Adam Reduces a Unique Form of Sharpness: Theoretical Insights Near the Minimizer Manifold

Xinghan Li*   Haodong Wen*   Kaifeng Lyu†
Institute for Interdisciplinary Information Sciences, Tsinghua University
*Equal contribution; alphabet ordering   †Corresponding author

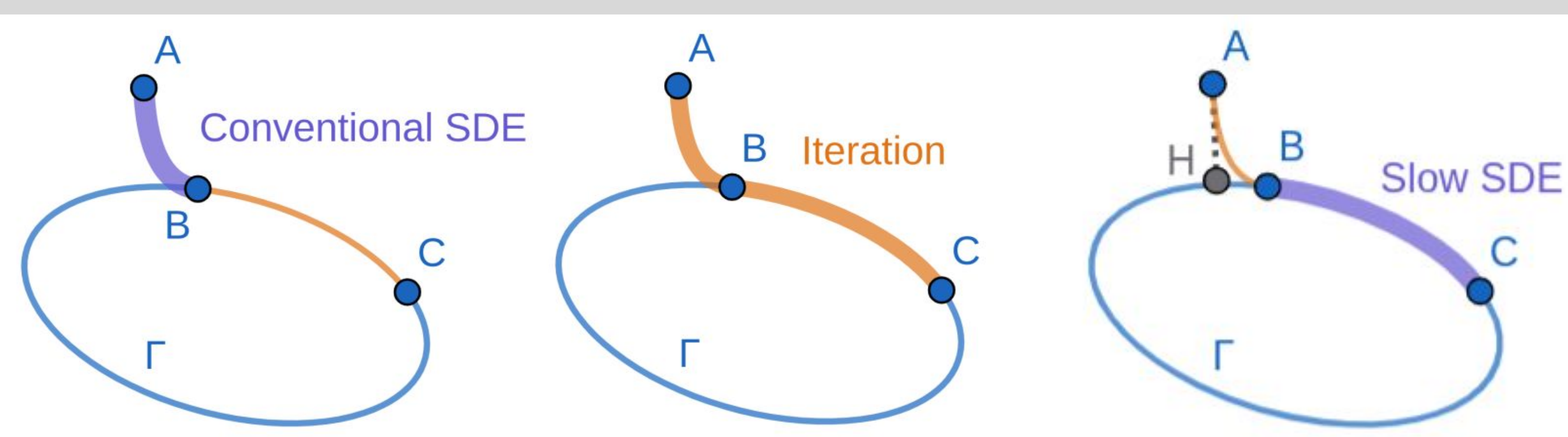
Key Contributions

- We capture the behaviour of Adam near the minimizer manifold for **up to $O(\eta^{-2})$ time** using SDE.
- Our SDE shows that Adam implicitly minimizes **a new kind of sharpness: $\mathrm{tr}(\mathrm{diag}(\mathbf{H})^{1/2})$** with label noise, instead of SGD's familiar tr(H). To highlight this discrepancy, we conducted two case studies: A sparse linear regression task with diagonal linear networks where Adam has provable generalization benefit over SGD, and one matrix factorization task where Adam generalizes worse.

Motivation

- SGD's implicit bias toward flatter minima is fairly well understood, but modern-day training uses Adam more.
- We lack a rigorous account of **what kind of sharpness Adam really pursues** and whether that helps or hurts generalization.

Method: Slow SDE



The Slow SDE **separates the slow implicit bias motion from the fast convergence motion,** and capture the slow motion (moving at speed η^2).

Main Results

A Class of Adaptive Gradient Methods

$$\begin{aligned} \boldsymbol{m}_{k+1} &:= \beta_1 \boldsymbol{m}_k + (1-\beta_1)\nabla \ell_k(\boldsymbol{\theta}_k) \\ \boldsymbol{v}_{k+1} &:= \beta_2 \boldsymbol{v}_k + (1-\beta_2) V\left(\nabla \ell_k(\boldsymbol{\theta}_k)\nabla \ell_k(\boldsymbol{\theta}_k)^\top\right) \\ \boldsymbol{\theta}_{k+1} &:= \boldsymbol{\theta}_k - \eta S(\boldsymbol{v}_{k+1})\boldsymbol{m}_{k+1}. \end{aligned}$$

Table 1: Examples of V, S functions for some optimizers in the AGM Framework.

