Does Adam Converge and When?

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, Zhi-Quan Luo

School of Data Science, The Chinese University of Hong Kong, Shenzhen

Shenzhen Research Institute of Big Data

What to expect from this talk?

- **For theorist:**
  - Convergence & divergence phase transition.
  - Problem-dependent bound v.s. Problem-independent bound.
  - A new method to analysis stochastic non-linear dynamic system.

- **For engineers:**
  - Is Adam a theoretically justified algorithm?
  - Shall we use it confidently?
  - How to tune hyperparameters?

Motivation

- **Adam** is one of the most popular algorithms in deep learning (DL).
  (It has received more than 110,000 citations)
- **Default** choice in many DL tasks:
  - NLP, GAN, RL, CV, GNN etc.
- Adam is also widely used in **SRIBD projects**:
  - Learning to Optimize (L2O);
  - Medical image segmentation;
  - 3D Reconstruction;
  - SRCON, etc.
- However, the behavior of Adam is **poorly understood** in theory.
- We aim to close the gap between theory and practice.

A Brief Introduction to Adam

- Consider $\min_{x} f(x) := \sum_{i=1}^{n} f_i(x)$.
- In DL tasks, n often stands for sample size; x denotes trainable parameters.
- Initialization $\nabla f(x_0)$, $m_0 = \nabla f(x_0)$
- In the k-th iteration: Randomly sample τ_k from $\{1,2,\ldots,n\}$

- SGD:
- $x_{k+1} = x_k - \eta_k \nabla \mathrm{f}_{\tau_k}(\mathrm{x}_k)$

- **SGD with momentum (SGDM):**
- $m_k = (1-\beta_1)\nabla \mathrm{f}_{\tau_k}(\mathrm{x}_k) + \beta_1 m_{k-1}$
- $x_{k+1} = x_k - \eta_k m_k$

A Brief Introduction to Adam

- Consider $\min_x f(x) \coloneqq \sum_{i=1}^{n} f_i(x)$.
- In the k-th iteration: Randomly sample τ_k from $\{1,2,\dots,n\}$

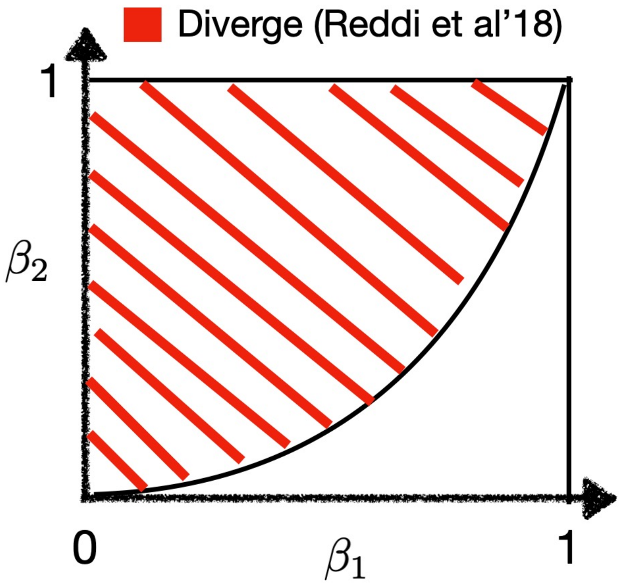
> - Adam:
> - $m_k = (1-\beta_1)\nabla \mathrm{f}_{\tau_\mathrm{k}}(\mathrm{x_k}) + \beta_1 m_{k-1}$
> - $v_k = (1-\beta_2)\nabla \mathrm{f}_{\tau_\mathrm{k}}(\mathrm{x}_k) \circ \nabla \mathrm{f}_{\tau_\mathrm{k}}(x_k) + \beta_2 v_{k-1}$
> - $x_{k+1} = x_k - \eta_k \frac{m_k}{\sqrt{v_k}}$

- Notations: $\circ, \sqrt{}$, and division are all element-wise operations.
- β_1: Controls the 1st-order momentum m_k. Default setting: $\beta_1 = 0.9$
- β_2: Controls the 2nd-order momentum v_k. Default setting: $\beta_2 = 0.999$

For a long time, Adam is criticized for its divergence issue

- Reddi et al. 2018 (ICLR Best paper) :

For any β_1, β_2 s.t. $\beta_1 < \sqrt{\beta_2}$, there exists a problem such that Adam diverges.

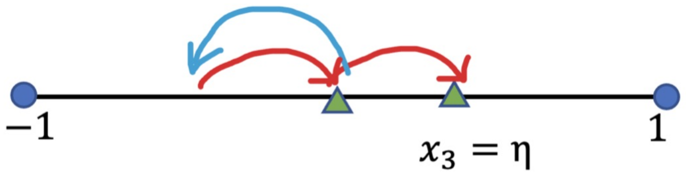


Why would Adam fail?

- Reddi et al. consider the toy problem $\min \sum_{i=1}^{3} f_i(x)$ where

- $f_i = \begin{cases} 100x, & if\ i = 1 \\ -x, & otherwise \end{cases}, x \in [-1,1]$

- The movement of Adam is not consistent with negative gradient direction.



