METHODS WITH PRECONDITIONING WITH WEIGHT DECAY REGULARIZATION

A PREPRINT

Kreinin M.
Department of Data Science
Moscow Institute of Physics and Technology
Moscow, Russia
kreinin.mv@phystech.edu

Babkin P.
Department of Data Science
Moscow Institute of Physics and Technology
Moscow, Russia
babkin.pk@phystech.edu

Statkevich E.
Department of Data Science
Moscow Institute of Physics and Technology
Moscow, Russia
statkevich.ei@phystech.edu

Beznosikov A.
Department of Data Science
Moscow Institute of Physics and Technology
Moscow, Russia
beznosikov.an@phystech.edu

April 2, 2023

ABSTRACT

Keywords First keyword · Second keyword · More

1 Difference from gradient descent

The main difference from gradient descent is that Adam's algorithm uses information about previous gradients and gradient squares of the function to update parameters of the model. The OASIS algorithm uses information about previous gradients, it also calculates the diagonal elements of the hessian (so it is a quasi-Newton method) that is used to update parameters of the model and the learning rate.

2 A unified description of the methods

Calculate the weighted value using data(?) from current and previous steps. These methods use second-order(?) information: the square of the gradient elements in the Adam's algorithm and the Hessian's diagonal in the OASIS case. The next step of the algorithm is calculated like Newton's method, using the gradient and the inverse of the second-order value.

3 Algorithms

Algorithm 1 Gradient Descent

Require: θ_0, α, f
  $t = 0$
  while θ not converged **do**
    $\theta_t = \theta_{t-1} - \alpha \cdot \nabla f(\theta_{t-1})$
  end while

Algorithm 2 Adam

Require: $\alpha, \beta_1, \beta_2, \epsilon, f$
$m_0 = 0$ – 1-st moment vector
$v_0 = 0$ – 2-nd moment vector
$t = 0$ – timestep
while θ not converged **do**
$t = t + 1$
$g_t = \nabla_\theta f_t(\theta_{t-1})$
$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
$\hat{m_t} = \frac{m_t}{1-\beta_1^t}$
$\hat{v_t} = \frac{v_t}{1-\beta_2^t}$
$\theta_t = \theta_{t-1} - \alpha \cdot \frac{\hat{m_t}}{\sqrt{v_t}+\epsilon}$
end while

Algorithm 3 AdamW

Require: $\alpha, \beta_1, \beta_2, \epsilon, f$
$m_0 = 0$ – 1-st moment vector
$v_0 = 0$ – 2-nd moment vector
$t = 0$ – timestep
while θ not converged **do**
$t = t + 1$
$g_t = \nabla_\theta f_t(\theta_{t-1})$
$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
$\hat{m_t} = \frac{m_t}{1-\beta_1^t}$
$\hat{v_t} = \frac{v_t}{1-\beta_2^t}$
$\theta_t = \theta_{t-1} - \alpha \cdot \frac{\hat{m_t}}{\sqrt{v_t}+\epsilon} - \lambda\theta_t$
end while

Algorithm 4 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**
$D_k = \beta D_{k-1} + (1 - \beta_2) \cdot diag\left(z_k \odot \nabla^2 F(w_k) z_k\right)$
$(\hat{D_k})_{ii} = max\{|D_k|_{i,i}; \alpha\}, \forall i = \overline{1, d}$
$\eta_k = min\{\sqrt{1 + \theta_{k-1}} \cdot \eta_{k-1}; \frac{||w_k - w_{k-1}||_{\hat{D_k}}}{2||\nabla F(w_k) - \nabla F(w_{k-1})||^*_{\hat{D_k}}}\}$
$\theta_k = \frac{\eta_k}{\eta_{k-1}}$
end for

4 Examples of problems where algorithms are more effective than gradient descent

These methods are examples of the last stage of development of gradient descent. They are faster due to keeping information about previous iterations, because of which they choose a more efficient direction of descent and will work quicker in the case of a ill-conditioned quadratic problem. Also, simple gradient descent will "wobble" in case of, for example, a saddle point, and these methods, thanks to the hessian in the case of OASIS and the gradient square in the case of Adam, will be able to hit it.

5 Our evaluations

We try to proof convergence for preconditioned methods with weight decay.

Some assumption that we have:

Assumption 1 (Convex). *The function f is convex, i.e.* $\forall y, x \in \mathbb{R}^d$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{1}$$

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* $\forall x, y \in \mathbb{R}^d$*,*

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

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* $\forall x, y \in \mathbb{R}^d$*,*

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

5.1 Proof of convergence

To proof this algorigthm we use simplified form of AdamW algorithm

Algorithm 5 AdamW

Require: r, ε, f
  while w not converged **do**
    $t = t + 1$
    $D_t = \text{diag}(|\nabla f(w_t)|_i)$
    $w_t = w_{t-1} - \eta \cdot \nabla f(w_t) D_t^{-1} - \lambda \nabla r(w_t)$
  end while

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. 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

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

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.

