

Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise

Pierre BRAS and Gilles PAGÈS, presented by Pierre BRAS

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

Presented at the 14th conference on Monte Carlo Methods and Applications.

June 28, 2023



Optimization problem

Let $V : \mathbb{R}^d \rightarrow \mathbb{R}$ be \mathcal{C}^1 , coercive (i.e. $V(x) \rightarrow +\infty$ as $|x| \rightarrow \infty$).

Objective : find $\operatorname{argmin}(V)$.

• Example : Regression as an optimization problem

- $\{\Phi_x : x \in \mathbb{R}^{d'}\}$ family of functions $\Phi_x : \mathbb{R}^{d'} \rightarrow \mathbb{R}$ 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 , $\Phi_x(u_i) \approx v_i$

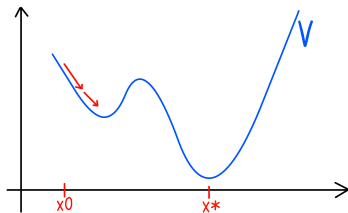
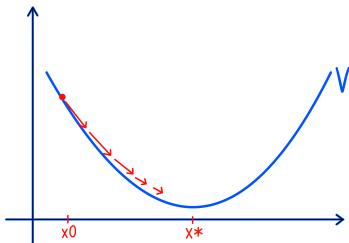
$$\implies \text{Find } \min_{x \in \mathbb{R}^{d'}} \frac{1}{N} \sum_{i=1}^N (\Phi_x(u_i) - v_i)^2 =: \min_{x \in \mathbb{R}^{d'}} V(x).$$

- 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 and explore:

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

\implies SGLD algorithms (Stochastic Gradient Langevin Dynamics)

- The continuous version becomes:

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

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

- Its invariant measure is the **Gibbs measure**

$$\nu_\sigma(x) dx \propto e^{-2(V(x) - V^*)/\sigma^2} dx.$$

- For small σ , ν_σ is concentrated around $\operatorname{argmin}(V)$:

Solve the Langevin equation \implies approximation of $\nu_\sigma \implies$ approximation of $\operatorname{argmin}(V)$.

- Another possibility : make $\sigma \rightarrow 0$ while iterating the algorithm :

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

where $a(t)$ is decreasing and $a(t) \xrightarrow[t \rightarrow 0]{} 0$.

The continuous version becomes :

Langevin-Simulated Annealing Equation

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

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

- 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.

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

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

- Correction term so that $\nu_{a(t)} \propto \exp(-2V(x)/a^2(t))$ is still the "instantaneous" invariant measure
- (\tilde{Y}_t) : the corresponding discrete scheme.

Objectives and assumptions

- ν^* : limit measure of ν_σ as $\sigma \rightarrow 0$.
- We use the L^1 -Wasserstein distance:

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

and we show that

$$\mathcal{W}_1(Y_t, \nu^*) \rightarrow 0 \quad \text{and} \quad \mathcal{W}_1(\bar{Y}_t, \nu^*) \rightarrow 0.$$

- We have

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

The convergence is limited by the slowness of $a(t)$ as

$\mathcal{W}_1(\nu_{a(t)}, \nu^*) \asymp a(t) \asymp \log^{-1/2}(t)$. In fact we also prove

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

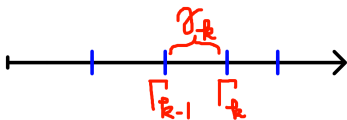
for every $\alpha < 1$.

- **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 $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 < \infty$,
 $\Gamma_n := \gamma_1 + \dots + \gamma_n$.

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

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



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

$$\forall t \geq t_0, \mathcal{W}_1(X_t^x, X_t^y) \leq Ce^{-\rho t}|x - y| \quad \text{and} \quad \mathcal{W}_1(X_t^x, \pi^*) \leq Ce^{-\rho t}(1 + |x|).$$

[Pagès-Panloup 2020]

- Problems before applying the domino strategy: non-homogeneous Markov chain + the ellipticity parameter fades away in $a(t)$.

