Nostalgic Adam

Weighting more of the past gradients
when designing the adaptive learning rate

Haiwen Huang, Chang Wang, Bin Dong

Peking University

2019.8

the

MELODY

of

Deep Learning

M Model

E Evaluation

L Loss

O Optimization

D Dataset


M
Model
E
Evaluation
L
Loss
O
Optimization
D
Dataset

CONTENTS

- Introduction to stochastic optimization methods
- Non-Convergence Issue
- Our solution: NosAdam
- Why Nostalgic: a landscape approach
- Some more experiments
- Further Discussion

- **Introduction to stochastic optimization methods**
- Non-Convergence Issue
- Our solution: NosAdam
- Why Nostalgic: a landscape approach
- Some more experiments
- Further Discussion

- Along with the rise of deep learning, various first-order stochastic optimization methods emerged.
- The most basic one is stochastic gradient descent(SGD):

$$x_{t+1} = x_t - \alpha_t g^{(B)}(x_t),$$

B: batch size, $g^{(B)}(x_t)$: gradient estimate on one batch

- A well-known acceleration scheme is Nesterov's Acclerated Gradient method, and now people often use Momentum method.

- Later, many adaptive stochastic optimization methods emerged. Like AdaGrad(2010), RMSProp(2012), AdaDelta(2012), Adam(2014)……
- A general form of such algorithms can be written as:

$$x_{t+1} = x_t - \frac{\alpha_t}{\psi(g_1, g_2, \ldots, g_t)} \phi(g_1, g_2, \ldots g_t),$$

Adaptive learning rate

Gradient estimation

ADAM

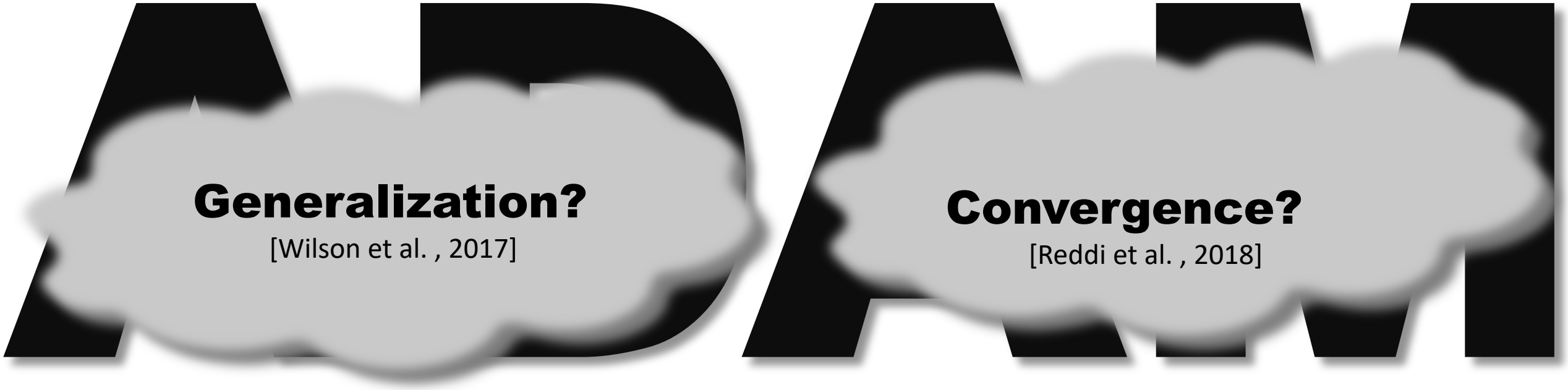
- ADAM update: $x_{t+1} = x_t - \frac{\alpha_t}{\sqrt{v_t}} m_t$,

  where $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$, $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$

  (with bias correction)

- The estimation technique of m_t and v_t is called EMA (Exponential Moving Average).

Two clouds obscured the prevalence of ADAM



“Dissecting ADAM” [Balles and Hennig, 2018] dissected ADAM into two parts: sign-based direction and variance adaption magnitude.
They pointed out that generalization is mainly determined by the sign effect rather than the adaptive learning rate, and the sign effect is problem-dependent.

- Introduction to stochastic optimization methods
- **Non-Convergence Issue**
- Our solution: NosAdam
- Why Nostalgic: a landscape approach
- Some more experiments
- Further Discussion

$$f_t(x) = \begin{cases} Cx & \text{t mod 3=1} \\ -x & \text{otherwise} \end{cases},$$ with C slightly larger than 2

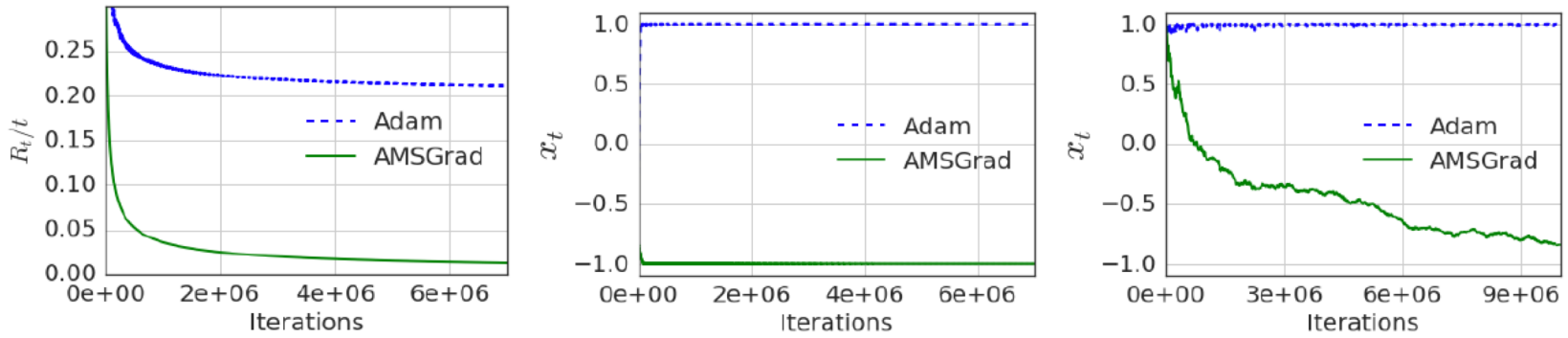


Figure 1: Performance comparison of ADAM and AMSGRAD on synthetic example on a simple one dimensional convex problem inspired by our examples of non-convergence. The first two plots (left and center) are for the online setting and the the last one (right) is for the stochastic setting.

ON THE CONVERGENCE OF ADAM AND BEYOND [Reddi et al. , 2018]

- They point out that in the convex setting, using EMA to calculate v_t will make $\Gamma_t := \frac{\sqrt{v_t}}{\alpha_t} - \frac{\sqrt{v_{t-1}}}{\alpha_{t-1}}$ not positive definite, thus making the algorithm diverge.

