Adam

John Gay Bill Huang Yushen Huang Alex Mankowski

December 3, 2020

Table of Contents

1. Introduction
2. Adam Analysis
3. Experiments

Table of Contents

1 Introduction

2 Adam Analysis

3 Experiments

Adam

- Adam [2] is a stochastic optimization algorithm using only first order gradients
- Adam: Adaptive moment estimation
- Published in 2015
- Authors: Diederik Kingma and Jimmy Ba
- Draws from AdaGrad [1] (2011) and RMSProp (2012)

Preliminary Knowledge: Regret Analysis

- Regret is the difference in payoff between the action taken and the other available actions
- The action with the highest payoff is the one that should have been taken (optimal w^*)
- We only care about positive regret
- Negative or zero regret means we took a better action than the one in comparison

Regret Formally

- Inputs:
  - Loss function $\ell(w)$
  - weights w
  - optimal weights w^*
- Output: Regret R

$$R = \ell(w) - \ell(w^*)$$

Problem Address

The Adam method is mainly a method which deals with the smooth convex stochastic optimization problem:

$$\min_{\theta \in \Theta} f(\theta) := \mathbb{E}[F(\theta, \boldsymbol{\xi})]$$

where $F(., \boldsymbol{\xi})$ is a smooth convex function, $F(\theta, .)$ is a measurable function, the first moment always exists, and Θ is a convex and compact set. This regulation will guarantee the the function $f(x)$ is well defined and the $f(x)$ is convex and smooth. In practice the random variable θ is usually the data we generate. (ie, $\theta = (x, y)$). We note that it is usually expensive to calculate the expectation of this problem, hence we usually do not have an explicit expression of $f(\theta)$.

Adam's Update

Recall that the Adam's performs the following update:

- $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
- $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
- $\hat{m}_t = \frac{m_t}{(1-\beta_1^t)}$, and $\hat{v}_t = \frac{\hat{v}_t}{(1-\beta_2^t)}$
- $\theta_t = \theta_{t-1} - \alpha_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$

Table of Contents

1. Introduction

2. Adam Analysis

3. Experiments

Adam Main Convergence Result

Now suppose $\{\theta_t\}_{t=1}^{\infty}$ is the weight generated by the Adam method and we make the following assumption: Assume that $F(\theta, \boldsymbol{\xi})$ has uniformly bounded gradient: $\|\nabla_\theta F(\theta, \xi)\|_2 \leq G$, because in $\mathbb{R}^n$, the L^2 norm equivalent to L^∞, we can also assume $\|\nabla_\theta F(\theta, \xi)\|_\infty \leq G_\infty$. Additionally, the distance between any $\{\theta_t\}_{t=1}^{\infty}$ is bounded, $\|\theta_n - \theta_m\|_2 \leq D$ and $\|\theta_n - \theta_m\|_\infty \leq D_\infty$ for any integer n and m. We also assume that $\beta_1, \beta_2 \in [0, 1)$ satisfy $\frac{\beta_1^2}{\beta_2} < 1$.Let $\alpha_t = \frac{\alpha}{\sqrt{t}}$ and $\beta_{1,t} = \beta_1 \lambda^{t-1}$, $\lambda \in (0, 1)$. We will have the following result:

$$\frac{R(T)}{T} = O(\frac{1}{\sqrt{T}})$$

where the regret $R(T)$ is defined by:

$$R(T) = \sum_{t=1}^{T} F(\theta_t, \xi_t) - F(\theta^*, \xi_t)$$

where $\theta^* = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} F(\theta, \xi_t)$

Intuition of the Result

- We notice that the main result of Adam actually is not done yet. Recall that we need to show that $\mathbb{E}[f(\theta_t) - f(\theta^*)] \to 0$ and the reason why we take the expectation is that θ_t is random... Before we use the main result of the Adam Paper to show the convergence of the method, let's give an intuitive explanation of those result and why Adam is a good method.
- In computer science perspective,the regret can actually be understood as a dynamical programming with some randomness,because at each time step you will have a scenario ξ_t, we will make a decision θ_t and compare with the optimal fixed decision.
- In a math perspective, you can understand regret as a stochastic process and note that you can make a decision based on the previous information and those previous information can be understood as $\sigma - Algebra$ the and the decision you make is somehow similar to calculate the conditional expectation...