Optimizer	Function V	Function S	Remarks
Adam	$V(\boldsymbol{M}) = \mathrm{diag}(\boldsymbol{M})$	$S(\boldsymbol{v}) = \mathrm{Diag}(1/(\sqrt{\boldsymbol{v}}+\epsilon))$	
Adam-mini	$V(\boldsymbol{M})_i = \frac{1}{\lvert B(i)\rvert}\sum_{j\in B(i)} M_{jj}$	$S(\boldsymbol{v}) = \mathrm{Diag}(1/(\sqrt{\boldsymbol{v}}+\epsilon))$	Parameters partitioned; i belongs to block $B(i)$.
Adalayer	$V(\boldsymbol{M})_i = \frac{1}{\lvert L(i)\rvert}\sum_{j\in L(i)} M_{jj}$	$S(\boldsymbol{v}) = \mathrm{Diag}(1/(\sqrt{\boldsymbol{v}}+\epsilon))$	i belongs to layer $L(i)$ in the model.
AdamE-λ	$V(\boldsymbol{M}) = \mathrm{diag}(\boldsymbol{M})$	$S(\boldsymbol{v}) = \mathrm{Diag}(1/(\boldsymbol{v}^{\odot\lambda}+\epsilon))$	

Slow SDE for AGMs:

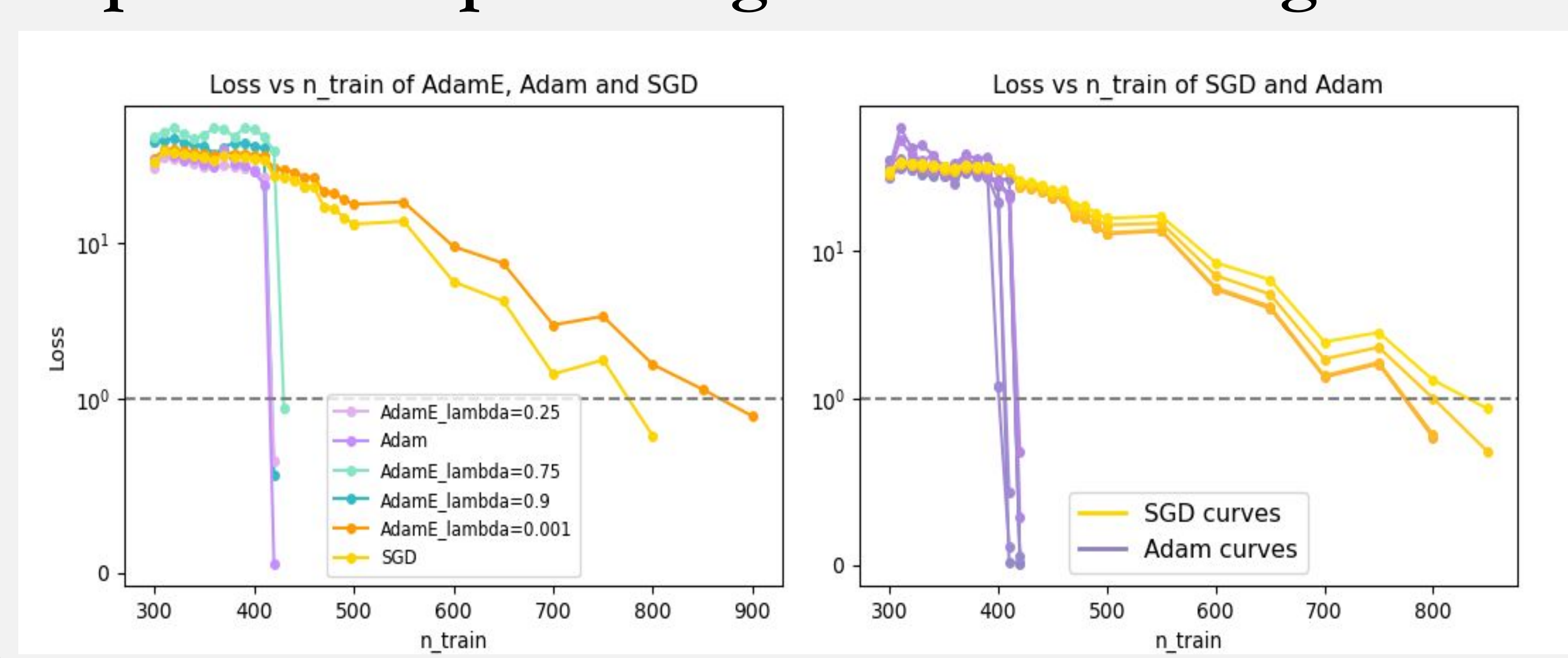
$$\begin{cases} \mathrm{d}\boldsymbol{\zeta}(t) = P_{\boldsymbol{\zeta},\boldsymbol{S}(t)}\Big(\underbrace{\boldsymbol{\Sigma}_{\parallel}^{1/2}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))\mathrm{d}\boldsymbol{W}_t}_{diffusion} \underbrace{-\frac{1}{2}\boldsymbol{S}(t)\nabla^3\mathcal{L}(\boldsymbol{\zeta})\left[\boldsymbol{\Sigma}_{\diamond}(\boldsymbol{\zeta}(t);\boldsymbol{S}(t))\right]\mathrm{d}t}_{drift}\Big), \\ \mathrm{d}\boldsymbol{v}(t) = \underbrace{c\left(V(\boldsymbol{\Sigma}(\boldsymbol{\zeta})) - \boldsymbol{v}\right)\mathrm{d}t}_{Preconditioner\ drift}. \qquad \left(\frac{1-\beta_2}{\eta^2} = c, \quad \zeta_0 \in \boldsymbol{\Gamma}\right) \end{cases}$$

