Methods with preconditioning and weight decaying

Matvei Kreinin

Moscow Institute of Physics and Technology

Scientific supervisor: A. Beznosikov PhD

2023

Problem statement

Minimize function:

$$\min_{w \in \mathbb{R}^d} f(w) \tag{1}$$

The classic approach to the solution of minimize function:

$$w_t = w_{t-1} - \eta \nabla f(w_t).$$

Preconditioned algorithms:

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

where g_t is an unbiased stochastic gradient, D_t is a matrix of preconditioning.

Different ways of matrix with preconditioning

AdaGrad:

$$D_t = diag\left\{\sqrt{\sum_{i=0}^{t} g_i \odot g_i}\right\}$$

RMSProp and Adam:

$$D_t^2 = \beta D_{t-1}^2 + (1-\beta) diag\{g_t \odot g_t\}$$

OASIS:

$$D_t = diag\{z \odot \nabla^2 f(w_t) z\}$$

where z is a random vector from Randamaher distribution.

New minimizing function

$$\min_{w \in \mathbb{R}^d} F(w) := f(w) + r(w)$$

where $r(w)$ is a the regularization function.

Algorithm 1 Different ways of using preconditioning for regularized problem

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

Again new target function

Put D_t^{-1} out of brackets and get new target function:

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

A new regularization function $\nabla \tilde{r}(w) = D_t \nabla r(w)$.
New target function:

$$\min_{w \in \mathbb{R}^d} \tilde{F}(w) := f(w) + \tilde{r}(w)$$

,where $\tilde{F}(w)$ changes every time step.

Assumptions

Assumption (Regularizer structure)

Regularizer r is separable, i.e. it can be viewed in the form:

$$r(w) = \sum_{i=1}^{d} r_i(w_i).$$

Assumption (Preconditioner structure)

Preconditioner D_t can be viewed in the following form:

$$D_t = diag\left\{d_t^1 \ldots, d_t^d\right\}.$$

Assumptions

Assumption (L-smoothness)

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

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

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

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

Assumptions

Assumption (PL–condition)

There exists $\mu > 0$*, such that* $\forall w \in \mathbb{R}^d$

$$||\nabla f(w)|| \geq 2\mu(f(w) - f^*).$$

Assumption (Preconditioner)

Restrictions on preconditioner D_t

$$\alpha I \preccurlyeq D_t \preccurlyeq \Gamma I \Leftrightarrow \frac{I}{\alpha} \preccurlyeq D_t^{-1} \preccurlyeq \frac{I}{\Gamma}.$$

Assumption (Expectations)

Restrictions on D_t *and* g_t *are unbiased, i.e.*

$$\mathbb{E}\left[D_t\right] = D_t \text{ and } \mathbb{E}\left[g_t\right] = \nabla f(w_t), \mathbb{E}\left[||g_t - \nabla f||^2\right] \leq \sigma^2.$$

Lemmas

Lemma (Existence of $\widetilde{r}$)

Suppose the Assumptions 1, 2 hold, the function $\widetilde{r}$ exists and has following form:

$$\widetilde{r}(w) = \sum_{i=1}^{d} d_t^i r_i(w_i)$$

Lemma (L-smoothness of $\widetilde{r}$)

Suppose the Assumptions 1, 2, 3 hold, The gradient of $\widetilde{r}$ is $L_{\tilde{r}}$-continuous, i.e. there exists a constant $L_{\tilde{r}} > 0$ such that $\forall x, y \in \mathbb{R}^d$,

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

where $L_{\tilde{r}} = ||D_t|| L_r$

Theorems

Theorem (1)

Suppose the Assumptions 3, 5 hold, let $\varepsilon > 0$ and let the step-size satisfy, where $L_f, L_{\tilde{r}}$ - lipschitz constants of functions f and $\tilde{r}$, $\alpha I \preccurlyeq D_t \preccurlyeq \Gamma$

$$\eta < \frac{2\alpha}{L_f + \Gamma L_{\tilde{r}}\alpha}.$$

Then, the number of iterations performed by algorithms with preconditioning and weight decaying, 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 (1) can be bounded by

$$T = \mathcal{O}\left(\frac{2\Delta_0\Gamma\alpha}{\left(2\alpha - \left(L_f + \Gamma L_{\tilde{r}}\alpha\right)\eta\right)\eta\varepsilon}\right).$$

