

Converge or Diverge? A Story of Adam

Ruoyu Sun



深圳国际工业与应用数学中心
SHENZHEN INTERNATIONAL CENTER FOR INDUSTRIAL AND APPLIED MATHEMATICS



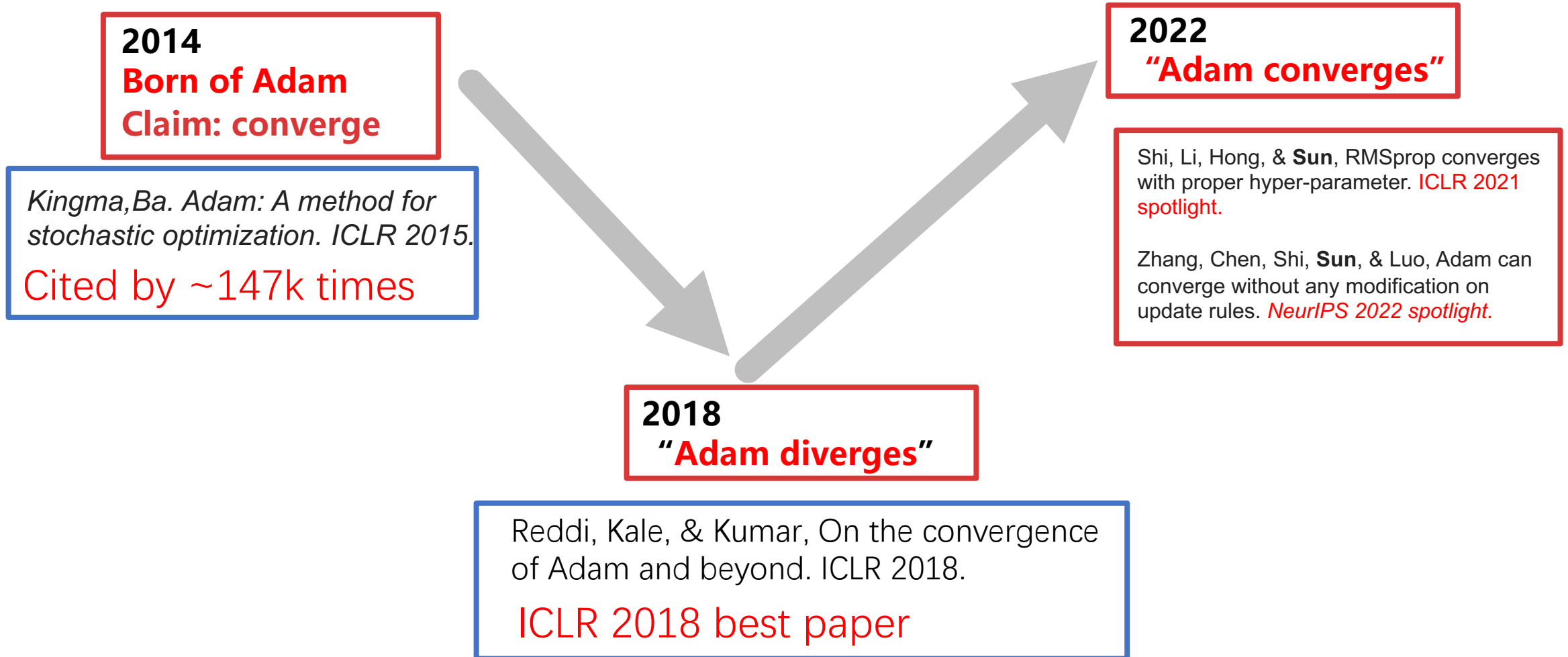
深圳市大数据研究院
Shenzhen Research Institute of Big Data



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

冯康科学计算青年研讨会
Aug 02, 2023

Overview of Story



Our work is tweeted by Dr. Kingma (an author of Adam)

Dr. Kingma: inventor of Adam and VAE;
co-founder of OpenAI; now a leader of Google Brain

← 主题帖

 **Durk Kingma**
@dpkingma

“Adam can converge without any modification on update rules”
arxiv.org/abs/2208.09632

Proves that (vanilla) Adam is theoretically justified without any modification. Presented at NeurIPS'22.
[翻译推文](#)



arxiv.org
Adam Can Converge Without Any Modification On Update ...
Ever since Reddi et al. 2018 pointed out the divergence issue of Adam, many new variants have been designed to obtain ...

上午7:55 · 2022年12月3日

79 转推 3 引用推文 406 喜欢次数

 **Durk Kingma** @dpkingma · 12月3日 ...
回复 @dpkingma
By Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, Zhi-Quan Luo.
1 5

 **Durk Kingma** @dpkingma · 12月3日 ...
Also provides suggestions for tuning hyperparameters beta1 and beta2.
1 12

 **Durk Kingma** @dpkingma · 12月3日 ...
Author handles:
[@ericzhang0410](#)
[@RuoyuSun_UI](#)
(Twitter, implement tweet edit functionality please? kthx)

 **Kevin Patrick Murphy** @... · 2022/12/4 ...
I'll need to add this ref to my book!
15

What to expect from this talk?

- Adam: To converge or not? That is a question.
- For practitioners & non-optimizers:
 - Story of Adam: born, popularity, convergence
 - how to tune hyperparameters of Adam
- For theory researchers:
 - different meanings of “algorithm convergences”
 - new method to analyze stochastic non-linear dynamic system

Empirical Guidance: Hyperparameter Tuning

- We prove that Adam **can converge** without ANY modification.
- **Hyperparameter tuning:**
 - First, tune up β_2 .
Then, try different β_1 with $\beta_1 < \sqrt{\beta_2}$
 - **Detailed suggestions:** end of talk

Tip for professors:

If DL experiments failed, one more question:
have you tuned Adam hyperparameters?

(many think Adam is tuning-free)

Contents

1. Story of Adam
2. Main Results
3. Proof Ideas
4. Experiments and Summary

Contents

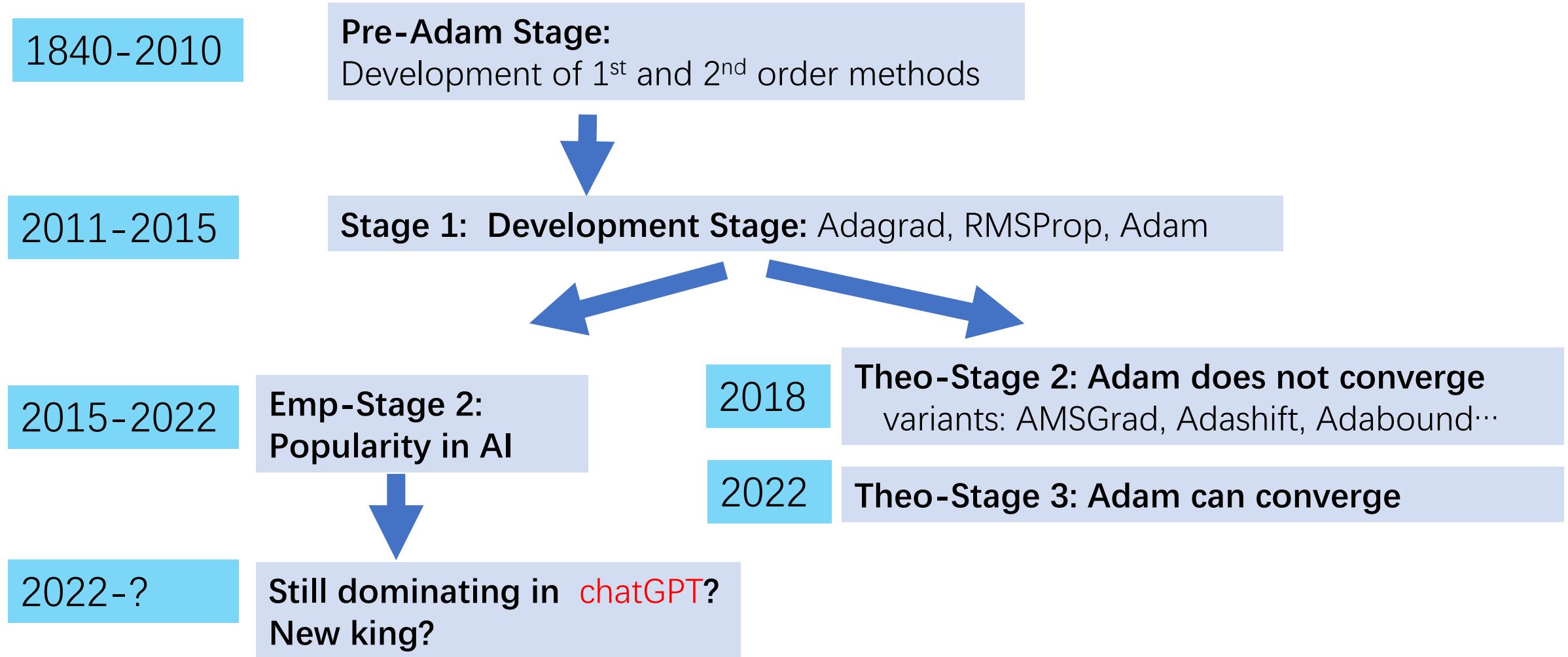
1. Story of Adam

2. Main Results

3. Proof Ideas

4. Experiments and Summary

Story of Adam: More Complete Version



Pre-Adam Stage: Classical Algorithms (1840-2010)

