Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning

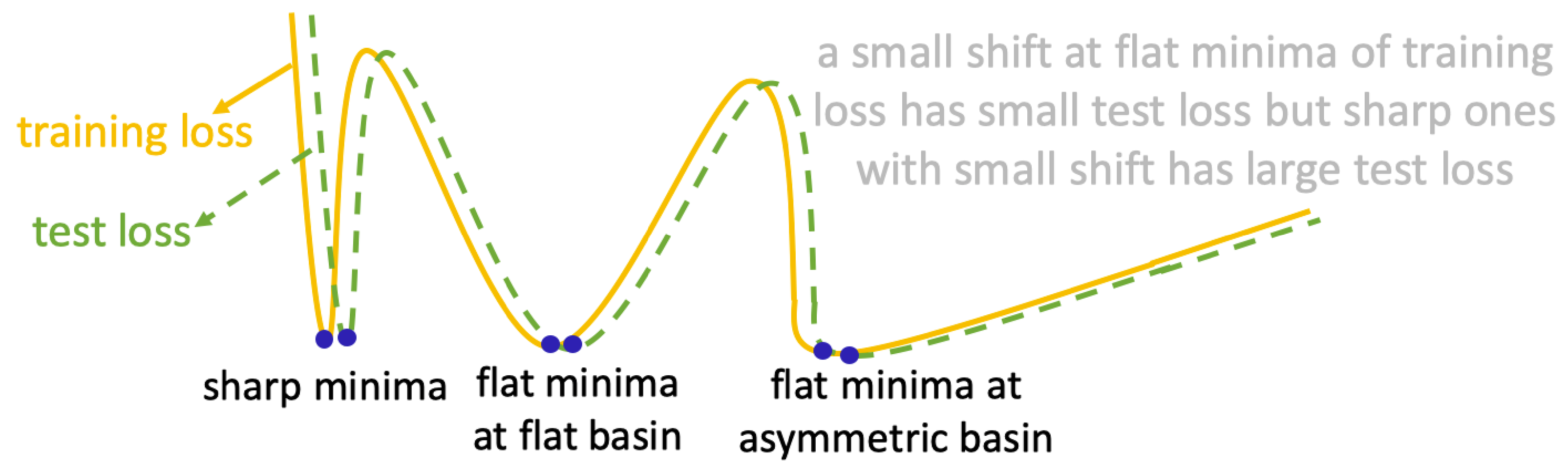
Pan Zhou*, Jiashi Feng†, Chao Ma‡, Caiming Xiong*, Steven HOI*, Weinan E‡

*Salesforce Research, † National University of Singapore, ‡ Princeton University

Problem Setup

Observations: SGD often generalizes better than adaptive gradient algorithms, e.g. ADAM?

Empirical explanation: adaptive gradient algorithms often converge to sharp minima, while SGD prefers to find flat minima at the flat or asymmetric basins/valleys.



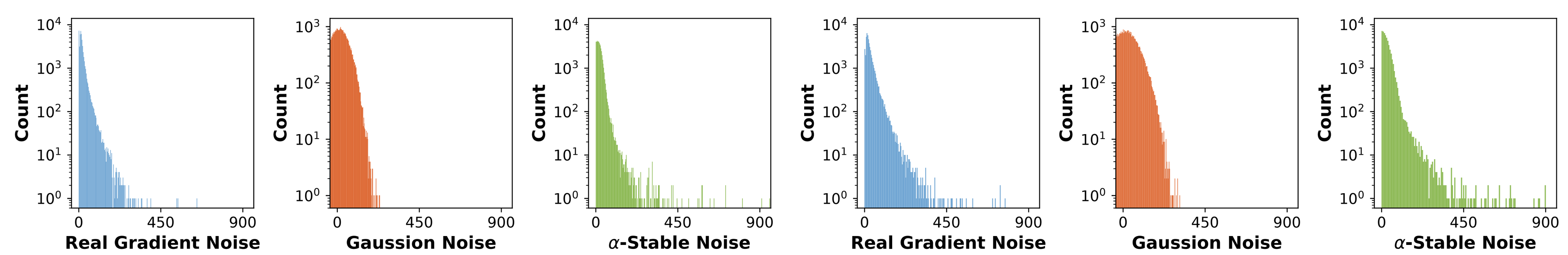
Problem: why SGD often converges to flat minima, while adaptive gradient algorithms, e.g. ADAM, do not?

