Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise

Pierre BRAS and Gilles PAGÈS

Sorbonne Université

December 9, 2021

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 $\text{argmin}(V) := \{x \in \mathbb{R}^d : \ V(x) = \min_{\mathbb{R}^d} V\}$.

Objective : find argmin(V).

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

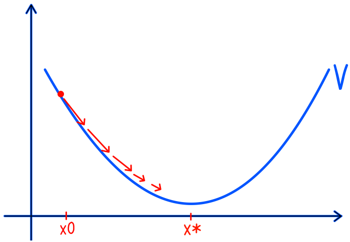
– $\{\Phi_x : \ x \in \mathbb{R}^d\}$ family of functions $\Phi_x : \mathbb{R}^{d'} \to \mathbb{R}$ parametrized by $x \in \mathbb{R}^d$ (e.g. Φ_x is a neural function).
– for $1 \leq i \leq N$, $(u_i, v_i) \in \mathbb{R}^{d'} \times \mathbb{R}$: data associated to a regression problem
– We want to find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

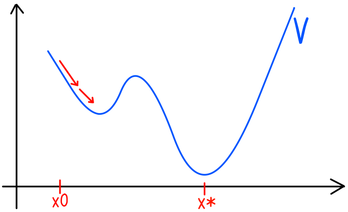
$$\Longrightarrow \quad \text{Find } \min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} (\Phi_x(u_i) - v_i)^2 =: \min_{x \in \mathbb{R}^d} V(x).$$

• 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**

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

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

• Exogenous noise σdW_t added to escape local minima ('traps') and explore the state space.

• For small σ, ν_σ is concentrated around argmin(V):
Solve the Langevin equation $\Longrightarrow$ approximation of ν_σ $\Longrightarrow$ approximation of argmin(V).

- We have $\nu_\sigma \underset{\sigma\to 0}{\longrightarrow} \text{argmin}(V)$ 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 argmin(V)
- Schedule $a(t) = A\log^{-1/2}(t)$ then $X_t \underset{t\to\infty}{\longrightarrow} \text{argmin}(V)$ 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.

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

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

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

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

  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 $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 < \infty$, $\Gamma_n := \gamma_1 + \cdots + \gamma_n$.

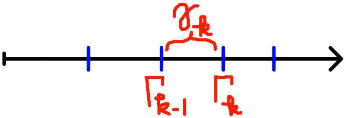
Domino strategy

- ([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 $a(t)$.
$\implies$ 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.

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

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:

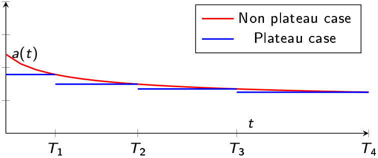
$$\mathcal{W}_1(X_{T_{n+1}}, \nu_{a_{n+1}} \ |X_{T_n}) \le Ce^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \mathbb{E}\left[|\nu_{a_{n+1}} - X_{T_n}| \ |X_{T_n}\right]$$

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

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

And we iterate:

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



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

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

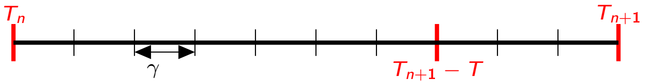
$$\mathcal{W}_1([X_{T_{n+1}}^{x_0}], \nu_{a_{n+1}}) \leq C(1 + |x_0|)\mu_n a_n,$$

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 $> T$:

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

$$= |\mathbb{E}P^{X,n}_{T_{n+1}-T_n-k\gamma} f(X^{x,n}_\gamma) - \mathbb{E}P^X_{T_{n+1}-T_n-k\gamma,n} f(Y^x_{\gamma,T_n+(k-1)\gamma})|$$

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

$$\leq Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n-k\gamma)}[f]_{\text{Lip}}\sqrt{\gamma}(a_n - a_{n+1})$$

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

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

using Taylor formula up to order 4.

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

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

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

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

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

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

$\implies$ 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:

$$\begin{aligned} \mathcal{W}_1([Y_{T_{n+1}}^{x_0}],\nu_{a_{n+1}}) &\le \mathcal{W}_1([Y_{T_{n+1}}^{x_0}],[X_{T_{n+1}}^{x_0}]) + \mathcal{W}_1([X_{T_{n+1}}^{x_0}],\nu_{a_{n+1}}) \\ &\lessapprox CV(x_0)n^{-(1-(\beta+1)(C_1+C_2)/A^2)} \\ \mathcal{W}_1([Y_{T_{n+1}}^{x_0}],\nu^\star) &\le \mathcal{W}_1([Y_{T_{n+1}}^{x_0}],[X_{T_{n+1}}^{x_0}]) + \mathcal{W}_1([X_{T_{n+1}}^{x_0}],\nu^\star) \lessapprox CV(x_0)a_n \end{aligned}$$

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

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 $\mathcal{W}_1(X, \bar{Y})$.

Thank you for your attention !