⇒ 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

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

Then

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

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

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:

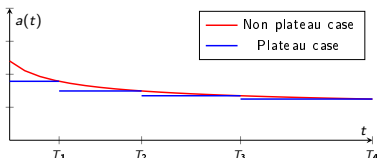
$$\mathcal{W}_1(X_{T_{n+1}}, \nu_{a_{n+1}} | X_{T_n}) \leq C e^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \mathbb{E} [|\nu_{a_{n+1}} - X_{T_n}| | X_{T_n}]$$

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

$$\begin{aligned} \mathcal{W}_1(X_{T_{n+1}}^{\text{x0}}, \nu_{a_{n+1}}) &\leq C e^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \mathcal{W}_1(X_{T_n}^{\text{x0}}, \nu_{a_{n+1}}) \\ &\leq C e^{C_1/a_{n+1}^2} e^{-\rho_{a_{n+1}}(T_{n+1}-T_n)} \left(\mathcal{W}_1(X_{T_n}^{\text{x0}}, \nu_{a_n}) + \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) \right). \end{aligned}$$

And we iterate:

$$\begin{aligned} \mathcal{W}_1(X_{T_{n+1}}^{\text{x0}}, \nu_{a_{n+1}}) &\leq \mu_{n+1} \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) + \mu_{n+1} \mu_n \mathcal{W}_1(\nu_{a_{n-1}}, \nu_{a_n}) + \cdots \\ &\quad + \mu_{n+1} \cdots \mu_1 \mathcal{W}_1(\nu_{a_0}, \nu_{a_1}) + \mu_{n+1} \cdots \mu_1 \mathcal{W}_1(\delta_{x_0}, \nu_{a_0}), \\ \mu_n &:= C e^{C_1/a_n^2} e^{-\rho_{a_n}(T_n - T_{n-1})}. \end{aligned}$$



$$\begin{aligned}\mathcal{W}_1(X_{T_{n+1}}^{x_0}, \nu_{a_{n+1}}) &\leq \mu_{n+1} \mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) + \mu_{n+1} \mu_n \mathcal{W}_1(\nu_{a_{n-1}}, \nu_{a_n}) + \cdots \\ &\quad + \mu_{n+1} \cdots \mu_1 \mathcal{W}_1(\nu_{a_0}, \nu_{a_1}) + \mu_{n+1} \cdots \mu_1 \mathcal{W}_1(\delta_{x_0}, \nu_{a_0}), \\ \mu_n &= C e^{C_1/a_n^2} e^{-\rho_{a_n}(T_n - T_{n-1})}, \quad \rho_{a_n} = e^{-C_2/a_n^2}\end{aligned}$$

We use (technical)

$$\mathcal{W}_1(\nu_{a_n}, \nu_{a_{n+1}}) \leq C(a_n - a_{n+1}), \quad \mathcal{W}_1(\nu_{a_n}, \nu^*) \leq C a_n.$$

We now choose

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

yielding

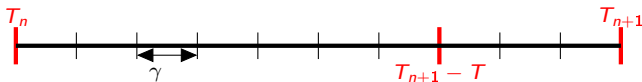
$$\mathcal{W}_1(X_{T_{n+1}}^{x_0}, \nu_{a_{n+1}}) \leq C(1 + |x_0|) \mu_n a_n,$$

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

$$\mathcal{W}_1(X_{T_{n+1}}^{x_0}, \nu^*) \leq \mathcal{W}_1(X_{T_{n+1}}^{x_0}, \nu_{a_{n+1}}) + \mathcal{W}_1(\nu_{a_{n+1}}, \nu^*) \leq C(1 + |x_0|) a_n.$$

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 $T > 0$:

$$\begin{aligned} & \left| \mathbb{E}f(X_{T_{n+1}-T_n}^{X,n}) - \mathbb{E}f(Y_{T_{n+1}-T_n}^{X,n}) \right| \\ & \leq \sum_{k=1}^{\lfloor (T_{n+1}-T_n-T)/\gamma \rfloor} \left| P_{(k-1)\gamma, T_n}^Y \circ (P_{\gamma, T_n+(k-1)\gamma}^Y - P_{\gamma}^{X,n}) \circ P_{T_{n+1}-T_n-k\gamma}^{X,n} f(x) \right| \\ & \quad + \sum_{k=\lfloor (T_{n+1}-T_n-T)/\gamma \rfloor + 1}^{\lfloor (T_{n+1}-T_n)/\gamma \rfloor} \left| P_{(k-1)\gamma, T_n}^Y \circ (P_{\gamma, T_n+(k-1)\gamma}^Y - P_{\gamma}^{X,n}) \circ P_{T_{n+1}-T_n-k\gamma}^{X,n} f(x) \right| \end{aligned}$$