Intuition of the Result

- The reason why the Adam has faster than SGD convergence is that it is an adaptive version of SGD and it contains some second order information or the curvature information. Additionally, it does not require a large memory footprint which makes it faster than the stochastic gradient descent.
- We should also notice that : $\theta^* = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} F(\theta, \xi_t) = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} F(\theta, \xi_t)$ and we notice that by Strong Law of large number $\frac{1}{T} \sum_{t=1}^{T} F(\theta, \xi_t) \to \mathbb{E}(F(\theta, \xi)$ as $T \to \infty$ when $\xi_1, \cdots, \xi_T$ is iid random variable. hence $\frac{R(T)}{T}$ can be approximated by:

$$\frac{R(T)}{T} \approx \frac{1}{T} \sum_{t=1}^{T} (F(\theta_t, \xi_t) - \mathbb{E}[F(\theta^*, \xi)])$$

Hence intuitively we should expect the convergence of regret implies the convergence of the Adam.

From Regret to Convergence

- The original Adam paper conducts theoretical analysis through the online learning framework. Adam achieves an asymptotic regret of $R(T) = O(\sqrt{T})$.
- **Question: what does this mean in terms of Adam's convergence?**
- The framework for convergence analysis: let $\{(x_i, y_i)\}$ denote our data, $l(x, y)$ be the loss function, $h(\cdot, \theta)$ be our model parametrized by θ. We seek to solve the optimization problem:

$$\theta^* = \arg\min_{\theta} f(\theta) = \arg\min_{\theta} \mathbb{E}_{(x,y)}[l(h(x, \theta), y)]$$

- Recall from class: we have the following convergence result for SGD, where B is the bound on the variance of f's gradient, and μ is the coefficient for f's strong convexity.

$$\mathbb{E}[f(\bar{\theta}) - f(\theta^*)] \leq \frac{B^2}{2\mu^2 T} \log(T), \quad \bar{\theta} = \frac{1}{T} \sum_{t=1}^{T} \theta_{t-1}$$

- So if we let $T \to \infty$, the expected error can be made as small as desired.

Adam's Convergence

- Let $h_t(x) = h(x_t, \theta_t)$. Let $l_t(h_t) = l(h(x_t, \theta_t), y_t)$. Assume l_t are i.i.d from D, define $L(h, D) = \mathbb{E}_{l_t \sim D} l_t(h)$ for any $t = 1, ..., T$.
- For Adam, the following is true [1]:

$$\frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} L(h_t, D)] \leq \min_{h' \in H} \mathbb{E}[L(h', D)] + \frac{R(T)}{T} \qquad (*)$$

- Using (*), we now show Adam converges to the expected optimal value at a rate of $O(\frac{1}{\sqrt{T}})$.

Adam's Convergence, cont.

- Denote $\bar{h} = \frac{1}{T}\sum_{t=1}^{T} h_t, \bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$ as the averaged model and parameters. Denote $H_t = \{h_1, l_1, ..., h_{t-1}, l_{t-1}, h_t\}$ to be the entire history up to iterate t. By Jensen's inequality, we have from (*):

$$LHS : \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} L(h_t, D)] \geq \mathbb{E}[L(\bar{h}, D)] = \mathbb{E}_{H_t}\mathbb{E}_{(x_t,y_t)}[l(h(x_t, \bar{\theta}), y_t)] = \mathbb{E}_{H_t}[f(\bar{\theta})]$$

$$RHS : \min_{h' \in H} \mathbb{E}[L(h', D)] + \frac{R(T)}{T} = \min_{\theta} f(\theta) + \frac{R(T)}{T} = f(\theta^*) + \frac{R(T)}{T}$$

- Putting it together, we have:

$$\mathbb{E}_{H_t}[f(\bar{\theta})] \leq f(\theta^*) + \frac{R(T)}{T}$$
$$\mathbb{E}_{H_t}[f(\bar{\theta}) - f(\theta^*)] \leq \frac{R(T)}{T} = O(\frac{1}{\sqrt{T}})$$

- Therefore, Adam converges to the expected optimal value at a rate of $O(\frac{1}{\sqrt{T}})$.

