Adaptive Gradient Tracking In Stochastic Optimization

Zhanhong Jiang, Xian Yeow Lee*, Sin Yong Tan*, Aditya Balu*, Young M Lee, Chinmay Hegde†, Soumik Sarkar*

Johnson Controls, Iowa State University*, New York University†

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Adaptive learning rates often cause adaptive gradient descent algorithms such as Adam to underperform when compared with SGD in terms of generalization due to large variance. In this work, we develop AdaTrack, which uses the adaptive gradient tracking to control the degree of penalty throughout the optimization process. We present the theoretical analysis to show the sublinear regret bound. Empirically from experiments, AdaTrack compares favorably with current state-of-the-art such as RAdam by reducing the variance with fewer intermediate parameters and outperforms AdaBound and Adam by improving the training performance significantly.

Introduction

- Adaptive learning rates have been proposed as alternatives to SGD to accelerate convergence of gradient descent algorithms However, generalization capabilities of these adaptive learning algorithms can be poor when model and data are complex.
- We consider the problem from the perspective of *gradient tracking*, which tracks the difference between two consecutive steps of (stochastic) gradients to reduce the gradient's variance.
- We propose AdaTrack, an adaptive gradient descent algorithm that leverages the exponential moving average (EMA) of gradient tracking, to penalize significant variations of gradients during optimization, thus enhancing generalization capability.
- We propose another variant, RAT, which incorporates adaptive gradient tracking into RAdam and empirically compare AdaTrack and RAT to RAdam, AdaBound, Adam, and SGD on four image classification datasets.

Algorithm

Algorithm 1: Adaptive Gradient Tracking Descent

Input: $\alpha_t, \beta_1, \beta_2, \beta_3, \epsilon, x_0, \nabla f_{-1}(x_{-1}, \zeta_{-1})$ ▷ Input params

$m_0, v_0, y_0, t = 0$ ▷ Initializations

while $t \leq T$ **do**

$m_{t+1} = \beta_1 m_t + (1-\beta_1)\nabla f_t(x_t, \zeta_t)$ ▷ Approximate first moment

$v_{t+1} = \beta_2 v_t + (1-\beta_2)\nabla f_t^2(x_t, \zeta_t)$ ▷ Approximate second moment

$y_{t+1} = \beta_3 y_t + (1-\beta_3)(\nabla f_t(x_t, \zeta_t) - \nabla f_{t-1}(x_{t-1}, \zeta_{t-1}))$ ▷ EMA of gradient tracking

$\hat{m}_{t+1} = \frac{m_{t+1}}{1-\beta_1^t}$ ▷ Bias-correction for first moment

$\hat{v}_{t+1} = \frac{v_{t+1}}{1-\beta_2^t}$ ▷ Bias-correction for second moment

$\hat{y}_{t+1} = \frac{y_{t+1}}{1-\beta_3^t}$ ▷ Bias-correction for adaptive gradient tracking

$x_{t+1} = x_t - \alpha_t \frac{\hat{m}_{t+1}}{\sqrt{\hat{v}_{t+1}+\epsilon}} - \alpha_t \hat{y}_{t+1}$ ▷ Update using adaptive gradient tracking descent

$t = t + 1$

return x_T

Theoretical Analysis

- Table 1 shows the comparisons between different stochastic recursive gradient schemes based on usage of different variables $\left(\nabla f_t(x_t, \zeta_t)\ \&\ \nabla f_{t-1}(x_{t-1}, \zeta_{t-1})\ \&\ \nabla f_t(x_{t-1}, \zeta_t)\ \&\ x_{t-1}\right)$ used in the scheme.

Table 1

Method	$\nabla f_t(x_t, \zeta_t)$	$\nabla f_{t-1}(x_{t-1}, \zeta_{t-1})$	$\nabla f_t(x_{t-1}, \zeta_t)$	x_{t-1}	Computation
SARAH	✓	✗	✓	✓	$\mathcal{O}(2dT)$
SPIDER	✓	✗	✓	✓	$\mathcal{O}(2dT)$
ROOT-SGD	✓	✗	✓	✓	$\mathcal{O}(2dT)$
Gradient Tracking	✓	✓	✗	✗	$\mathcal{O}(dT)$

- To investigate the convergence of AdaTrack, we present the regret bound instead of the static error bound. The regret analysis is based on the online learning framework given an arbitrary unknown sequence of convex loss functions, $\{f_0(x), f_1(x), ..., f_T(x)\}$. Specifically, the regret is expressed as:

$$\mathcal{R}_T^S := \sum_{t=0}^{T} [f_t(x_t) - f_t(x^*)], \tag{1}$$

where $x^* = \text{argmin}_{x \in \mathcal{X}} \sum_{t=0}^{T} f_t(x)$. We present the following informal result for AdaTrack: Assume that f_t is Lipschitz continuous and that $\mathcal{X}$ is compact. Let $\beta_1, \beta_2, \beta_3 \in [0, 1)$ satisfy $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$ and $\beta_{1,t} = \beta_1 \lambda^{t-1}, \lambda \in (0, 1)$. Thus, for all $T \geq 0$, when the learning rate $\alpha_t = \mathcal{O}(\frac{1}{\sqrt{t+1}})$, AdaTrack has the sublinear regret, i.e., $\mathcal{R}_T^S = \mathcal{O}(\sqrt{T})$.