- Note $\Gamma_t \geq 0$ literally means “non-increasing” adaptive learning rate ($\frac{\sqrt{v_t}}{\alpha_t} \geq \frac{\sqrt{v_{t-1}}}{\alpha_{t-1}}$).

Algorithm 2 AMSGRAD

Input: $x_1 \in \mathcal{F}$, step size $\{\alpha_t\}_{t=1}^{T}$, $\{\beta_{1t}\}_{t=1}^{T}$, β_2
Set $m_0 = 0$, $v_0 = 0$ and $\hat{v}_0 = 0$
for $t = 1$ **to** T **do**
$g_t = \nabla f_t(x_t)$
$m_t = \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$
$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
$\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ and $\hat{V}_t = \mathrm{diag}(\hat{v}_t)$
$x_{t+1} = \Pi_{\mathcal{F}, \sqrt{\hat{V}_t}}(x_t - \alpha_t m_t / \sqrt{\hat{v}_t})$
end for

- AdaShift: Use g_{t-n}^2 to replace g_t^2.

Algorithm 1 AdaShift: Temporal Shifting with Block-wise Spatial Operation

Input: $n, \beta_1, \beta_2, \phi, \theta_0, \{f_t(\theta)\}_{t=1}^{T}, \{\alpha_t\}_{t=1}^{T}, \{g_{-t}\}_{t=0}^{n-1}$,

set $v_0 = 0$
for $t = 1$ **to** T **do**
  $g_t = \nabla f_t(\theta_t)$
  $m_t = \sum_{i=0}^{n-1} \beta_1^i g_{t-i} / \sum_{i=0}^{n-1} \beta_1^i$
  for $i = 1$ **to** M **do**
    $v_t[i] = \beta_2 v_{t-1}[i] + (1-\beta_2)\phi(g_{t-n}^2[i])$
    $\theta_t[i] = \theta_{t-1}[i] - \alpha_t / \sqrt{v_t[i]} \cdot m_t[i]$
  end for
end for
// We ignore the bias-correction, epsilon and other misc for the sake of clarity

- Note how this shifting strategy is similar to AMSGrad (will be discussed later).

- Introduction to stochastic optimization methods
- Non-Convergence Issue
- **Our solution: NosAdam**
- Why Nostalgic: a landscape approach
- Some more experiments
- Further Discussion

- Our motivation comes from the convergence proof.
- Recall: $\Gamma_t = \frac{\sqrt{v_t}}{\alpha_t} - \frac{\sqrt{v_{t-1}}}{\alpha_{t-1}} \succcurlyeq 0$ is a very important sufficient condition for convergence.
- Recall ADAM: $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \qquad x_{t+1} = x_t - \frac{\alpha_t}{\sqrt{v_t}} m_t$
- Without loss of generality, we let $0 \leq \beta_{2,t} = \frac{B_{t-1}}{B_t} \leq 1$, here $\{B_t\}$ is an arbitrary series, then we can prove:

Lemma 3.1. *The positive semi-definiteness of* $\frac{V_t}{\alpha_t^2} - \frac{V_{t-1}}{\alpha_{t-1}^2}$ *is satisfied if and only if* $\frac{B_t}{t}$ *is non-increasing.*

Lemma 3.1. *The positive semi-definiteness of* $\frac{V_t}{\alpha_t^2} - \frac{V_{t-1}}{\alpha_{t-1}^2}$ *is satisfied if and only if* $\frac{B_t}{t}$ *is non-increasing.*

Proof.

$$
\begin{aligned}
\frac{V_t}{\alpha_t^2} =& \frac{t}{\alpha^2} \sum_{j=1}^{t} \Pi_{k=1}^{t-j} \beta_{2,t-k+1} (1-\beta_{2,j}) g_j^2 \\
=& \frac{t}{\alpha^2} \sum_{j=1}^{t} \frac{B_{t-1}}{B_t} \cdots \frac{B_j}{B_{j+1}} \frac{B_j - B_{j-1}}{B_j} g_j^2 \\
=& \frac{t}{B_t \alpha^2} \sum_{j=1}^{t} b_j g_j^2 \geq \frac{t-1}{B_{t-1}\alpha^2} \sum_{j=1}^{t-1} b_j g_j^2 \\
=& \frac{V_{t-1}}{\alpha_{t-1}^2}
\end{aligned}
$$

□

- Again, without loss of generality, let $B_t = \sum_{j=1}^{t} b_j$, if b_j is monotonous, we can see that $\frac{B_t}{t}$ is non-increasing $\Leftrightarrow$ b_j is non-increasing

- Note that $v_t = \sum_{j=1}^{t} g_j^2 \frac{b_j}{B_t}$, we can see that the sufficient condition for convergence is that

*in the weighted average v_t , **the weights of gradients should be non-increasing** ---- Nostalgic!*

Algorithm 2 Nostalgic Adam Algorithm

Input: $x \in F, m_0 = 0, V_0 = 0$

for $t = 1$ **to** T **do**

$g_t = \nabla f_t(x_t)$

$\beta_{2,t} = B_{t-1}/B_t$, where $B_t = \sum_{k=1}^{t} b_k$ for $t \geq 1$, and $B_0 = 0$

$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$

$v_t = \beta_{2,t} v_{t-1} + (1 - \beta_{2,t}) g_t^2$, and $V_t = diag(v_t)$

$\hat{x}_{t+1} = x_t - \alpha_t m_t / \sqrt{V_t}$

$x_{t+1} = \mathcal{P}_{\mathcal{F}, \sqrt{V_t}}(\hat{x}_{t+1})$

end for

- NosAdam-HH: $b_k = k^{-\gamma}$

Theorem 3.2 (Convergence of NosAdam). *Let B_t and b_k be the sequences defined in Algorithm 1, $\alpha_t = \alpha/\sqrt{t}$, $\beta_{1,1} = \beta_1, \beta_{1,t} \leq \beta_1$ for all t. Assume that $\mathscr{F}$ has bounded diameter D_∞ and $||\nabla f_t(x)||_\infty \leq G_\infty$ for all t and $x \in \mathscr{F}$. Furthermore, let $\beta_{2,t}$ be such that the following conditions are satisfied:*

