Accelerated Hamiltonian Monte Carlo Algorithms for Stochastic Optimization and Their Application in Finance

Zigan Wang

joint work with
Luxu Liang (Tsinghua), Ariel Neufeld (NTU), and Ying Zhang (HKUSTGZ)

Overview

1. Background
2. Algorithm
3. Numerical results
4. Takeaways

Background

AI Innovation in Finance

In today's world, AI technologies have been rapidly revolutionizing all aspects of the financial research and industry. Some examples include

- **Asset pricing:** Han et al. (2018); Becker et al. (2019); Gu et al. (2020, 2021); Chen et al. (2024a); Bryzgalova et al. (2025).
- **Insurance and Risk Management:** Tsang and Wong (2020); Jin et al. (2021); Fernandez-Arjona and Filipović (2022); Almeida et al. (2023); Chen et al. (2024b).
- **Portfolio optimization and Algorithmic trading:** Jaimungal (2022); Cartea et al. (2023); Jaimungal et al. (2023); Ni et al. (2025).

AI Innovation in finance

- **Main idea:** Approximate the target function $F(\cdot)$ with a neural network rather than a fixed parametric family:

$$F(\cdot) \approx \begin{cases} F^{\text{simple}}(\cdot;\theta), & \text{(traditional approach)}, \\ F(\cdot;\theta), & \text{(AI approach)}. \end{cases}$$

  $F^{\text{simple}}(\cdot;\theta)$: (piecewise) linear, quadratic, or otherwise analytically tractable parametric function—including those derived from *elegant mathematical models*
  $F(\cdot;\theta)$: a neural network parameterized by θ.

- Existing results[1] justify deep neural networks via the universal approximation theorem: for any $\epsilon > 0$, there exists a network with parameters θ^* such that

$$|F(\cdot) - F(\cdot;\theta^*)| < \epsilon.$$

[1] See, e.g., Buehler et al. (2019), Tsang and Wong (2020).

Challenges

Advantages: scalability, data-driven (vs. model-driven), universal approximation.

Challenge 1: Actually, it's not trivial to find the optimal parameters **:(**

- Its high-dimensional, non-convex landscape—with many local minima and flat regions—makes the optimization process inherently slow and challenging.

Challenges

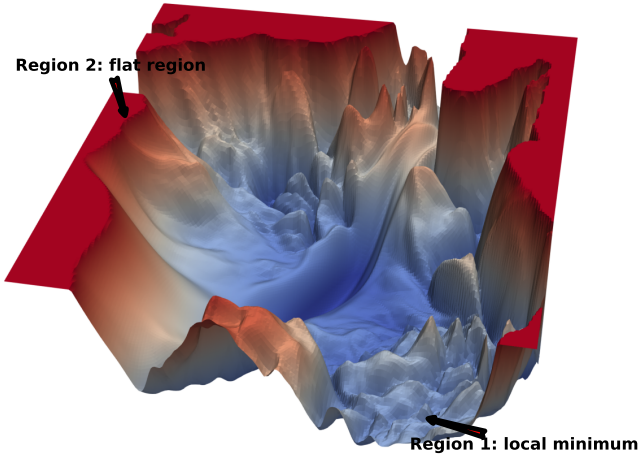


Figure 1: Highlighted local minima (region 1) and flat region (region 2) in the 3D landscape.

Challenges

Challenge 2: In financial applications, we frequently face cases where the stochastic gradients are **discontinuous**, e.g.,

- Credit scoring and lending decisions, threshold rules—e.g., approving a loan if the credit score exceeds a set value (Einav et al., 2013).
- Progressive taxes, withholding rules, and tiered bonuses are piecewise functions with naturally discontinuous gradients at their boundaries (Chetty, 2012).
- Marketing budget allocation and advertising, saturation effects often occur (Gordon et al., 2019), and the response is frequently modeled as a piecewise-linear function (Ghose and Todri-Adamopoulos, 2016).

Challenges

Challenge 2: In financial applications, we frequently face cases where the stochastic gradients are **discontinuous**, e.g.,

- Quantile estimation and Conditional Value at Risk (CVaR) (Takeuchi et al., 2006; Bardou et al., 2009).
- Vector Quantization (VQ) techniques in option pricing, hedging, and market simulation (Pagès et al., 2004).
- Regularized optimization problems involving ReLU neural networks (Safran and Shamir, 2018).