Stochastic Differential Equation (SDE) Based Analysis

Observation: stochastic gradient noise $\boldsymbol{u}_t$ in SGD and ADAM are heavy-tailed and approximately obey symmetric α-stable ($\mathcal{S}\alpha\mathcal{S}$) distribution

$$\boldsymbol{u}_t = \nabla \boldsymbol{F}\theta_t - \nabla f_{\mathcal{S}_t}(\theta_t), \qquad (1)$$

where $\boldsymbol{F}(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$ is the objective function, $f_{\mathcal{S}_t}(\theta_t) = \frac{1}{S}\sum_{i\in\mathcal{S}_t} f_i(\theta_t)$ is a mini-batch loss.



(a) ADAM (b) SGD

Assumption: Assume $\boldsymbol{u}_t$ obeys $\mathcal{S}\alpha\mathcal{S}$ distribution with a time-dependent covariance matrix Σ_t

$$\Sigma_t = \frac{1}{S}\Big[\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta_t)\nabla f_i(\theta_t)^T - \nabla \boldsymbol{F}(\theta_t)\nabla \boldsymbol{F}(\theta_t)^T\Big].$$

to better characterize the underlying structure in the gradient noise $\boldsymbol{u}_t$.

Levy-driven SDE of SGD:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t) \xrightarrow{\textbf{SGD}} \mathrm{d}\boldsymbol{\theta}_t = -\nabla \boldsymbol{F}(\boldsymbol{\theta}_t) + \varepsilon\Sigma_t \mathrm{d}L_t, \qquad (2)$$

where $\varepsilon = \eta^{(\alpha-1)/\alpha}$. The i-th entry $L_{t,i}$ in Lévy motion $L_t \in \mathbb{R}^d$ obeys the $\mathcal{S}\alpha\mathcal{S}(1)$ distribution, i.e. $\mathbb{E}[\exp(i\omega x)] = \exp(-\sigma^\alpha|\omega|^\alpha)$ if $x \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$. $\mathcal{S}\alpha\mathcal{S}$ distribution has decay density like $1/|x|^{1+\alpha}$.

Levy-driven SDE of ADAM:

$$\begin{cases}\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\boldsymbol{m}_t/(1-\beta_1^t)/\left(\sqrt{\boldsymbol{v}_t/(1-\beta_2^t)}+\epsilon\right),\\ \boldsymbol{m}_t = \beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t),\\ \boldsymbol{v}_t = \beta_2\boldsymbol{v}_{t-1} + (1-\beta_2)[\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t)]^2,\end{cases} \xrightarrow{\textbf{ADAM}} \begin{cases}\mathrm{d}\boldsymbol{\theta}_t = -\mu_t\boldsymbol{Q}_t^{-1}\boldsymbol{m}_t + \varepsilon\boldsymbol{Q}_t^{-1}\Sigma_t\mathrm{d}L_t,\\ \mathrm{d}\boldsymbol{m}_t = \beta_1(\nabla\boldsymbol{F}(\boldsymbol{\theta}_t) - \boldsymbol{m}_t),\\ \mathrm{d}\boldsymbol{v}_t = \beta_2([\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t)]^2 - \boldsymbol{v}_t),\end{cases} \qquad (3)$$

where $\varepsilon=\eta^{(\alpha-1)/\alpha}$, $\boldsymbol{Q}_t=\mathrm{diag}\left(\sqrt{\omega_t\boldsymbol{v}_t}+\epsilon\right)$, $\mu_t=1/(1-e^{-\beta_1 t})$ and $\omega_t = 1/(1-e^{-\beta_2 t})$ are two constants to correct the bias in $\boldsymbol{m}_t$ and $\boldsymbol{v}_t$.