$$1. \frac{B_t}{t} \leq \frac{B_{t-1}}{t-1}$$

$$2. \frac{B_t}{tb_t^2} \geq \frac{B_{t-1}}{(t-1)b_{t-1}^2}$$

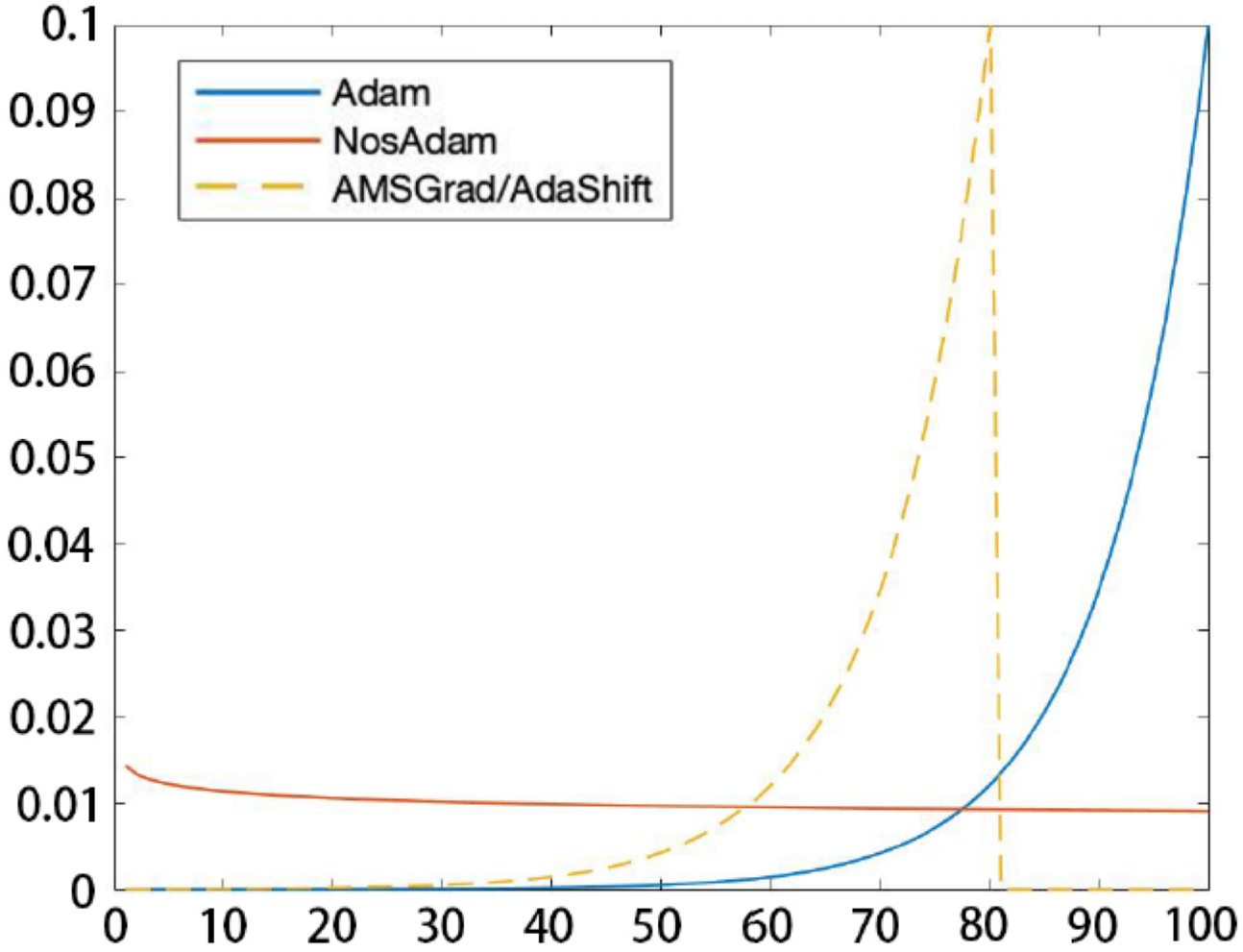
Then for $\{x_t\}$ generated using NosAdam, we have the following bound on the regret

$$R_T \leq \frac{D_\infty^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T} v_{T,i}^{\frac{1}{2}} + \frac{D_\infty^2}{2(1-\beta_1)} \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{\beta_{1,t} v_{t,i}^{\frac{1}{2}}}{\alpha_t}$$
$$+ \frac{\alpha\beta_1}{(1-\beta_1)^3} \sum_{i=1}^{d} \sqrt{\frac{B_T}{T} \frac{\sum_{t=1}^{T} b_t g_{t,i}^2}{b_T^2}}$$

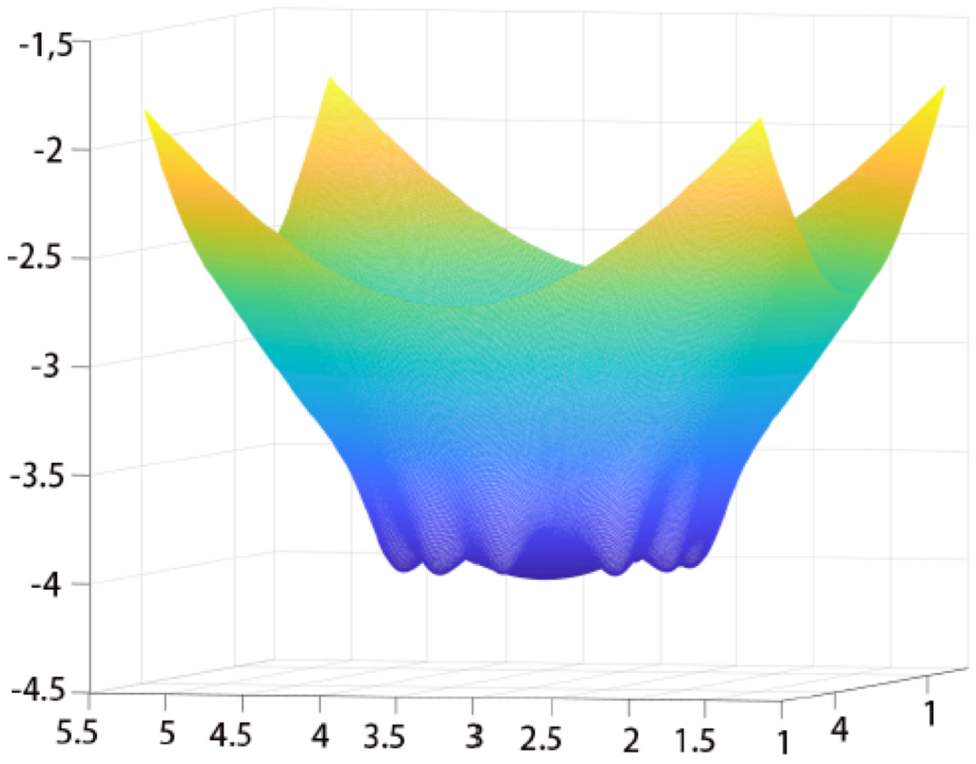
- Introduction to stochastic optimization methods
- Non-Convergence Issue
- Our solution: NosAdam
- **Why Nostalgic: a landscape approach**
- Some more experiments
- Further Discussion

- Weight Comparison:

