and ν be two probability distributions on $\mathbb{R}^d$ with densities f and g respectively with finite moments of order p. Assume that there exists $M \geq 1$ such that $f \leq Mg$. Then

$$\mathcal{W}_p(\mu, \nu)^p \leq \mathbb{E}|X - Y|^p - \frac{1}{M}\mathbb{E}|X - \tilde{X}|^p,$$

where X and $\tilde{X} \sim \mu$, $Y \sim \nu$ and X, $\tilde{X}$ and Y are mutually independent.

Proof: Let $X \sim \mu$, $Y \sim \nu$, $U \sim \mathcal{U}([0,1])$ independent and

$$X' := Y\mathbb{1}\{U \leq f(Y)/(Mg(Y))\} + X\mathbb{1}\{U > f(Y)/(Mg(Y))\}.$$

Then $X' \sim \mu$ and

$$\begin{aligned}
\mathbb{E}|X' - Y|^p &= \mathbb{E}|Y - X|^p \mathbb{1}\{U > f(Y)/(Mg(Y))\} \\
&= \int_{(\mathbb{R}^d)^2} |y - x|^p \left(\int_0^1 \mathbb{1}\{u > f(y)/(Mg(y))\} du \right) f(x)g(y)dxdy \\
&= \int_{(\mathbb{R}^d)^2} |y - x|^p f(x)g(y)dxdy - \frac{1}{M}\int_{(\mathbb{R}^d)^2} |y - x|^p f(x)f(y)dxdy \\
&= \mathbb{E}|X - Y|^p - \frac{1}{M}\mathbb{E}|X - \tilde{X}|^p.
\end{aligned}$$

We have

$$\frac{\nu_{a_{n+1}}(x)}{\nu_{a_n}(x)} = \frac{\mathcal{Z}_{a_{n+1}}}{\mathcal{Z}_{a_n}} e^{-2(V(x)-V^\star)(a_{n+1}^{-2}-a_n^{-2})} \leq \frac{\mathcal{Z}_{a_{n+1}}}{\mathcal{Z}_{a_n}} =: M_n.$$

Assuming that $\mathrm{argmin}(V^\star) = \{x^\star\}$ (or $\{x_1^\star, \ldots, x_\ell^\star\}$) with $\nabla^2 V(x^\star) > 0$, we have

$$\mathcal{Z}_a^{-1} = \int e^{-2(V(x)-V^\star)/a^2} ds \underset{a\to 0}{\sim} a^d \int e^{-x^\top \nabla^2 V(x^\star) x} dx$$

using that

$$\forall \varepsilon > 0, \quad \nu_a\{V \geq V^\star + \varepsilon\} \underset{a\to 0}{\longrightarrow} 0.$$

Using the convexity inequality

$$\left|e^{-2z/a_n^2} - e^{-2z/a_{n+1}^2}\right| \leq 4e^{-2z/a_n^2}\frac{z}{a_{n+1}^2}\frac{(a_n - a_{n+1})}{a_n},$$

we get

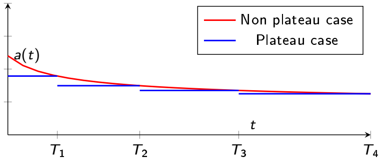
$$\begin{aligned}
\mathcal{Z}_{a_n}^{-1} - \mathcal{Z}_{a_{n+1}}^{-1} &= a_{n+1}^d \int \left(e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} - e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_{n+1}^2}\right) dx \\
&\leq 4a_{n+1}^{d-1}(a_n - a_{n+1}) \int e^{-2(V(a_{n+1}x+x_i^\star)-V^\star)/a_n^2} \frac{V(a_{n+1}x + x_i^\star) - V^\star}{a_{n+1}^2} dx \\
&\sim 2a_{n+1}^{d-1}(a_n - a_{n+1}) \int e^{x^\top \nabla^2 V(x^\star)x}(x^\top V(x^\star)x)dx
\end{aligned}$$

Using similar Taylor expansions we obtain

$$\mathcal{W}_1(a_n, a_{n+1}) \leq \mathbb{E}|\nu_{a_{n+1}} - \nu_{a_n}| - \frac{1}{M_n}\mathbb{E}|\nu_{a_{n+1}} - \widetilde{\nu}_{a_{n+1}}| \leq C(a_n - a_{n+1}).$$