- Central issue in (unconstrained) nonlinear optimization:
information v.s. computation

1st order methods: gradient descent (1847, Cauchy)

Second order methods: Newton method

Quasi-2nd order methods:

BFGS (1970s), LBFGS (1980s), BB (1980s)

SGD and SGD with Momentum

- Consider $\min_x f(x) := \sum_{i=1}^n f_i(x)$.
 - n : number of samples (or mini-batches of samples)
 - x : trainable parameters
- In the k -th iteration: Randomly sample τ_k from $\{1, 2, \dots, n\}$

SGD (Stochastic gradient descent): $x_{k+1} = x_k - \eta_k \nabla f_{\tau_k}(x_k)$

SGD with momentum (SGDM):

$$m_k = (1 - \beta_1) \nabla f_{\tau_k}(x_k) + \beta_1 m_{k-1}$$

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



1st order momentum



Iterate update

Stage 1: Development of Adam (2011-2015)

2011: Adagrad, JMLR

Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research* 12.7 (2011).

2012: RMSProp, Lecture notes

[CITATION] Lecture 6.5-**rmsprop**: Divide the gradient by a running average of its recent magnitude

[T Tieleman, G Hinton](#) - COURSERA: Neural networks for machine learning, 2012

☆ Save  Cite Cited by 6438 Related articles

2015: Adam, ICLR

Kingma, Ba. Adam: A method for stochastic optimization. ICLR 2015.

Adagrad

In 2011, in an intern at Google, to handle language problems, Duchi proposed Adagrad.

$$g_t = \nabla F_i(\theta_t),$$

Compute stochastic gradient

$$v_t = \frac{t-1}{t} v_{t-1} + \frac{1}{t} g_t \circ g_t,$$

2nd order momentum

$$\theta_{t+1} = \theta_t - \alpha_t \frac{g_t}{\sqrt{v_t} + \epsilon}, \quad t = 0, 1, 2, \dots,$$

Iterate update

Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research* 12.7 (2011).

RMSProp

AdaGrad: it treats all samples equally.

RMSprop: use EMA (exponential moving average) to define v_t .

RMSProp: At the t -th iteration, randomly pick i and compute $g_t = \nabla F_i(\theta_t)$, then:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t \circ g_t,$$
$$\theta_{t+1} = \theta_t - \alpha_t \frac{g_t}{\sqrt{v_t} + \epsilon}.$$



2nd order momentum



Iterate update

PyTorch default: $\beta_2 = 0.99, \epsilon = 10^{-8}$ (together with $\alpha_t = 0.01$)

Common choice: $\beta_2 \in \{0.9, 0.99, 0.999\}$.

Adam

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

- Adam (Kingma and Ba'15):

- $m_k = (1 - \beta_1) \nabla f_{\tau_k}(x_k) + \beta_1 m_{k-1}$

- $v_k = (1 - \beta_2) \nabla f_{\tau_k}(x_k) \circ \nabla f_{\tau_k}(x_k) + \beta_2 v_{k-1}$

- $x_{k+1} = x_k - \eta_k \frac{\sqrt{1 - \beta_2^k}}{1 - \beta_1^k} \frac{m_k}{\sqrt{v_k}}$



1st order momentum



2nd order momentum



Iterate update

- 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: $\beta_2 = 0.999$

Emp-Stage 2: Popularity in AI

- **Adam** becomes the most popular algorithms in deep learning (DL). (>147,000 citations, in July 2023)
- **Default in LLM (large language models)**

```
optimizer = optim.Adam(net.parameters(), lr=args.lr, betas=(args.beta1, args.beta2), eps=1e-08,|  
                        weight_decay=args.weightdecay, amsgrad=False)
```

Empirical fact (sad?): Adam is the choice of chatGPT.

--New algorithms, including Sophie, Lion, **cannot beat Adam** on 100-billion-parameter models.

Misconception on Adam

- well-tuned SGD (with momentum) outperforms Adam

[Wilson et al.'2017] “Marginal values of adaptive gradient methods” claimed:

- Criticism of Adam [Wilson2017] did not stop Adam. Why?
- I asked: **who used Adam?** Check top papers citing Adam paper.

1000x Deep learning
I Goodfellow, Y Bengio, A Courville - 2016 - synapse.koreamed.org
Kwang Gi Kim https://doi.org/10.4258/hir.2016.22.4.351 ing those who are beginning their careers in deep learning and artificial intelligence research. The other target audience consists of software engineers who may not have a background in machine learning or ...
☆ 99 Cited by 27202 Related articles All 24 versions 80

Transformer, NLP
Attention is all you need
A Vaswani, N Shazeer, N Parmar, J Uszkoreit, ... - arXiv preprint arXiv:1706.03762, 2017 - arxiv.org
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We ...
☆ 99 Cited by 23024 Related articles All 22 versions 80

GAN
Unsupervised representation learning with deep convolutional generative adversarial networks
A Radford, L Metz, S Chintala - arXiv preprint arXiv:1511.06434, 2015 - arxiv.org
In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help bridge the gap between the ...
☆ 99 Cited by 9551 Related articles All 2 versions 80

GAN
Image-to-image translation with conditional adversarial networks
P Isola, JY Zhu, T Zhou, ... - Proceedings of the IEEE ..., 2017 - openaccess.thecvf.com
We investigate conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to ...
☆ 99 Cited by 9448 Related articles All 15 versions 80

GAN
Unpaired image-to-image translation using cycle-consistent adversarial networks
JY Zhu, T Park, P Isola, AA Efros - Proceedings of the IEEE ..., 2017 - openaccess.thecvf.com
Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. However, for many tasks, paired training data will not be available. We ...
☆ 99 Cited by 8857 Related articles All 14 versions 80

GNN
Semi-supervised classification with graph convolutional networks
TN Kipf, M Welling - arXiv preprint arXiv:1609.02907, 2016 - arxiv.org
We present a scalable approach for semi-supervised learning on graph-structured data that is based on an efficient variant of convolutional neural networks which operate directly on graphs. We motivate the choice of our convolutional architecture via a localized first-order ...
☆ 99 Cited by 8675 Related articles All 15 versions 80

NLP: Vision
Show, attend and tell: Neural image caption generation with visual attention
K Xu, J Ba, R Kiros, K Cho, A Courville, ... - International ..., 2015 - proceedings.mlr.press
Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard ...
☆ 99 Cited by 7254 Related articles All 28 versions 80

GAN
Wasserstein generative adversarial networks
M Arjovsky, S Chintala, L Bottou - conference on machine ..., 2017 - proceedings.mlr.press
We introduce a new algorithm named WGAN, an alternative to traditional GAN training. In this new model, we show that we can improve the stability of learning, get rid of problems like mode collapse, and provide meaningful learning curves useful for debugging and ...
☆ 99 Cited by 7289 Related articles All 10 versions 80

NLP
Deep contextualized word representations
ME Peters, M Neumann, M Iyyer, M Gardner, ... - arXiv preprint arXiv ..., 2018 - arxiv.org
We introduce a new type of deep contextualized word representation that models both (1) complex characteristics of word use (eg, syntax and semantics), and (2) how these uses vary across linguistic contexts (ie, to model polysemy). Our word vectors are learned functions of ...
☆ 99 Cited by 7086 Related articles All 20 versions 80

