Vanilla SGD with Momentum Survives Heavy-Tailed Noise: Convergence Analysis without Gradient Clipping or Normalization

uai2026

Ryusei Yamada*†, Naoki Sato*, Hideaki Iiduka (Meiji University, Japan) †:pecokai0204@gmail.com

明治大学 MEIJI UNIVERSITY

Key Idea: Employing **Hölder smoothness** instead of L-smoothness guarantees the convergence of vanilla SGDM under heavy-tailed noise.

I. Overview

【Background & Motivation】

▷ Heavy-tailed noise frequently emerges in recent LLM training. However, theoretical understanding of optimizers without normalization or clipping remains insufficient.

▷ The failure to guarantee vanilla SGD convergence stems from the inherent incompatibility between classical L-smoothness and heavy-tailed noise. We hypothesized that adopting Hölder smoothness would resolve this fundamental clash, thereby guaranteeing the convergence of vanilla SGDM.

【Contribution】

▷ We provide the first theoretical guarantee that vanilla SGD with momentum converges under heavy-tailed noise for strongly convex, convex, and nonconvex functions.

▷ We establish that vanilla SGD converges in expectation under heavy-tailed noise across all three classes of objective functions.

▷ We demonstrate through experiments that the condition $\nu + 1 \leq \mathfrak{p}$ is crucial for stable convergence, establishing Hölder smoothness as a more appropriate framework than standard L-smoothness.

II. Problem Setting

We minimize the empirical risk over strongly convex / convex / nonconvex.

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}),$$

where each f_i $(i \in [n])$ is differentiable.

Algorithm 1 Vanilla SGD with Momentum

Require: $\boldsymbol{x}_1 \in \mathbb{R}^d, \eta_t > 0, b \in [n], \beta \in [0,1), \boldsymbol{m}_0 := \boldsymbol{0}, T \in \mathbb{N}$
  for $t = 1$ to T **do**
    $\boldsymbol{m}_t := \beta \boldsymbol{m}_{t-1} + (1-\beta)\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$
    $\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \eta_t \boldsymbol{m}_t$
  end for
  return $\boldsymbol{x}_{T+1}$

【Assumption】

(A1) Hölder smoothness of f: $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x}-\boldsymbol{y}\|^{\nu}$. $(\nu \in (0,1])$

(A2) Unbiased gradient: $\mathbb{E}_\xi[\mathsf{G}_\xi(\boldsymbol{x})] = \nabla f(\boldsymbol{x})$.

(A3) Bounded $\mathfrak{p}$-th moment: $\mathbb{E}_\xi[\|\mathsf{G}_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{\mathfrak{p}}] \leq \sigma^{\mathfrak{p}}$. $(\mathfrak{p} \in (1,2])$

(A4) Bounded sequence: (i) $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \leq D_1$, (ii) $\|\boldsymbol{z}_t - \boldsymbol{x}^\star\| \leq D_2$. $(D_1, D_2 > 0)$

III. Main Results (Convergence analysis of vanilla SGD with momentum)

Theorem 3.2 (Strongly convex case).
Under (A1)-(A3) and $\nu + 1 \leq \mathfrak{p}$,

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f^\star\right] \leq \frac{f(\boldsymbol{x}_1) - f^\star}{2}\left(1 - \frac{\eta E_2}{4}\right)^T + \frac{E_3}{1-\beta} = \mathcal{O}\left(\left(1 - \frac{\eta E_2}{4}\right)^T + \eta^{\nu+1}\right),$$

where E_2 and E_3 are non-negative constants.

Theorem 3.4 (Convex case).
Under (A1)-(A4) and $\nu + 1 \leq \mathfrak{p}$,

$$\min_{1\leq t\leq T}\mathbb{E}\left[f(\boldsymbol{x}_t) - f^\star\right] \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^{\nu+1}}{\eta(\nu+1)D_2^{\nu-1}T} + E_6\eta + \frac{2^{1-\nu}E_5}{(\nu+1)D_2^{\nu-1}}\eta^\nu = \mathcal{O}\left(\frac{1}{\eta T} + \eta^\nu\right),$$

where E_5 and E_6 are non-negative constants.

Theorem 3.6 (Nonconvex case).
Under (A1)-(A3) and $\nu + 1 \leq \mathfrak{p}$,

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|^2\right] \leq \frac{4(f(\boldsymbol{x}_1) - f^\star)}{\eta T} + \frac{LE_0\eta^\nu}{4(\nu+1)} + E_8\eta^{2\nu} = \mathcal{O}\left(\frac{1}{\eta T} + \eta^\nu\right),$$

where E_0 and E_8 are non-negative constants.