Then if $\sum_n \mathcal{W}_1(\nu_{a_n} - \nu_{a_{n+1}}) < +\infty$ the *Cauchy sequence* ν_{a_n} converges for $\mathcal{W}_1$ with

$$\mathcal{W}_1(\nu_{a_n}, \nu^\star) \leq Ca_n.$$



$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu_{a_{n+1}}) \le \mu_{n+1}\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) + \mu_{n+1}\mu_n\mathcal{W}_1(\nu_{a_{n-1}}, \nu_{a_n}) + \cdots$$
$$+ \mu_{n+1}\cdots\mu_1\mathcal{W}_1(\nu_{a_0}, \nu_{a_1}) + \mu_{n+1}\cdots\mu_1\mathcal{W}_1(\delta_{x_0}, \nu_{a_0}),$$
$$\mu_n = Ce^{C_1/a_n^2}e^{-\rho_{a_n}(T_n - T_{n-1})}, \quad \rho_{a_n} = e^{-C_2/a_n^2}$$

We now choose

$$T_{n+1} - T_n = Cn^\beta, \beta > 0, \quad a_n = \frac{A}{\sqrt{\log(T_n)}}, \quad A > 0 \text{ large enough}$$

yielding

$$a_n - a_{n+1} \asymp (n\log^{3/2}(n))^{-1}, \quad \sum_n (a_n - a_{n+1}) < \infty,$$
$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu_{a_{n+1}}) \le C(1 + |x_0|)\mu_n a_n,$$

where $\mu_n = O(\exp(-Cn^\eta))$. And

$$\mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu^\star) \le \mathcal{W}_1([X^{x_0}_{T_{n+1}}], \nu_{a_{n+1}}) + \mathcal{W}_1(\nu_{a_{n+1}}, \nu^\star).$$

One word on the degenerate case $\nabla^2 V(x^\star) \not> 0$

We assume instead that $\text{argmin}(V) = \{x^\star\}$ (or $\{x_1^\star, \dots, x_\ell^\star\}$) and that $x^\star$ is a strict polynomial minimum i.e.

$$\exists r > 0,\ \forall h \in \mathcal{B}(x^\star, r) \setminus \{0\},\ \sum_{k=2}^{2p} \frac{1}{k!} \nabla^k V(x^\star) \cdot h^k > 0.$$

Following [Bras 2021], under some conditions we obtain the following Central Limit theorem:

$$(a^{-2\alpha_1}, \dots, a^{-2\alpha_d}) * (B \cdot (Z_a - x^\star)) \to Z \quad \text{in law}$$

where $\alpha_i \in (0, 1/2]$, B an orthogonal base, $Z_a \sim \nu_a$ and Z a certain non-degenerate random vector.
Then proceeding to similar Taylor expansions but replacing the changes of variables in $x \mapsto ax$ by $x \mapsto B^{-1} \cdot (a^{2\alpha_1}, \dots, a^{2\alpha_d}) * x$, we obtain

$$\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) \le C(n \log^{1+\alpha_{\min}}(n))$$

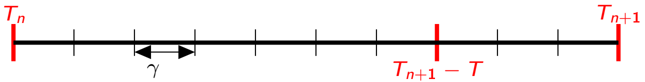
and $\sum_n \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}})$ is still a convergence Bertrand series with

$$\mathcal{W}_1(\nu_{a_n}, \nu^\star) \le C a_n^{2\alpha_{\min}}.$$

• This case is not only theoretical, degenerate minima were observed for over-parametrized neural networks [Sagun-Bottou-LeCun 2016].

Convergence of Y_t with continuously decreasing $(a(t))$

- We apply *domino strategy* to bound $\mathcal{W}_1(X_t, Y_t)$:



- for f Lipschitz-continuous and fixed $T > 0$:

$$\left|\mathbb{E}f(X^{x,n}_{T_{n+1}-T_n}) - \mathbb{E}f(Y^x_{T_{n+1}-T_n,T_n})\right|$$
$$\leq \sum_{k=1}^{\lfloor (T_{n+1}-T_n-T)/\gamma \rfloor} \left| P^Y_{(k-1)\gamma,T_n} \circ (P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_\gamma) \circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x) \right|$$
$$+ \sum_{k=\lfloor (T_{n+1}-T_n-T)/\gamma \rfloor+1}^{\lfloor (T_{n+1}-T_n)/\gamma \rfloor} \left| P^Y_{(k-1)\gamma,T_n} \circ (P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_\gamma) \circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x) \right|$$