Experimental Results: Benchmark Functions

- Figure 1 illustrates the convergence trajectories of different optimizers for benchmark Rastrigin and Rosenbrock functions. Green dots signify the global optima.

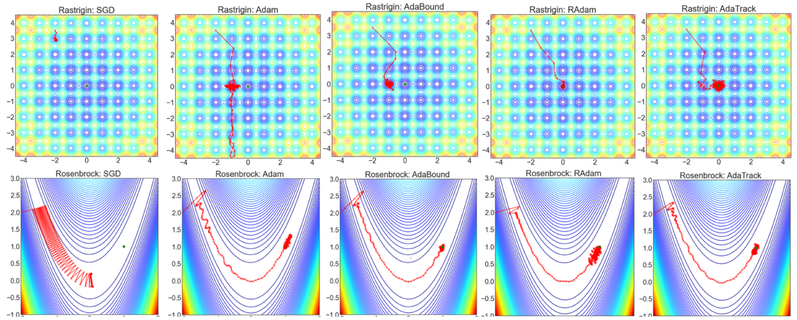

Figure 1

Experimental Results: Image Classification

- Figure 2 compares of training loss trends for six optimizers across four benchmark image classification datasets.

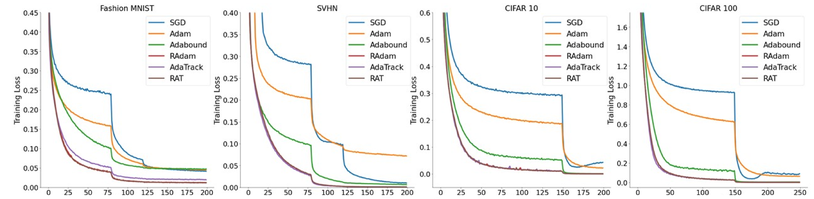

Figure 2

- Table 2 tabulated the resting accuracies of different optimizers for image datasets. A multi-layer CNN network architecture was used for Fashion MNIST, VGG19 was used for SVHN and ResNet34 was used for both CIFAR datasets.

Table 2

	Fashion MNIST	SVHN	CIFAR 10	CIFAR 100
	Test Acc.(%)	Test Acc.(%)	Test Acc.(%)	Test Acc.(%)
SGD	93.4 ± 0.17	95.8 ± 0.12	92.9 ± 0.30	71.9 ± 0.63
Adam	93.5 ± 0.09	95.6 ± 0.13	92.9 ± 0.17	71.5 ± 0.22
Adabound	93.2 ± 0.10	95.8 ± 0.10	94.9 ± 0.17	76.6 ± 0.23
RAdam	93.6 ± 0.15	96.0 ± 0.09	94.6 ± 0.24	74.4 ± 0.13
AdaTrack	93.5 ± 0.12	96.0 ± 0.11	94.3 ± 0.12	72.5 ± 0.84
RAT	93.6 ± 0.21	96.0 ± 0.04	94.4 ± 0.10	74.1 ± 0.30

Conclusions

- This work presented a new stochastic optimizer, AdaTrack, by leveraging adaptive gradient tracking developed to reduce the gradient's variance.
- We discussed the difference among different stochastic recursive gradient schemes and presented the analytical results which enabled a decent sublinear regret bound for convex loss functions.
- Empirical results demonstrated that AdaTrack outperformed SGD and Adam and is competitive when with state-of-the-art optimizers by introducing a different way to reduce variance.
- Future research directions include: 1) applications to different tasks, e.g., natural language processing and reinforcement learning; 2) the combination with momentum-based methods.

Self-aware Complex Systems Laboratory

DICE Lab

Johnson Controls

References: [1] Jiang, Z., Lee, X.Y., Tan, S.Y., Balu, A., Lee Y.M, Hegde, C. , Sarkar, S., Adaptive Gradient Tracking In Stochastic Optimization, OPT 2020 workshop

Acknowledgements: This work was partly supported by the National Science Foundation under grant number CAREER-1845969.