An Example on Discontinuous Stochastic Gradients

We aim to compute VaR and CVaR. Following Bardou et al. (2008); Chow et al. (2018),

$$\min_{\theta} u(\theta) := \min_{\theta} \left(\mathbb{E}[\theta + \frac{1}{1-q}(g(X) - \theta)_+] + \lambda_r |\theta|^2 \right),$$

where $0 < q < 1$, g may model complex payoffs, and X can accommodate a broad class of asset distributions, including stochastic/local volatility models. Then

$$\mathrm{VaR}_q(g(X)) = \arg\min_{\theta} u(\theta), \quad \mathrm{CVaR}_q(g(X)) = \min_{\theta} u(\theta).$$

The stochastic gradient is

$$H(\theta, x) = 1 - \frac{1}{1-q} \mathbb{1}_{\{g(x) \geq \theta\}} + 2\lambda_r \theta = -\frac{q}{1-q} + \frac{1}{1-q} \mathbb{1}_{\{g(x) < \theta\}} + 2\lambda_r \theta,$$

which is discontinuous. Later, we will show that our assumptions cover these scenarios.

Limitations of Existing Methods

Limitations of popular optimization methods (SGD and its variants) in finance:

- SGD and its variants: lack global convergence guarantees and, in non-convex settings, are only assured to reach a **stationary point**.
- Adam can become **unstable** when training deep models with large batch sizes (Molybog et al., 2023), yet remains widely used in practice.
- They typically assume smooth gradients—an assumption often violated in finance, where stochastic gradients are often **discontinuous**.
  - *Subgradient* or *gradient-free* methods suffer from slow convergence, high variance, and step-size sensitivity (Bottou et al., 2018; Nesterov and Spokoiny, 2017).
- A gap remains between theory and practice, especially with discontinuous stochastic gradients.

Core Research Question

Can we develop an **optimizer** that enjoys global convergence guarantees in non-convex settings with discontinuous stochastic gradients, while outperforming SGD and Adam in real-world financial applications?

Our Contribution

To this end, we

- introduce a new stochastic optimization algorithm, **A**ccelerated **S**tochastic **G**radient **H**amiltonian **M**onte **C**arlo (**A-SGHMC**), leveraging **exponential discretization** and a **boosting function** to achieve superior empirical performance,
- develop a theoretical framework for optimization with **discontinuous** stochastic gradients, giving upper bounds on Wasserstein-1/2 distances and expected excess risk, which can be controlled arbitrarily small,
- and demonstrate the strong **empirical performance** in insurance claim prediction.

Comparison with Related SGHMC Work

Table 1: Comparison with the most related SGHMC literature.

Paper	Lipschitz need?	Discontinuous allowed?	Non-asymptotic W_2?	Boosting function?	Exponential discretization?
Chau and Rásonyi (2022)	Global Lipschitz	×	✓ (cannot be arbitrarily small)	×	×
Gao et al. (2022)	Global Lipschitz	×	✓ (cannot be arbitrarily small)	×	✓
Akyildiz and Sabanis (2024)	Local Lipschitz	×	✓	×	×
This paper	Lipschitz-in-average	✓	✓	✓	✓

Algorithm

Optimization Problem

- Recall that we consider the following unconstrained stochastic optimization problem:

$$\min_{\theta \in \mathbb{R}^d} u(\theta) := \mathbb{E}[U(\theta, X)].$$

- Let $h := \nabla u$ and define the stochastic gradient of u, denoted by $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$, as an unbiased estimator of h, i.e., $h(\theta) = \mathbb{E}\left[H\left(\theta, X_0\right)\right]$.
- We consider the **second-order** Langevin Stochastic Differential Equation (SDE):

$$\begin{aligned} \mathrm{d}V_t &= -\gamma V_t \, \mathrm{d}t - h\left(\theta_t\right) \mathrm{d}t + \sqrt{2\gamma\beta^{-1}} \mathrm{d}B_t, \\ \mathrm{d}\theta_t &= V_t \, \mathrm{d}t, \end{aligned} \tag{1}$$

where $(\theta_t, V_t)_{t \in \mathbb{R}_+}$ are the position and momentum processes, $h = \nabla u$, $\beta > 0$ the inverse temperature, $\gamma > 0$ the friction coefficient, and B_t a d-dimensional Brownian motion.

Langevin-based Algorithms

Why consider the second-order Langevin SDE over the first-order?

- For the continuous-time second-order dynamics, Eberle et al. (2019) show that the second-order SDE converges to its stationary distribution faster than the best-known convergence rate of the first-order SDE in the 2-Wasserstein metric under certain assumptions, even when $u(\cdot)$ is non-convex.
- Gao et al. (2022) demonstrate that for a class of non-convex problems, SGHMC improves upon the (vanilla) SGLD algorithm in terms of gradient complexity, i.e., the total number of stochastic gradients required to reach a global minimum.
- Extensive experiments indicate that SGHMC outperforms SGLD dynamics across various applications (Chen et al., 2015; Li et al., 2019; Wang et al., 2023).