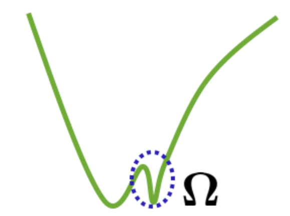
Main Results

Define the **escaping time** Γ from Ω as (constant $\gamma > 0$ satisfies $\lim_{\varepsilon\to 0}\varepsilon^\gamma = 0$)

$$\Gamma = \inf\{t \geq 0 \mid \boldsymbol{\theta}_t \notin \Omega^{-\varepsilon^\gamma}\}$$

Define the **escaping set** $\mathcal{W}$ at the basin Ω as

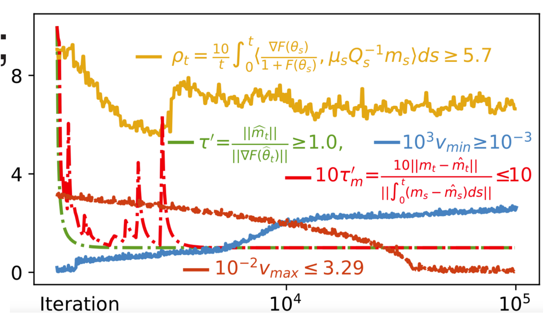
$$\mathcal{W} = \{\boldsymbol{y}\in\mathbb{R}^d \mid \boldsymbol{Q}_{\theta^*}^{-1}\Sigma_{\theta^*}\boldsymbol{y} \notin \Omega^{-\varepsilon^\gamma}\},$$



where $\Sigma_{\theta^*} = \lim_{\theta_t\to\theta^*}\Sigma_t$ for both SGD and ADAM, and $\boldsymbol{Q}_{\theta^*} = \boldsymbol{I}$ for SGD and $\boldsymbol{Q}_{\theta^*} = \lim_{\theta_t\to\theta^*}\boldsymbol{Q}_t$ for ADAM.

Assumption 1: For ADAM and SGD, the objective $\boldsymbol{F}(\boldsymbol{\theta})$ is upper-bounded non-negative, and is locally μ-strongly convex and ℓ-smooth in the basin Ω.

Assumption 2: For ADAM, a) $\int_0^\Gamma\left\langle\frac{\nabla\boldsymbol{F}(\boldsymbol{\theta}_s)}{1+\boldsymbol{F}(\boldsymbol{\theta}_s)}, \mu_s\boldsymbol{Q}_s^{-1}\boldsymbol{m}_s\right\rangle \mathrm{d}s \geq 0$ a.s.;
b) $\|\boldsymbol{m}_t - \widehat{\boldsymbol{m}}_t\| \leq \tau_m\|\int_0^{t-}(\boldsymbol{m}_s - \widehat{\boldsymbol{m}}_s)\mathrm{d}s\|$ and $\|\widehat{\boldsymbol{m}}_t\| \geq \tau\|\nabla\boldsymbol{F}(\widehat{\boldsymbol{\theta}}_t)\|$ where $\widehat{\boldsymbol{m}}_t$ and $\widehat{\boldsymbol{\theta}}_t$ are obtained by Eqn. (3) with $\varepsilon = 0$;
c) the i-th entry $\boldsymbol{v}_{t,i}$ of $\boldsymbol{v}_t$ obeys $v_{\min} \leq \sqrt{\boldsymbol{v}_{t,i}} \leq v_{\max}$ ($\forall i, t$);
d) $\beta_1 \leq \beta_2 \leq 2\beta_1$.