- $\boldsymbol{\Sigma}(\boldsymbol{\zeta})$ is the covariance at $\boldsymbol{\zeta}$, $\mathbf{S}(t) := \mathbf{S}(\mathbf{v}(t))$, $P_{\zeta,\mathbf{S}(t)}$: $\boldsymbol{\zeta}$'s projection on manifold, defined by gradient flow pre-conditioned with $\mathbf{S}(t)$. $\boldsymbol{\Sigma}_{\parallel}(\boldsymbol{\zeta};\mathbf{S})$ and $\boldsymbol{\Sigma}_{\diamond}(\boldsymbol{\zeta};\mathbf{S})$ are matrices related to $\boldsymbol{\Sigma}(\boldsymbol{\zeta})$, $\boldsymbol{\zeta}$ and $\mathbf{S}$.
- The drift term in Slow SDE can be interpreted as **adaptive semi-gradient descent** minimizing $\mu(\zeta,\mathbf{v}) := \langle \nabla^2\mathcal{L}(\zeta), \boldsymbol{\Sigma}_{\diamond}(\zeta(\mathbf{t});\mathbf{S}(\mathbf{t}))\rangle$

Adam's Generalization Benefit with Label Noise

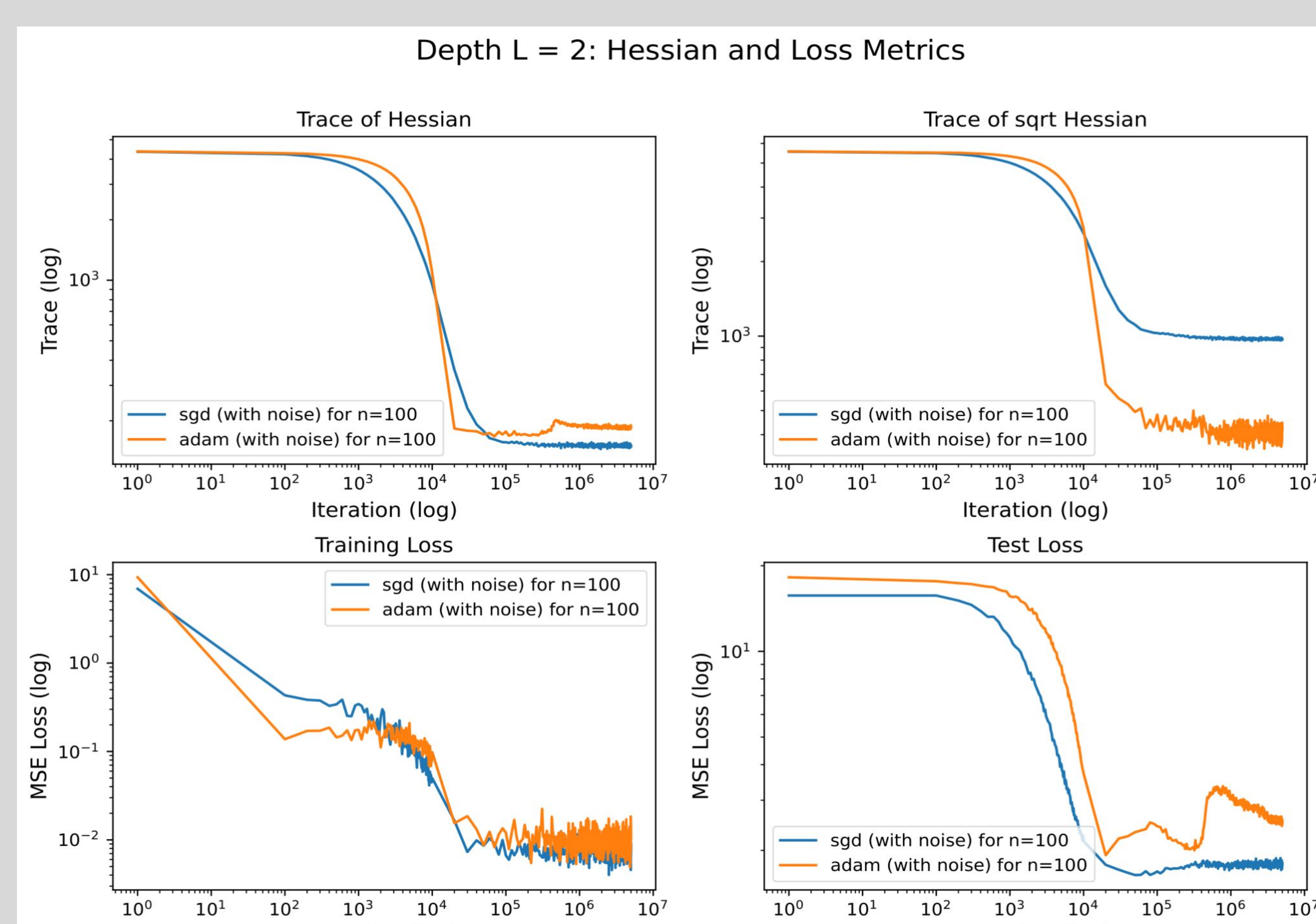
- Label noise $\boldsymbol{\Sigma} \equiv \alpha\nabla^2\mathcal{L}$, SGD ODE: $d\zeta(t) = -\frac{\alpha}{4}P_{\parallel}(\zeta)\nabla^3\mathcal{L}(\zeta)[\mathbf{I}]\,dt.$