Reinforcement learning
Continuous control with deep reinforcement learning
TP Lillicrap, JJ Hunt, A Pritzel, N Heess, T Erco, ... - arXiv preprint arXiv ..., 2015 - arxiv.org
We adapt the ideas underlying the success of Deep Q-Learning to the continuous action domain. We present an actor-critic, model-free algorithm based on the deterministic policy gradient that can operate over continuous action spaces. Using the same learning algorithm ...
☆ 99 Cited by 5994 Related articles All 14 versions 80

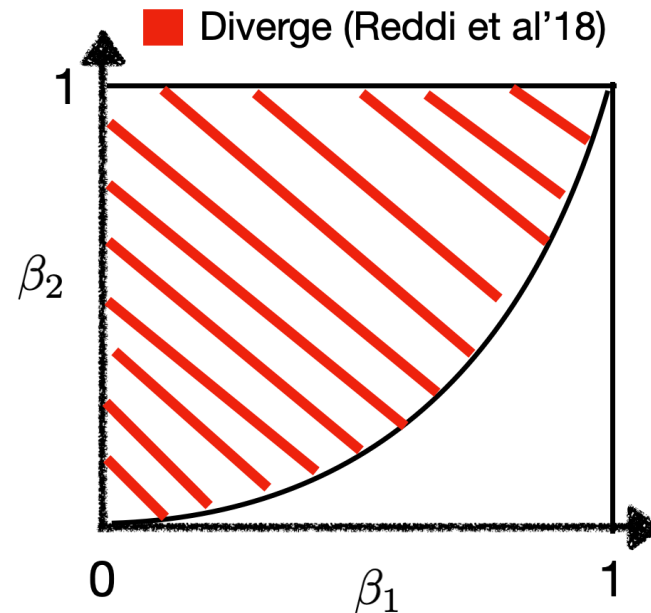
10 top cited papers in 2017:
4 GANs, 3 NLP, 1 RL, 1 GNN.

Finding:
people use Adam in GAN, RL, NLP,
etc; not in image classification

Theo-Stage 2: “Adam does not converge”

Reddi et al.18 (ICLR Best paper):

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



Implication on “convergence issue”

ICLR’18 paper reader’s comment:

[-] **Is the problem with Adam, or with the theoretical framework used to analyse it?**

Jeremy Bernstein

26 Apr 2018 (modified: 26 Apr 2018) ICLR 2018 Conference Paper807 Public

“My claim is that...for any problem, a properly tuned-Adam will converge at least as well as SGD and signSGD”

ICLR’18 paper authors reply:

[-] **TL;DR : Its with the algorithm** 

ICLR 2018 Conference Paper807 Authors

01 Jun 2018 ICLR 2018 Conference Paper807 Official Comment Readers:

 Everyone

Comment: Dear Jeremy,

Thank you for your interest in the paper.

To answer your question "Is the problem with Adam?" : Our paper shows that the algorithm defined in the Adam paper (<https://arxiv.org/pdf/1412.6980.pdf>, Algorithm 1) (including one with decreasing step size alpha) has convergence issues. Specifically, for any setting of the Adam parameters (beta_1, beta_2, minibatch size, epsilon, etc) there is a convex optimization setting where Adam will not converge to the optimal solution, even if decreasing learning rates are used. This is in contrast to algorithms like SGD which, with decreasing learning rates, is guaranteed to converge.

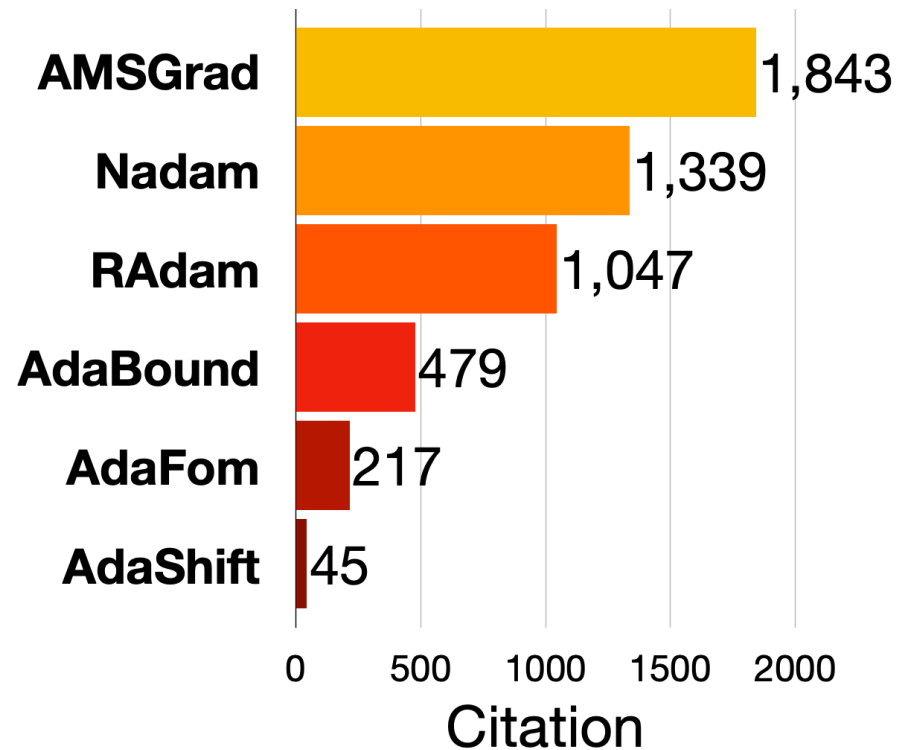
Our paper shows that the algorithm defined in the Adam paper has convergence issues.

To Overcome Divergence, ...

- Modify Adam
 - **AMSGrad, AdaFom** [Reddi et al.'18, Chen et al.'18]: keep $v_k \geq v_{k-1}$
 - **Slow convergence** [Zhou et al.'18]
 - **AdaBound** [Luo et al.'19]: Impose constraint: $v_k \in [C_l, C_u]$
 - **Need to tune two extra hyperparameters**

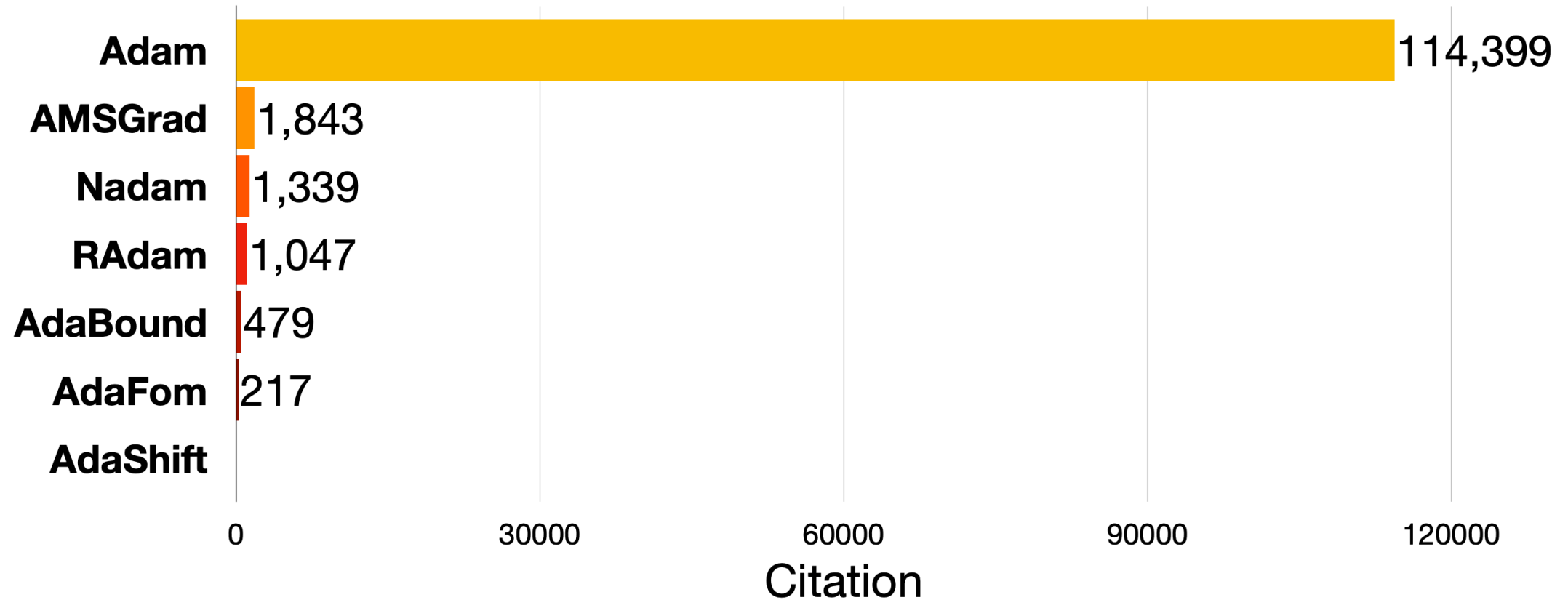
However, vanilla Adam works well for most practical applications!