Escaping time of SGD and ADAM. Suppose Assumptions 1 and 2 hold. Let $\rho_0 = \frac{1}{16(1+c_2\kappa_1)}$ and $\ln\left(\frac{2\Delta}{\mu\varepsilon^{1/3}}\right) \leq \kappa_2\varepsilon^{-\frac{1}{3}}$ with $\Delta = \boldsymbol{F}(\boldsymbol{\theta}_0) - \boldsymbol{F}(\boldsymbol{\theta}^*)$ and a constant c_2. Then for any $\boldsymbol{\theta}_0 \in \Omega^{-2\varepsilon^\gamma}$, $\varepsilon\in(0,\varepsilon_0]$, $\gamma\in(0,\gamma_0]$ and $\rho\in(0,\rho_0]$ satisfying $\varepsilon^\gamma \leq \rho_0$ and $\lim_{\varepsilon\to 0}\rho = 0$, SGD in (2) and ADAM in (3) obey

$$\Gamma = \mathcal{O}\left(\frac{1}{m(\mathcal{W})\Theta}\right),$$

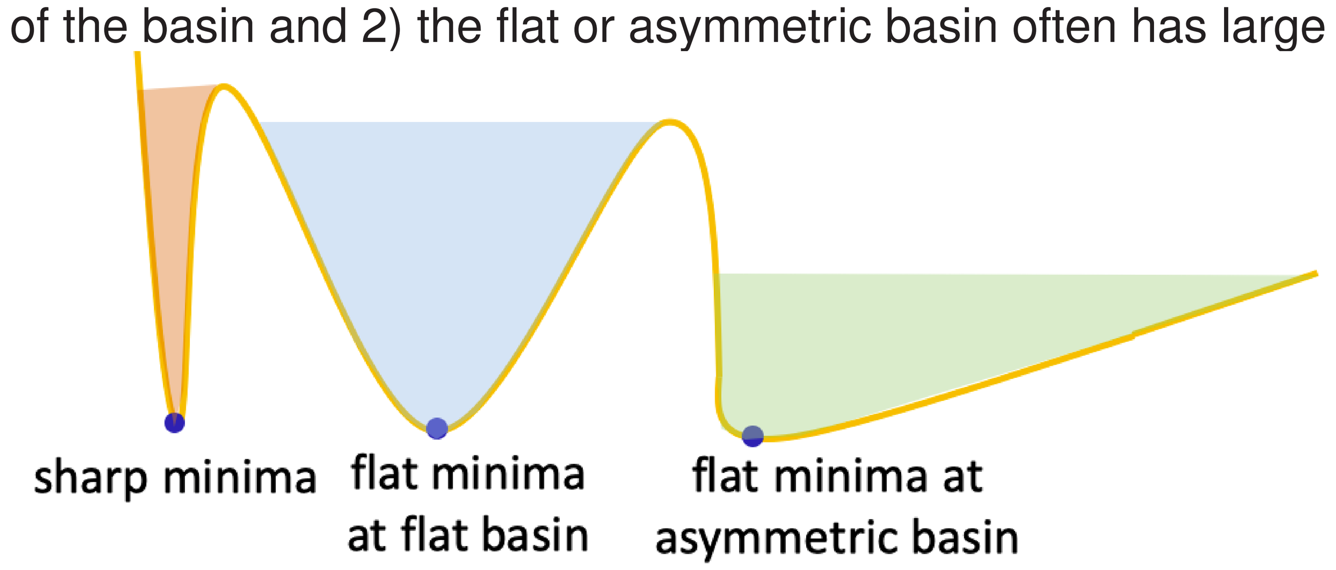
where $m(\mathcal{W})$ denotes the non-zero Radon measure of the escaping set $\mathcal{W}$ in SGD and ADAM, $\Theta = \frac{2}{\alpha}\varepsilon^\alpha$ in which in which α is the tail index of stochastic gradient noise.

Result 1: Preference to Flat Minima

Definition of "flat" minima: A minimum $\boldsymbol{\theta}^*\in\Omega$ is said to be flat if its basin Ω has large nonzero Radon measure.

Due to $m(\mathcal{W})$, both ADAM and SGD have large escaping time Γ at the "flat" minima.