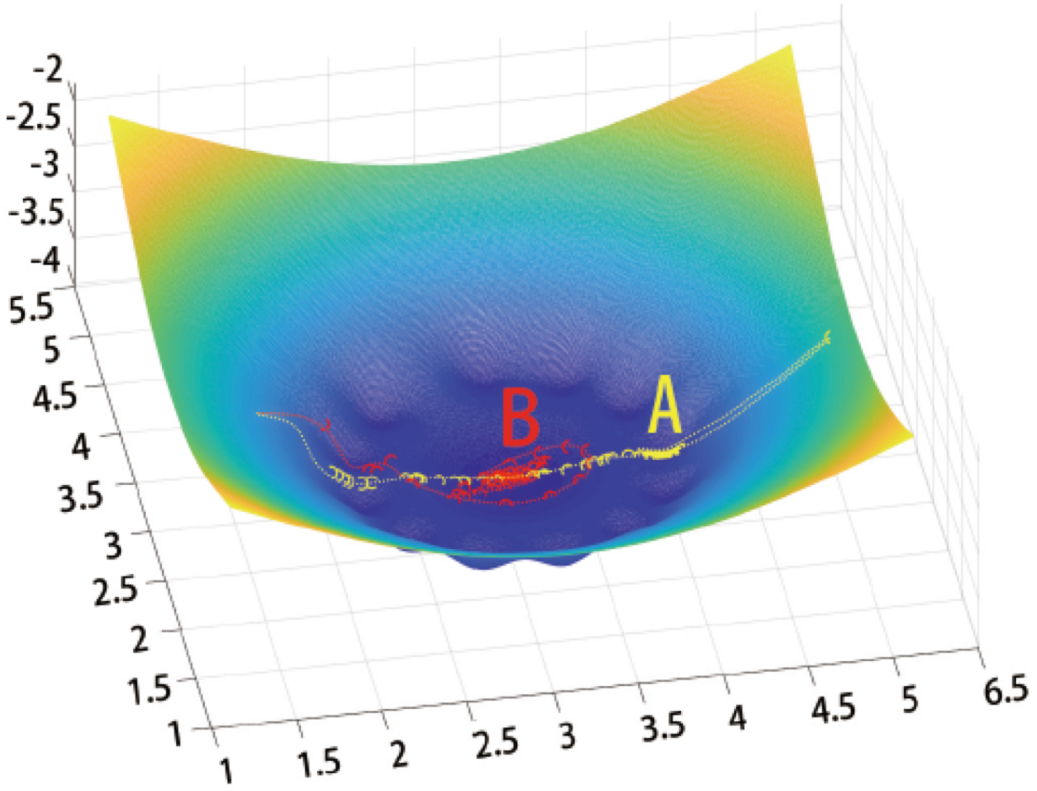
- $v_t^{(\text{Adam})} = \sum_{k=1}^{t}(1-\beta_2)\beta_2^{t-k}g_k^2$
- $v_t^{(\text{NosAdam})} = \sum_{k=1}^{t}\frac{b_k}{B_t}g_k^2$
- $v_t^{(\text{AMSGrad})} = v_{t-n(t)}^{(\text{Adam})}$
- $v_t^{(\text{AdaShift})} = v_{t-n}^{(\text{Adam})}$



- Flat global minima, shallow local minimas



(a) Bowl-shaped Landscape



(b) Trajectories of NosAdam (red) and Adam (yellow)

- “max” operation in AMSGrad can be problematic, since it is vulnerable to large gradient

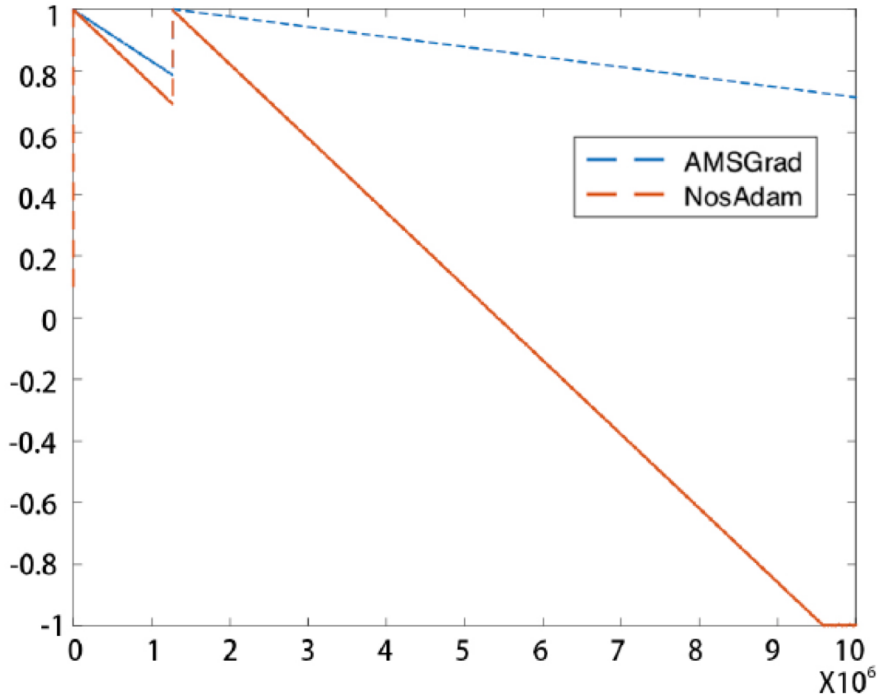
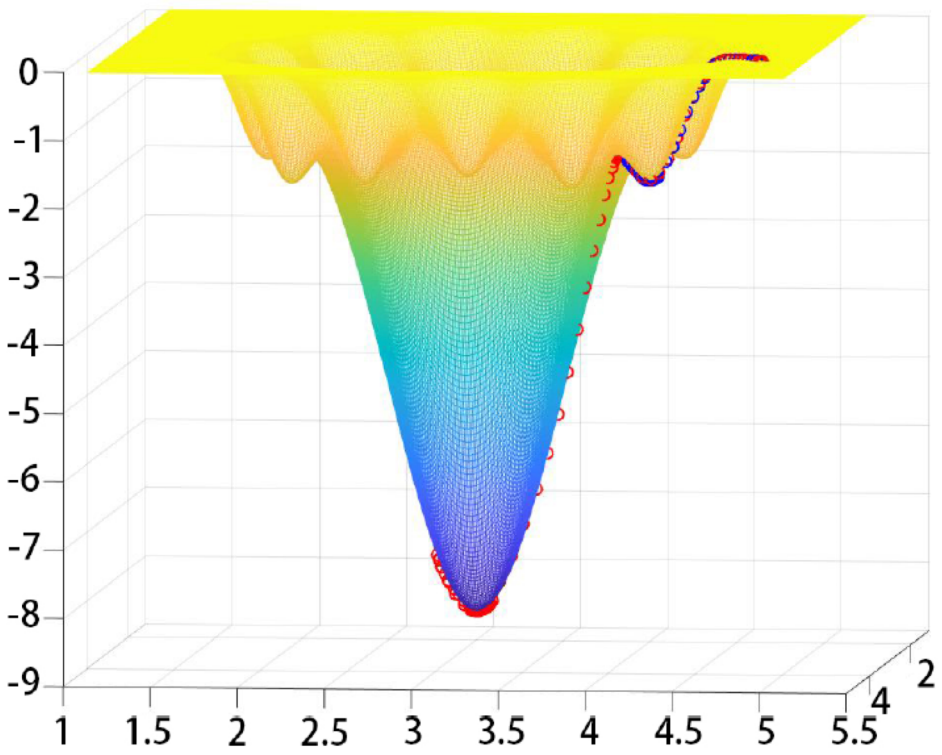