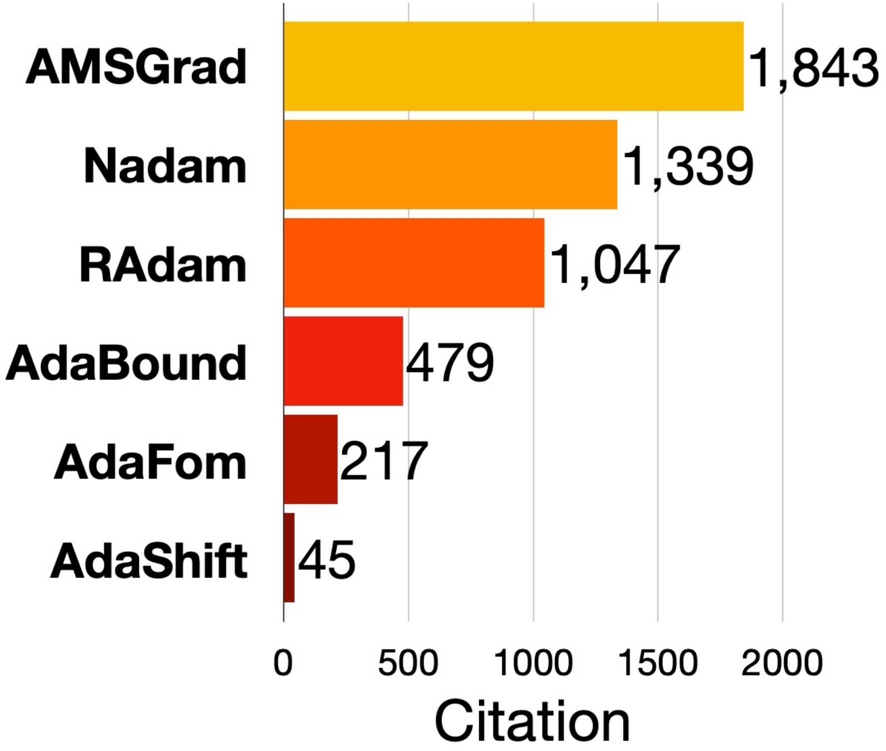
- Reasons of divergence: 1. About the function: f_i differs a lot from each other.

  2. About the algorithm: v_k distorts the update direction.

How to ensure convergence?

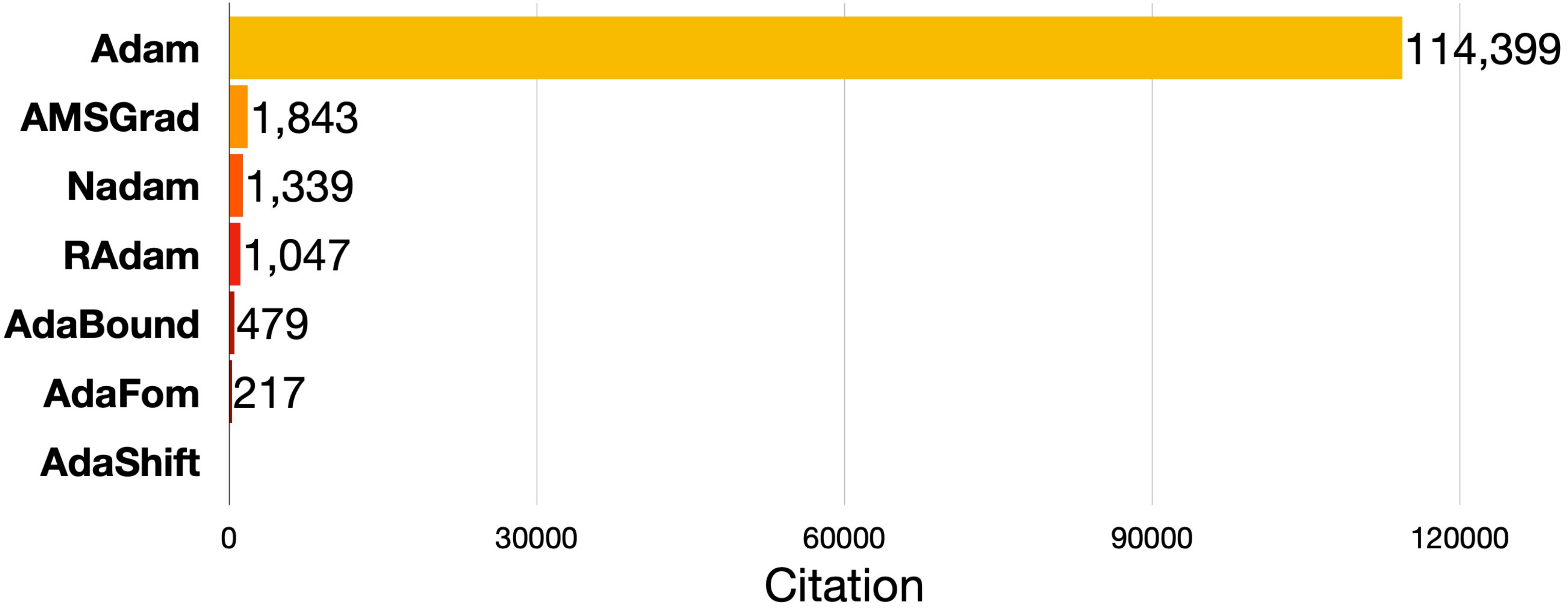
- A popular line of work: Modify the algorithm! For instance:
  - **AMSGrad, AdaFom** [Reddi et al. 18, Chen et al.18]: keep $v_k \geq v_{k-1}$
  - **AdaBound** [Luo et al. 19]: Impose constraint: $v_k \in [C_l, C_u]$
  - etc.
- Although these Adam-variants fix the divergence issue, they often bring new issues. For instance:
  - **AMSGrad and AdaFom** are reported to be slow [Zhou et al. 18].
  - **AdaBound** introduces 2 extra hyperparameters.
- On the other hand, **Adam remains exceptionally popular. It works well in practice!** (either under default setting, or after proper tuning).

