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 methods with preconditioning with weight decays.
- Research the convergence on practical tasks.
- Create and investigate new optimization algorithm
- Compare it with the others

Notation

- Minimization problem:
$$\min_{x \in \mathbb{R}^d} f(x)$$
- $r(x)$ – regularization function, $r(x) = \frac{\lambda}{2}||x||_2^2$
- methods with preconditioning
$$w_t = w_{t-1} - \eta \cdot D_t^{-1} g_t,$$
- 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* $\forall x, y \in \mathbb{R}^d$,

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

Differen ways of regularization

Algorithm 1 Different ways of regularization

Require: η, f
  while w not converged **do**
    $t = t + 1$
    $g_t \leftarrow$ stochastic gradient
    $g_t \leftarrow g_t + \nabla r(w_t)$ standart regularization
    $D_t \leftarrow$ preconditioning matrix
    $w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1}(g_t + \nabla r(w_t)) - \eta \cdot \nabla r(w_t)$ hessian weight decay,
  weight decay
  end while

Theorem №1

Theorem (1)

Suppose the Assumption 1, 3 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 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 problem 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.

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

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)

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

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

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

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

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, 3 and let $\varepsilon > 0$ and let the step-size satisfy

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

Then, the number of iterations performed by 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 problem 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:

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

AdamW

Algorithm 2 Adam

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

while θ not converged **do**

$t = t + 1$

$g_t = \nabla f(w_{t-1}) + \nabla r(w_{t-1})$ AdamL2

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

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

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

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

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

end while

OASIS

Algorithm 3 OASIS

Require: $w_0, \eta_0, D_0, \theta_0 = +\infty$

$w_1 = w_0 - \eta \hat{D_0}^{-1} \nabla f(w_0)$

for $k = 1, 2, \ldots$ **do**

$g_k = \nabla f(w_k) + \nabla r(w_{t-1})$ OASISL2

$D_k = \beta D_{k-1} + (1 - \beta_2) \cdot diag\left(z_k \odot \nabla^2 \left(f(w_k) + r(w_k)\right) z_k\right)$ OASISWH

$(\hat{D}_k)_{ii} = max\{|D_k|_{i,i}; \alpha\}, \forall i = \overline{1, d}$

$\eta_k = min\{\sqrt{1 + \theta_{k-1}} \cdot \eta_{k-1}; \frac{||w_k - w_{k-1}||_{\hat{D}_k}}{2||\nabla f(w_k) - \nabla f(w_{k-1})||^*_{\hat{D}_k}}\}$

$w_{k+1} = w_k - \eta_k g_k D_k^{-1} - \eta \nabla r(w_{t-1})$ OASISW

$\theta_k = \frac{\eta_k}{\eta_{k-1}}$

end for

Experiment

Model - logistic regression

Dataset - mushrooms

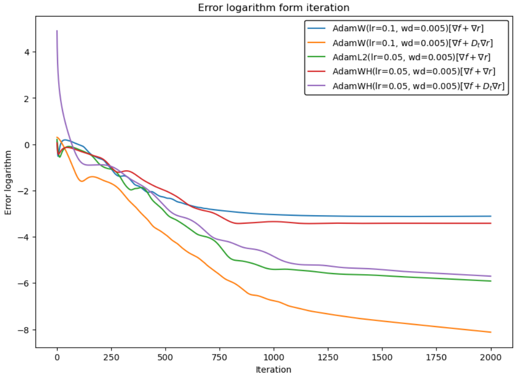

Figure: Adam on dataset mushrooms

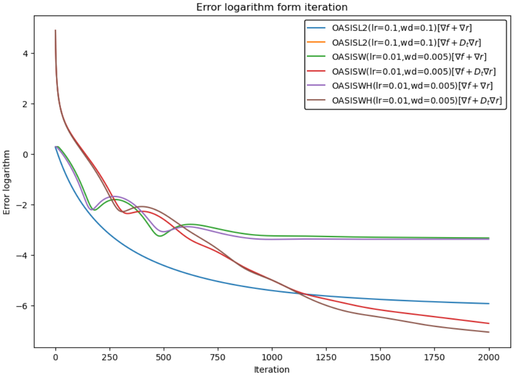

Figure: OASIS on dataset mushrooms

Experiment

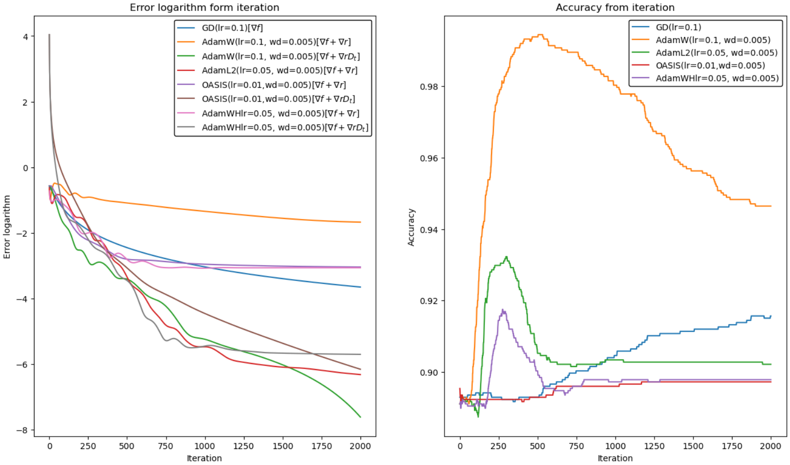

Figure: Compare different optimization algorithms on dataset: mushrooms

Experiment

Model - ResNet18, dataset - CIFAR10

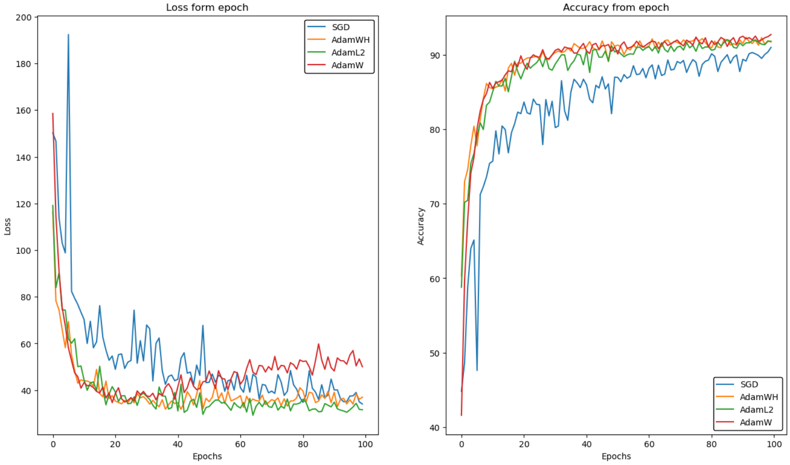

Figure: Different optimization algorithms on dataset: CIFAR10

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).

Conclusion:

- Proofed the convergence of algorithms with preconditioning with weight decay.
- The optimization algorithms AdamW, AdamL2, AdamWH, OASISW, OASISL2, OASISWH, GD on a real problem are investigated on neural networks
- Create new optimization algorithm AdamWH.
- The optimization algorithms AdamW, AdamL2, AdamW, OASISW, OASISL2, OASISWH, GD on a real problem are investigated on logistic regression.
- Investigate optimal learning rates and weight decay.
- Choose weight decays less then learning rates.