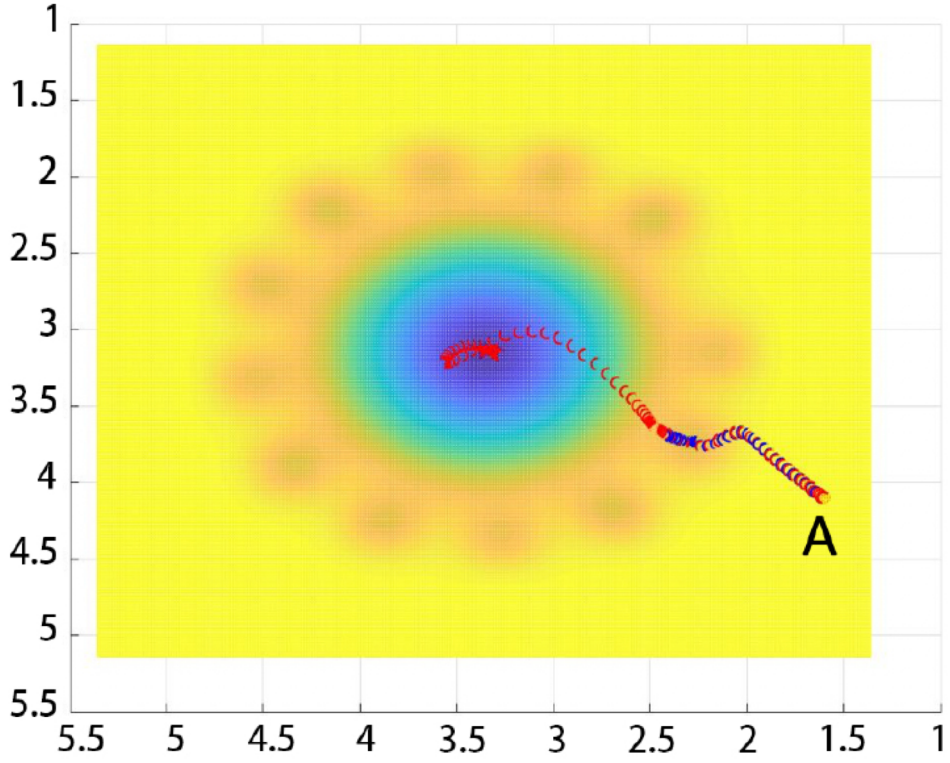
Figure 3: Appearance of a large gradient at around 10^6 step. The y-axis shows the value x, and the x-axis shows the number of iterations. The figure shows AMSGrad is greatly slowed down after encountering a large gradient.

Algorithm 2 AMSGRAD

Input: $x_1 \in \mathcal{F}$, step size $\{\alpha_t\}_{t=1}^{T}$, $\{\beta_{1t}\}_{t=1}^{T}$, β_2
Set $m_0 = 0$, $v_0 = 0$ and $\hat{v}_0 = 0$
for $t = 1$ **to** T **do**
$g_t = \nabla f_t(x_t)$
$m_t = \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$
$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
$\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ and $\hat{V}_t = \mathrm{diag}(\hat{v}_t)$
$x_{t+1} = \Pi_{\mathcal{F}, \sqrt{\hat{V}_t}}(x_t - \alpha_t m_t / \sqrt{\hat{v}_t})$
end for