Comparison: Adam vs its variants



- *Disclaimer: contribution is not proportional to citation.
  But citation might reflect the popularity among practitioners.

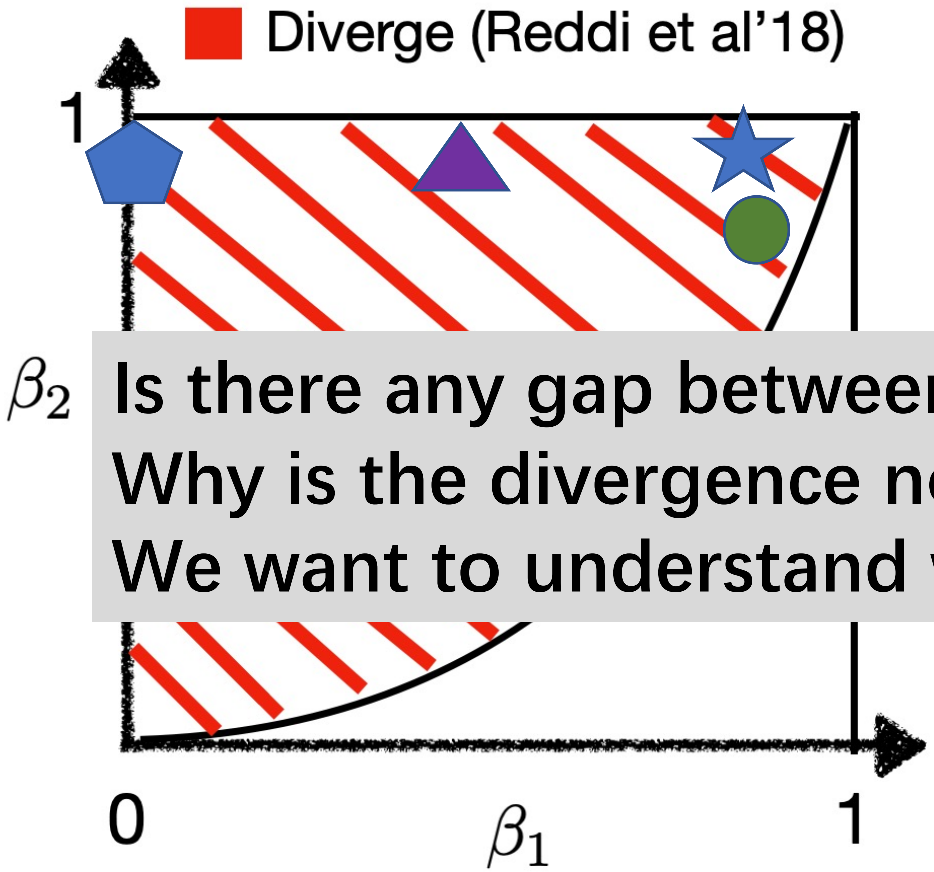
However, Adam remains overwhelmingly popular



- The attention Adam received is astonishing!
- Partially because many variants bring new issues (e.g., slow)

The divergence theory does not go well with practice

We find that the reported (β_1, β_2) in the successful applications **actually satisfy the divergence condition** $\beta_1 < \sqrt{\beta_2}$!



Most deep learning tasks (e.g. RL, NLP, CV, GAN, etc.):
$\beta_1 = 0.9, \beta_2 = 0.999$

l, DCGAN, etc:
.999

uage models (e.g. GPT-3)
.95

First-order GAN, MSG-GAN:
$\beta_1 = 0, \beta_2 = 0.999$

Is there any gap between theory and practice?
Why is the divergence not observed?
We want to understand why.

Why is the divergence not observed?

- Reddi et al. 18 consider $\min_{\mathbf{x}} f(x) \coloneqq \sum_{i=1}^{n} f_i(x)$

Proof(Reddi et al. 18):

For any fixed β_1, β_2 s.t. $\beta_1 < \sqrt{\beta_2}$, we can find an $\boldsymbol{n}$ to construct the divergence example $f(x)$

- An important (but often ignored) feature: Reddi et al. fix β_1, β_2 **before picking the problem** (n is changing).
- While in optimization field, parameters are often **problem-dependent** (e.g. the step size for GD). As such, the divergence is hardly surprising.
- **Conjecture 1: Adam might converge under fixed problem (or fixed $\boldsymbol{n}$.)**

Why is the divergence not observed?

- Reddi et al. 18 consider $\min_{\mathrm{x}} f(x) \coloneqq \sum_{i=1}^{n} f_i(x)$

$$f_i = \begin{cases} nx, & if\ i = 1 \\ -x, & otherwise \end{cases}, x \in [-1,1]$$

- **In this example:** f_i differs a lot from each other.
  In DL applications: all f_i come from a certain underlying non-linear groundtruth mapping (up to certain noise).
  As such, f_i are supposed to be "similar"!
- **Conjecture 2: Adam can converge when f_i are "similar".**
- We will verify **Conjecture 1 and 2.**

Assumptions

- Consider $\min_{x} f(x) := \sum_{i=1}^{n} f_i(x)$
- **A1:** Assume $\nabla f_i(x)$ are L-Lipschitz continuous.
- **A2:** $\sum_{i=1}^{n} || \nabla f_i(\mathrm{x})||_2^2 \leq D_1 || \nabla f(\mathrm{x})||_2^2 + D_0$.