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)

Why is divergence not observed?

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

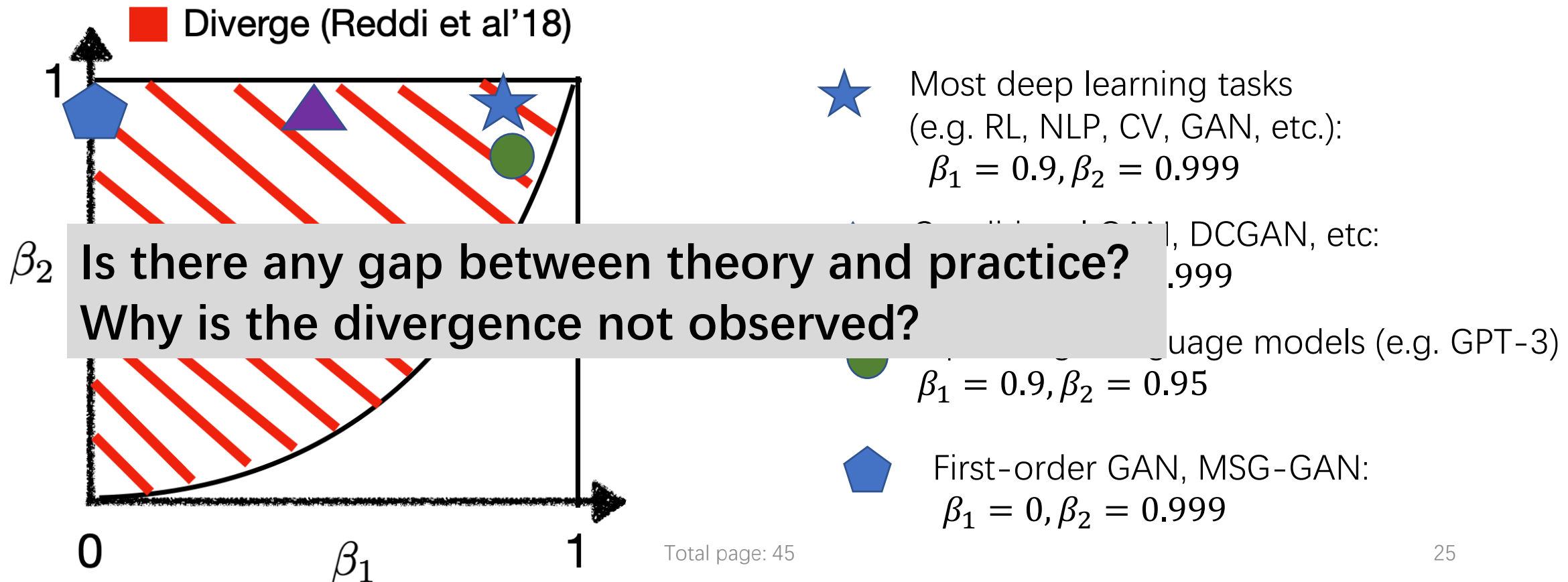
Proof(Reddi et al. 18):

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

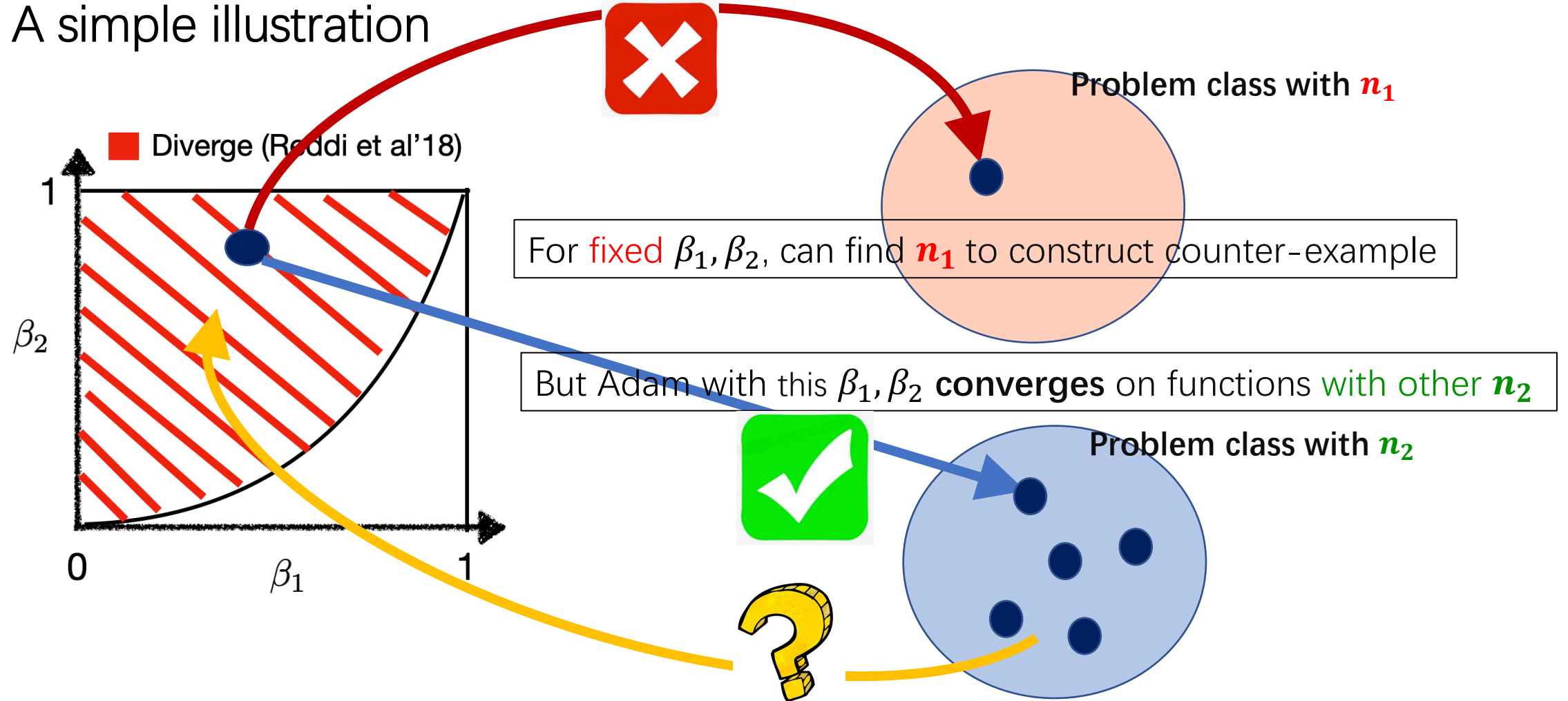
- An important (but often ignored) feature: Reddi et al. fix β_1, β_2 **before picking the problem** (change n according to β_1, β_2)
- While in practice, parameters are often **problem-dependent** (e.g. tuning lr for GD)
- **Conjecture: Adam might converge for fixed problem (or fixed n)**

Divergence theory does not match practice

Observation: the reported (β_1, β_2) **actually satisfy divergence condition** $\beta_1 < \sqrt{\beta_2}$!



A simple illustration



Question: Does Adam converge for fixed problem class (fixed n)?