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

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

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

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

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

$$\mathcal{W}_1(X_{T_{n+1}-T_n}^{x,n}, Y_{T_{n+1}-T_n, T_n}^x) \leq \underbrace{Ce^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) \rho_{n+1}^{-1}}_{=: \lambda_{n+1}} V(x).$$

With $x_n := X_{T_n}^{x_0}$, $y_n = Y_{T_n}^{y_0}$:

$$\begin{aligned} \mathcal{W}_1(X_{T_{n+1}}^{x_0}, Y_{T_{n+1}}^{y_0}) &= \mathcal{W}_1(X_{T_{n+1}-T_n}^{x_n,n}, Y_{T_{n+1}-T_n, T_n}^{y_n}) \\ &\leq \mathcal{W}_1(X_{T_{n+1}-T_n}^{x_n,n}, X_{T_{n+1}-T_n}^{y_n,n}) + \mathcal{W}_1(X_{T_{n+1}-T_n}^{y_n,n}, Y_{T_{n+1}-T_n, T_n}^{y_n}) \\ &\leq \underbrace{Ce^{C_1 a_{n+1}^{-2}} e^{-\rho_{n+1}(T_{n+1}-T_n)}}_{\mu_{n+1}} \mathcal{W}_1(X_{T_n}^{x_0}, Y_{T_n}^{y_0}) + \underbrace{Ce^{C_1 a_{n+1}^{-2}} (a_n - a_{n+1}) \rho_{n+1}^{-1}}_{\lambda_{n+1}} \mathbb{E} V(Y_{T_n}^{y_0}), \end{aligned}$$

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

$$a_n \simeq \frac{A}{\sqrt{\log(T_n)}}$$

$$T_{n+1} \simeq Cn^{\beta+1}$$

$$a_n - a_{n+1} \asymp \frac{1}{n \log^{3/2}(n)}$$

$$e^{C_1 a_{n+1}^{-2}} \simeq n^{(\beta+1)C_1/A^2}$$

$$\rho_n^{-1} = e^{C_2 a_{n+1}^{-2}} \simeq n^{(\beta+1)C_2/A^2}$$

$$\implies \lambda_n \asymp n^{-(1-(\beta+1)(C_1+C_2)/A^2)}$$

and we choose A large enough such that

$$1 - (\beta + 1)(C_1 + C_2)/A^2 > 0.$$

Then:

$$\mathcal{W}_1(Y_{T_{n+1}}^{x_0}, \nu_{a_{n+1}}) \leq \mathcal{W}_1(Y_{T_{n+1}}^{x_0}, X_{T_{n+1}}^{x_0}) + \mathcal{W}_1(X_{T_{n+1}}^{x_0}, \nu_{a_{n+1}})$$

$$\lesssim CV(x_0) n^{-(1-(\beta+1)(C_1+C_2)/A^2)}$$

$$\mathcal{W}_1(Y_{T_{n+1}}^{x_0}, \nu^*) \leq \mathcal{W}_1(Y_{T_{n+1}}^{x_0}, X_{T_{n+1}}^{x_0}) + \mathcal{W}_1(X_{T_{n+1}}^{x_0}, \nu^*) \lesssim CV(x_0) a_n$$

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

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

$$\forall x, \mathbb{E}[\zeta_n(x)] = 0 \quad (\text{mini-batch noise}).$$

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

- We compare the performances of adaptive Langevin-Simulated Annealing algorithms versus vanilla SGLD.
- We train a fully connected feedforward neural network, two hidden layers, ReLU activation, 128 units each on the MNIST dataset.
- We choose σ as a diagonal matrix which is the square root of the preconditioner.