Comparison with [Liu and Zhou, 2025]

Liu and Zhou [2025] established that normalized SGD with momentum achieves the following convergence rates for nonconvex functions under heavy-tailed noise:

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{3\mathfrak{p}-2}}\right) \text{ (with known } \mathfrak{p}\text{)},$$

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{2\mathfrak{p}}}\right) \text{ (with unknown } \mathfrak{p}\text{)}.$$

In contrast, our Theorem 3.6 implies that vanilla SGD with momentum achieves:

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{2\mathfrak{p}}}\right) \text{ (with known } \mathfrak{p}\text{)},$$

$$\min_{1\leq t\leq T}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t)\|\right] = \mathcal{O}\left(T^{-\frac{\mathfrak{p}-1}{4}}\right) \text{ (with unknown } \mathfrak{p}\text{)}.$$

▷ For $\mathfrak{p} \in (1,2]$, we have $\frac{\mathfrak{p}-1}{3\mathfrak{p}-2} \geq \frac{\mathfrak{p}-1}{2\mathfrak{p}} \geq \frac{\mathfrak{p}-1}{4}$.

▷ Note that vanilla with known $\mathfrak{p}$ matches normalized with unknown $\mathfrak{p}$.

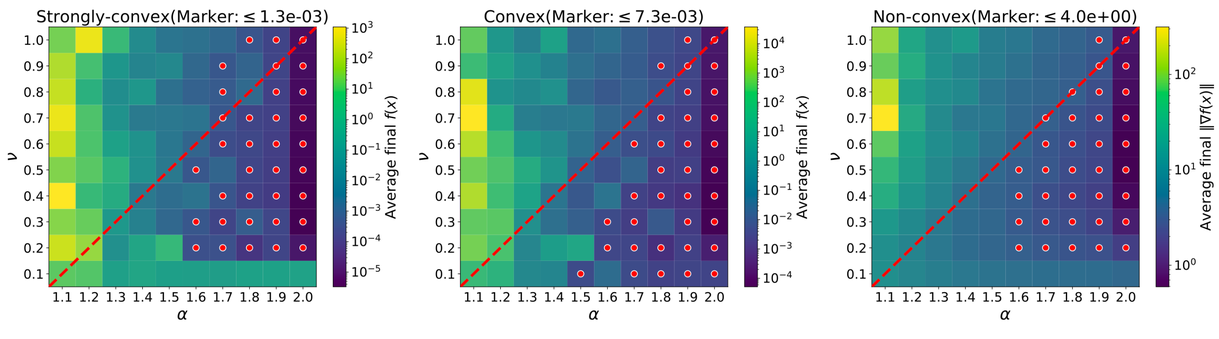
▷ While its normalized SGDM achieves an optimal rate, the rate of vanilla SGDM is sub-optimal.

▷ Even without gradient control, vanilla SGDM still has sublinear convergence. This is the baseline that more complex algorithms must beat.

[**Liu and Zhou, 2025**] Zijian Liu and Zhengyuan Zhou, "*Nonconvex Stochastic Optimization under Heavy-Tailed Noises: Optimal Convergence without Gradient Clipping*," International Conference on Learning Representations, 2025.

IV. Numerical Experiments

Figure 1: Convergence performance across various (ν, α) pairs. The heatmaps show the average final objective values (for strongly convex and convex cases) and gradient norms (for nonconvex cases) after 4,000 steps of vanilla SGD with momentum. The top 35 performing combinations are indicated by red markers. Below the red dotted line, the theoretical condition $\nu + 1 \leq \alpha$ is satisfied, ensuring convergence guarantees.



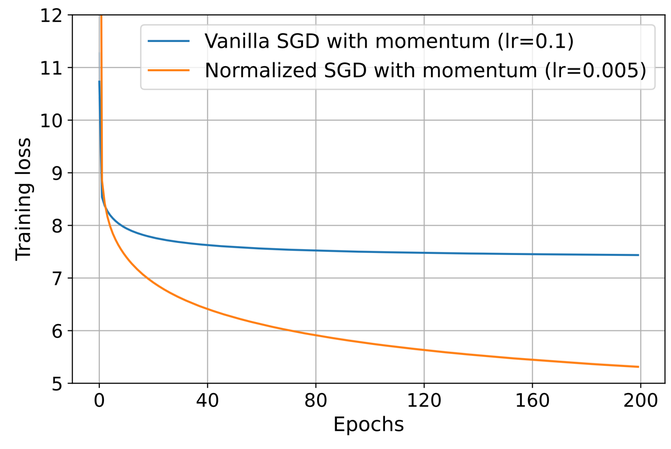


Figure 3: Loss function value for training versus the number of epochs in training Transformer-XL on the WikiText-2 dataset. The solid lines represent the mean value, and the shaded areas represent the maximum and minimum values over three runs.