Contents

1. Story of Adam

2. Main Results

3. Proof Ideas

4. Experiments and Summary

Assumptions

- Consider $\min_x f(x) := \sum_{i=1}^n f_i(x)$
- **A1 (L-smooth)**: assume $\nabla f_i(x)$ are L-Lipschitz continuous.
- **A2 (Affine Variance)**: $\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x) - \frac{1}{n} \sum_{i=1}^n \nabla f_i(x)\|_2^2 \leq D_1 \|\nabla f(x)\|_2^2 + D_0$
 - A2** is quite general:
 - When $D_1 = 0$, A2 becomes bounded variance, commonly used in SGD analysis
 - **A2** allows $D_1 > 0$ and thus it is weaker than bounded variance.
 - When $D_0 = 0$, **A2** becomes "Strong Growth Condition (SGC)" [Vaswani et al., 19]
 - **Intuition**: When $\|\nabla f(x)\|=0 \Rightarrow$ we have $\|\nabla f_i(x)\|=0$.
 - SGC holds for overparametrized networks (Vaswani et al.19)
- We do NOT need the following assumptions, which are common in the literature
- ~~**A3** (bounded gradient): $\|\nabla f(x)\| < C$~~
- ~~**A4** (bounded 2nd-order momentum): $v_k \in [C_t, C_u]$~~
- To our knowledge, **A1+ A2** are the mildest assumption set so far.

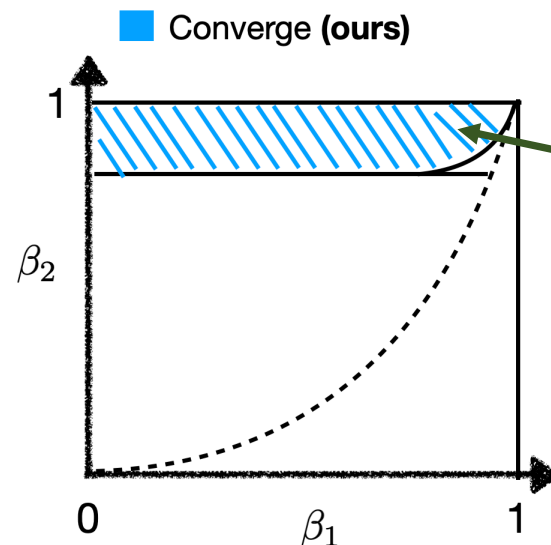
Convergence results for large β_2

- **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:

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

- **RK:** When $D_0 = 0$ (e.g., for overparameterized models): Adam converges to stationary points



We identify a safe region! (UNKNOWN BEFORE!)

Remark: Convergence to neighborhood

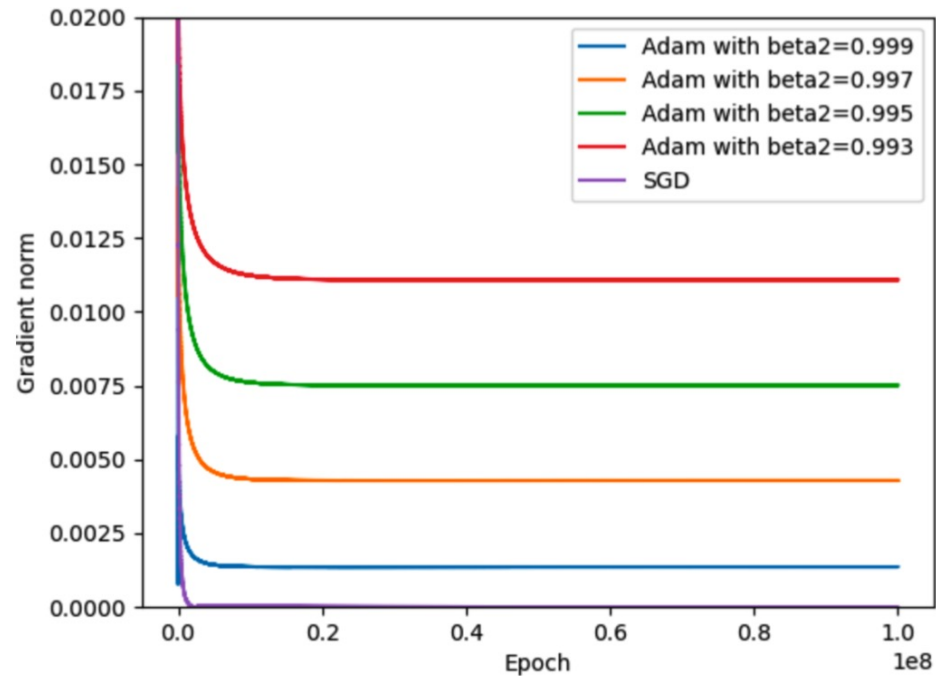
- When $D_0 > 0$: **converges to a neighborhood of stationary points** with size $O((1 - \beta_2)D_0)$. (a.k.a. “ambiguity zone”).
- **This** is common for
 - constant-step SGD [Yan et al., 2018; Yu et al., 2019]
 - diminishing-lr adaptive gradient methods [Zaheer et al., 2018; Shi et al., 2020]:

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

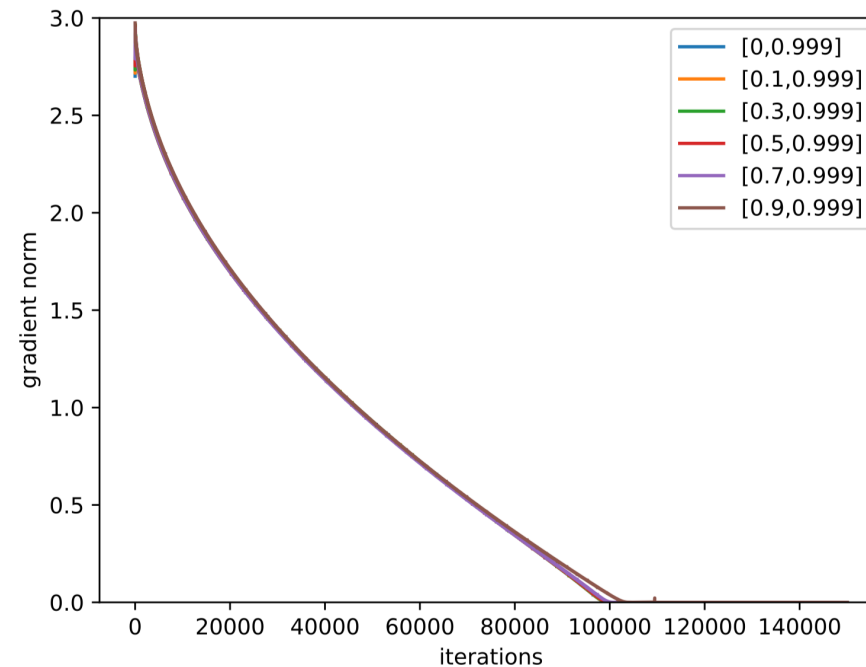
Intuition: Although η_k is diminishing, $\frac{\eta_k}{\sqrt{v_k}}$ may not decrease.

Remark: Convergence to neighborhood.

Left: An example with $D_0 > 0$



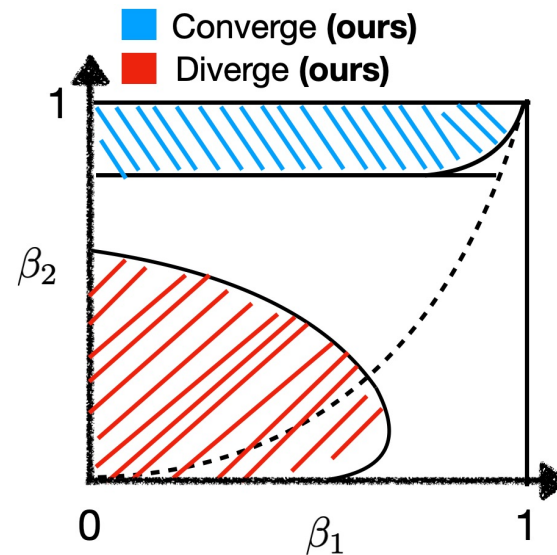
Right: An example with $D_0 = 0$



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

How does Adam behave in the rest of the region?