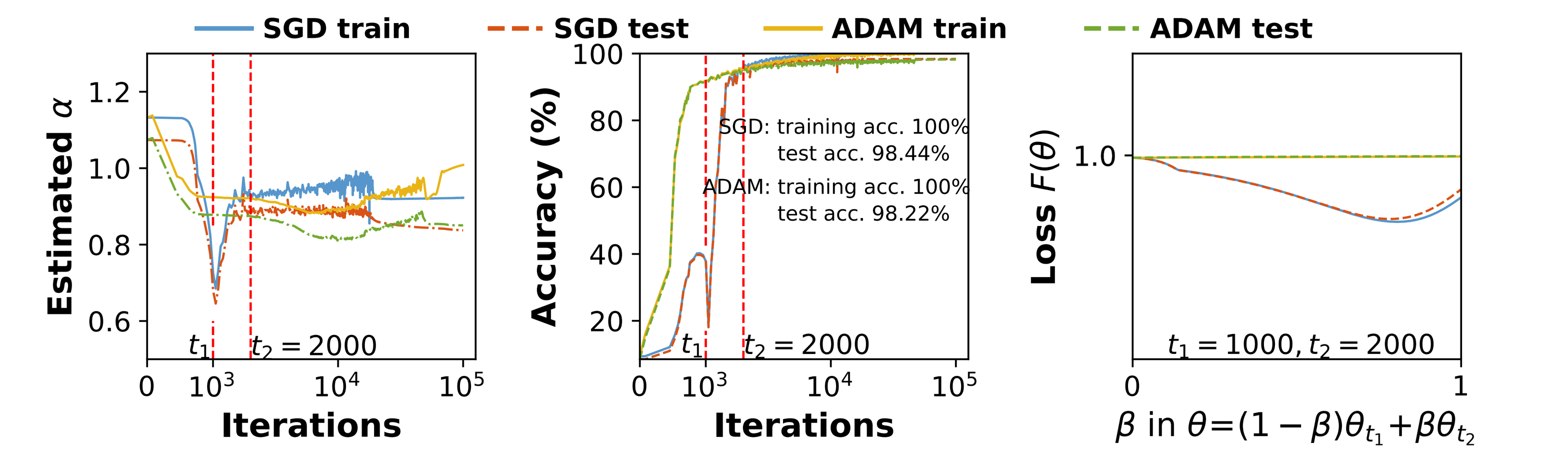
- Define complementary set $\mathcal{W}^c$ of $\mathcal{W}$ as
$$\mathcal{W}^c = \{\boldsymbol{y}\in\mathbb{R}^d \mid \boldsymbol{Q}_{\theta^*}^{-1}\Sigma_{\theta^*}\boldsymbol{y}\in\Omega^{-\varepsilon^\gamma}\}.$$
large measure of $\Omega \longrightarrow$ large $m(\mathcal{W}^c) \longrightarrow$ small $m(\mathcal{W})$ ($m(\mathcal{W}\cup\mathcal{W}^c)$=constant) $\longrightarrow$ large escaping time Γ.
- Minima with large Radon measure often refers to the conventional flat minima or the minima at the asymmetric basin, since 1) Radon measure positively rely on the volume of the basin and 2) the flat or asymmetric basin often has large volume.



