Variants of RMSProp and Adagrad with Logarithmic Regret Bounds

Mahesh Chandra Mukkamala[1,2], Matthias Hein[1].

[1]Saarland University, [2] Max Planck Institute for Informatics.
Saarbrücken, Germany.

Contributions

Motivation:

- Use **RMSProp** (Hinton et al., 2012) in Online Learning.
- Train deep nets with optimal algorithms for strongly convex problems.

Main Contributions:

- **Variant of RMSProp** for online convex problems.
- **Equivalence of RMSProp and Adagrad** (Duchi et al. [2010]).
- Proposed **SC-Adagrad** and **SC-RMSProp** with $\log T$-type regret bounds (Hazan et al. [2007]) for strongly convex problems.
- Better test accuracy on various **Deep Nets**.

Online convex optimization

Let C be a convex set.

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

- We predict action $\theta_t \in C$.
  Note: Like weights of your model.
- Adversary chooses $f_t : C \to \mathbb{R}$ (continuous convex loss).
  Note: Like obtaining a sample from a training dataset.
- We suffer loss $f_t(\theta_t)$.
- We update θ_t, using (sub)-gradient $g_t \in \partial f_t(\theta_t)$.

Online convex optimization (Contd...)

- **Cumulative loss:** $\sum_{t=1}^{T} f_t(\theta_t)$.
- Best possible prediction in hindsight:

$$\theta^* = \arg\min_{\theta \in C} \sum_{t=1}^{T} f_t(\theta).$$

- The adversarial regret at time $T \in \mathbb{N}$:

$$R(T) = \sum_{t=1}^{T} (f_t(\theta_t) - f_t(\theta^*)).$$

Goal: Obtain bounds on $R(T)$ (to perform well w.r.t. θ^*) .

Overview of Algorithms

Online (Projected) Gradient Descent (OGD):

$$\theta_{t+1} = P_C\left(\theta_t - \alpha_t g_t\right)$$

Optimal Regret Bounds:

Convex Problems: $O(\sqrt{T})$	**Strongly Convex Problems:** $O(\log T)$
OGD with $\alpha_t = \frac{\alpha}{\sqrt{t}}$ (Zinkevich [2003])	OGD with $\alpha_t = \frac{\alpha}{t}$ (Hazan et al. [2007])
Adagrad (Duchi et al. [2011])	SC-Adagrad (Ours)
RMSProp (Ours)	SC-RMSProp (Ours)

Notation

- Let A be a symmetric, positive definite matrix.
- **Weighted projection** of z onto a convex set $C \subset \mathbb{R}^d$ is

$$P_C^A(z) = \arg\min_{x \in C} \left\{ \|x - z\|_A^2 = \langle x - z, A(x - z) \rangle \right\}$$

- $(a \odot b)_i = a_i b_i \; \forall i \in [d], \forall a, b \in \mathbb{R}^d$.
- Note: $\mathbf{0} \in \mathbb{R}^d$

Adagrad vs SC-Adagrad

Adagrad

Input: $\theta_1 \in C, \delta > 0, v_0 = \mathbf{0}$
for $t = 1$ **to** T **do**
$g_t \in \partial f_t(\theta_t)$
$v_t = v_{t-1} + (g_t \odot g_t)$
$A_t = \mathrm{diag}(\sqrt{v_t}) + \delta \mathbb{I}$
$\theta_{t+1} = P_C^{A_t}\left(\theta_t - \alpha A_t^{-1} g_t\right)$
end for

- Effective step-size is $O\left(\frac{1}{\sqrt{t}}\right)$.
- Proposed in Duchi et al. [2010] with $\sqrt{T}$-type regret bounds for **convex problems**.
- One of the most popular algorithms to train deep neural networks.
- $\delta = 10^{-8}$ for numerical stability.

SC-Adagrad

Input: $\theta_1 \in C, \delta_0 > \mathbf{0}, v_0 = \mathbf{0}$
for $t = 1$ **to** T **do**
$g_t \in \partial f_t(\theta_t)$
$v_t = v_{t-1} + (g_t \odot g_t)$
Set $\mathbf{0} < \delta_t \leq \delta_{t-1}$ element wise.
$A_t = \mathrm{diag}(v_t) + \mathrm{diag}(\delta_t)$
$\theta_{t+1} = P_C^{A_t}\left(\theta_t - \alpha A_t^{-1} g_t\right)$
end for

- Effective step-size is $O\left(\frac{1}{t}\right)$.
- Duchi et al. [2010] $(\delta_t = \delta)$ with $\log T$-type regret bounds for **strongly convex (SC) problems**.
- Regret bounds: $\delta_t \in \mathbb{R}^d$, $\delta_t \leq \delta_{t-1}$?
- Better test accuracy on deep nets, δ_t should start large, decay with v_t.