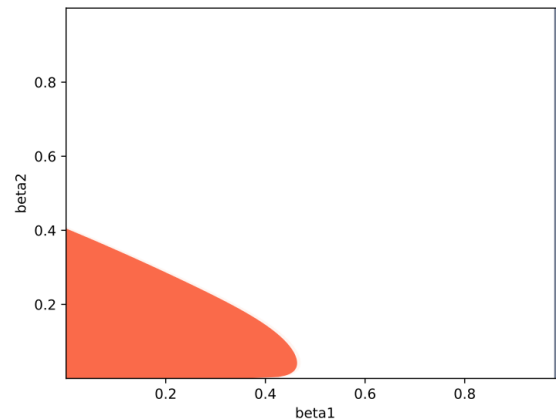
- When β_2 is large: we have shown a positive result.



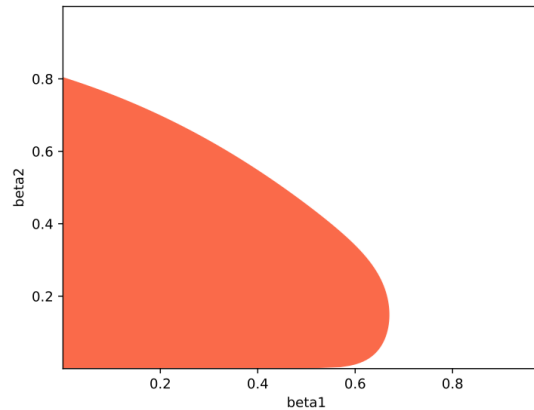
- When β_2 is small: we will show that Adam can still diverge! (even if the problem class is fixed)

Divergence can happen when β_2 is small

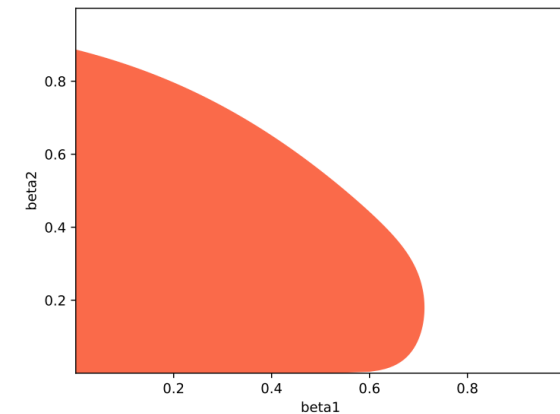
- **Thm 2:** For any fixed n , there exists a function $f(x)$ satisfying **A1** and **A2**, s.t. when (β_1, β_2) are chosen in the orange region (size depends on n), s.t. Adam's iterates and function values **diverge to infinity**
- The region is **precisely drawn** (plotted by solving some analytic conditions)
- region enlarges with n



(b) $n = 10$



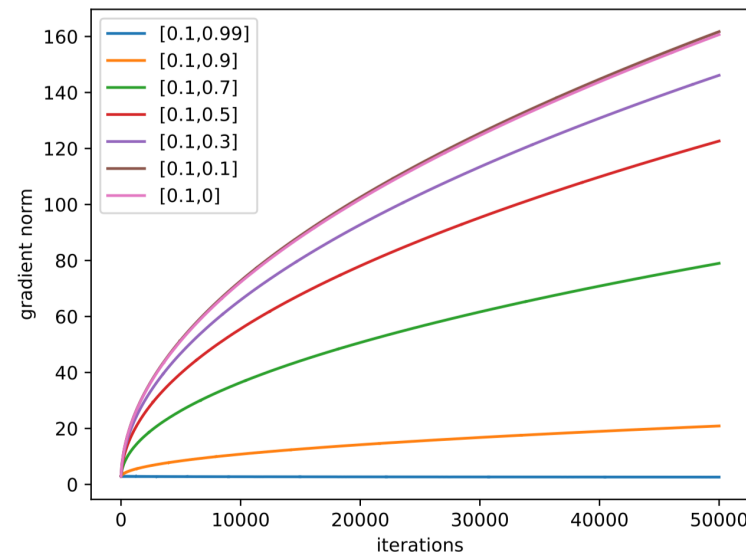
(c) $n = 50$



(d) $n = 100$

Some remarks on the divergence theorem

- **Remark 2:** For Adam, it is important to remove the bounded gradient assumption ($\|\nabla f(x)\| < C$) !!!
- In practice, the gradient of iterates can be unbounded.



(d) $n = 20$

Total page: 45

Implication to practitioners

- **Case study:** Bob is using Adam to train NNs. However, Adam with default hyperparameter 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 tuning:
In one sentence: First, tune up β_2 . Then, try different β_1 with $\beta_1 < \sqrt{\beta_2}$

Contents

1. Story of Adam

2. Main Results

3. Proof Ideas

4. Experiments and Summary

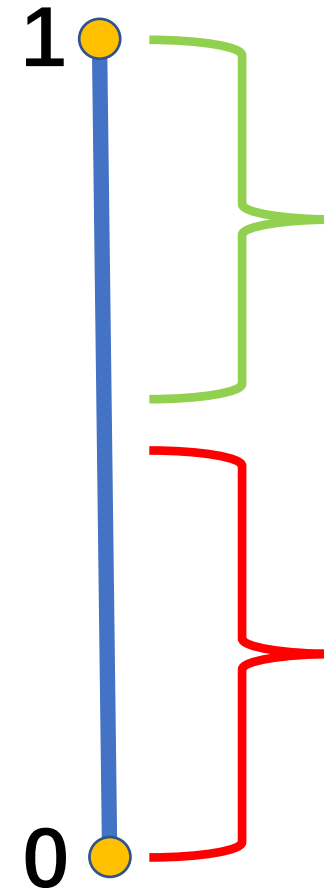
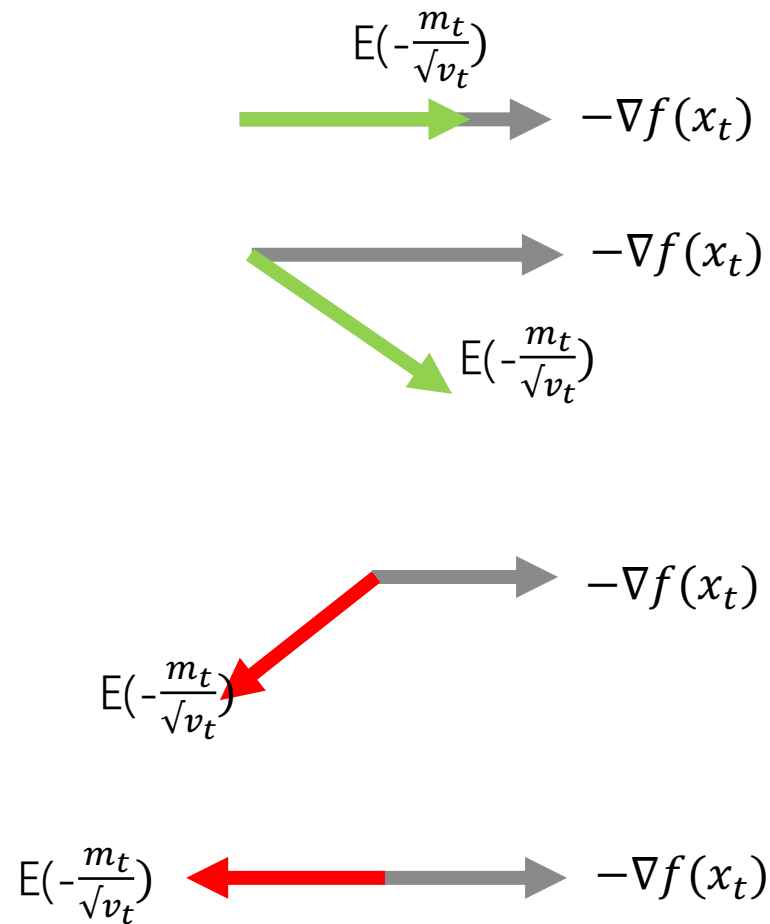
Intuition behind convergence and divergence

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

$\beta_2 = 1$



$\beta_2 = 0$



Converge

Diverge

Proof Ideas for Convergence Analysis: An Overview

$$\text{Want to show: } \mathbb{E} \left\langle \nabla f(x_k), \frac{m_k}{\sqrt{v_k}} \right\rangle = \mathbb{E} \left\langle \nabla f(x_k), \frac{(1-\beta_1)\nabla f_{\tau_k}(x_k) + \beta_1 m_{k-1}}{\sqrt{(1-\beta_2)\nabla f_{\tau_k}(x_k) \circ \nabla f_{\tau_k}(x_k) + \beta_2 v_{k-1}}} \right\rangle > 0$$

Want to prove: $\mathbb{E}(g, m/u) > 0$, where $u = \sqrt{v}$.