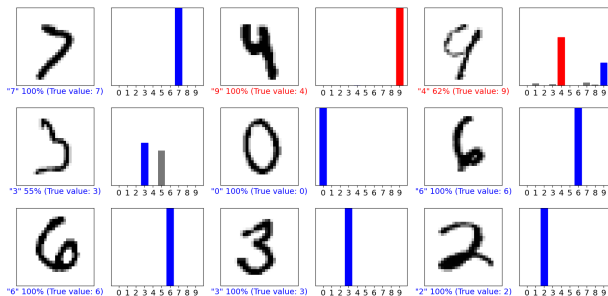
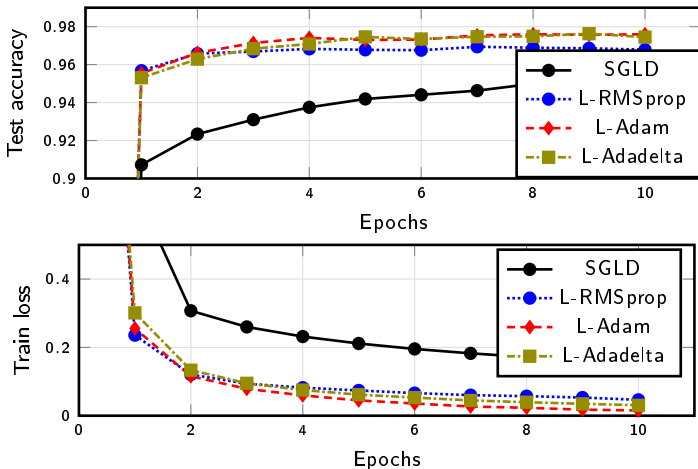


Figure: MNIST dataset

- We compare the performances of adaptive Langevin-Simulated Annealing algorithms versus vanilla SGLD.
- We train a fully connected feedforward neural network, two hidden layers, ReLU activation, 128 units each on the MNIST dataset.
- We choose σ as a diagonal matrix which is the square root of the preconditioner.



- In an other paper [Bras 2022], we investigate the benefits of Langevin optimizers with their respective non-Langevin counterparts optimizers and we show that the deeper the neural network is, the greater are the gains provided by Langevin algorithms.

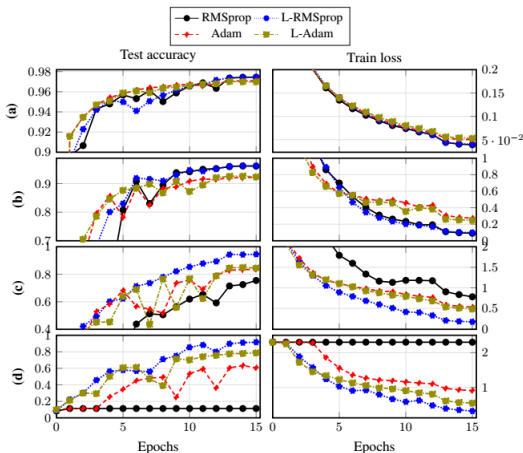


Figure: Training of neural networks of various depths on the MNIST dataset. (a): 3 hidden layers, (b): 20 hidden layers, (c): 30 hidden layers, (d): 40 hidden layers.

Simulations

- In another paper [Bras-Pagès 2023], we apply Langevin algorithms to deep stochastic control with gains.

Stochastic Optimal Control by Neural Networks

$$\begin{aligned} \min_u J(u) &:= \mathbb{E} \left[\int_0^T G(X_t) dt + F(X_T) \right], \\ dX_t &= b(X_t, u_t) dt + \sigma(X_t, u_t) dW_t, \quad t \in [0, T], \\ u_t &= u_\theta(t, X_t), \quad u_\theta \text{ neural network.} \end{aligned}$$

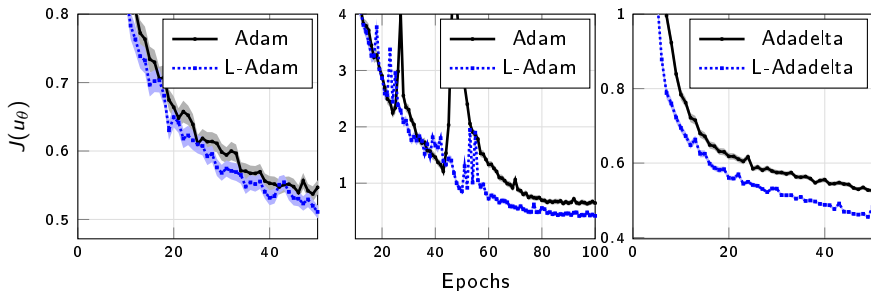


Figure: Training for the deep hedging control problem [Buehler et al. 2019] on a Heston model with $N = 30, 50, 50$ discretization times.

Thank you for your attention !