Langevin-based Algorithms

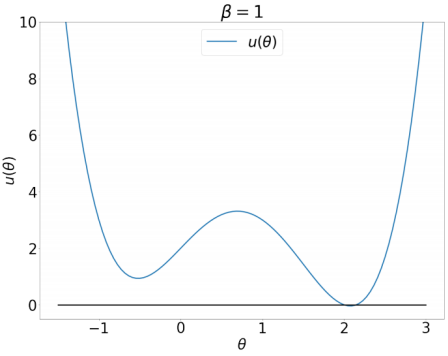
- **Fact:** Under suitable assumptions, SDE (1) converges to its invariant measure (see, e.g., Pavliotis (2014))

$$\bar{\pi}_\beta(\theta, v) \propto \exp\left(-\beta\left(u(\theta) + \frac{1}{2}|v|^2\right)\right). \tag{2}$$

  Importantly, the marginal distribution of (2) with respect to θ coincides with the target distribution π_β. Hence, sampling from (2) in the extended space and discarding the momentum coordinate (i.e., retaining only θ) is equivalent to sampling directly from π_β.

- Note that $\bar{\pi}_\beta$ concentrates on the minimizer of $u(\theta)$ when β is large enough. Therefore, the optimization problem is closely linked to the problem of sampling from $\bar{\pi}_\beta$, see, e.g., Hwang (1980).

Langevin-based Algorithms



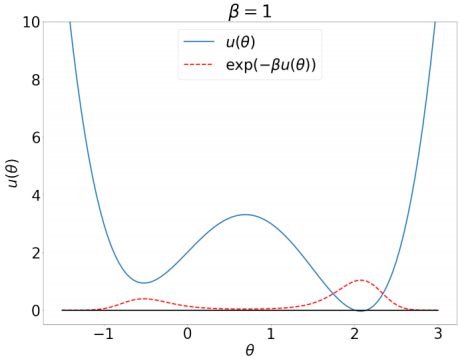


Figure 2: Our objective function and target distribution when $\beta = 1$.

Langevin-based Algorithms

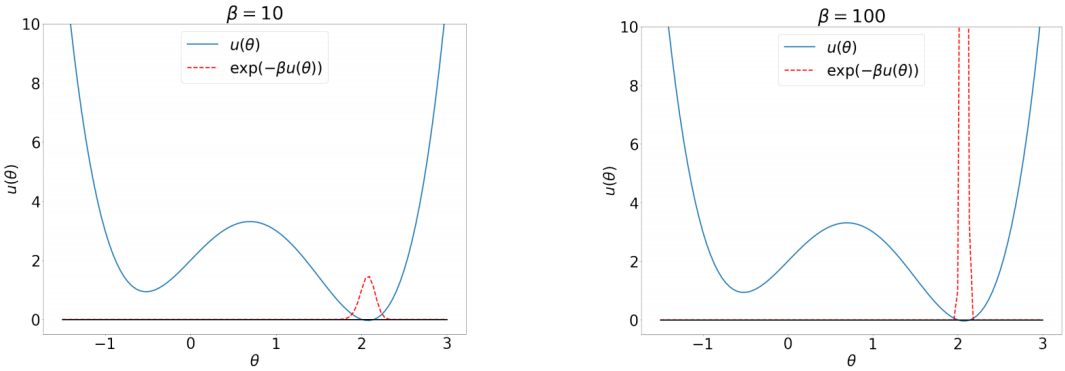


Figure 3: Our objective function and target distribution when $\beta = 10$ and $\beta = 100$.

Langevin-based Algorithms

- Discretizing (1) yields the Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014) algorithm: for $n \in \mathbb{N}$,

$$V_0^\eta := V_0, \quad V_{n+1}^\eta = V_n^\eta - \eta\left[\gamma V_n^\eta + H(\theta_n^\eta, X_{n+1})\right] + \sqrt{2\gamma\eta\beta^{-1}}\xi_{n+1},$$
$$\theta_0^\eta := \theta_0, \quad \theta_{n+1}^\eta = \theta_n^\eta + \eta V_n^\eta,$$

  where η is the step size, $(X_i)_{n\in\mathbb{N}}$ is an i.i.d. sequence, and ξ_n is a standard Gaussian noise in $\mathbb{R}^d$.
  - The Euler approximation of the second-order Langevin SDE.
  - h may be infeasible $\Rightarrow$ use the stochastic gradient, which is an unbiased estimator of h :