- **A2 says that:** ∇f_i are similar up to certain noise.
- **This is motivated from practical DL tasks.**
- **A2** is quite general. When $D_1 = \frac{1}{n}$, A2 becomes bounded variance:
- $\frac{1}{n}\sum_{i=1}^{n} || \nabla f_i(\mathrm{x}) - \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \nabla f_i(x) ||_2^2 \leq \frac{D_0}{n}$.
- **A2** allows $D_1 \neq \frac{1}{n}$ and thus it is weaker than bounded variance.
- When $D_0 = 0$, **A2** becomes ``Strong Growth Condition (SGC)" [Vaswani et al., 19]
- When $|| \nabla f(\mathrm{x})||=0 \implies$ we have $|| \nabla f_i(\mathrm{x})||=0$.

Related works

- [Zaheer et al. 18, De et al. 18, Shi et al. 20]:
  RMSProp(simplified version of Adam with $\beta_1 = 0$) can converge.
- It is important to study **Adam** rather than **RMSprop**:
  - **Numerically:** Adam outperforms RMSprop on complicated tasks:
  - On **Atari**, the mean rewarded is improved from 88% to 110% [Agarwal et al.20; Jiang 20]
  - **Theoretically:** CANNOT reveal the interaction between β_1 and β_2.
- [Huang et al. 21] and [Guo et al. 22]: Under large β_1, Adam can converge. They require:
- ~~A3: Assume~~ $\cancel{||\nabla f(x)|| < C}$.
- ~~A4: Assume~~ $\cancel{v_k \in [C_l, C_u]}$.
- ~~This line of work is restricted to **AdaBound**.~~
- To our knowledge, **A1+ A2** are the mildest assumption so far.

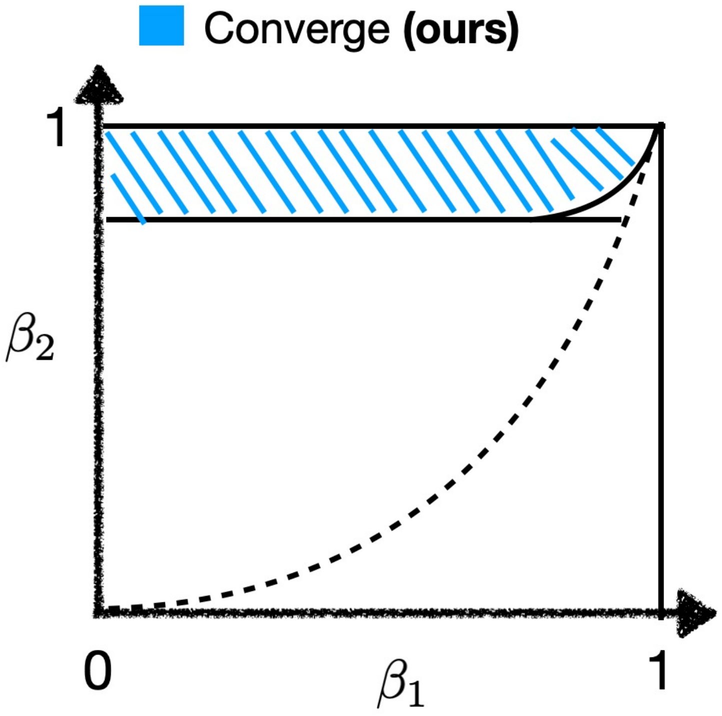
Convergence results for large $\boldsymbol{\beta_2}$

- **Theorem 1:** Consider the previous setting.
  When $\beta_2 \geq 1 - O\left(\frac{1-\beta_1^n}{n^{3.5}}\right)$ and $\beta_1 < \sqrt{\beta_2} < 1$, Adam with stepsize $\eta_k = \frac{1}{\sqrt{k}}$ converges to the neighborhood of stationary points.

- $\min_{k \in [1,T]} \mathbb{E} \| \nabla f(x_k) \|_2^2 = O(\frac{\log T}{\sqrt{T}} + (1-\beta_2) D_0)$

Remark: When $D_0 = 0$: Adam converges to stationary points.

Convergence results for large $\boldsymbol{\beta_2}$



- $m_k = (1-\beta_1)\nabla f_{\tau_k}(x_k) + \beta_1 m_{k-1}$
- $v_k = (1-\beta_2)\nabla f_{\tau_k}(x_k) \circ \nabla f_{\tau_k}(x_k) + \beta_2 v_{k-1}$
- $x_{k+1} = x_k - \eta_k \frac{m_k}{\sqrt{v_k}}$

Intuition: Large β_2 weaken the movement of v_k

Previously, the majority of the region is claimed to be dangerous. (when fixing β_2 first)
While we successfully identify a safe region (when choosing β_2 according to n).

Our result helps explain why Adam works well in practice (e.g. NLP, CV applications):
They choose hyperparameters in the safe region.

Remark: Convergence to neighborhood.