Major challenge 1: Random u appears in the denominator, may blow up.

Major challenge 2: momentum m contains history information.

Solution Step 1 (Special case of $m = g$, or $\beta_1 = 0$, i.e. RMSProp)

Prove $\mathbb{E}\langle g, g/u \rangle > \text{constant}$. [Shi et al.'2021]

Solution Step 2 (general m , true Adam):

Prove $\mathbb{E}\langle g, m/u \rangle \approx \mathbb{E}\langle g, g/u \rangle$

i.e. $\mathbb{E}\langle g, (g - m)/u \rangle \approx 0$

i.e. $\mathbb{E}\langle g - m, g/u \rangle \approx 0$ 

Step 2.1: $\mathbb{E}(g - m) \approx 0$.

Step 2.2: $\mathbb{E}\langle g - m, g/u \rangle \approx 0$.

Proof Ideas for Convergence Analysis: An Overview

Want to show: $\mathbb{E} \left\langle \nabla f(x_k), \frac{m_k}{\sqrt{v_k}} \right\rangle = \left\langle \nabla f(x_k), \frac{(1-\beta_1)\nabla f_{\tau_k}(x_k) + \beta_1 m_{k-1}}{\sqrt{(1-\beta_2)\nabla f_{\tau_k}(x_k) \circ \nabla f_{\tau_k}(x_k) + \beta_2 v_{k-1}}} \right\rangle > 0$

Challenge 1: m_k contains heavy history

Challenge 2: v_k brings non-linear perturbation.

Key Idea: Established a new property of Adam's momentum under random permutations.

Step 1: Show the periodical property of momentum

Step 2: Control the perturbation when β_2 is large

Lemma 5.1. (Informal) Consider Algorithm 1. For every $l \in [d]$ and any $\beta_1 \in [0, 1)$, we have the following result under Assumption 2.1

$$\delta(\beta_1) := \mathbb{E} \sum_{i=0}^{n-1} (m_{l,k,i} - \partial_l f_{\tau_k,i}(x_{k,0})) = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right),$$

where $\partial_l f(x_{k,0})$ is the l -th component of $\nabla f(x_{k,0})$; $m_{l,k,i} = (1 - \beta_1)\partial_l f_{\tau_k,i}(x_{k,i}) + \beta_1 m_{l,k,i-1}$.

Lemma 5.2. (Informal) Under Assumption 2.1 and 2.2 consider Algorithm 1 with large β_2 and $\beta_1 < \sqrt{\beta_2}$. For those l with gradient component larger than certain threshold, we have:

$$\left| \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} - \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} \right| = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right); \quad (5)$$

$$\mathbb{E} \left(\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}} \sum_{i=0}^{n-1} (m_{l,k,i} - \partial_l f_{\tau_k,i}(x_{k,0})) \right) = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right). \quad (6)$$

Tackling Challenge 1: $\mathbf{m}_k$ contains heavy history

- **Want To Show:** $\mathbb{E}(\mathbf{g} - \mathbf{m}) = \mathbb{E}(\nabla f(x_k) - \mathbf{m}_k) \approx 0$
- **What is the math problem here?** Estimate difference of two sum's.
- Understanding Step (i): **Batch case**
 - $\mathbf{g}$ = gradient of $f(x) := \sum_{i=1}^n f_i(x)$ is a **sum**.
 - $\mathbf{m}$ = weighted average of past gradients = **double sum** (weighted average of sum's)
- Understanding Step (ii): **Stochastic case**
 - $\mathbf{g}$ = gradient of $f(x) := \sum_{i \in B} f_i(x)$ is a (partial) **sum**.
 - $\mathbf{m}$ = weighted average of past gradients = **double sum** (weighted average of sum's)

Tackling Challenge 1: m_k contains heavy history

- **Want To Show:** $\mathbb{E}(\nabla f(x_k) - m_k) \approx 0$
- **Solution:** we construct a simple model as follows.
- **Color-ball of the 1st kind:** consider a box contains two balls labeled c_0 and c_1 . In each round (epoch), we randomly sample balls from the box without replacement, then we put them both back. We denote the 1st sampled label in the k -th epoch as a_k and the 2nd sampled one as b_k .
- We define the following quantities:

$$m_1 = \underbrace{b_k + \beta a_k}_{m_{1,k}} + \underbrace{\beta^2 b_{k-1} + \beta^3 a_{k-1}}_{m_{1,k-1}} + \cdots + \underbrace{\beta^{2(k-1)} b_1 + \beta^{2(k-1)+1} a_1}_{m_{1,1}};$$

$$m_0 = \underbrace{a_k}_{m_{0,k}} + \underbrace{\beta^1 b_{k-1} + \beta^2 a_{k-1}}_{m_{0,k-1}} + \cdots + \underbrace{\beta^{2(k-1)-1} b_1 + \beta^{2(k-1)} a_1}_{m_{1,1}};$$

Tackling Challenge 1: m_k contains heavy history

- **Want To Show:** $\mathbb{E}(\nabla f(x_k) - m_k) \approx 0$
- **Solution:** we construct a simple model as follows.
- **Color-ball of the 1st kind:** We further define the following quantities:

$$f_1 = c_1 \underbrace{(1 + \beta)}_{f_{1,k}} + \underbrace{\beta^2 + \beta^3 \cdots + \beta^{2(k-1)} + \beta^{2(k-1)+1}}_{f_{1,k-1}};$$

$$f_0 = c_0 \underbrace{(1 + \beta)}_{f_{0,k}} + \underbrace{\beta^2 + \beta^3 \cdots + \beta^{2(k-1)} + \beta^{2(k-1)+1}}_{f_{0,1}};$$

Lemma A.1. *In the color-ball model of the 1st kind, we have*

$$\left| \mathbb{E} \left[\sum_{i=0}^1 m_i - \sum_{i=0}^1 f_i \right] \right| = \boxed{\beta^{2(k-1)+1}} \left(\frac{c_0}{2} + \frac{c_1}{2} \right),$$

Surprising point:
No low-order terms!

Due to random permutation.