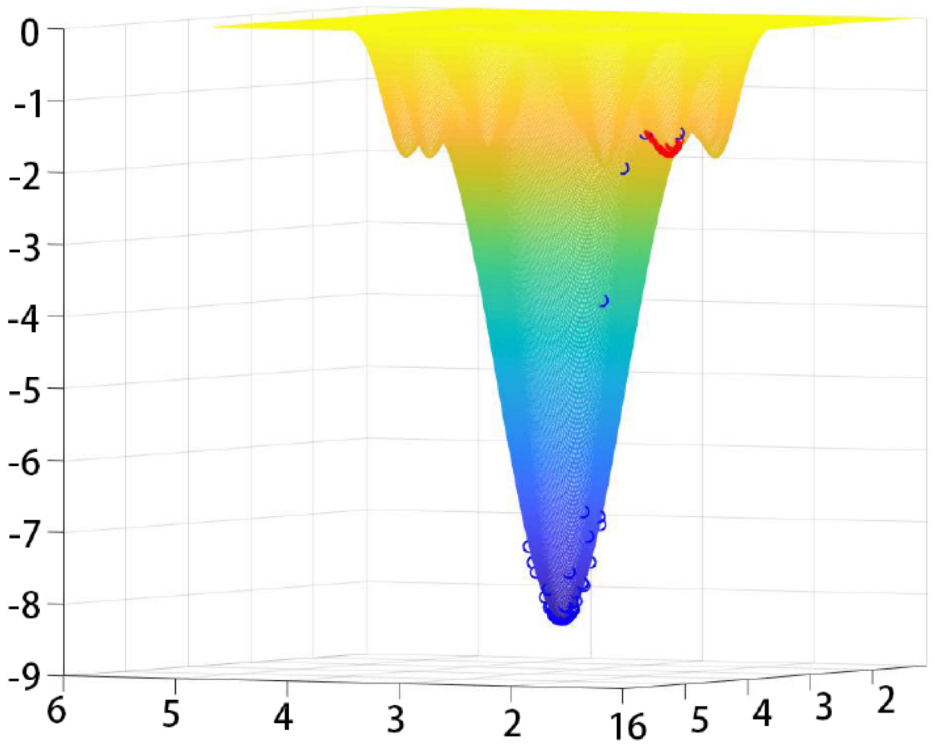


(a) Sharper Minima

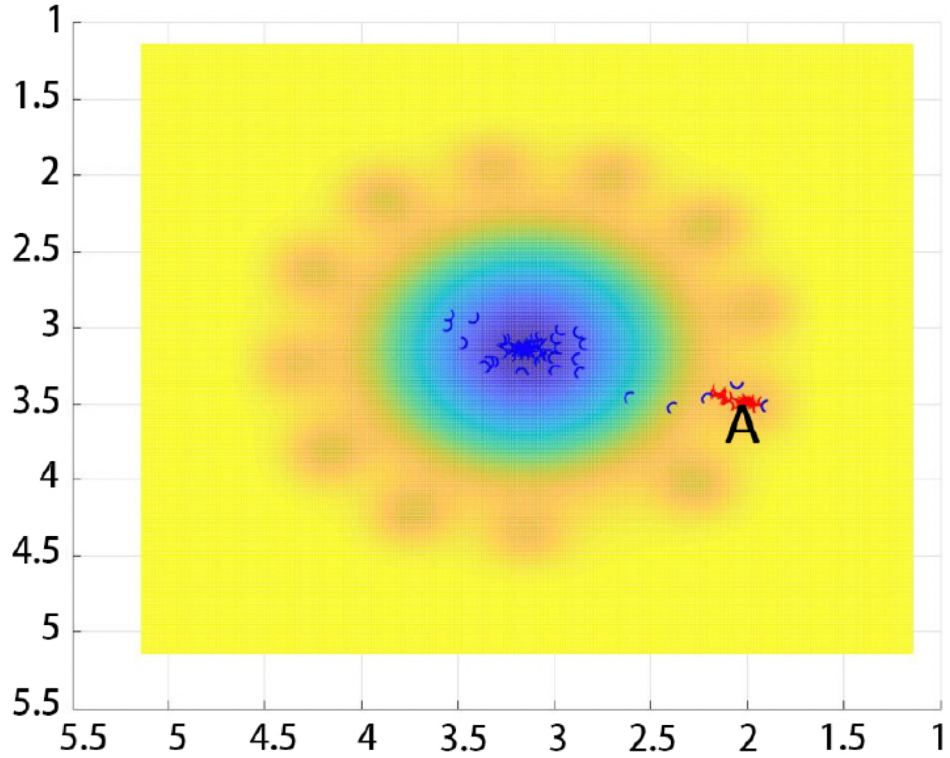


(b) Trajectories of AMSGrad (blue) and NosAdam (red)

- NosAdam can perform poorly when initialized close to a sharp valley

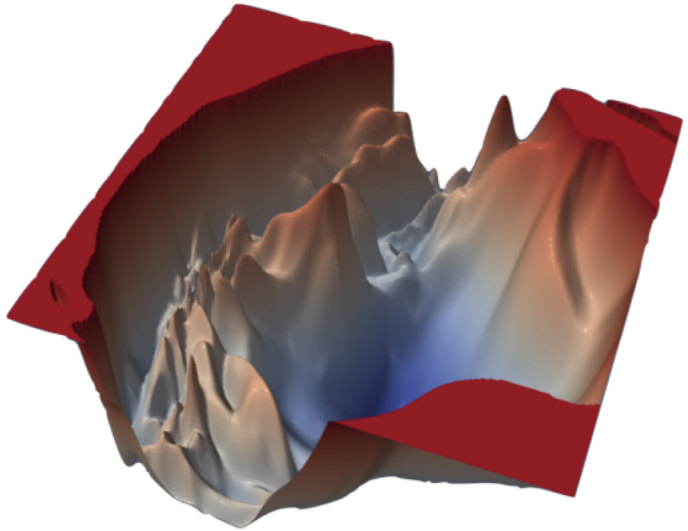


(a) Sharper Minima

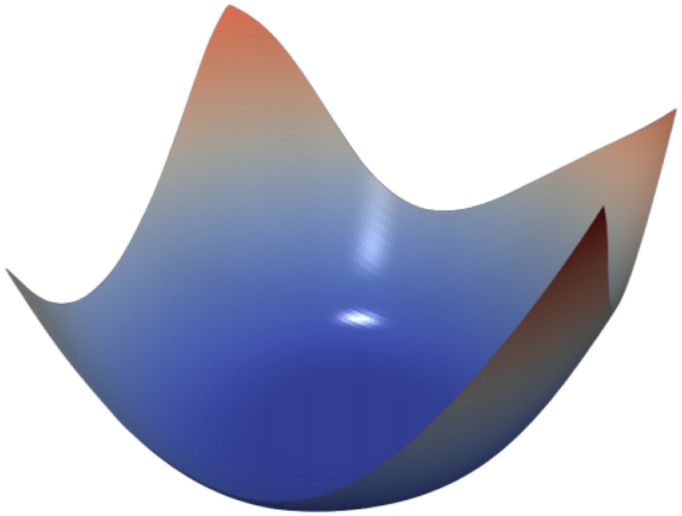


(b) Trajectories of Adam (blue) and NosAdam (red)