- When $D_0 = 0$: converges to stationary points.
- When $D_0 > 0$: converges to the neighborhood of stationary points with size $\mathrm{O}((1-\beta_2)D_0)$. (often called ``ambiguity zone").
- This is common for constant-step SGD [Yan et al., 2018; Yu et al., 2019]

and diminishing-step Adaptive gradient methods [Zaheer et al., 2018; Shi et al., 2020]:

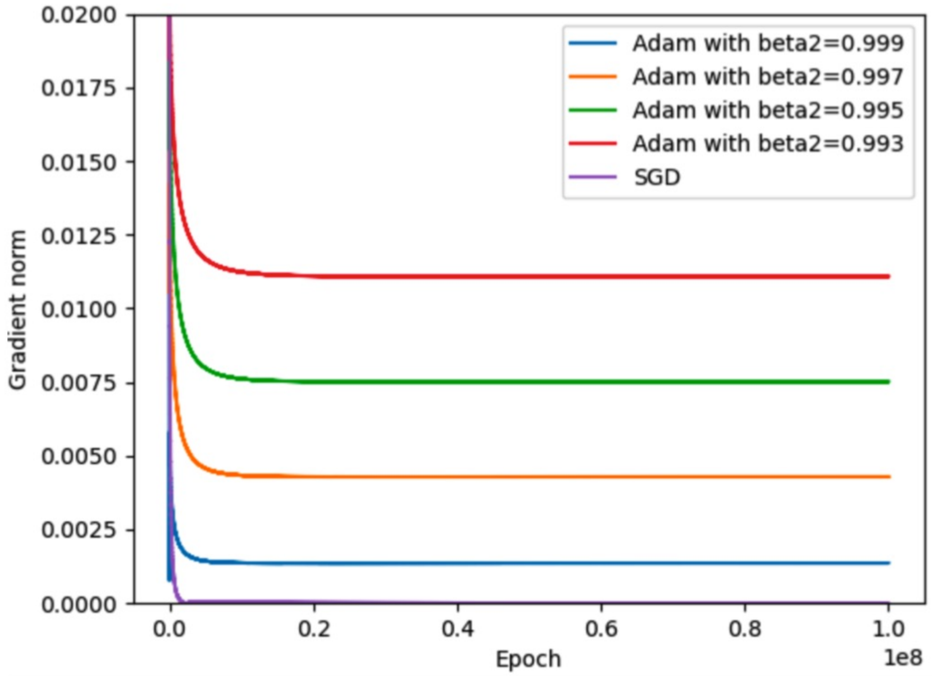
$$x_{k+1} = x_k - \frac{\eta_k}{\sqrt{v_k}} m_k$$

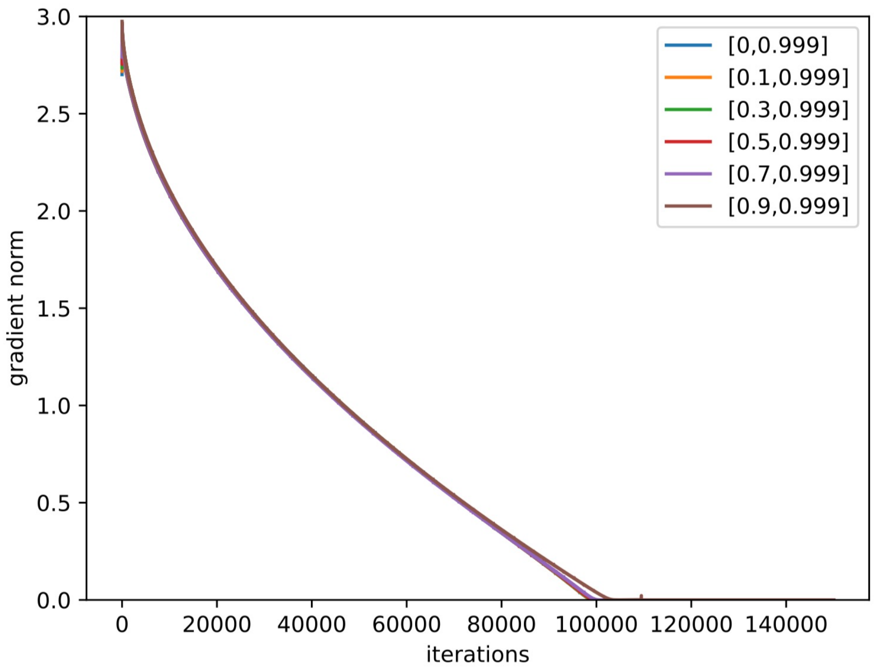
Although η_k is diminishing, $\frac{\eta_k}{\sqrt{v_k}}$ may not decrease.

Remark: Convergence to neighborhood.

Left: An example with $D_0 > 0$

Right: An example with $D_0 = 0$





Setting: Consider Adam & SGD with stepsize $\eta_k = \frac{1}{\sqrt{k}}$

Techniques for proving convergence

- To prove convergence, we want to show:
- $\mathbb{E}\left\langle \nabla f(x), \frac{m_k}{\sqrt{v_k}} \right\rangle = \left\langle \nabla f(x), \frac{(1-\beta_1)\nabla f_{\tau_k}(x_k)+\beta_1 m_{k-1}}{\sqrt{(1-\beta_2)\nabla f_{\tau_k}(x_k)\circ\nabla f_{\tau_k}(x_k)+\beta_2 v_{k-1}}} \right\rangle > 0$
- We point out there are at least two challenges
- We
- **Pr**

ranc

C

S

E

E

Challenge 2: non-linear perturbation. v_k makes the whole system a non-linear dynamics.