Contents

1. Story of Adam

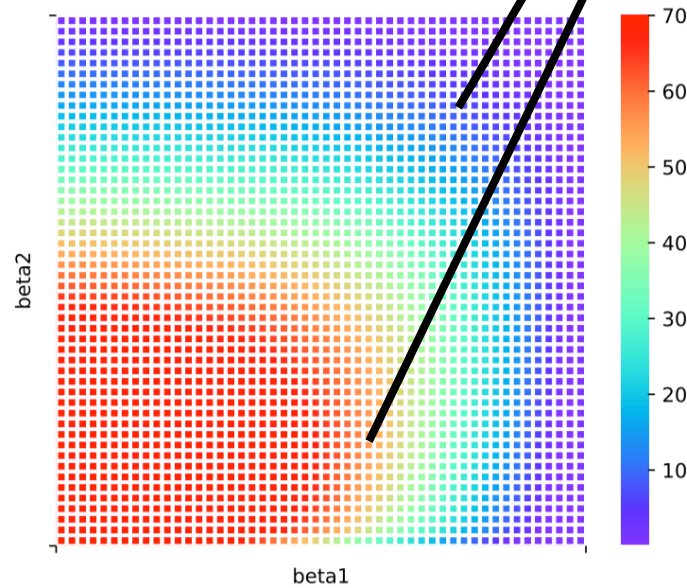
2. Main Results

3. Proof Ideas

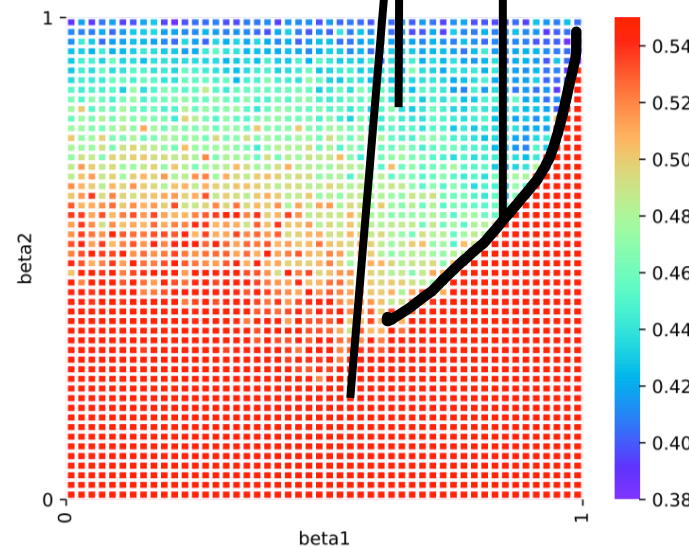
4. Experiments and Summary

Our theory is consistent with experiments

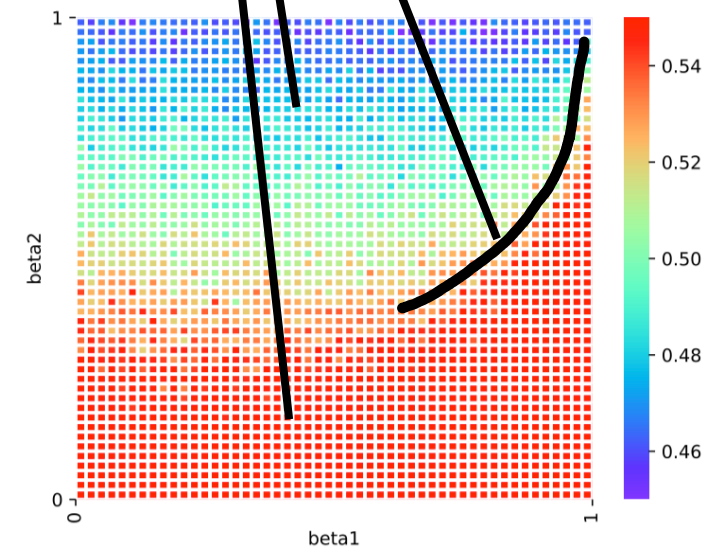
Optimization error is Smooth boundaries



(a) Function (2)

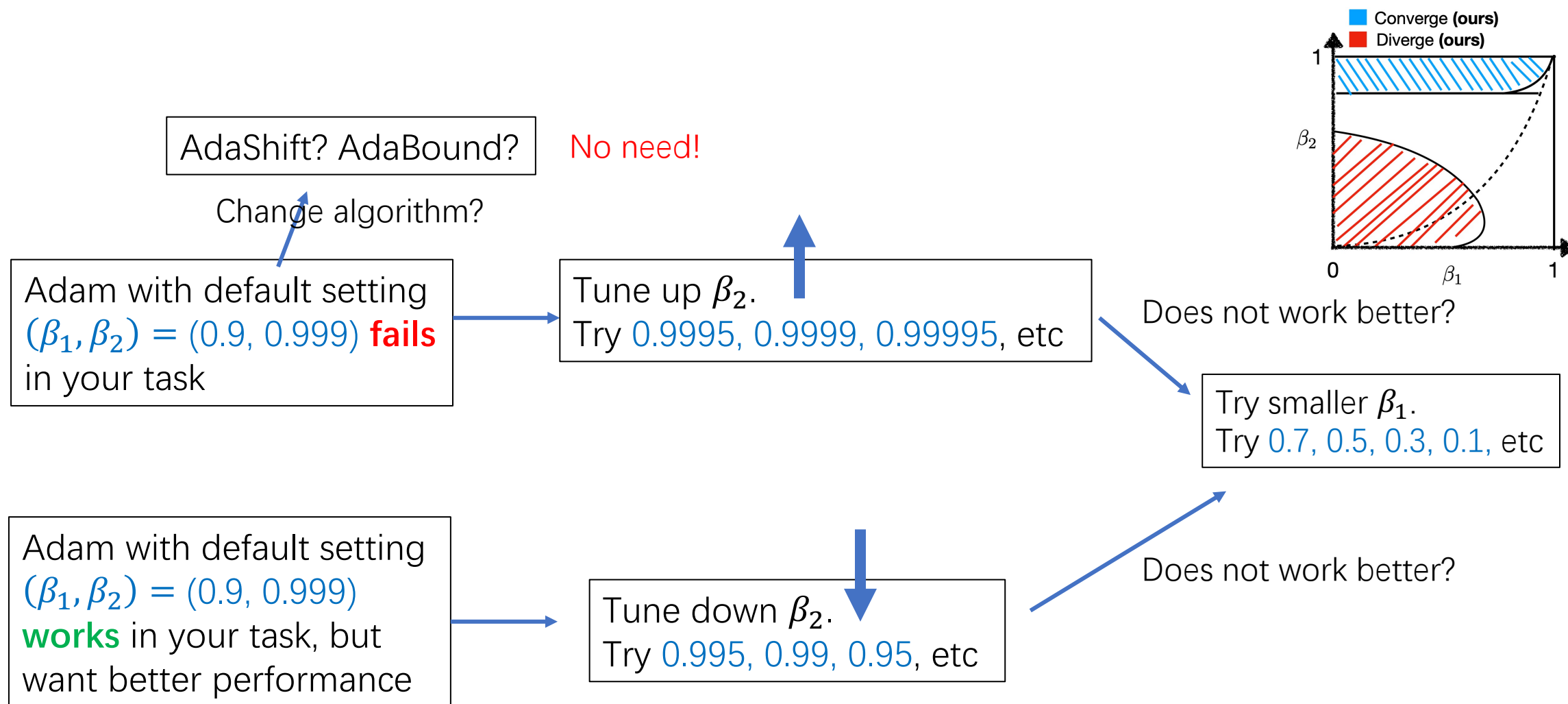


(b) MNIST

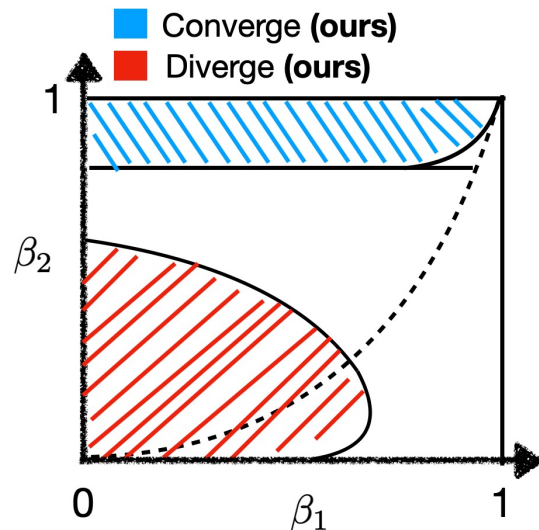


(c) CIFAR-10

Recipe for Adam hyperparameter-tuning



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)

Mainly based on:

- Shi, Li, Hong, & **Sun**, RMSprop converges with proper hyper-parameter. **ICLR 2021 spotlight.**
- Zhang, Chen, Shi, **Sun**, & Luo, Adam can converge without any modification on update rules. **NeurIPS 2022 spotlight.**

Thanks to all the collaborators!



References on Deep Learning Optimization

For a big picture on neural network optimization, see:

- **Book in preparation:**
R. Sun, “Lectures on deep learning optimization”, in preparation. To be published by **Springer Nature**.

Surveys:

- [S1] **R. Sun**, D. Li, S. Liang, T. Ding, R. Srikant, “The Global landscape of neural networks: An overview.” **IEEE Signal Processing Magazine**, 2020
- [S2] **R. Sun**, “Optimization for deep learning: an overview.” (survey paper). **Journal of Operations Research Society of China**, Special Issue on “Continuous Optimization: Past, Present and Future” (invited). 2020.

Thank You!

Email 邮箱: sunruoyu@cuhk.edu.cn

Website 个人网站: ruoyus.github.io

Interested in discussing chatGPT, machine learning,
optimization, communications, etc.

研究兴趣: ChatGPT、机器学习、优化、通信