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

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

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

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

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

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

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

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

Then:

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

□

5.2 Notes

But we need to remember that: $D_t = diag(\nabla f(w_t))$

$||\nabla f(w_t) + \nabla \tilde{r}(w_t)||^2 = ||\nabla f(w_t) + D_t \nabla r(w_t)||^2 = ||\nabla f(w_t) \cdot (I + \lambda w_t \nabla f(w_t))||^2 \leq ||\nabla f(w_t)||^2 \cdot ||I + \lambda w_t \nabla f(w_t)||^2$

Some additional information: $2\langle \nabla f(w_t)^2, \lambda w_t \rangle = \langle \nabla f(w_t)^2, \lambda \nabla f(w_t) w_t \rangle$

$$||\nabla f(w_t) + \nabla \tilde{r}(w_t)||^2 = ||\nabla f(w_t)||^2 + 2\langle \nabla f(w_t)^2, \lambda w_t \rangle + ||\lambda \nabla f(w_t) w_t||^2$$

$... = \langle \nabla f(w_t), \nabla f(w_t) \rangle + 2\langle \nabla f(w_t), \lambda \nabla f(w_t) w_t \rangle + \lambda^2 \langle \nabla f(w_t) w_t, \nabla f(w_t) w_t \rangle = \langle \nabla f(w_t)^2 + 2\lambda \nabla f(w_t) w_t + ... \rangle$

If life was a dream, then it will right:

$$\lambda \langle \nabla f(w_t)^2 (2 + \lambda w_t), w_t \rangle \geq 0$$

Assumption Function activation (B). *ut it's right for Neural Networks with function activation: sigmoid and tanh.*

From beginning:

$$||w_{t+1} - w_t||^2_{D_t} = ||\eta D_t^{-1}\nabla f(w_t) + \eta\nabla r(w_t)||^2_{D_t} = \eta^2||\nabla f(w_t) + \nabla\tilde{r}(w_t)||^2_{D_t^{-1}} \geq \frac{\eta^2}{\Gamma}||\nabla f(w_t) + \nabla\tilde{r}(w_t)||^2$$

To collapse this inequality, you need to choose a step so that the coefficient at the norm of the difference is negative, but we get nothing.

$$\tilde{F}(w_{t+1}) \leq \tilde{F}(w_t)$$

Now consider when the solution to the new problem will be the solution to the original problem.

$$||\nabla f(\tilde{w}_*) + \nabla\tilde{r}(\tilde{w}_*)|| \leq ||\nabla f(\tilde{w}_*)|| + ||\nabla\tilde{r}(\tilde{w}_*)|| \leq ||\nabla f(\tilde{w}_*)|| + \Gamma||\nabla r(w_*)|| \leq \varepsilon$$

If $\Gamma > 1$, then we get:

$$||\nabla f(\tilde{w}_*) + \nabla r(\tilde{w}_*)|| \leq ||\nabla f(\tilde{w}_*)|| + ||\nabla r(w_*)|| \leq ||\nabla f(\tilde{w}_*)|| + \Gamma||\nabla r(w_*)|| \leq \varepsilon$$

Our next proposal, it's changing in AdamW algorithm: consider, that $r(w) = \frac{\lambda}{2}||w||^2$

Algorithm 6 MyAdamW(λ)

Require: $\alpha, \beta_1, \beta_2, \epsilon, f$
  $m_0 = 0$ – 1-st moment vector
  $v_0 = 0$ – 2-nd moment vector
  $t = 0$ – timestep
  while θ not converged **do**
    $t = t + 1$
    $g_t = \nabla_\theta f_t(\theta_{t-1})$
    $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
    $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
    $v_{t,i} = \max(\lambda, v_{t,i})$, for each i
    $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
    $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
    $\theta_t = \theta_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{v_t}+\epsilon} - \alpha\theta_{t-1}$
  end while

Theorem 2. *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. Assume

$$\nabla\tilde{F} = \nabla f + \nabla\tilde{r} \tag{2}$$

$$L + ||D_t||l = \tilde{L} \tag{3}$$

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

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

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

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

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

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

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

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

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

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

Then:

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

□

comparison:

$$||\nabla f + D_t \nabla r|| = ||\nabla f + D_t \nabla r|| < \epsilon$$

$$||\nabla f + \nabla r|| = ||\nabla f + \nabla r \pm D_t \nabla r|| < \epsilon + ||D_t \nabla r - \nabla r|| \leq \epsilon + ||D_t - I||||\nabla r||$$