Table of Contents

1 Introduction

2 Adam Analysis

3 Experiments

Experiment Setup

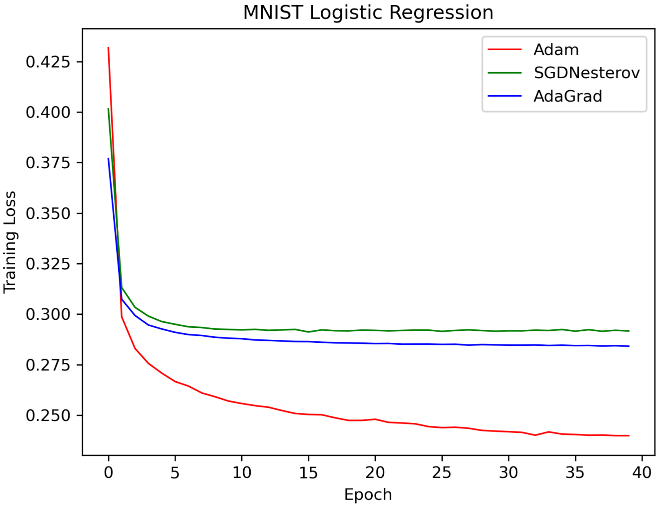
- Three experiments from the paper, one new application
- Comparing:
    - Adam
    - SGD with Nesterov Momentum
    - AdaGrad
- Optimize over the learning rates
- We only observe the training loss, however regularization is still used

MNIST Logistic Regression

- MNIST digits data [1]
- Single hidden layer neural network, with softmax activations
- Negative log-likelihood loss

[1] http://yann.lecun.com/exdb/mnist/

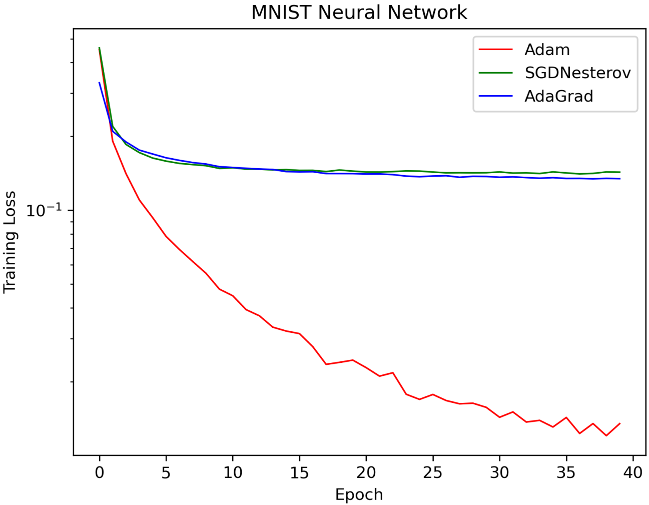
MNIST Logistic Regression



MNIST Neural Network

- Network with two hidden layers
- 1000 neurons for each layer
- Dropout is applied after each layer