$$h(\theta) = \mathbb{E}[H(\theta, x)].$$

Newly Proposed Algorithm: A-SGHMC

Update rule: For $n \in \mathbb{N}$

$$V_{n+1}^{\eta} = e^{-\gamma\eta} V_n^{\eta} - \gamma^{-1}\left(1 - e^{-\gamma\eta}\right) H_{\varepsilon,\eta}\left(\theta_n^{\eta}, X_{n+1}\right) + \sqrt{2\gamma\beta^{-1}}\xi_n^{V},$$
$$\theta_{n+1}^{\eta} = \theta_n^{\eta} + \gamma^{-1}\left(1 - e^{-\gamma\eta}\right) V_n^{\eta} - \gamma^{-2}\left(\gamma\eta + e^{-\gamma\eta} - 1\right) H_{\varepsilon,\eta}\left(\theta_n^{\eta}, X_{n+1}\right) + \sqrt{2\gamma\beta^{-1}}\xi_n^{\theta},$$

where $\eta > 0$ is the step-size, $\gamma > 0$ is the friction coefficient, $\beta > 0$ is the inverse temperature parameter, ξ_k^V, ξ_k^θ are normal distributed in $\mathbb{R}^d$ satisfying that:

$$\mathbb{E}\xi_n^V \left(\xi_n^V\right)^\top = \frac{1}{2\gamma}\left(1 - e^{-2\gamma\eta}\right) \cdot I_d,$$
$$\mathbb{E}\xi_n^\theta \left(\xi_n^\theta\right)^\top = \frac{1}{2\gamma^3}\left(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3\right) \cdot I_d,$$
$$\mathbb{E}\xi_n^\theta \left(\xi_n^V\right)^\top = \frac{1}{2\gamma^2}\left(1 - 2e^{-\gamma\eta} + e^{-2\gamma\eta}\right) \cdot I_d,$$

Newly Proposed Algorithm: A-SGHMC

$H_{\varepsilon,\eta}(\theta, x) = \left(H^{(1)}_{\varepsilon,\eta}(\theta, x), \cdots, H^{(d)}_{\varepsilon,\eta}(\theta, x)\right) : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is given by, for $i = 1, \cdots, d$,

$$H^{(i)}_{\varepsilon,\eta}\left(\theta^{\eta}_n, X_{n+1}\right) := H^{(i)}\left(\theta^{\eta}_n, X_{n+1}\right)\left(1 + \frac{\eta}{\varepsilon + \left|H^{(i)}\left(\theta^{\eta}_n, X_{n+1}\right)\right|}\right),$$

for every $\theta \in \mathbb{R}^d$, $X_n \in \mathbb{R}^m$, $0 < \varepsilon < 1$.

Features:

- Element-wise adaptive boosting function: to accelerate training speed and prevent the vanishing gradient.
- Exponential discretization: less discretization error, yield tigher upper bound.
- Scaled Gaussian noise: a consequence of the discretization of the Langevin SDE and helpful for the new algorithm to escape local minima.

Assumptions

Assumption 2.1 (Continuity in average for G and local Lipschitz continuity for F)

The function $H : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ takes the form of $H(\theta, x) = F(\theta, x) + G(\theta, x)$, where

1. *There exists a positive constant $L_G > 0$ such that, for all $\theta, \theta' \in \mathbb{R}^d$,*

$$\mathbb{E}\left[|G(\theta, X_0) - G(\theta', X_0)|\right] \leq L_G|\theta - \theta'|.$$

In addition, there exist a measurable function $K_G : \mathbb{R}^m \to (0, \infty)$ such that for any $(\theta, x) \in \mathbb{R}^d \times \mathbb{R}^m$,

$$|G(\theta, x)| \leq K_G(x).$$

2. *There exists a constant L_F and $\rho > 0$ such that for any $(\theta, x), (\theta', x') \in \mathbb{R}^d \times \mathbb{R}^m$,*

$$|F(\theta, x) - F(\theta', x')| \leq L_F(1 + |x| + |x'|)^\rho(|\theta - \theta'| + |x - x'|).$$

Remark. We draw inspiration from Fort et al. (2016); Chau et al. (2019), which focus on stochastic approximation; however, our setting/objectives differ fundamentally from theirs.

Assumptions

Assumption 2.2 (Initial condition on θ_0 and the data process X_n)

The process $\{X_n\}_{n\in\mathbb{N}_0}$ is i.i.d. with $\mathbb{E}\left[|X_0|^{4(\rho+1)}\right] < \infty$ and $\mathbb{E}\left[|K_G(X_0)|^4\right] < \infty$, where $\rho > 0$ and $K_G : \mathbb{R}^m \to (0,\infty)$ are introduced in Assumption 2.1.