- **AGM's slow SDE reduces to an ODE:** $\begin{cases} \mathrm{d}\mathbf{v}(t) = c\left(V(\Sigma(\boldsymbol{\zeta})) - \mathbf{v}\right)\mathrm{d}t, \\ \mathrm{d}\zeta(t) = -\frac{\alpha}{2}S(\mathbf{v})P_{\parallel,S(\mathbf{v})}(\boldsymbol{\zeta})S(\mathbf{v})\nabla^3\mathcal{L}(\boldsymbol{\zeta})[S(\mathbf{v})]\,\mathrm{d}t. \end{cases}$
- The fixed point of this ODE must satisfy $\nabla\mathbf{tr}(\mathbf{Diag}(\mathbf{H})^{1/2}) = \mathbf{0}$
- Changing Adam's sqrt to ^λ results in $\nabla\mathbf{tr}(\mathbf{Diag}(\mathbf{H})^{1-\lambda}) = \mathbf{0}$

Experiment: Sparse Regression with Diagonal Net



Takeaway: Adam's unique implicit bias aligns better with the sparsity requirement in this case (Adam-λ finds the optimum with minimal λ-norm), which arises from taking into consideration the 2nd order momentum compared to SGD.

Adam Loses in Matrix Factorization



Takeaway: Adam's implicit bias hurts generalizability in this case. Adam's implicit regularization differs qualitatively from SGD's, but the benignity of this difference depends.