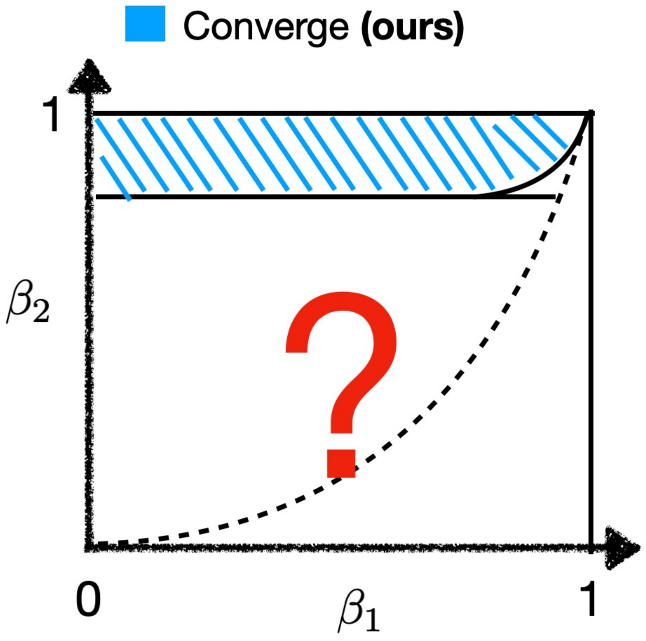
v_k also contains heavy history signal.

Further, v_k is statistically dependent with m_k. So $\mathbb{E}\left[\frac{m_k}{\sqrt{v_k}}\right] \neq \mathbb{E}[m_k]\mathbb{E}\left[\frac{1}{\sqrt{v_k}}\right]$

However, we are interested in any $\beta_1 \in [0,1)$

How does Adam behave in the rest of the region?

- When β_2 is large: we have shown a positive result.

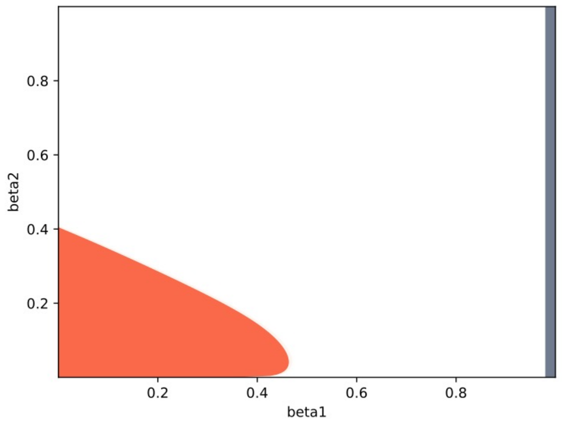


- When β_2 is small: we will show that Adam can still diverge! (even if the problem class is fixed)

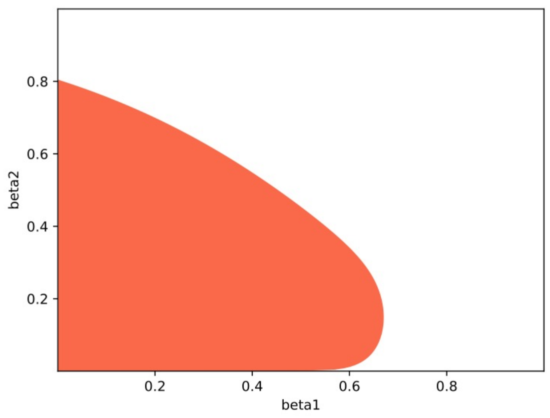
Divergence can happen when β_2 is small

- **Thm 2:** for any fixed n, there exists a function $f(x)$ satisfying **A1** and **A2**, s.t. when (β_1, β_2) are chosen in the orange region (size depends on n), Adam's iterates and function values diverge to infinity

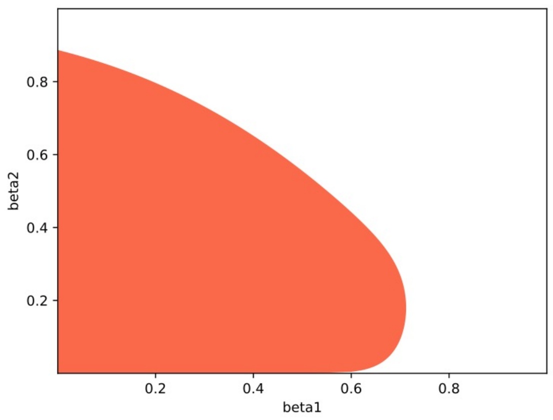
- The region is plotted by solving some analytic conditions.



(b) $n = 10$



(c) $n = 50$



(d) $n = 100$

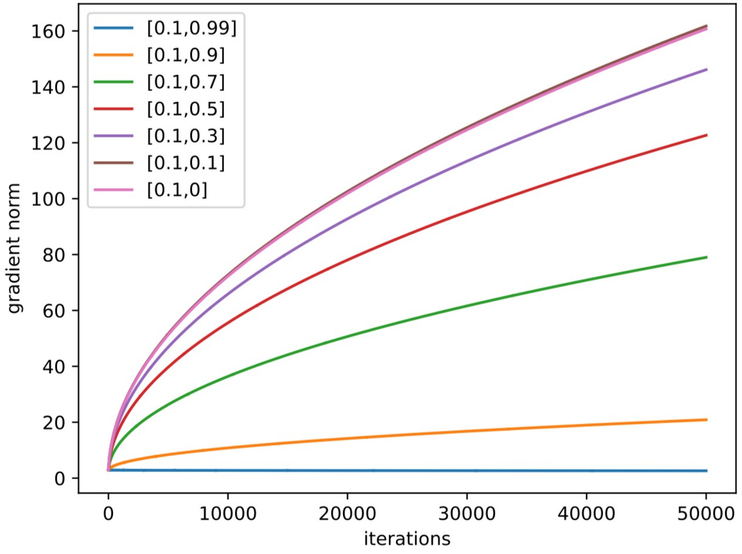
Some remarks on the divergence theorem

- **Remark 1.1:** The size of the orange region depends on n
- Our bound is **problem-dependent.**
- This is which is drastically different from (Reddi et al., 2018) which established the **problem-independent** worst-case choice of β_1 and β_2.