Theorems

Theorem (2)

Suppose the Assumptions 3, 4, 5 hold, let $\varepsilon > 0$ and let the step-size satisfy, where $\alpha I \preccurlyeq D_t \preccurlyeq \Gamma$, $L_{\tilde{F}} = L_f + \Gamma L_r$, and L_F, L_r - lipschitz constant of functions f and r,

$$\eta \leq \frac{2\alpha}{L_{\tilde{F}}}.$$

Let $\tilde{F}^*$ be a solution of the optimization function. Then, the number of iterations performed by algorithms with preconditioning and weight decaying, 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 (1) can be bounded by

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

Theorems

Theorem (3)

Suppose the Assumptions 3, 4, 5, 6 hold, let $\varepsilon > 0$ *and let the step-size satisfy*

$$\eta \approx \sqrt{\frac{\left(\tilde{F}(w_0) - \tilde{F}(w_*)\right)\alpha}{L\sigma^2}}.$$

Let $\tilde{F}^*$ *be a solution of the optimization function. Then, the number of iterations performed by algorithms with preconditioning and weight decaying, 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 (1) can be bounded by, where* $L_{\tilde{r}}, L_f$ *- lipschitz constant of functions* $\tilde{r}$ *and* f*, and* $L_{\tilde{F}} = L_f + \Gamma L_r$

$$T = \mathcal{O}\left(\frac{\Gamma\Delta_0}{\left(\frac{1}{\eta} - \frac{\Gamma L_{\tilde{r}}}{2} - \frac{\Gamma L_{\tilde{F}} L_f \eta^2}{2\alpha^2}\right)\varepsilon}\right).$$

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 $k = 1, 2, ...$ **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

Experiments

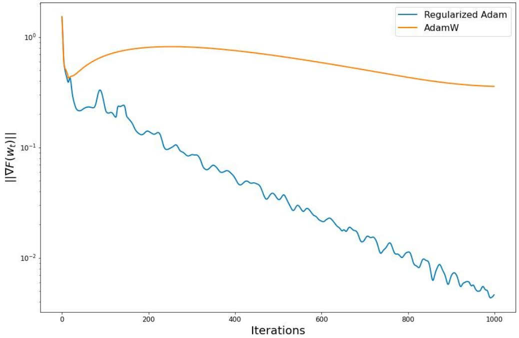


Figure: Adam and AdamW with basic criterion

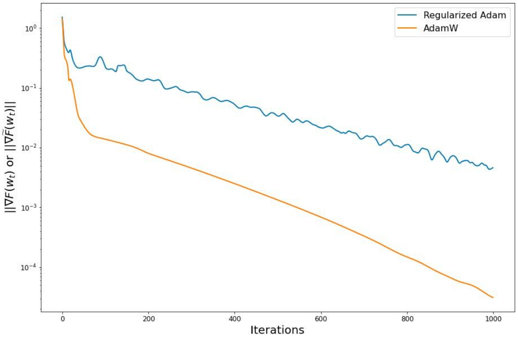
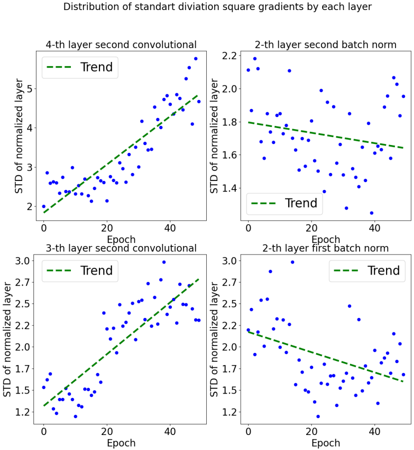


Figure: Adam and AdamW with modified criterion

Experiment



Deviation of the normalized weights in the convolutional layers has rising trend. Hence, difference between solutions of different problems is bounded below and methods converge to a different optimums.

Distribution of standard deviation of elements of matrix D_t over epochs.

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:

- Proposed novel approach how to apply weight decaying in algorithm.
- Theorethical analyse of convergency of methods with preconditioning and weight decaying.
- Create new optimization algorithm AdamWH.
- 3 theorems and 2 lemmas for estimating the convergence of methods with preconditioning and weight decaying are proved
- The further direction of analyzing the distribution of preconditioning elements in neural networks, another direction is to make a coordinate adam in neural networks.