Revisiting the Initial Steps in Adaptive Gradient Descent Optimization

Abulikemu Abuduweili and Changliu Liu
Robotics Institute, Carnegie Mellon University
abulikea@andrew.cmu.edu

CPAL-2025

Introduction

- Adaptive gradient methods like Adam are widely used but often struggle with poor generalization and unstable convergence compared to SGD.
- The default zero initialization of the second-order moment ($v_0 = 0$) is a key factor behind these issues.
- We propose a simple solution: initializing the second-order moment with non-zero values.
- Empirical results show that the proposed initialization stabilizes convergence and improves the performance of adaptive optimizers.

Instability of Adam optimizer

- **First step of Adam as sign descent.**

In its first step, Adam performs a pure sign-descent update due to the zero initialization of m_0, v_0.

$$\Delta\theta_1 = -\alpha\,\hat{m}_1/\sqrt{v_1} = -\alpha \cdot \mathrm{sign}(g_1)$$

- **Instability of Adam in Transformer training.**

When training Transformers, vanilla Adam without learning-rate warmup fails, due to large step sizes from its initial sign-descent behavior.

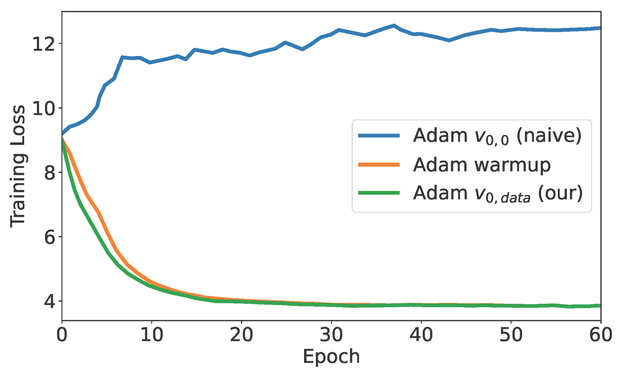

Fig 1. Training loss curve in Training Transformer.

- **Impact of sign descent and shrinking gradients.**

Neural networks often start on a flat loss landscape with small gradients. Adam's "sign descent" amplifies gradients, causing overly large updates.

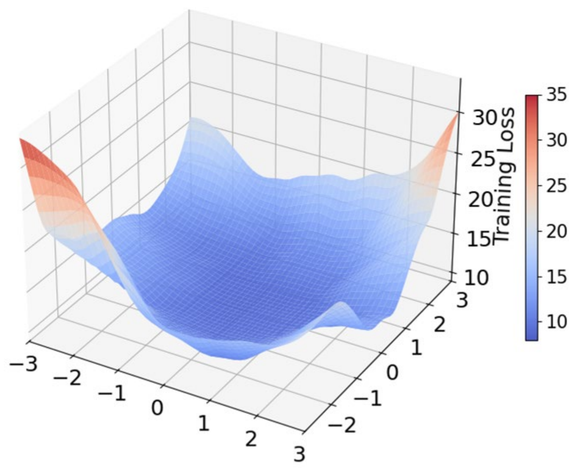

Fig 2. Initial loss landscape of Transformer.

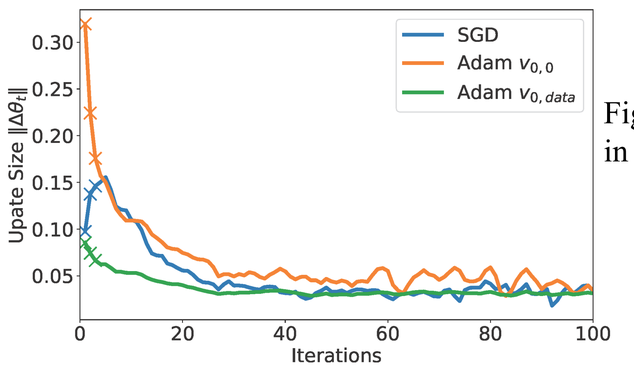

Fig 3. Update step v.s. iterations in Training Transformer.

- **Insights.**

Adam with non-zero initialization of v_0 avoids sign descent, stabilizes updates. Which also enables faster convergence without restrictive warmup constraints.

