Methods with preconditioning with weight decay regularization

Ekaterina Statkevich

Moscow Institute of Physics and Technology

Course: My first scientific paper (Strijov's practice)/Group 007

Expert: A. Beznosikov

2023

Goal of researches

Goal objectives: investigate theory and practical convergence of algorithms and create new method of optimization.

Problem:

- Prove the convergence of the method AdamW and OASIS.
- Research the convergence on practical tasks.
- Create and investigate new optimization algorithm
- Compare it with the others

Task

- Minimization problem:
$$\min_{x \in \mathbb{R}^d} f(x)$$
- $r(x)$ – regularization function, $r(x) = \frac{\lambda}{2}||x||_2^2$
- Objective function for AdamL2 method – $f(x) + r(x)$, where $r(x) = \frac{\lambda}{2}||x||_2^2$
- New regularization function $\tilde{r}(x) : \nabla \tilde{r}(x) = D_t \nabla r(x)$
- New objective function $\tilde{F}(x) = f(x) + \tilde{r}(x)$

Assumptions

Assumption 1 (Convex)

The function f is convex, i.e. $\forall x, y \in \mathbb{R}^d$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Assumption 2 (PL–condition)

If there exists $\mu > 0$, *such that* $||\nabla f(w)|| \geq 2\mu(f(w) - f^*)$, $\forall w \in \mathbb{R}^d$

Assumptions

Assumption 3 (L-l-smoothness)

The gradients of F are L-Lipschitz continuous $\forall w \in \mathbb{R}^d$*, i.e. there exists a constant* $L > 0$ *such that* $\forall x, y \in \mathbb{R}^d$,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}||x - y||^2$$

The gradient of r is l-Lipschitz continuous $\forall w \in \mathbb{R}^d$*, i.e. there exists a constant* $l > 0$ *such that* $\forall x, y \in \mathbb{R}^d$,

$$r(x) \leq r(y) + \langle \nabla r(y), x - y \rangle + \frac{l}{2}||x - y||^2$$

Algorithms

Algorithm General scheme for preconditions methods

Require: η, ϵ, f, r

while w not converged **do**

$t = t + 1$

$g_t = \nabla f_t(w_{t-1})$

$D_t = \text{diag}(\sqrt{g_t \odot g_t} + \varepsilon)$ [a]

$D_t = \mathbb{E}[z^T \nabla^2 f(w_{t-1}) z]$ [b]

$w_t = w_{t-1} - \eta \cdot g_t D_t^{-1}$

end while

[a] red - AdamW

[b] cyan - OASIS, where z in Rademacher distribution

Algorithm Adam(λ)

Require: $\eta, \beta_1, \beta_2, \epsilon, f, r$

while θ not converged **do**

$t = t + 1$

$g_t = \nabla f(w_{t-1}) + \nabla r(w_{t-1})$ [a]

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

$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$

$\hat{m}_t = \frac{m_t}{1-\beta_1^t} + \nabla r(w_{t-1})$ [b]

$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

$w_t = w_{t-1} - \eta \cdot \frac{\hat{m}_t}{\sqrt{v_t}+\epsilon}$ - $\eta \nabla r(w_{t-1})$ [c]

end while

[a] blue – AdamL2

[b] yellow – MyAdamW

Theorem №1

Theorem

Suppose the Assumption 1, 2, 4, 5 and let $\varepsilon > 0$ *and let the step-size satisfy*

$$\eta \leq \frac{2\alpha}{\tilde{L}}$$

Then, the number of iterations performed by AdamW algorithm, starting from an initial point $w_0 \in \mathbb{R}^d$ *with* $\Delta_0 = \tilde{F}(w_0) - \tilde{F}^*$*, required to obtain and* ε*-approximate solution of the convex problem (link here to problem 1) can be bounded by*