Assumption 2.3 (Local dissipativity condition)

There exist measurable functions $A : \mathbb{R}^m \to \mathbb{R}^{d\times d}$ and $B : \mathbb{R}^m \to \mathbb{R}$ such that the following hold: for any $(\theta, x) \in \mathbb{R}^d \times \mathbb{R}^m$,

1. *$\langle \theta, A(x)\theta\rangle \geq 0$ and $\langle F(\theta, x), \theta\rangle \geq \langle \theta, A(x)\theta\rangle - B(x)$,*
2. *The smallest eigenvalue of $\mathbb{E}[|A(X_0)|]$ is a positive real number $a > 0$ and $\mathbb{E}[|B(X_0)|] = b \geq 0$.*

Remark. Assumptions 2.1–2.3 hold in many applications, including quantile estimation, CVaR minimization, and regularized optimization with ReLU networks.

An Example on Discontinuous Stochastic Gradients

Following the VaR and CVaR example, the stochastic gradient is

$$H(\theta, X_0) = 1 - \frac{1}{1-q}\mathbb{1}_{\{g(X_0)\geq\theta\}} + 2\lambda_r\theta = -\frac{q}{1-q} + \frac{1}{1-q}\mathbb{1}_{\{g(X_0)<\theta\}} + 2\lambda_r\theta.$$

Taking $g(x) = x$, we have

$$\begin{aligned}
\mathbb{E}\big[\big|H(\theta, X_0) - H(\theta', X_0)\big|\big] &\leq 2\lambda_r|\theta - \theta'| + \frac{1}{1-q}\mathbb{E}\big[\big|\mathbb{1}_{\{X_0<\theta\}} - \mathbb{1}_{\{X_0<\theta'\}}\big|\big] \\
&\leq 2\lambda_r|\theta - \theta'| + \frac{1}{1-q}\left|\int_{\theta'}^{\theta} f_{X_0}(x)\,dx\right| + \frac{1}{1-q}\left|\int_{\theta}^{\theta'} f_{X_0}(x)\,dx\right| \\
&\leq 2(\lambda_r + \frac{1}{1-q}\|f_{X_0}\|_\infty)|\theta - \theta'|,
\end{aligned}$$

which verifies assumptions.

Main Theorems

Let

$$\eta_{\max} := \min\left\{1, \frac{1}{\gamma}, \frac{\lambda\gamma}{2K_1}, \frac{K_3}{K_2}, \frac{\lambda\gamma}{4K_4}, \frac{K_6}{K_5}, \frac{\lambda\gamma}{2\tilde{K}}\right\} > 0. \tag{3}$$

Theorem 1

Let Assumption 2.1-2.3 hold. Then, for any $\beta > 0$, there exist constants $C_1^\star, C_2^\star, C_3^\star, C_4^\star > 0$ such that, for every $n \in \mathbb{N}_0, 0 < \eta \leq \eta_{\max}$ with $\eta_{\max}$ defined in (3), we obtain

$$W_2\left(\mathcal{L}\left(\theta_n^\eta, V_n^\eta\right), \bar{\pi}_\beta\right) \leq \underbrace{C_1^\star \eta^{1/2} + C_2^\star \eta^{1/4}}_{\text{Discretization Error}} + \underbrace{C_3^\star e^{-C_4^\star \eta n}}_{\text{Invariant Measure Error}},$$

where $C_1^\star = \mathcal{O}\left(d^{1/2}\right), C_2^\star = \mathcal{O}\left(e^d\right), C_3^\star = \mathcal{O}\left(e^d\right)$, and $C_4^\star = \mathcal{O}\left(e^{-d}\right)$. In particular, for any $\epsilon > 0$, if we choose $\eta \leq \min\left\{\frac{\epsilon^2}{9C_1^{\star 2}}, \frac{\epsilon^4}{81C_2^{\star 4}}, \eta_{\max}\right\}$ and $n \geq \frac{\ln\left(3C_3^\star/\epsilon\right)}{C_4^\star \min\left\{\frac{\epsilon^2}{9C_1^{\star 2}}, \frac{\epsilon^4}{81C_2^{\star 4}}, \eta_{\max}\right\}}$, then we obtain $W_2\left(\mathcal{L}\left(\theta_n^\eta, V_n^\eta\right), \bar{\pi}_\beta\right) \leq \epsilon$.

Theorem 2 (Expected Excess Risk)

Let Assumption 2.1-2.3 hold. Then, for any $\beta > 0$, there exist constants $\bar{C}_1^\star, \bar{C}_2^\star, \bar{C}_3^\star, \bar{C}_4^\star > 0$ such that, for every $n \in \mathbb{N}_0, 0 < \eta \leq \eta_{\max}$ with $\eta_{\max}$ defined in (3), we obtain

$$\mathbb{E}\left[u\left(\theta_n^\eta\right)\right] - \inf_{\theta \in \mathbb{R}^d} u(\theta) \leq \underbrace{\bar{C}_1^\star \eta^{1/2} + \bar{C}_2^\star \eta^{1/4} + \bar{C}_3^\star e^{-\bar{C}_4^\star \eta n}}_{\text{Sampling Behavior}} + \underbrace{\frac{d}{2\beta} \log\left(\frac{8eL_h}{\gamma^2 \lambda(1-2\lambda)}\left(\frac{A_c}{d}+1\right)\right)}_{\text{Concentration Property}},$$