- **Remark 1.2:** The region expands to the whole region $[0,1]^2$ when n goes to infinity.
- When $n \to \infty$, our result recovers (actually stronger than) the problem-independent divergence result of (Reddi et al., 2018).
- We can view the divergence result of (Reddi et al., 2018) as **an asymptotic characterization** and our divergence result as a **non-asymptotic characterization** (for any fixed n).

Some remarks on the divergence theorem

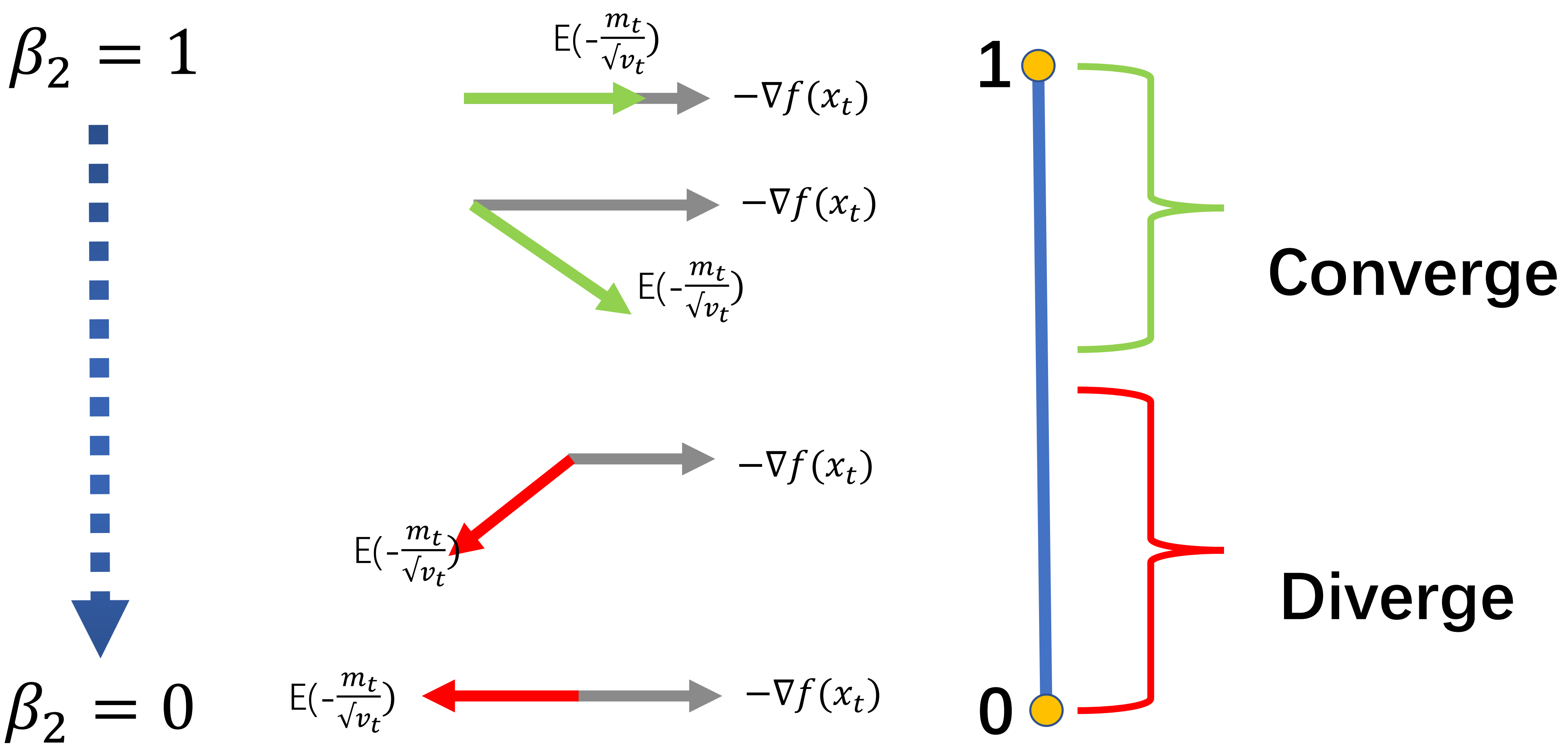
- **Remark 2:** For Adam, it is important to remove the bounded gradient assumption!!!
- In practice, the gradient of iterates can be unbounded.