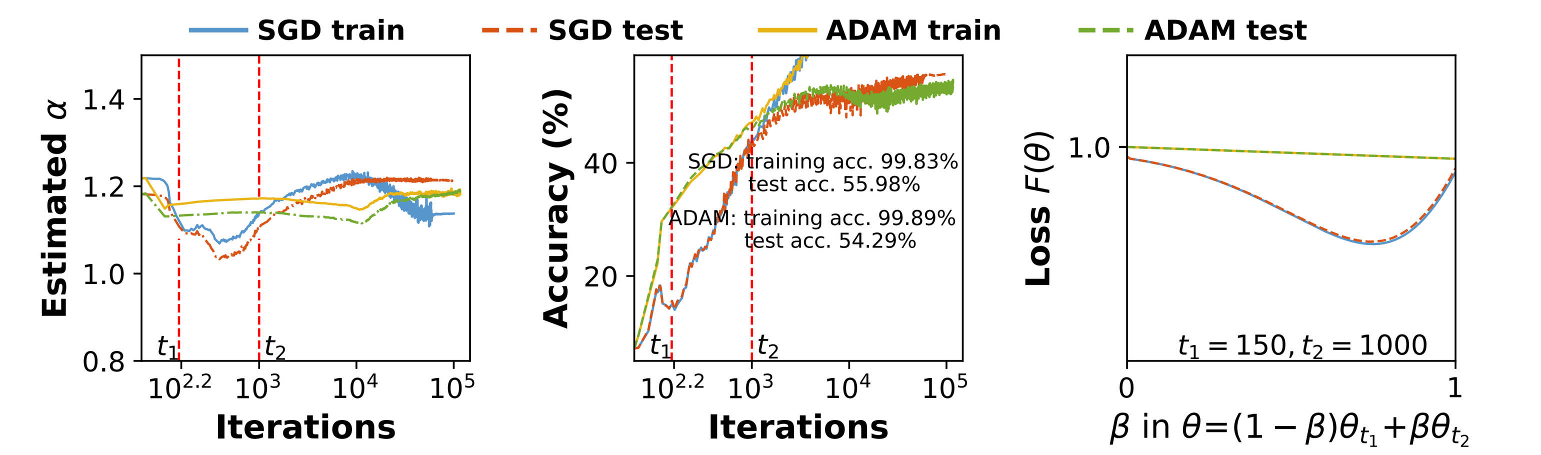
Result 2: Better Sharp Minima Escaping Ability of SGD over ADAM

Factor 1 $\Theta = \frac{2}{\alpha}\varepsilon^\alpha$

- with same learning rate ε in ADAM and SGD, the smaller tail index α, the smaller the escaping time Γ.
- for some iterations, SGD has smaller α , as exponential gradient average in ADAM smooths noise.



(a) MNIST (over-parameterized fully connected networks)



(b) CIFAR10 (over-parameterized fully connected networks)

Factor 2 $m(\mathcal{W})$ that positively depends on the volume of escaping set $\mathcal{W}$.

Approximating Ω as a quadratic basin with center $\boldsymbol{\theta}^*$, i.e.

$$\Omega = \{\boldsymbol{y} \mid \boldsymbol{F}(\boldsymbol{\theta}^*) + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}(\boldsymbol{\theta}^*)\boldsymbol{y} \leq h(\boldsymbol{\theta}^*)\}$$

with a basin height $h(\boldsymbol{\theta}^*)$ and Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta}^*)$ at $\boldsymbol{\theta}^*$.

Comparison of Escaping Sets of SGD and ADAM . Under the quadratic basin approximation, the escaping set $\mathcal{W}$ of ADAM is

$$\mathcal{W}_{\text{ADAM}} \approx \left\{\boldsymbol{y}\in\mathbb{R}^d \mid \boldsymbol{y}^T\boldsymbol{H}(\boldsymbol{\theta}^*)\boldsymbol{y} \geq S^2h_f^*\right\}.$$

and it satieties

$$m(\mathcal{W}_{\text{ADAM}}) < m(\mathcal{W}_{\text{SGD}})$$

where $\mathcal{W}_{\text{SGD}} = \left\{\boldsymbol{y}\in\mathbb{R}^d \,\middle|\, \boldsymbol{y}^T\bar{\Sigma}_{\theta^*}\boldsymbol{H}(\boldsymbol{\theta}^*)\bar{\Sigma}_{\theta^*}\boldsymbol{y} \geq S^2h_f^*\right\}$ is escaping set of SGD and $\bar{\Sigma}_{\theta^*} = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\boldsymbol{\theta}^*)\nabla f_i(\boldsymbol{\theta}^*)^T$.

Conclusion: SGD could better escape from sharp minima and converge to flatter minima, since

- from Factors 1 and 2, SGD has smaller escaping time and is much more unstable. For the same basin, ADAM could stuck in one basin, but SGD could not.
- both SGD and ADAM prefers to converge to flat minima