Methods with preconditioning with weight decay regularization

Matvei Kreinin

Moscow Institute of Physics and Technology

Course: My first scientific paper
(Strijov's practice)/Group 003

Expert: A. Beznosikov

2023

Goal of researches

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

Notation

- Minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x)$$

- Objective function $f(x)$
- $r(x)$ – regularization function, $r(x) = \frac{\lambda}{2}||x||_2^2$
- Objective function for AdamL2 method – $f(x) + r(x)$, where $r(x) = \frac{\lambda}{2}||x||_2^2$
- D_t – pseudo hessian on t-step, (AdamW: diagonal matrix of squares of gradients, OASIS: calculated only diagonal elements with Rademacher distribution)
- 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 (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* $\forall x, y \in \mathbb{R}^d$*,*

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

Assumptions

Assumption 3 (μ - strongly convex)

The function f is μ-strongly convex, i.e., there exists a constant $\mu > 0$ such that $\forall x, y \in \mathbb{R}^d$

$$f(x) \geq f(y) + \langle f(y), x - y \rangle + \frac{\mu}{2}||x - y||^2$$

Assumption 4 (PL–condition)

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

Assumption 5 (l-smoothness)

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 1 General scheme for preconditions methods

Require: η, ϵ, f, r
  while w not converged **do**
    $t = t + 1$
    $g_t = \nabla f_t(w_{t-1})$
    $D_t = \mathrm{diag}(\sqrt{g_t \odot g_t} + \varepsilon)$ [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 2 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
[c] red - AdamW

Theorem №1

Theorem (1)

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

$$\eta < \frac{2\alpha}{L + l \cdot \alpha}$$

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

$$T = \mathcal{O}\left(\frac{2\Delta_0 \Gamma \alpha}{(2\alpha - \tilde{L}\eta)\eta\varepsilon}\right)$$

Proof Theorem №1 (1/4)

Let's write first assumption for step t and $t+1$:

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L}{2}||w_{t+1} - w_t||^2,$$

Okay, by definition for our algorithm we have:

$$w_{t+1} - w_t = -\eta D_t^{-1} \nabla f(w_t) - \eta \nabla r(w_t),$$

and

$$\nabla f(w_t) = \frac{1}{\eta} D^t (w_t - w_{t+1}) - D^t \nabla r(w_t),$$

Proof Theorem №1 (2/4)

Okay, now let's replace $\nabla f(w_t)$ and $I \leq \frac{D_t}{\alpha}$

$$f(w_{t+1}) \leq f(w_t) + \langle \frac{1}{\eta} D_t(w_t - w_{t+1}) - D_t \nabla r(w_t), w_{t+1} - w_t \rangle + \frac{L}{2\alpha} ||w_{t+1} - w_t||^2_{D_t},$$

$$f(w_{t+1}) \leq f(w_t) + \left(\frac{L}{2\alpha} - \frac{1}{\eta} \right) ||w_{t+1} - w_t||^2_{D_t} - \langle D_t \nabla r(w_t), w_{t+1} - w_t \rangle,$$

Lets define new variable $\tilde{r} : \nabla \tilde{r} = D_t \nabla r(w_t)$. Then rewrite step using the variable and 5-th assumption.

$$\tilde{r}(w_{t+1}) \leq \tilde{r}(w_t) + \langle \tilde{r}(w_t), w_{t+1} - w_t \rangle + \frac{I}{2} (w_{t+1} - w_t)^T D_t (w_{t+1} - w_t),$$

Proof Theorem №1 (3/4)

$$f(w_{t+1}) \leq f(w_t) + \left(\frac{L}{2\alpha} - \frac{1}{\eta}\right) ||w_{t+1} - w_t||^2_{D_t} + \tilde{r}(w_t) - \tilde{r}(w_{t+1}) + \frac{l}{2}||w_{t+1} - w_t||^2_{D_t},$$

$\tilde{F}(w) = f(w) + \tilde{r}(w)$, $F(w) = f(w) + r(w)$, $(\tilde{L} = L + l\alpha)$, we get:

$$\tilde{F}(w_{t+1}) \leq \tilde{F}(w_t) + \left(\frac{\tilde{L}}{2\alpha} - \frac{1}{\eta}\right) ||w_{t+1} - w_t||^2_{D_t},$$

$$\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right) ||w_{t+1} - w_t||^2_{D_t} \leq \tilde{F}(w_t) - \tilde{F}(w_{t+1})$$

Proof Theorem №1 (4/4)

$$\frac{\eta^2(T+1)}{\Gamma}\left(\frac{1}{\eta}-\frac{\tilde{L}}{2\alpha}\right)\cdot\min_{k=0,T}||\nabla f(w_t)+\nabla\tilde{r}(w_t)||^2\leq$$

$$\leq\ \frac{\eta^2}{\Gamma}\left(\frac{1}{\eta}-\frac{\tilde{L}}{2\alpha}\right)\cdot\sum_{t=0}^{T}||\nabla f(w_t)+\nabla\tilde{r}(w_t)||^2\leq\tilde{F}(w_0)-\tilde{F}(w_*),$$

$$\min_{t=0,T}||\nabla f(w_t)+\nabla\tilde{r}(w_t)||^2\leq\frac{(\tilde{F}(w_0)-\tilde{F}(w_*))\Gamma}{(\frac{1}{\eta}-\frac{\tilde{L}}{2\alpha})\eta^2(T+1)}=\varepsilon,$$

$$T+1\geq\frac{\Delta_0\Gamma}{(\frac{1}{\eta}-\frac{\tilde{L}}{2\alpha})\eta^2\varepsilon}$$