Initialization for Second Moment v_0

- **Data-driven Initialization**, denoted as $v_{0,data}$.

$$v_0 = \sigma \cdot (\mathbb{E}[g(x_i, y_i)]^2 + \mathbb{VAR}[g(x_i, y_i)]), \qquad (x_i, y_i) \sim \mathcal{D}$$

- **Random Initialization**, denoted as $v_{0,rnd}$.

$$v_0 \sim \frac{\sigma}{\mathrm{fan_{in}} + \mathrm{fan_{out}}} \cdot \chi_1^2$$

- Gradient $g(x_i, y_i)$ is computed from sample $(x_i, y_i) \sim \mathcal{D}$. σ is a hyperparameter that controls the scale of v_0 . χ_1^2 is a chi-squared distribution with one degree of freedom. $\mathrm{fan_{in}}$ and $\mathrm{fan_{out}}$ correspond to the input and output sizes of the weight.

Experiments

- We tested our $v_{0,data}$, $v_{0,rnd}$ on various tasks with adaptive optimizers, using standard $v_0 = 0$ as the baseline.
- Both $v_{0,data}$ and v $v_{0,rnd}$ improved the performance of adaptive optimizers like Adam, AdamW, RAdam, AdaBound, and AdaBelief. With these strategies, Adam matched or exceeded the performance of many newer optimizer variants.

Table 1: Test accuracy ↑ (%) of ResNet-34 on CIFAR-10 dataset.

Optimization	Adam	AdamW	AdaBound	RAdam	AdaBelief
Vanilla $v_0 = 0$	95.25	95.36	95.3	95.61	95.94
$v_{0,rnd}$	95.87	95.94	95.80	95.83	96.11
$v_{0,data}$	96.02	95.95	95.96	95.90	**96.24**

Table 2: BLEU score ↑ of Transformer on IWSTL'14 DE-EN dataset.

Optimization	Adam	AdamW	RAdam	AdaBelief
Vanilla $v_0 = 0$	30.14	35.62	34.76	35.60
$v_{0,rnd}$	33.71	36.06	34.97	36.12
$v_{0,data}$	33.64	35.98	34.84	**36.18**

Table 3: FID score ↓ of DCGAN on CIFAR-10 dataset dataset.

Optimization	Adam	AdamW	AdaBound	RAdam	AdaBelief
Vanilla $v_0 = 0$	54.22	52.39	118.75	48.24	47.25
$v_{0,rnd}$	48.60	46.94	92.36	47.70	45.91
$v_{0,data}$	47.02	45.25	85.45	47.84	**45.02**

- Adam $v_{0,rnd}$ has a flatter loss landscape than Vanilla Adam, which often means better generalization.

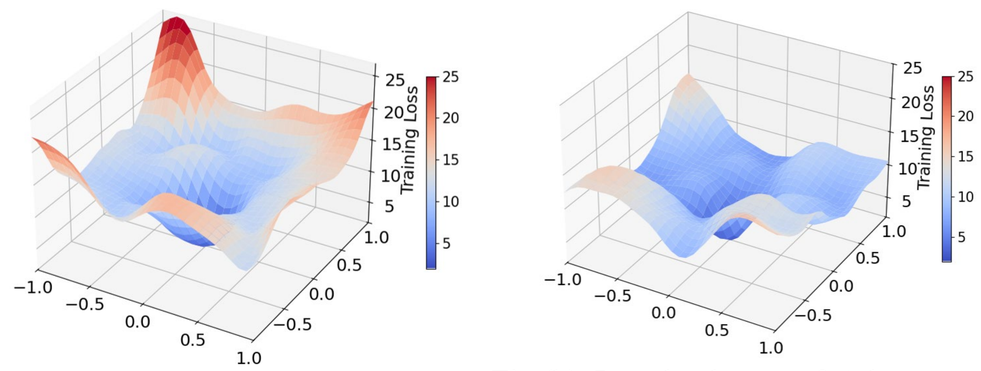

Fig (a). Loss landscape of Adam .

Fig (b). Loss landscape of Adam $v_{0,rnd}$

Conclusions

- We revisited the Adam's instability caused by the sign-descent behavior during early iterations.
- We proposed two simple yet effective approaches: data-driven initialization and random initialization of the second-moment estimate v_0
- Our empirical results demonstrate that these initialization strategies significantly enhance the performance and stability of several adaptive gradient optimization methods.