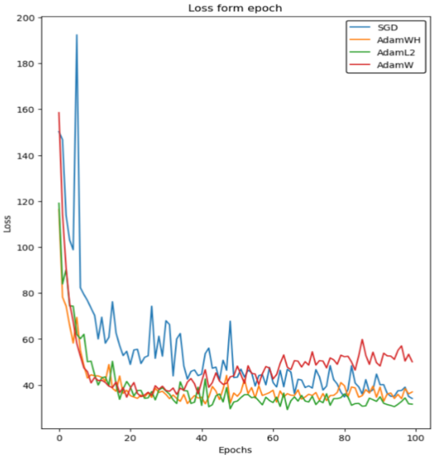
$$T = \mathcal{O}\left(\frac{\ln \frac{\Delta_0}{\epsilon}}{2\mu\eta^2(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})}\right)$$

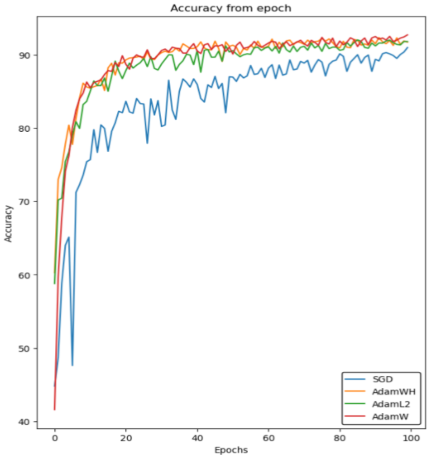
Experiment 1

Experimental conditions

- Model: ResNet18 (100 epoch, batch size 128)
- Grid of learning rates = [0.01, 0.005, 0.0005], weight decays = [0.005, 0.0005, 0.00005]
- Data set: CIFAR10.

Experiment 1: result





Experiment 2

Experimental conditions

- Model: Logistic regression
- Optimizers: AdamW, AdamL2, MyAdamW, OASIS.
- Data set: mushrooms

Experiment 2: result

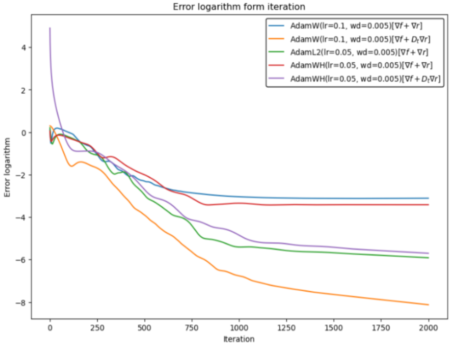


Figure 1: Adam on dataset mushrooms

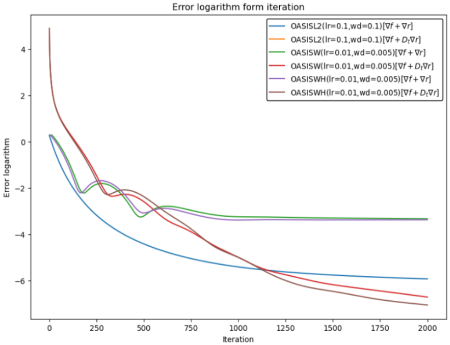


Figure 2: OASIS on dataset mushrooms

Conclusion:

THEORETICAL:

- Proofed the convergence of AdamW algorithm with different assumptions.
- Create new optimization algorithm MyAdamW.

EXPERIMENTAL:

- The optimization algorithms AdamW, AdamL2, MyAdamW, GD on a real problem are investigated on logistic regression and neural network
- Found optimal learning rates and weight decayes.

Publications:

- Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization."arXiv preprint arXiv:1412.6980 (2014).
- Jahani, Majid, et al. "Doubly adaptive scaled algorithm for machine learning using second-order information."arXiv preprint arXiv:2109.05198 (2021).
- Sadiev, Abdurakhmon, et al. "Stochastic gradient methods with preconditioned updates."arXiv preprint arXiv:2206.00285 (2022).
- Beznosikov, Aleksandr, et al. "On scaled methods for saddle point problems."arXiv preprint arXiv:2206.08303 (2022).
- Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization."arXiv preprint arXiv:1711.05101 (2017).
- Xie, Zeke, Issei Sato, and Masashi Sugiyama. "Stable weight decay regularization."(2020).