ζ-strongly convex function

Definition: Let C be a convex set. We say $f : C \to \mathbb{R}$ is **ζ-strongly convex**, if there exists $\zeta > 0$ such that for all $x, y \in C$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \zeta \|y - x\|^2$$

Quadratic lower bound of the function.

Logarithmic Regret bounds

Theorem: If

- $f_t : C \to \mathbb{R}$ is ζ-strongly convex function ($\zeta > 0$) with $g_t \in \partial f_t(\theta_t)$.
- $\sup_{t\geq 1} \|g_t\|_\infty \leq G_\infty$, $\sup_{t\geq 1} \|\theta_t - \theta^*\|_\infty \leq D_\infty$.
- $\alpha \geq \frac{G_\infty^2}{2\zeta}$.

then regret bound of SC-Adagrad for $T \geq 1$ is

$$R(T) \leq \frac{D_\infty^2 \operatorname{tr}(\operatorname{diag}(\delta_1))}{2\alpha} + \frac{\alpha}{2}\sum_{i=1}^{d} \log\left(\frac{v_{T,i} + \delta_{T,i}}{\delta_{1,i}}\right) + \frac{1}{2}\sum_{i=1}^{d} \inf_{t\in\{1,\ldots,T\}} \left(\frac{(\theta_{t,i} - \theta_i^*)^2}{\alpha} - \frac{\alpha}{v_{t,i} + \delta_{t,i}}\right)(\delta_{T,i} - \delta_{1,i})$$

Data-dependent logarithmic regret bounds for SC problems.

RMSProp (Hinton et al., 2012)

Most popular adaptive gradient method used in deep learning.

Idea: Moving average of second order gradients.

Can we use RMSProp for Online learning?

RMSProp vs SC-RMSProp

RMSProp (Ours)

$\theta_1 \in C, \delta > 0, v_0 = \mathbf{0}, \alpha > 0, 0 < \gamma \leq 1$

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

$g_t \in \partial f_t(\theta_t)$

$\beta_t = 1 - \frac{\gamma}{t}$

$v_t = \beta_t v_{t-1} + (1 - \beta_t)(g_t \odot g_t)$

Set $\epsilon_t = \frac{\delta}{\sqrt{t}}$ and $\alpha_t = \frac{\alpha}{\sqrt{t}}$

$A_t = \mathrm{diag}(\sqrt{v_t}) + \epsilon_t I$

$\theta_{t+1} = P_C^{A_t}\left(\theta_t - \alpha_t A_t^{-1} g_t\right)$

end for

SC-RMSProp

$\theta_1 \in C, \delta_0 > \mathbf{0}, v_0 = \mathbf{0}, \alpha > 0, 0 < \gamma \leq 1$

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

$g_t \in \partial f_t(\theta_t)$

$\beta_t = 1 - \frac{\gamma}{t}$

$v_t = \beta_t v_{t-1} + (1 - \beta_t)(g_t \odot g_t)$

Set $\mathbf{0} < \delta_t \leq \delta_{t-1}$ element wise.

Set $\epsilon_t = \frac{\delta_t}{t}$ and $\alpha_t = \frac{\alpha}{t}$

$A_t = \mathrm{diag}(v_t + \epsilon_t)$

$\theta_{t+1} = P_C^{A_t}\left(\theta_t - \alpha_t A_t^{-1} g_t\right)$

end for

- Original RMSProp with $\beta_t = 0.9, \alpha_t = \alpha > 0, \epsilon_t = \delta > 0$.
- Achieves $\sqrt{T}$-type regret bounds for **convex problems**.

- Effective step-size is $O\left(\frac{1}{t}\right)$.
- Achieves $\log T$-type regret bounds for **strongly convex problems**.

Interesting Phenomenon

Choose $\beta_t = 1 - \frac{1}{t}$ we have

RMSProp (Ours) $\equiv$ Adagrad

SC-RMSProp $\equiv$ SC-Adagrad

Example: Decay Scheme

Choose $\xi_1, \xi_2 > 0$

SC-Adagrad: $\delta_t = \xi_2 e^{-\xi_1 \, v_t}$, **SC-RMSProp:** $\delta_t = \xi_2 e^{-\xi_1 \, t \, v_t}$

Pros:

- Enhanced adaptivity as $\delta_t \in \mathbb{R}^d$.
- Stabilizes the quadratic growth of v_t in g_t.
- Exponential decay in v_t.

Rule of Thumb:

- $\xi_1 = 0.1, \xi_2 = 1$ for convex problems.
- $\xi_1 = 0.1, \xi_2 = 0.1$ for **deep learning**.

Open question: What is the optimal decay scheme?