- Visualizing the loss landscapes of neural nets, Li, et al

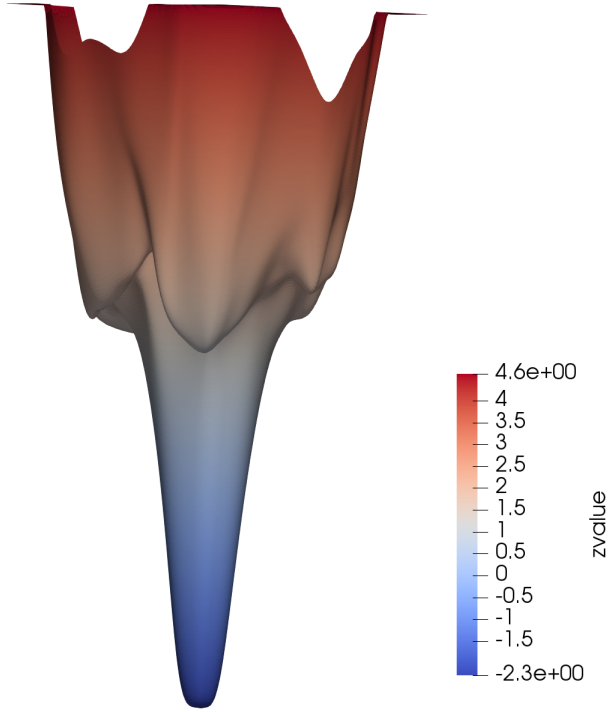


(a) ResNet-110, no skip connections



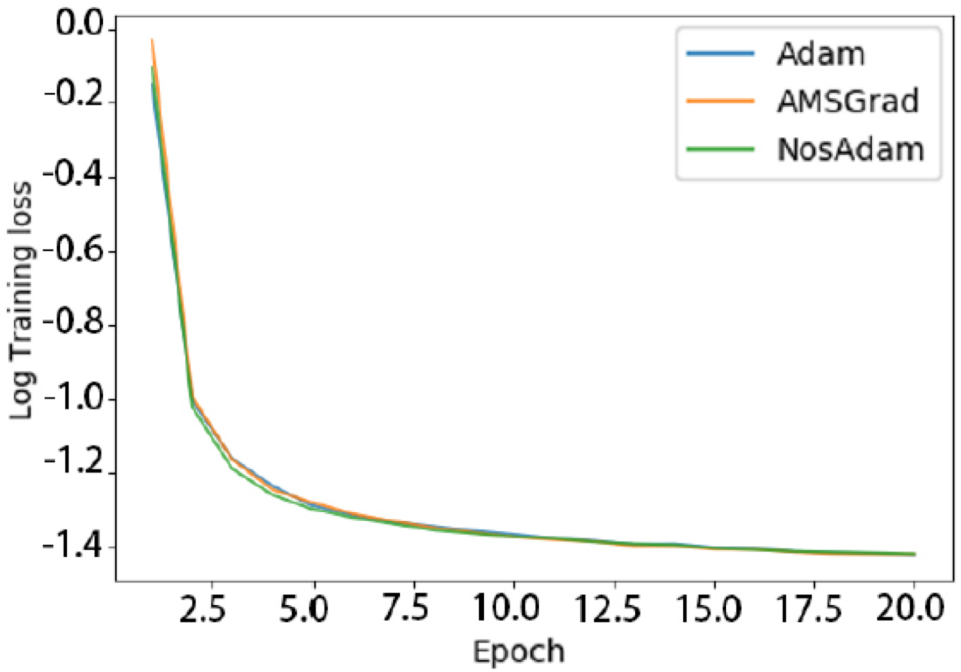
(b) DenseNet, 121 layers

Figure 4: The loss surfaces of ResNet-110-noshort and DenseNet for CIFAR-10.

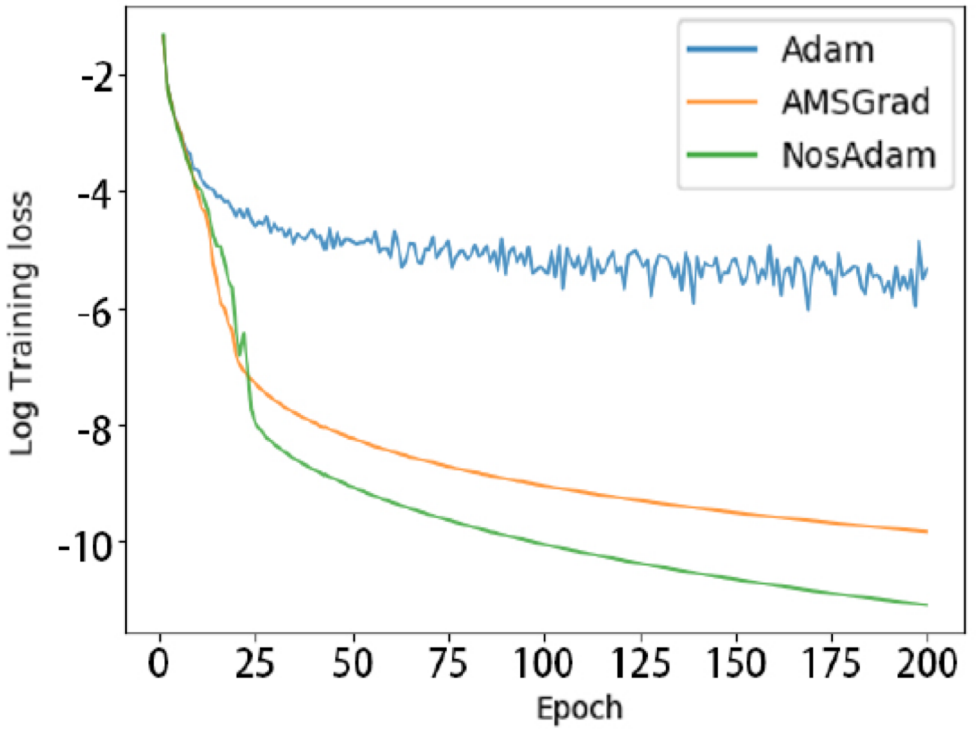


ResNet56 on CIFAR-10

- Introduction to stochastic optimization methods
- Non-Convergence Issue
- Our solution: NosAdam
- Why Nostalgic: a landscape approach
- **Some more experiments**
- Further Discussion