(d) $n = 20$

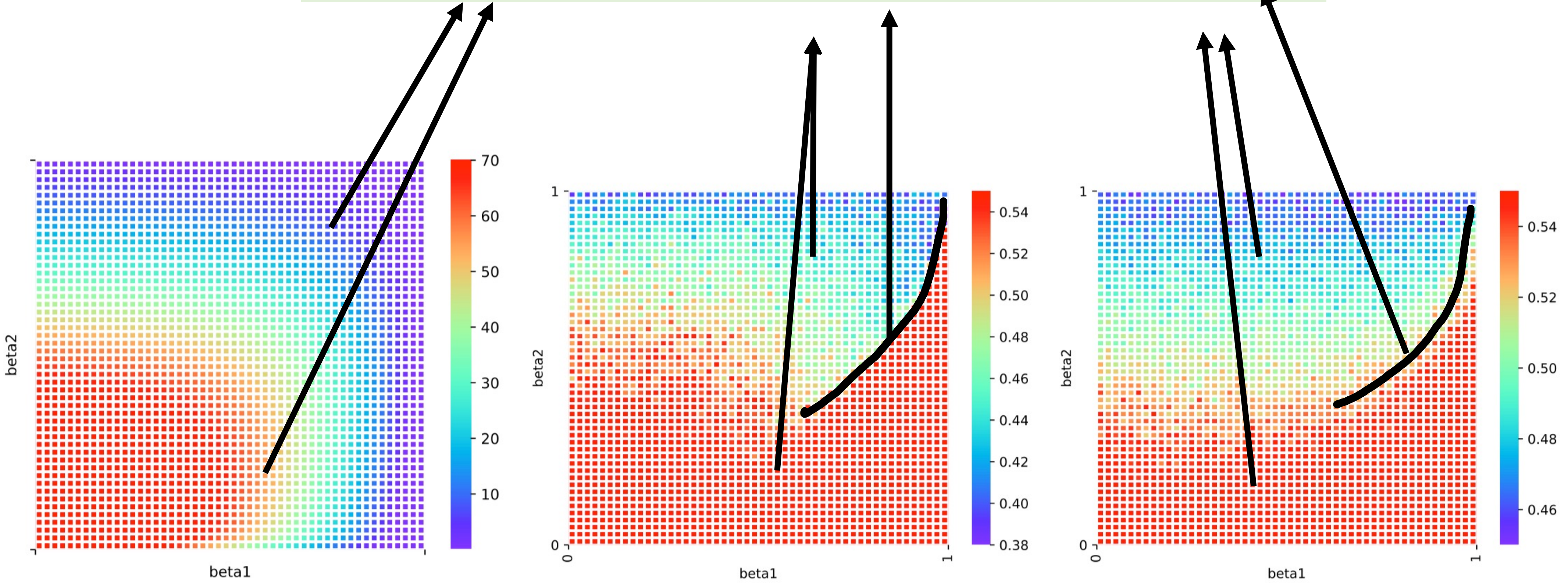
Intuition behind convergence and divergence

Adam: $\boldsymbol{x^{t+1} = x^{t} - \eta_t \frac{m_t}{\sqrt{v_t}}}$



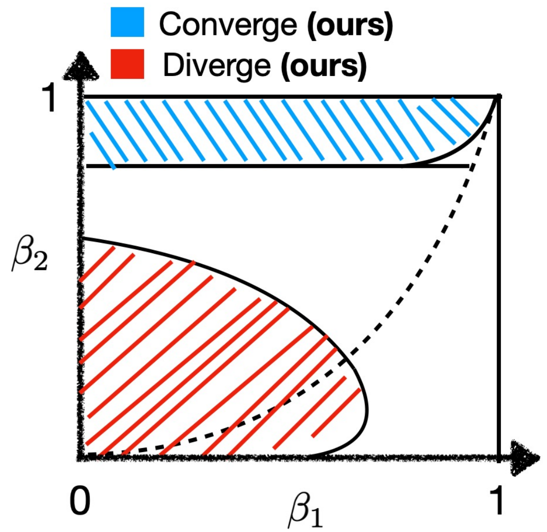
Our theory is consistent with experiments

Optim Optimization error is Smooth boundaries ion



(a) Function (2)   (b) MNIST   (c) CIFAR-10

Summary: the behavior of Adam changes dramatically under different hyperparameters



When increasing β_2:
There is a phase transition from divergence to convergence.

Setting	Hyperparameters	Adam's behavior
$\forall f(x)$ under **A1** and **A2** with $D_0 = 0$	β_2 is large and $\beta_1 < \sqrt{\beta_2}$	Converges to stationary points **(Ours)**
$\forall f(x)$ under **A1** and **A2** with $D_0 > 0$	β_2 is large and $\beta_1 < \sqrt{\beta_2}$	Converges to the neighborhood of stationary points **(Ours)**
$\exists\ f(x)$ under **A1** and **A2**	β_2 is small and a wide range of β_1	Diverges to infinity **(Ours)**

Implication to practitioners

- **Case study:** Bob is using Adam to train NNs. However, Adam with default hyperparamter fails in his tasks.
- Bob heard there is a well-known result that Adam can diverge.
- So he wonders: shall I keep tuning hyperparameter to make it work?
- Or shall I just give up and switch to other algorithms like AdaBound (which has 2 extra hyperparameters)?

Our suggestions:

1. Adam is still a theoretically justified algorithm. **Please use it confidently!**
2. Suggestions for hyperparameter tunning:
   First, tune up β_2. Then, try different β_1 with $\beta_1 < \sqrt{\beta_2}$

Mainly based on:

[1] Adam Can Converge Without Any Modification on Update Rules (Under review) .

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, Zhi-Quan Luo

Thanks to all the collaborators!