where $\bar{C}_1^\star = \mathcal{O}(d^{3/2})$, $\bar{C}_2^\star = \mathcal{O}(e^d)$, $\bar{C}_3^\star = \mathcal{O}(e^d)$, and $\bar{C}_4^\star = \mathcal{O}(e^{-d})$. In particular, for any $\epsilon > 0$, if we first choose $\beta \geq \max\left\{\frac{16d^2}{\epsilon^2}, \frac{4d}{\epsilon}\log\left(\frac{8eL_h}{\gamma^2\lambda(1-2\lambda)}\left(\frac{\lambda u(0)+\lambda|h(0)|+b'/2}{d}+1\right)\right)\right\}$, then choose $\eta \leq \min\left\{\frac{\epsilon^2}{16\bar{C}_1^{\star 2}}, \frac{\epsilon^4}{256\bar{C}_2^{\star 4}}, \eta_{\max}\right\}$ and $n \geq \frac{\ln\left(4\bar{C}_3^\star/\epsilon\right)}{\bar{C}_4^\star \min\left\{\frac{\epsilon^2}{16\bar{C}_1^{\star 2}}, \frac{\epsilon^4}{256\bar{C}_2^{\star 4}}, \eta_{\max}\right\}}$, then we obtain $\mathbb{E}\left[u\left(\theta_n^\eta\right)\right] - \inf_{\theta \in \mathbb{R}^d} u(\theta) \leq \epsilon$.

Comparison with Related SGHMC Work

Table 2: Comparison with the most related SGHMC literature.

Paper	Lipschitz need?	Discontinuous allowed?	Non-asymptotic W_2?	Boosting function?	Exponential discretization?
Chau and Rásonyi (2022)	Global Lipschitz	×	✓ (cannot be arbitrarily small)	×	×
Gao et al. (2022)	Global Lipschitz	×	✓ (cannot be arbitrarily small)	×	✓
Akyildiz and Sabanis (2024)	Local Lipschitz	×	✓	×	×
This paper	Lipschitz-in-average	✓	✓	✓	✓

Numerical results

Numerical experiment: Network-based Non-linear Gamma Regression

- We study non-linear Gamma regression, which extends the classical linear model by replacing its linear predictor with a neural network to capture non-linear relationships among inputs. This approach is widely used in the insurance industry to predict claim sizes (Frees, 2014; Garrido et al., 2016; Yang et al., 2018).
- Here, we provide an example of a non-linear Gamma regression model based on neural networks which can be used to predict a target variable $Y \in (0, \infty)$ given an input variable $Z \in \mathbb{R}^m$. Under the assumption that Y follows a certain Gamma distribution, its logarithmic mean function can be estimated by minimizing the negative log-likelihood (NLL) function associated with its density function (Garrido et al., 2016). We then train a neural network to approximately solve this minimization problem.

Numerical experiment: Network-based Non-linear Gamma Regression

- We assume that Y follows the Gamma distribution with mean $\mu \in (0, \infty)$ and log-dispersion $\phi \in \mathbb{R}$. Denote by $f_Y : (0, \infty) \to (0, \infty)$ the probability density function of Y given explicitly by

$$f_Y(y; \mu, \phi) \equiv f_Y(y) := \frac{1}{y\Gamma(\exp(-\phi))} \left(\frac{y}{\mu \exp(\phi)} \right)^{\exp(-\phi)} e^{-\frac{y \exp(-\phi)}{\mu}}, \quad y \in (0, \infty),$$

where $\Gamma(\exp(-\phi))$ denotes the gamma function evaluated at $\exp(-\phi)$. Then, we model the logarithmic mean function $\widehat{\mu} : \mathbb{R}^d \times \mathbb{R}^m \to (0, \infty)$ of Y by

$$\log \widehat{\mu}(\theta, z) = \log \mathbb{E}[Y \mid Z = z, \theta] := \mathfrak{N}(\theta, z), \tag{4}$$

where $\mathfrak{N}(\theta, z)$ is a ReLU neutal network and z is the input.

