

Is the Optimal Hyperparameter for Adam Optimizer $\beta_1 = \beta_2$?

2026-02-04

Abstract

Recently, the author came across the paper [Why Adam Works Better with \$\beta_1 = \beta_2\$: The Missing Gradient Scale Invariance Principle](#), which claims that Adam performs better when $\beta_1 = \beta_2$. A colleague reminded the author that similar views were expressed in last year's paper [In Search of Adam's Secret Sauce](#), and a new paper released yesterday, [The Effect of Mini-Batch Noise on the Implicit Bias of Adam](#), also reached similar conclusions. This article explores the theoretical benefits of setting $\beta_1 = \beta_2$ in Adam.

1 Adam Optimizer

The Adam optimizer is defined as:

$$\text{AdamW} := \begin{cases} \mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t \\ \mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2 \\ \hat{\mathbf{m}}_t = \mathbf{m}_t / (1 - \beta_1^t) \\ \hat{\mathbf{v}}_t = \mathbf{v}_t / (1 - \beta_2^t) \\ \mathbf{u}_t = \hat{\mathbf{m}}_t / (\sqrt{\hat{\mathbf{v}}_t} + \epsilon) \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t (\mathbf{u}_t + \lambda_t \boldsymbol{\theta}_{t-1}) \end{cases} \quad (1)$$

2 Online Estimation

According to the paper [In Search of Adam's Secret Sauce](#), when $\beta_1 = \beta_2 = \beta$, $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{v}}_t$ can be viewed as online estimates of the first and second moments of the gradient. Expanding these terms, we get:

$$\hat{\mathbf{m}}_t = \frac{1 - \beta}{1 - \beta^t} \sum_{k=1}^t \beta^{t-k} \mathbf{g}_k, \quad \hat{\mathbf{v}}_t = \frac{1 - \beta}{1 - \beta^t} \sum_{k=1}^t \beta^{t-k} \mathbf{g}_k^2 \quad (2)$$

It can be shown that the sum of the coefficients is always 1, meaning $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{v}}_t$ are weighted averages of $\mathbf{g}_t$ and $\mathbf{g}_t^2$, respectively. Further expansion shows:

$$\begin{aligned} \hat{\mathbf{v}}_t &= \hat{\mathbf{m}}_t^2 + \frac{1 - \beta}{1 - \beta^t} \sum_{k=1}^t \beta^{t-k} (\mathbf{g}_k - \hat{\mathbf{m}}_t)^2 \\ &= \hat{\mathbf{m}}_t^2 + \hat{\boldsymbol{\sigma}}_t^2 \end{aligned} \quad (3)$$

This reveals the relationship between the second moment and the mean and variance of the gradient.

3 Signal-to-Noise Awareness

With these notations, the update step of Adam can be written (assuming $\epsilon = 0$) as:

$$\mathbf{u}_t = \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t}} = \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{m}}_t^2 + \hat{\boldsymbol{\sigma}}_t^2}} = \frac{\text{sign}(\hat{\mathbf{m}}_t)}{\sqrt{1 + \hat{\boldsymbol{\sigma}}_t^2 / \hat{\mathbf{m}}_t^2}} \quad (4)$$

This form of the update step has several advantages. First, each component is bounded between $[-1, 1]$, preventing the risk of exploding gradients. Second, the term $\hat{\boldsymbol{\sigma}}_t^2 / \hat{\mathbf{m}}_t^2$ resembles the reciprocal of the signal-to-noise ratio, leading to the interpretation of this update as "signal-to-noise aware steepest descent."

4 First-Order Expansion

The paper [Why Adam Works Better with \$\beta_1 = \beta_2\$: The Missing Gradient Scale Invariance Principle](#) attempts to continuous Adam into an ODE. However, for Mini-Batch optimization, gradient noise cannot be ignored, making ODE modeling questionable. Instead, the author proposes a first-order expansion approach:

$$\hat{\mathbf{v}}_t \approx \hat{\mathbf{m}}_t^2 + 2\hat{\mathbf{m}}_t \left(\frac{1 - \beta_2}{1 - \beta_2^t} \sum_{k=1}^t \beta_2^{t-k} \mathbf{g}_k - \hat{\mathbf{m}}_t \right) \quad (5)$$

To ensure $\hat{\mathbf{v}}_t$ is close to $\hat{\mathbf{m}}_t^2$, the first-order term should vanish, leading to $\beta_2 = \beta_1$. This ensures $\mathbf{u}_t = \hat{\mathbf{m}}_t / \sqrt{\hat{\mathbf{v}}_t}$ is closer to $\text{sign}(\hat{\mathbf{m}}_t)$, which is bounded and more robust to gradient scale changes.

5 Dual Optimization

Both papers share the idea of bounding $\mathbf{u}_t$ to improve training stability. This leads to the following dual optimization problem:

$$\max_{\beta_2} \min_{\mathbf{g}_1, \dots, \mathbf{g}_t} \hat{\mathbf{v}}_t, \quad \text{s.t.} \quad \hat{\mathbf{m}}_t = \frac{1 - \beta_1}{1 - \beta_1^t} \sum_{k=1}^t \beta_1^{t-k} \mathbf{g}_k \quad (6)$$

Using Cauchy-Schwarz inequality, we find that the optimal β_2 is β_1 , ensuring $\hat{\mathbf{v}}_t \geq \hat{\mathbf{m}}_t^2$.

6 Related Work

The original Adam paper [Adam: A Method for Stochastic Optimization](#) recommends $\beta_1 = 0.9$, $\beta_2 = 0.999$. Earlier analysis in [From Hessian Approximation to Adaptive Learning Rate Optimizers](#) also suggests $\beta_2 > \beta_1$. These conclusions are not contradictory but reflect different contexts. In small models and small batch sizes, larger β_2 values help stabilize $\hat{\mathbf{v}}_t$, making Adam behave like SGD. In contrast, in large models and large batch sizes, $\beta_1 = \beta_2$ provides better stability, aligning with recent trends in LLM training, where default Adam parameters are shifting toward $\beta_1 = 0.9$, $\beta_2 = 0.95$.

7 Conclusion

This article analyzes the β_1, β_2 parameters of the Adam optimizer, showing from a stability perspective that $\beta_1 = \beta_2$ is often an optimal choice. This can be interpreted as steepest descent under signal-to-noise awareness.