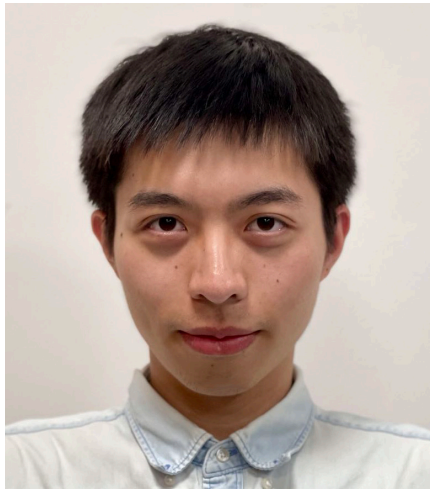
Adam Can Converge Without Any Modification On Update Rules

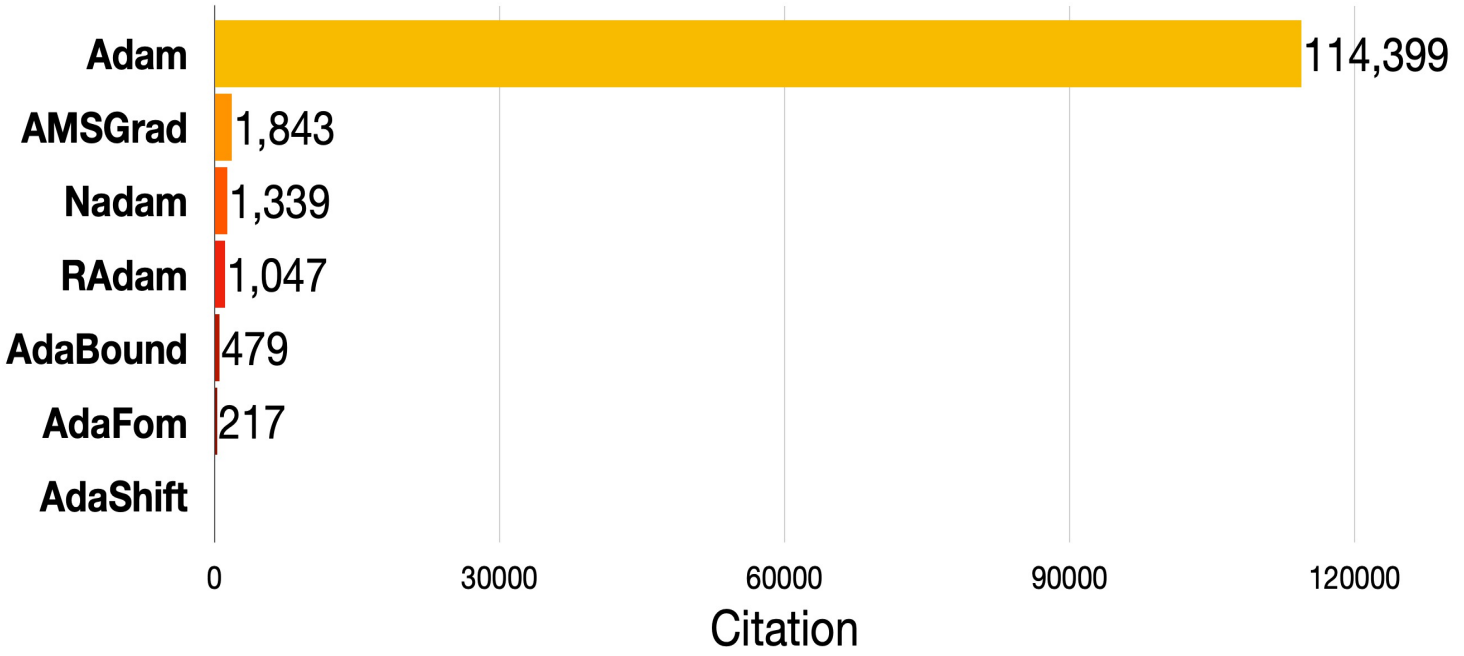
Yushun Zhang[1], **Congliang Chen**[1], **Naichen Shi**[2], **Ruoyu Sun**[1], **Zhi-Quan Luo**[1]

1: The Chinese University of Hong Kong, Shenzhen, China; 2: University of Michigan, US

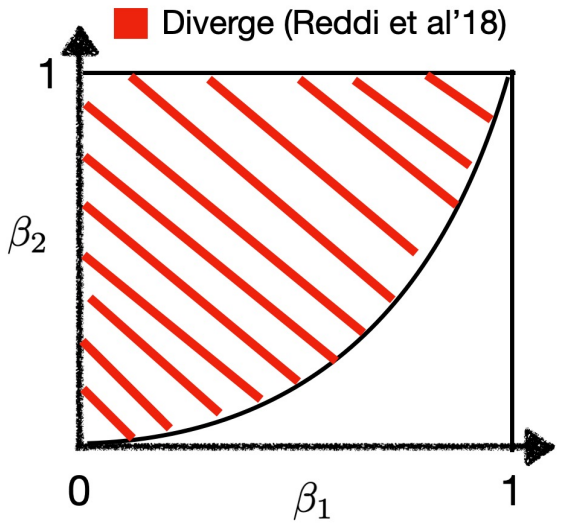
Adam is popular in practice

Adam is one of the most popular algorithms in deep learning (DL).
(It has received more than 110,000 citations)



In theory, Adam is known to have divergence issue

Reddi et al.18 **(ICLR Best paper):**
For any **fixed β_1, β_2 (the hyperparameter of Adam)** s.t. $\beta_1 < \sqrt{\beta_2}$, there exists a problem such that Adam diverges



An important (but often ignored) feature: Reddi et al. fix β_1, β_2 before picking the problem

While in optimization field, parameters are often **problem-dependent** (e.g. the step size for GD < 2/L)

Conjecture: Adam might converge under fixed problem.

Our Results: Adam Can Converge Without Any Modification

Theorem 1: Given fixed problem, we prove that:

when $\beta_2 \geq 1 - O\left(\frac{1-\beta_1^n}{n^{3.5}}\right)$, $\beta_1 < \sqrt{\beta_2} < 1$, **Adam converges with rate $O(\frac{log\ k}{\sqrt{k}})$.***

Theorem 2: Consider the same setting as above, $\exists f(x)$, s.t., when (β_1, β_2) lies in the red region, the sequence $\{x_k\}$ and $\{f(x_k)\}$ of Adam diverges to ∞.

Implication: Adam is still theoretically justified!
please use it confidently!

Suggestions for hyperparameter-tuning:
1. Increase β_2 until convergence
2. Try different $\beta_1 < \sqrt{\beta_2}$ for better performance