Experimental Setup

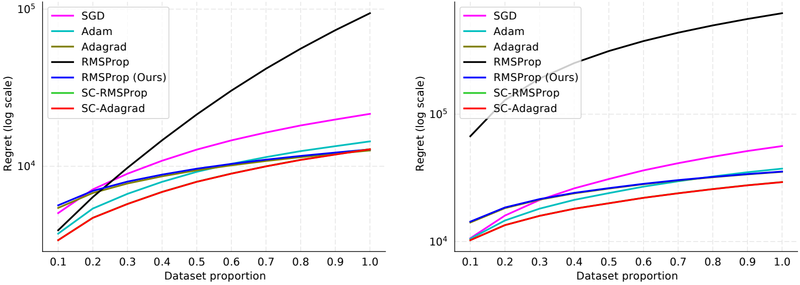
Algorithms:

- **SGD** (Bottou [2010]) (step-size is $O\left(\frac{1}{t}\right)$ for SC problems).
- **Adam** (Kingma and Bai [2015]) (step-size is $O\left(\frac{1}{\sqrt{t}}\right)$ for SC problems).
- **Adagrad** (Duchi et al. [2011]).
- **RMSProp** (Hinton et al. [2012]) with $\beta = 0.9$.
- **RMSProp (Ours)** with $\beta_t = 1 - \frac{\gamma}{t}$ and $\gamma = 0.9$.
- **SC-RMSProp** with $\gamma = 0.9$ and $\delta_t = \xi_2 e^{-\xi_1 t v_t}$.
- **SC-Adagrad** with $\delta_t = \xi_2 e^{-\xi_1 v_t}$.

Only one varying parameter: the stepsize α from $\{1, 0.1, 0.01, 0.001, 0.0001\}$.

All deep learning experiments in batch setting.

Online L2-Regularized Softmax Regression

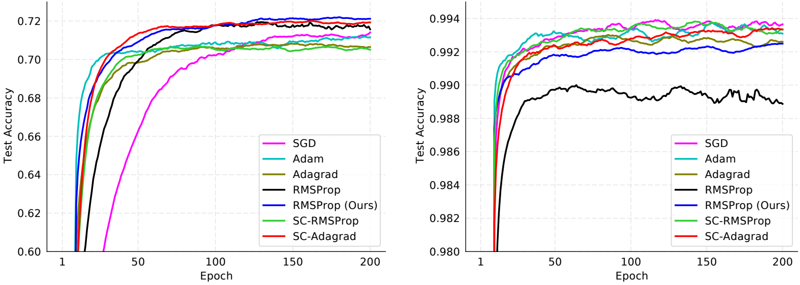


(a) CIFAR10 (b) CIFAR100

Figure : Regret (log scale) vs Dataset Proportion

Lower regret for SC-Adagrad and SC-RMSProp

Experiments: Convolutional Neural Networks

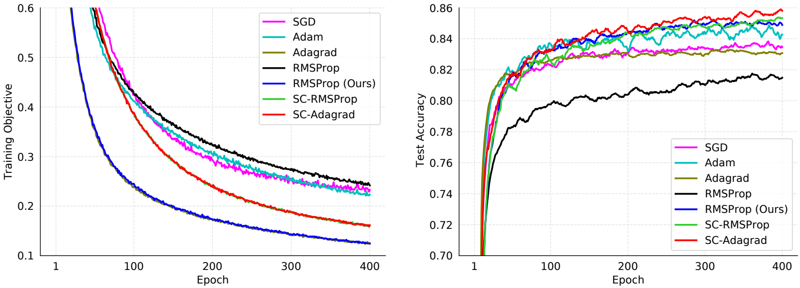


(a) CIFAR10 (b) MNIST

Figure : Test Accuracy vs Number of Epochs for 4-layer CNN

SC-Adagrad is competitive on CIFAR10 and MNIST

Experiments: Residual Networks



(a) Training Objective (b) Test Accuracy

Figure : Plots of ResNet-18 (He et al. [2016]) on CIFAR10 dataset

High test accuracy on CIFAR10 dataset by SC-Adagrad

Also check our paper for experiments on convex problems, multilayer perceptron.

Conclusion

SC-Adagrad is competitive on various deep nets.

Open question: Why does it work for Deep Learning?

CODE: `github.com/mmahesh`

POSTER: Tuesday (Tomorrow), Gallery 28

Thank you . . .

References I

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *COLT*, page 257, 2010.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *ML*, 69(2-3): 169–192, 2007.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

References II

G. Hinton, N. Srivastava, and K. Swersky. Lecture 6d - a separate, adaptive learning rate for each connection. Slides of Lecture Neural Networks for Machine Learning, 2012.

D. P. Kingma and J. L. Bai. Adam: a method for stochastic optimization. *ICLR*, 2015.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.