Then:

$$T=\mathcal{O}\left(\frac{2\Delta_0\Gamma\alpha}{(2\alpha-\tilde{L}\eta)\eta\varepsilon}\right)$$

Theorem №2

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

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

Proof Theorem №2

Assume

$$\nabla \tilde{F} = \nabla f + \nabla \tilde{r}$$

$$L + ||D_t||I = \tilde{L}$$

$$w_{t+1} - w_t = -\eta D_t^{-1} \nabla r(w_t) - \eta \nabla r(w_t) = -\eta D_t^{-1} (\nabla f + \nabla \tilde{r})(w_t) = -\eta D_t^{-1} \nabla \tilde{F}(w_t)$$

Then we write $\tilde{L}$-smoothness for $\tilde{F}$

$$\tilde{F}(w_{t+1}) - \tilde{F}(w_t) \leq \langle \nabla \tilde{F}(w_t), w_{t+1} - w_t \rangle + \frac{\tilde{L}}{2} ||w_{t+1} - w_t||^2$$

Proof Theorem №2

$$\tilde{F}(w_{t+1}) - \tilde{F}(w_t) \leq -\langle \frac{1}{\eta} D_t(w_{t+1} - w_t), w_{t+1} - w_t \rangle + \frac{\tilde{L}}{2} ||w_{t+1} - w_t||^2 =$$

$$= (\frac{\tilde{L}}{2\alpha} - \frac{1}{\eta}) ||w_{t+1} - w_t||^2_{D_t} = (\frac{\tilde{L}}{2\alpha} - \frac{1}{\eta}) || - \eta D_t^{-1} \nabla \tilde{F}(w_t)||^2_{D_t} \leq$$

$$\leq (\frac{\tilde{L}}{2\alpha} - \frac{1}{\eta}) \eta^2 ||\nabla \tilde{F}(w_t)||^2_{D_t^{-1}}$$

Then we use PL-condition for the function $\tilde{F}$:

$$||\nabla \tilde{F}(w_t)||^2_{D_t^{-1}} \geq 2\mu(\tilde{F}(w_t) - \tilde{F}^*)$$

Proof Theorem №2

$$\tilde{F}(w_t) - F^* \geq \tilde{F}(w_{t+1}) - \tilde{F}^* + (\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})\eta^2 2\mu(\tilde{F}(w_t) - \tilde{F}^*) =$$

$$= \left(1 + 2\mu\eta^2(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})\right)(\tilde{F}(w_{t+1}) - \tilde{F}^*),$$

$$\epsilon \geq \Delta_0 \left(1 + 2\mu\eta^2(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})\right)^{-T} \geq (\tilde{F}(w_T) - \tilde{F}^*)$$

$$T = \frac{\ln \frac{\Delta_0}{\epsilon}}{\ln(1 + 2\mu\eta^2(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}))} \approx \frac{\ln \frac{\Delta_0}{\epsilon}}{2\mu\eta^2(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})}$$

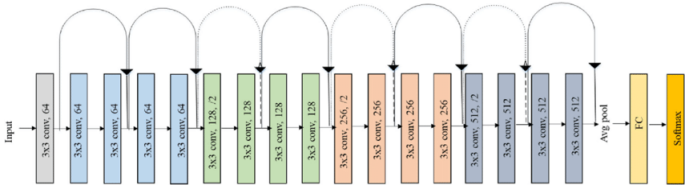
Then:

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

Experiments

Experimental conditions

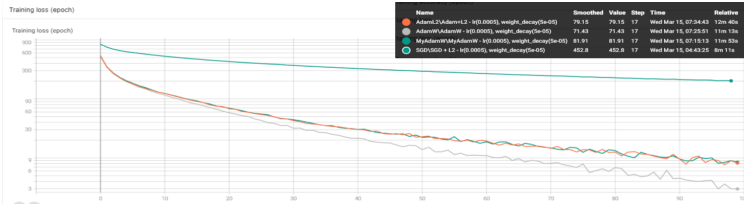
- Model: ResNet18 (100 epoch),



- CosineAnnealingLR scheduler:
  $\eta_t = \eta_{min} + \frac{1}{2}(\eta_{min} + \eta_{max})\left(1 + cos\left(\frac{T_{cur}}{T_{max}\pi}\right)\right)$
- Grid of learning rates = [0.01, 0.005, 0.0005], weight decays = [0.005, 0.0005, 0.00005]
- Data set: CIFAR10. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

Experiment

Training loss (epoch)



Testing accuracy (loss)