- for $k = 1, \ldots, (T_{n+1} - T_n - T)/\gamma$, the kernel $P^{X,n}_{T_{n+1}-T_n-k\gamma}$ has an exponential contraction effect on time $> T$:

$$|(P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_\gamma) \circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x)|$$
$$= |\mathbb{E}P^{X,n}_{T_{n+1}-T_n-k\gamma} f(X^{x,n}_\gamma) - \mathbb{E}P^X_{T_{n+1}-T_n-k\gamma,n} f(Y^x_{\gamma,T_n+(k-1)\gamma})|$$
$$\leq Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n-k\gamma)} [f]_{\text{Lip}} \mathbb{E}|X^{x,n}_\gamma - Y^x_{\gamma,T_n+(k-1)\gamma}|$$
$$\leq Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n-k\gamma)} [f]_{\text{Lip}} \sqrt{\gamma}(a_n - a_{n+1})$$

• Bounds for the error on time intervals no longer than T:

$$|(P^Y_{\gamma,T_n+(k-1)\gamma} - P^{X,n}_\gamma)\circ P^{X,n}_{T_{n+1}-T_n-k\gamma} f(x)| \leq C a_{n+1}^{-2}(a_n - a_{n+1})[f]_{\text{Lip}} \frac{\gamma V(x)}{\sqrt{T_{n+1}-T_n-k\gamma}}$$

using Taylor expansion. Indeed,

$$\mathbb{E}[g(Y^x_{\gamma,u}) - g(X^{x,n}_\gamma)] = \langle \nabla g(x), \mathbb{E}[Y^x_{\gamma,u} - X^{x,n}_\gamma]\rangle + \mathbb{E}[\langle \nabla g(X^{x,n}_\gamma) - \nabla g(x), Y^x_{\gamma,u} - X^{x,n}_\gamma\rangle] + \int_0^1 (1-s)\mathbb{E}\left[\nabla^2 g(sY^x_{\gamma,u} + (1-s)X^{x,n}_\gamma)(Y^x_{\gamma,u} - X^{x,n}_\gamma)^{\otimes 2}\right] ds$$

and we apply to $g = P^{X,n}_t f : x \mapsto \mathbb{E} f(X^{x,n}_t)$ and using Bismuth-Elworthy-Li formula:

$$\nabla_x \mathbb{E} f(X^{x,n}_t) = \mathbb{E}\left[f(X^{x,n}_t)\frac{1}{t}\int_0^t (a_n \sigma^{-1}(X^{x,n}_s) Z^{(x),n}_s)^\top dW_s\right]$$

where $Z^{(x),n}$ is the tangent process (also for higher derivatives...)

• On the other side we have

$$dV^p(X_t^{x,n}) = p\nabla V(X_t^{x,n})^\top \cdot V^{p-1}(X_t^{x,n}) \left(-\sigma\sigma^\top(X_t^{x,n})\nabla V(X_t^{x,n}) + a_{k+1}^2 \Upsilon(X_t^{x,n})\right) dt$$
$$+ p\nabla V(X_t^{x,n})^\top \cdot V^{p-1}(X_t^{x,n}) a_{k+1}\sigma(X_t^{x,n}) dW_t$$
$$+ \frac{p}{2}\left(\nabla^2 V(X_t^{x,n}) V^{p-1}(X_t^{x,n}) + (p-1)|\nabla V(X_t^{x,n})|^2 \cdot V^{p-2}(X_t^{x,n})\right) a_{k+1}^2 \sigma\sigma^\top(X_t^{x,n}) dt$$

and using that ∇V coercive, $|\nabla V| \leq CV^{1/2}$, $\sigma\sigma^\top \geq \sigma_0 I_d$ and σ is bounded, the dominant term in dt is $\sim -CV^{p-1}(\nabla V)^2(X_t^{x,n}) \leq 0$ so that

$$\sup_{t\geq 0} \mathbb{E} V^p(X_t^{x,n}) \leq CV^p(x) \quad \text{and} \quad \sup_{t\geq 0} \mathbb{E} V^p(Y_{t,u}^x) \leq CV^p(x).$$

- We apply on each time interval $[T_n, T_{n+1})$ and obtain the recursive inequality

$$\mathcal{W}_1([X^{x,n}_{T_{n+1}-T_n}],[Y^{x}_{T_{n+1}-T_n,T_n}]) \le Ce^{C_1 a_{n+1}^{-2}}(a_n - a_{n+1})\rho_{n+1}^{-1}V(x).$$