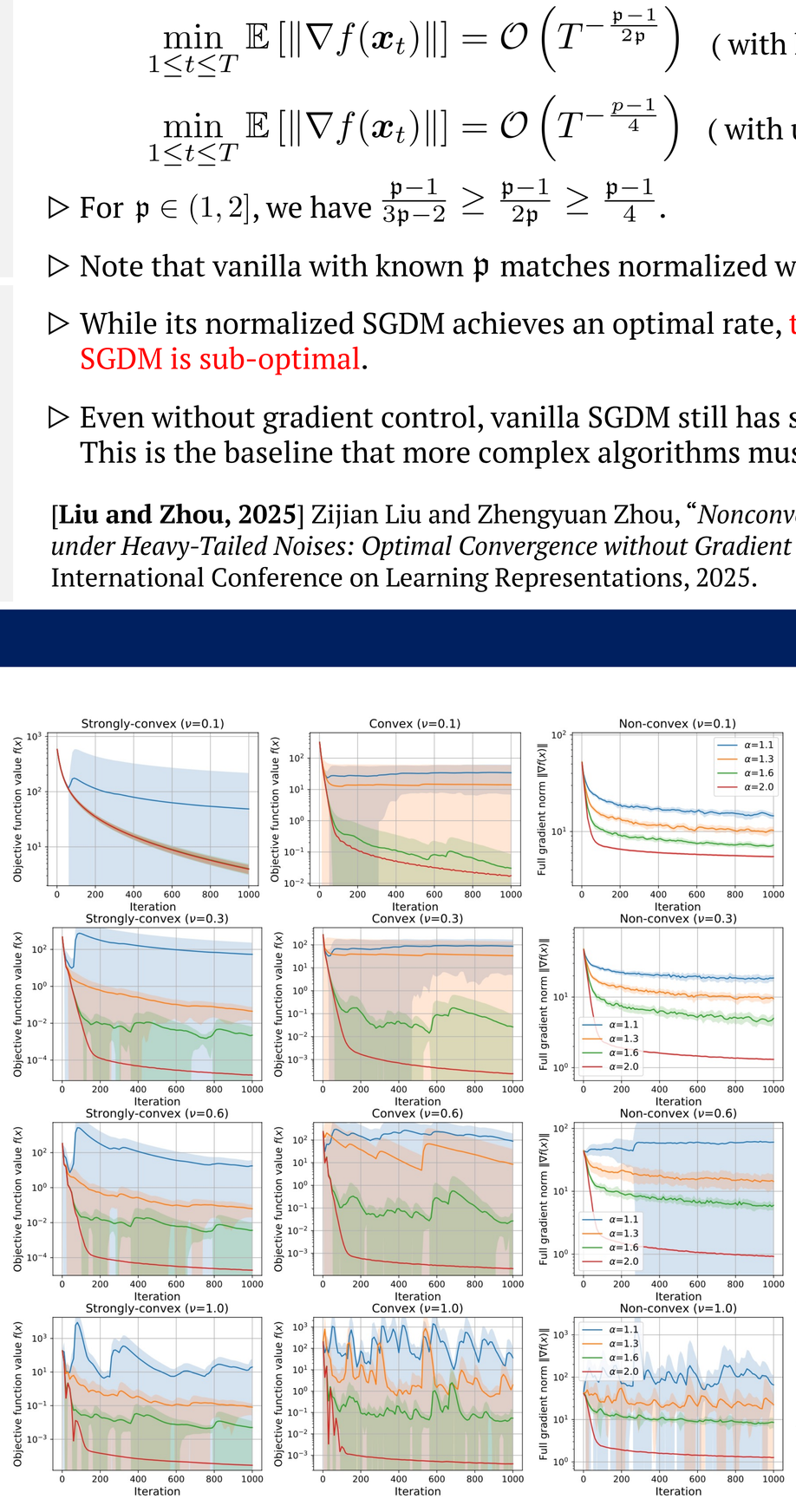


Figure 2: Convergence trajectories of vanilla SGD with momentum. The plots illustrate the evolution of objective values (for strongly convex and convex cases) and gradient norms (for non-convex cases) over 1,000 steps for the functions defined in Eq.(2). Solid lines represent the average of 20 independent trials, while the lightly shaded regions indicate the range between the minimum and maximum values.