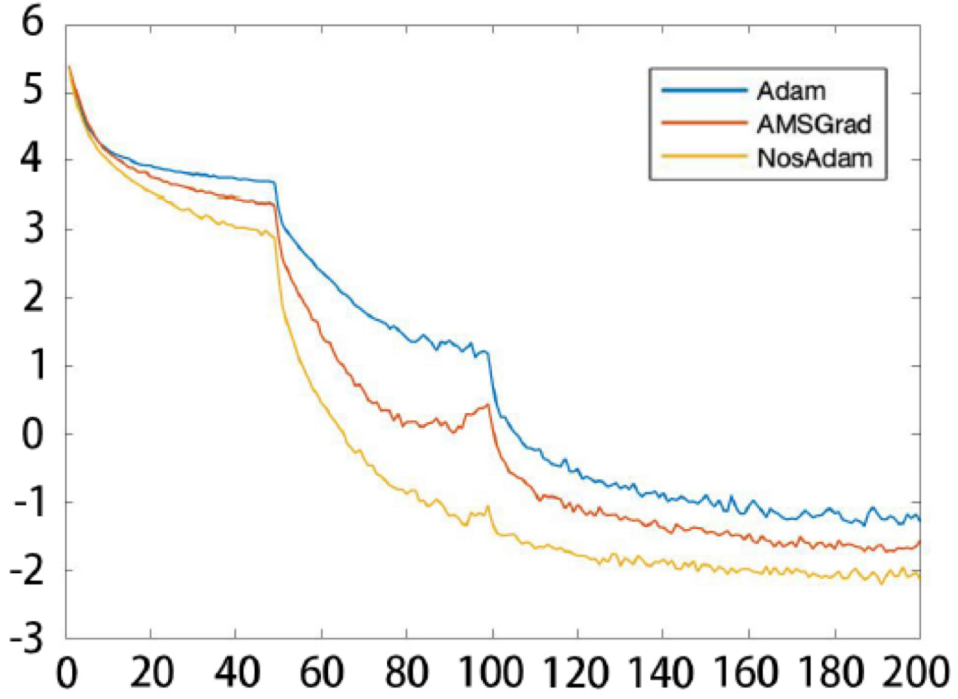


(a) Logistic Regression

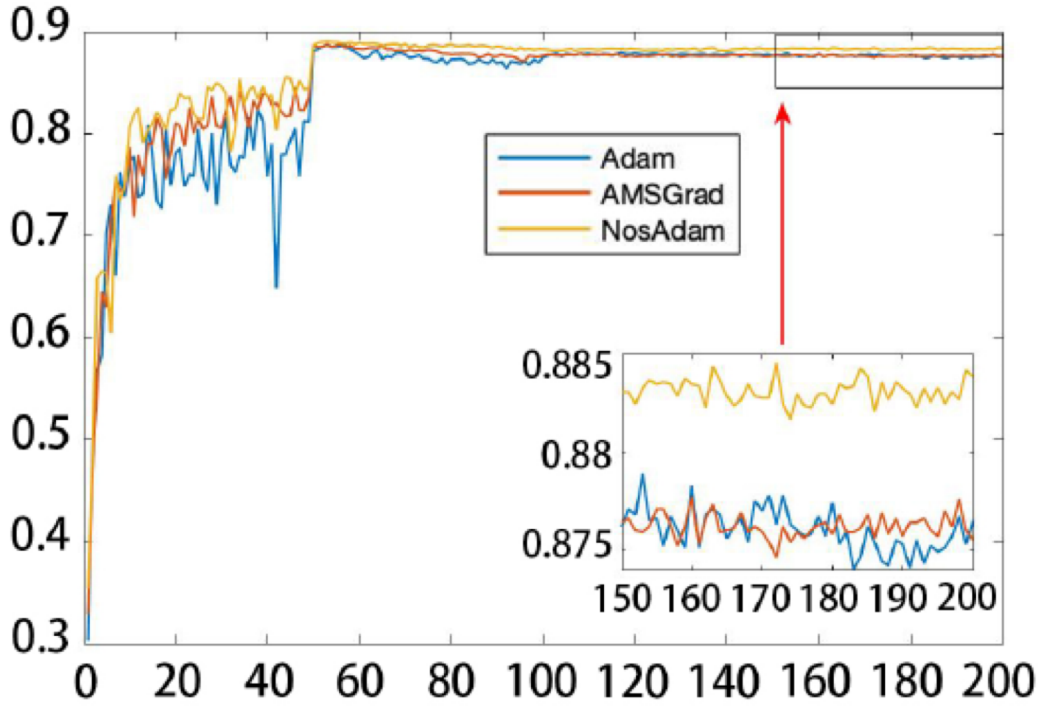


(b) Multi-layer Fully Connected Neural Network

Figure 6: Experiments of logistic regression and multi-layer fully connected neural network on MNIST.



(a) Log Training Loss



(b) Test Accuracy

Figure 7: Experiments of Wide ResNet on CIFAR-10.

- Introduction to stochastic optimization methods
- Non-Convergence Issue
- Our solution: NosAdam
- Why Nostalgic: a landscape approach
- Some more experiments
- **Further Discussion (if time permits)**

Role of v_t ?

Role of v_t ?

- Dissecting ADAM assumes v_t to be second moment estimate of $E[g_t^2]$.
- This is questionable:
- Padam: $v_{\mathrm{t}} = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \;\; x_{t+1} = x_t \; - \frac{\alpha_t}{v_t^{1/p}} m_t$
- We also prove:

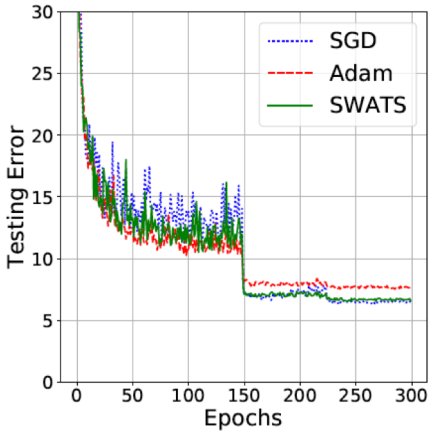
$$v_{\mathrm{t}}^{nos} = \beta_{2,t} v_{t-1} + \left(1-\beta_{2,t}\right) g_t^p, \qquad x_{t+1} = x_t \; - \frac{\alpha_t}{v_t^{1/p}} m_t$$

Role of v_t ?

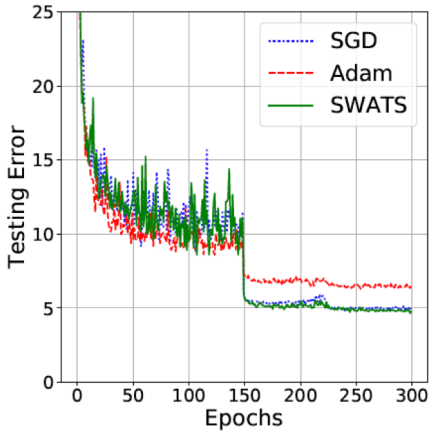
- AdaGrad: *"the adaptation allows us to find needles in haystacks in the form of very predictive but rarely seen features."*
- v_t is balancing the update speeds of different features according to their abundance in the data set ?

Role of v_t ?

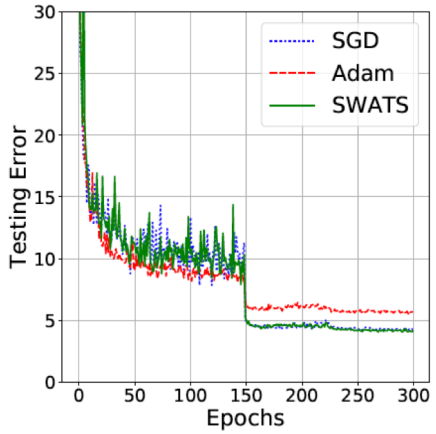
- SWATS: uses Adam for earlier epochs and then fix the re-scaling term v_t for later epochs.
- Does there exist an optimal re-scaling term (learning rate)? (Once achieved, use forever)



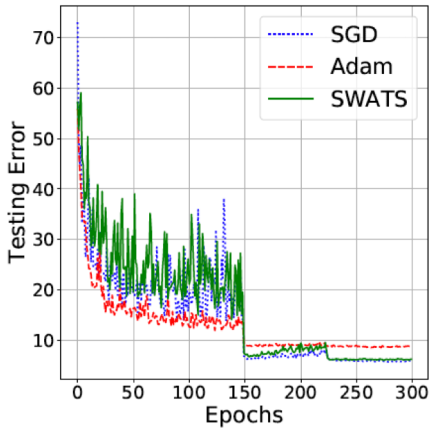
(a) ResNet-32 — CIFAR-10



(b) DenseNet — CIFAR-10



(c) PyramidNet — CIFAR-10



(d) SENet — CIFAR-10

Take-home Message!

- It's a good idea to *weight more of the past gradients* when designing the calculation of v_t (both theoretically and empirically)
- Understanding optimizers based on landscape investigation is an interesting direction.
- The understanding of the mysterious role of v_t can be a next breakthrough.

Thanks for Listening!
Questions?