Numerical experiment: Network-based Non-linear Gamma Regression

- The mean function $\widehat{\mu}$ defined in (4) can be estimated by minimizing the NLL function $\ell : (0,\infty) \times \mathbb{R}^m \times \mathbb{R}^{d+1} \to \mathbb{R}$ given by

$$\ell(y,z,\Theta) := -\log f_Y(y;\widehat{\mu}(\theta,z),\phi)$$
$$= \log\left(y\Gamma\left(\frac{1}{\exp(\phi)}\right)\right) - \frac{1}{\exp(\phi)}\left(\log\left(\frac{y}{\exp(\phi)}\right) - \widehat{\mathfrak{N}}(\theta,z)\right) + \frac{y}{\exp(\phi)}\exp(-\widehat{\mathfrak{N}}(\theta,z)).$$

The associated regularized stochastic optimization problem is

$$\text{minimize} \quad \mathbb{R}^{d+1} \ni \theta \mapsto u(\theta) := \mathbb{E}[\ell(Y,Z,\theta)] + \lambda_r|\theta|^2.$$

Numerical experiment: Setting

- For the numerical experiments, we consider the auto-insurance claim data from "freMTPL2sev" in the R package "CASdatasets" (Dutang and Charpentier, 2019), which contains $\widetilde{N} = 24,944$ observations. Its i-th observation, $i = 1, \ldots, \widetilde{N}$, consists of a target variable, denoted by $y_i \in (0, \infty)$, indicating the average claim size for one year and an input vector, denoted by $\mathbf{z}_i \in \mathbb{R}^m$, containing relevant quantities including, e.g., driver's age, vehicle's age, and region. More precisely, in this case, for each i, the input vector $\mathbf{z}_i \in \mathbb{R}^m$ with $m = 65$ contains 4 continuous variables and 7 categorical variables.
- For the training and testing purposes, we split the dataset such that the training set contains 70% of the observations and the test set contains 30% of the observations.

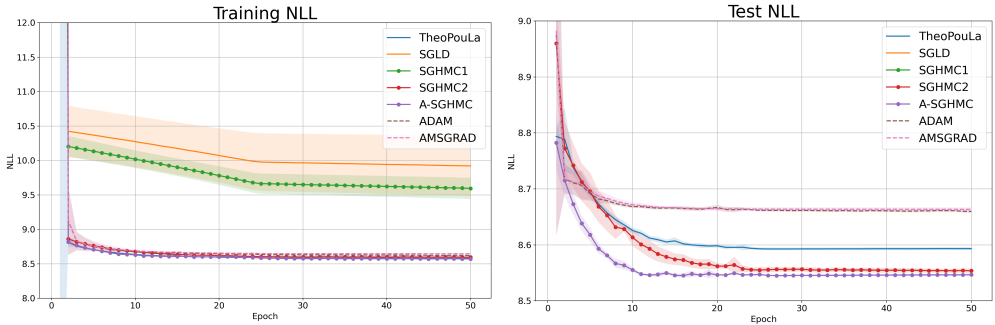


Figure 4: Negative log-likelihood (NLL) curves on the training and test sets. The shaded areas represent the mean $\pm$ standard deviation for each algorithm.

Takeaways

Takeaways

Based on our experiments, we observe the following:

- Adam is excellent when a reasonably accurate solution is required quickly.
- **Without** extensive hyperparameter tuning, Adam already performs reasonably well using its default settings. However, even with significant effort devoted to tuning, achieving further improvements is often difficult. Moreover, Adam lacks solid theoretical guarantees and can exhibit instability in practice.
- On the other hand, Langevin-based algorithms can **outperform** Adam if careful attention is given to hyperparameter tuning. Additionally, they are supported by theoretical results.
- Therefore, when theoretical guarantees on convergence are critical and complex deep learning architectures are involved, consider using our algorithm over Adam.

References I

Akyildiz, O. D. and Sabanis, S. (2024). Nonasymptotic analysis of stochastic gradient hamiltonian monte carlo under local conditions for nonconvex optimization. *Journal of Machine Learning Research*, 25(113):1–34.

Almeida, C., Fan, J., Freire, G., and Tang, F. (2023). Can a machine correct option pricing models? *Journal of Business & Economic Statistics*, 41(3):995–1009.

Bardou, O., Frikha, N., et al. (2009). Computing var and cvar using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods & Applications*, 15(3).

Bardou, O. A., Frikha, N., et al. (2008). Computation of var and cvar using stochastic approximations and unconstrained importance sampling. *arXiv preprint arXiv:0812.3381*.

Becker, S., Cheridito, P., and Jentzen, A. (2019). Deep optimal stopping. *Journal of Machine Learning Research*, 20(74):1–25.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311.

Bryzgalova, S., Pelger, M., and Zhu, J. (2025). Forest through the trees: Building cross-sections of stock returns. *The Journal of Finance*, 80(5):2447–2506.

References II

Buehler, H., Gonon, L., Teichmann, J., and Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8):1271–1291.

Cartea, Á., Jaimungal, S., and Sánchez-Betancourt, L. (2023). Reinforcement learning for algorithmic trading. *Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices. Cambridge University Press*.

