Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise in L^1-Wasserstein distance

Pierre BRAS and Gilles PAGÈS

Sorbonne Université

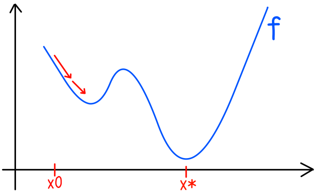
19 octobre 2021

The Langevin Equation in $\mathbb{R}^d$

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

- $V : \mathbb{R}^d \to \mathbb{R}^+$: coercive function to be minimized
- $\sigma > 0$: noise parameter
- Invariant Gibbs measure : $\nu_\sigma(dx) \propto \exp\left(-2V(x)/\sigma^2\right) dx$

- Solve optimization problem with gradient descent type algorithm : $\min_{x\in\mathbb{R}^d} V(x)$.
- Exogenous noise σ added to escape local minima ('traps') and explore the state space (SGLD algorithms)
- For small σ, ν_σ is concentrated around $\mathrm{argmin}(V)$.



The Langevin-Simulated Annealing Equation

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

- $V : \mathbb{R}^d \to \mathbb{R}^+$: coercive function to be minimized
- $\sigma > 0$: noise parameter
- $a : \mathbb{R}^+ \to \mathbb{R}^+$ non-increasing with $a(t) \underset{t\to\infty}{\longrightarrow} 0$.

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

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}(\nabla V(\bar{X}_n) + \zeta_{n+1}) + a(\gamma_1 + \cdots + \gamma_n)\sigma\sqrt{\gamma_{n+1}}\mathcal{N}(0, I_d),$$

(γ_n) : decreasing step sequence,

(ζ_n) : noise of SGD with $\mathbb{E}[\zeta_n] = 0$.

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

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

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

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

- Prove the convergence in $\mathcal{W}_1$ of Y_t and $\bar{Y}_t$ to $\nu^\star$ (supported by argmin(V))
- $$\mathcal{W}_1(Y_t,\nu^\star) \le \mathcal{W}_1(Y_t,\nu_{a(t)}) + \mathcal{W}_1(\nu_{a(t)},\nu^\star)$$
  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
  $$\mathcal{W}_1(Y_t^{x_0},\nu_{a(t)}) \le C_\alpha \max(1+|x_0|, V(X_0))t^{-\alpha}$$
  $$\mathcal{W}_1(\bar{Y}_t^{x_0},\nu_{a(t)}) \le C_\alpha \max(1+|x_0|, V^2(X_0))t^{-\alpha}$$
  for every $\alpha < 1$.
- **Assumptions :**
  1. V is strongly convex outside some compact set
  2. σ is bounded and elliptic : $\sigma\sigma^\top \ge \sigma_0 I_d$, $\sigma_0 > 0$.
  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 for $\mathcal{W}_1$ (and d_{TV})
- Recall :

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

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

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

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] :

$$\begin{aligned} &\mathcal{W}_1(X^x_t, X^y_t) \leq Ce^{-\rho t}|x - y| \\ \implies &\mathcal{W}_1(X^x_t, \nu) \leq Ce^{-\rho t}(1 + |x|). \end{aligned}$$

• 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.

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

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 :

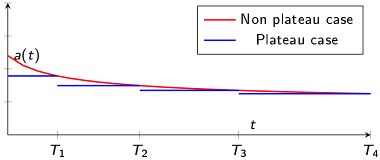
$$\mathcal{W}_1(X_{T_{n+1}}, \nu_{a_{n+1}} \ |X_{T_n}) \leq Ce^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \mathbb{E}\left[|\nu_{a_{n+1}} - X_{T_n}| \ |X_{T_n}\right]$$

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

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

And we iterate :

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



$$\mu_n := Ce^{C_1/a_n^2}e^{-\rho_{a_n}(T_n - T_{n-1})}, \quad \rho_{a_n} = e^{-C_2/a_n^2}.$$

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

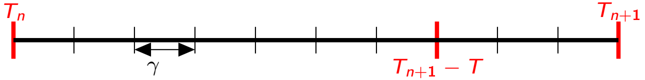
$$\mathcal{W}_1([X_{T_{n+1}}^{x_0}], \nu_{a_{n+1}}) \leq C(1 + |x_0|)\mu_n a_n,$$

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

$$\mathcal{W}_1([X_{T_{n+1}}^{x_0}], \nu^\star) \leq \mathcal{W}_1([X_{T_{n+1}}^{x_0}], \nu_{a_{n+1}}) + \mathcal{W}_1(\nu_{a_{n+1}}, \nu^\star) \leq Ca_n(1 + |x_0|).$$

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 $> T$:

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

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

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

using Taylor formula up to order 4.

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

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

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

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

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

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

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

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

$$\gamma_{n+1} \text{ decreasing to } 0, \quad \sum_n \gamma_n = \infty, \quad \sum_n \gamma_n^2 < \infty,$$

$$\Gamma_n = \gamma_1 + \cdots + \gamma_n.$$

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

Thank you for your attention !