MNIST Neural Network

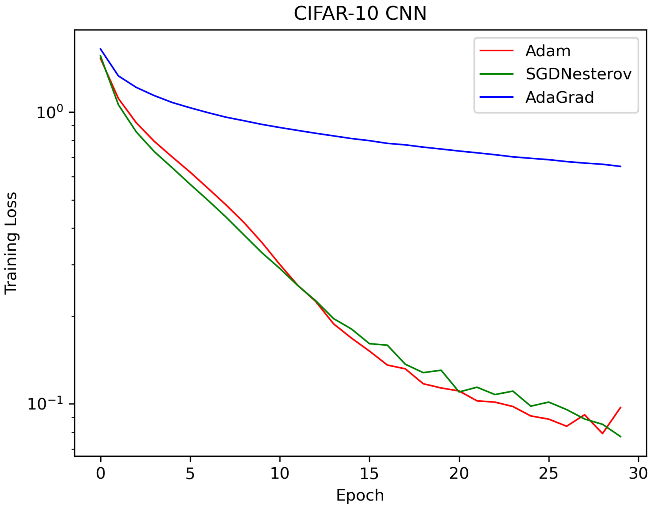


Cifar-10 CNN

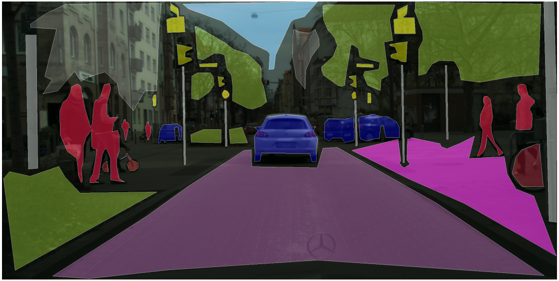
- 50,000 training 32x32 color images with 10 classes [2]
  - Airplanes, frogs, automobiles, and more!
- (5x5 convolutions → 3x3 max pooling)*3 → fully connected
- Dropout is applied to the input and fully connected layers

[2] https://www.cs.toronto.edu/ kriz/cifar.html

Cifar-10 CNN


CIFAR-10 CNN
Adam
SGDNesterov
AdaGrad
Training Loss
10^0
10^{-1}
0
5
10
15
20
25
30
Epoch

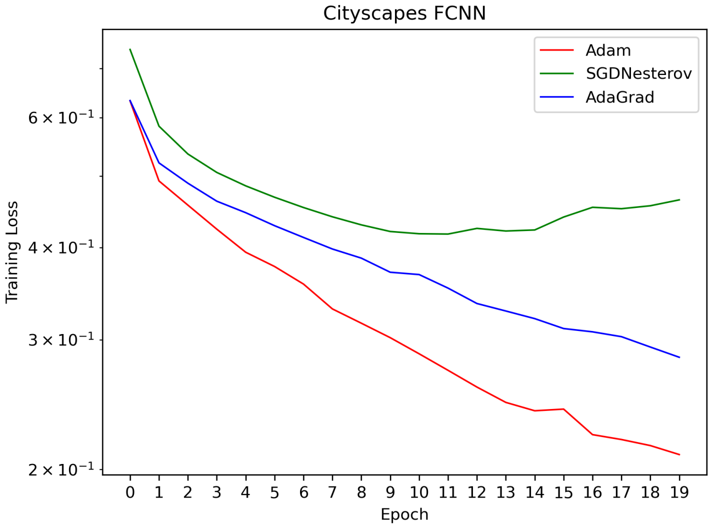
Cityscape FCNN



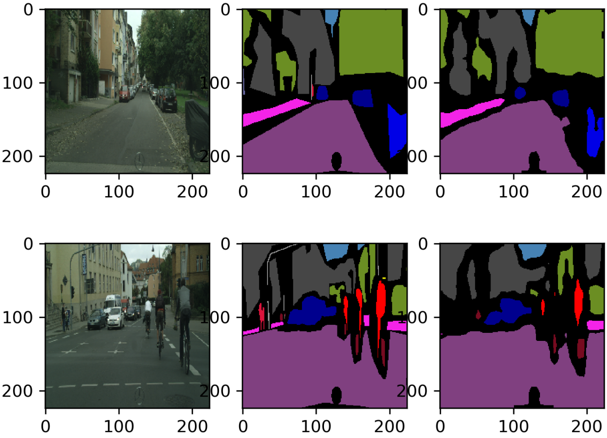
- Coarse semantic segmentation, 20 classes, 2975 images [3]
- Fully convolutional neural network [3]
  - Resnet-50 backbone, fully-connected layers replaced with upsampling

[3] https://www.cityscapes-dataset.com/

Cityscapes FCNN



Cityscapes FCNN



References I

Duchi, John, Hazan, Elad, and Singer, Yoram.
Adaptive subgradient methods for online learning and stochastic optimization.
The Journal of Machine Learning Research, 12:2121–2159, 2011.

D. P. Kingma and J. L. Ba
Adam: A Method for stochastic Optimization.
San Diego: The International Conference on Learning Representations (ICLR), 2015.

E. Shelhamer, J. Long and T. Darrell
Fully Convolutional Networks for Semantic Segmentation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, 1 April 2017

References II

C. Zhang, “Csc 665: Online to batch conversion,” December 17 2019.