Chau, H. N., Kumar, C., Rásonyi, M., and Sabanis, S. (2019). On fixed gain recursive estimators with discontinuity in the parameters. *ESAIM: Probability and Statistics*, 23:217–244.

Chau, H. N. and Rásonyi, M. (2022). Stochastic gradient hamiltonian monte carlo for non-convex learning. *Stochastic Processes and their Applications*, 149:341–368.

Chen, C., Ding, N., and Carin, L. (2015). On the convergence of stochastic gradient mcmc algorithms with high-order integrators. *Advances in neural information processing systems*, 28.

Chen, L., Pelger, M., and Zhu, J. (2024a). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Chen, T., Fox, E., and Guestrin, C. (2014). Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR.

References III

Chen, Z., Lu, Y., Zhang, J., and Zhu, W. (2024b). Managing weather risk with a neural network-based index insurance. *Management Science*, 70(7):4306–4327.

Chetty, R. (2012). Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply. *Econometrica*, 80(3):969–1018.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. (2018). Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51.

Dutang, C. and Charpentier, A. (2019). Casdatasets: insurance datasets. *R package version*, pages 1–0.

Eberle, A., Guillin, A., and Zimmer, R. (2019). Couplings and quantitative contraction rates for langevin dynamics.

Einav, L., Jenkins, M., and Levin, J. (2013). The impact of credit scoring on consumer lending. *The RAND Journal of Economics*, 44(2):249–274.

Fernandez-Arjona, L. and Filipović, D. (2022). A machine learning approach to portfolio pricing and risk management for high-dimensional problems. *Mathematical Finance*, 32(4):982–1019.

Fort, G., Moulines, E., Schreck, A., and Vihola, M. (2016). Convergence of markovian stochastic approximation with discontinuous dynamics. *SIAM Journal on Control and Optimization*, 54(2):866–893.

References IV

Frees, E. W. (2014). Frequency and severity models. *Predictive modeling applications in actuarial science*, 1:138–64.

Gao, X., Gürbüzbalaban, M., and Zhu, L. (2022). Global convergence of stochastic gradient hamiltonian monte carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947.

Garrido, J., Genest, C., and Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205–215.

Ghose, A. and Todri-Adamopoulos, V. (2016). Toward a digital attribution model. *MIS quarterly*, 40(4):889–910.

Gordon, B. R., Zettelmeyer, F., Bhargava, N., and Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.

References V

Han, J., Jentzen, A., and E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510.

Hwang, C.-R. (1980). Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182.

Jaimungal, S. (2022). Reinforcement learning and stochastic optimisation. *Finance and Stochastics*, 26(1):103–129.

Jaimungal, S., Saporito, Y. F., Souza, M. O., and Thamsten, Y. (2023). Optimal trading in automatic market makers with deep learning. *arXiv preprint arXiv:2304.02180*.

Jin, Z., Yang, H., and Yin, G. (2021). A hybrid deep learning method for optimal insurance strategies: Algorithms and convergence analysis. *Insurance: Mathematics and Economics*, 96:262–275.

Li, Z., Zhang, T., Cheng, S., Zhu, J., and Li, J. (2019). Stochastic gradient hamiltonian monte carlo with variance reduction for bayesian inference. *Machine Learning*, 108(8):1701–1727.

Molybog, I., Albert, P., Chen, M., DeVito, Z., Esiobu, D., Goyal, N., Koura, P. S., Narang, S., Poulton, A., Silva, R., et al. (2023). A theory on adam instability in large-scale machine learning. *arXiv preprint arXiv:2304.09871*.

References VI

Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566.

Ni, C., Li, Y., and Forsyth, P. (2025). Optimal multi-period leverage-constrained portfolios: a neural network approach. *Journal of Economic Dynamics and Control*, page 105127.

Pagès, G., Pham, H., and Printems, J. (2004). Optimal quantization methods and applications to numerical problems in finance. In *Handbook of computational and numerical methods in finance*, pages 253–297. Springer.

Pavliotis, G. A. (2014). Stochastic processes and applications. *Texts in applied mathematics*, 60.

Safran, I. and Shamir, O. (2018). Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pages 4433–4441. PMLR.

Takeuchi, I., Le, Q., Sears, T., Smola, A., et al. (2006). Nonparametric quantile estimation.

Tsang, K. H. and Wong, H. Y. (2020). Deep-learning solution to portfolio selection with serially dependent returns. *SIAM Journal on Financial Mathematics*, 11(2):593–619.

Wang, Z., Chen, Y., Song, Q., and Zhang, R. (2023). Enhancing low-precision sampling via stochastic gradient hamiltonian monte carlo. *arXiv preprint arXiv:2310.16320*.

Yang, Y., Qian, W., and Zou, H. (2018). Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470.