With $x_n := X^{x_0}_{T_n}$, $y_n = Y^{x_0}_{T_n}$:

$$\begin{aligned}
\mathcal{W}_1([X^{x_0}_{T_{n+1}}],[Y^{x_0}_{T_{n+1}}]) &= \mathcal{W}_1([X^{x_n,n}_{T_{n+1}-T_n}],[Y^{y_n}_{T_{n+1}-T_n,T_n}]) \\
&\le \mathcal{W}_1([X^{x_n,n}_{T_{n+1}-T_n}],[X^{y_n,n}_{T_{n+1}-T_n}]) + \mathcal{W}_1([X^{y_n,n}_{T_{n+1}-T_n}],[Y^{y_n}_{T_{n+1}-T_n,T_n}]) \\
&\le \underbrace{Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n)}}_{\mu_{n+1}} \mathcal{W}_1([X^{x_0}_{T_n}],[Y^{x_0}_{T_n}]) + \underbrace{Ce^{C_1 a_{n+1}^{-2}}(a_n - a_{n+1})\rho_{n+1}^{-1}}_{\lambda_{n+1}}\mathbb{E}V(Y^{x_0}_{T_n}),
\end{aligned}$$

The convergence is controlled by

$$\lambda_{n+1} := Ce^{C_1 a_{n+1}^{-2}}(a_n - a_{n+1})\rho_{n+1}^{-1}$$

with

$$a_n \simeq \frac{A}{\sqrt{\log(T_n)}}$$
$$T_{n+1} \simeq Cn^{\beta+1}$$
$$a_n - a_{n+1} \asymp \frac{1}{n\log^{3/2}(n)}$$
$$e^{C_1 a_{n+1}^{-2}} \simeq n^{(\beta+1)C_1/A^2}$$
$$\rho_n^{-1} = e^{C_2 a_{n+1}^{-2}} \simeq n^{(\beta+1)C_2/A^2}$$

$\implies$ Choosing $A > 0$ large enough yields the convergence to 0 of $\mathcal{W}_1([X^{x_0}_{T_{n+1}}],[Y^{x_0}_{T_{n+1}}])$ at rate $n^{-(1-(\beta+1)(C_1+C_2)/A^2)}$. Then:

$$\mathcal{W}_1([Y^{x_0}_{T_{n+1}}],\nu_{a_{n+1}}) \le \mathcal{W}_1([Y^{x_0}_{T_{n+1}}],[X^{x_0}_{T_{n+1}}]) + \mathcal{W}_1([X^{x_0}_{T_{n+1}}],\nu_{a_{n+1}})$$
$$\lessapprox CV(x_0)n^{-(1-(\beta+1)(C_1+C_2)/A^2)}$$
$$\mathcal{W}_1([Y^{x_0}_{T_{n+1}}],\nu^\star) \le \mathcal{W}_1([Y^{x_0}_{T_{n+1}}],[X^{x_0}_{T_{n+1}}]) + \mathcal{W}_1([X^{x_0}_{T_{n+1}}],\nu^\star) \lessapprox CV(x_0)a_n$$

$$\bar{Y}^{x_0}_{\Gamma_{n+1}} = \bar{Y}_{\Gamma_n} + \gamma_{n+1}\left(b_{a(\Gamma_n)}(\bar{Y}^{x_0}_{\Gamma_n}) + \zeta_{n+1}(\bar{Y}^{x_0}_{\Gamma_n})\right) + a(\Gamma_n)\sigma(\bar{Y}^{x_0}_{\Gamma_n})(W_{\Gamma_{n+1}} - W_{\Gamma_n})$$

γ_{n+1} decreasing to 0, $\quad \sum_n \gamma_n = \infty, \quad \sum_n \gamma_n^2 < \infty, \quad \Gamma_n = \gamma_1 + \cdots + \gamma_n,$

$\forall x, \ \mathbb{E}[\zeta_n(x)] = 0.$

We adopt the same strategy of proof to bound $\mathcal{W}_1(X, \bar{Y})$.

• $\mathrm{d}_{TV_1}(\pi_1, \pi_2) = \sup\left\{\int_{\mathbb{R}^d} f(x)(\pi_1 - \pi_2)(dx) :\ f : \mathbb{R}^d \to \mathbb{R},\ \|f\|_\infty = 1\right\}.$

• We follow a similar *domino strategy*; the main difference is the short term bound

$$\left|P^Y_{T_{n+1}-\gamma-T_n,T_n} \circ (P^Y_{\gamma,T_{n+1}-\gamma} - P^{X,n}_{\gamma})f(x)\right|.$$

• [Pagès-Panloup 2020] modify the *domino strategy* by regrouping the last terms and establishing *domino* Malliavin bounds using the regularizing effect of the kernel.

• In our case we propose a simpler method allowing to track the dependency in $a(t)$. We establish total variation bounds of d_{TV} between an SDE and its one-step Euler-Maruyama scheme in small time (or more generally between two SDEs with close coefficients), giving rate $\sim \gamma^{1/2}$ ([Bras-Pages-Panloup 2021]).

Thank you for your attention !