METHODS WITH PRECONDITIONING WITH WEIGHT DECAY REGULARIZATION

A PREPRINT

Kreinin M.
Department of Intelligent Systems
MIPT
Moscow, Russia
kreinin.mv@phystech.edu

Beznosikov A.
Department of Data Science
MIPT
Dolgoprudny, Russia
beznosikov.an@phystech.edu

ABSTRACT

This paper examines the convergence rate of adaptive gradient methods when a regularization function is added to the target function. This is an area of research, since many machine learning problems use the heuristic, heuristic regularization, and we investigate the theoretical and practical convergence of adaptive gradient methods.

Keywords Adam · OASIS · Regularization · ADAHessian · Weight Decay

1 Introduction

In machine learning we consider unconstrained optimization problem

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

Problems of the form (1) cover a plethora of applications, including empirical risk minimization, deep learning, and supervised learning tasks such as regularized least squares or logistic regression. This minimization problem can be difficult to solve, particularly when the number of training samples n, or problem dimension d, is large, or if the problem is nonconvex.

There are second order preconditioners methods for solving this problem, such as Adam [Kingma and Ba, 2014] , OASIS [], AdaHessian.

Throughout this work we assume that each f : $\mathbb{R}^d \to \mathbb{R}$ is twice differentiable and also L-smooth. This is formalized in the following assumption.

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

$$f(w) \geq f(w') + \langle \nabla F(w'), w - w' \rangle \tag{1}$$

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* $\forall w, w' \in \mathbb{R}^d$*,*

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

Assumption 3 (Twice differentiable). *The function f is twice continuously differentiable.*

Assumption 4 (μ - strongly convex). *The function f is μ-strongly convex, i.e., there exists a constant $\mu > 0$ such that* $\forall w, w' \in \mathbb{R}^d$

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

2 Problem statement

We consider the unconstrained optimization problem

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

Problems of the form cover a plethora of applications, including empirical risk minimization, deep learning, and supervised learning tasks such as regularized least squares or logistic regression.

But we can add regulirazation $r(w)$ – regularization, and solve the unconstrained optimization problem.

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

For solving these problems we always scaled adaptive gradients algorithms like ADAM, OASIS, ADAHessian. We investigate how the convergence rate will change when the regularization function is added to these algorithms.

In the convergence algorithm that we study, we will investigate algorithms of the following two kinds. In the first one, the regularization function is taken out separately in the recalculation of model weights, and in the second one, the function is dominated by the inverse "hessian".

Algorithm 1 Based algorithm

```
$w_0, D_0, \eta_0$
  for $k = 1, 2, ...$ do
    $D_k$ - update by using information of $f(w)$
    $\eta_k$ - update
    Set $w_{k+1} = w_k - \eta_k D_k^{-1} \nabla f(w_k) - \eta \nabla r(w)$
    or like that
    Set $w_{k+1} = w_k - \eta_k D_k^{-1} (\nabla f(w_k) + \nabla r(w))$
  end for
```

3 Experiment

We will conduct an experiment in which we compare Adam with added L2 regularization with AdamW, at different learning rate and decoupled weight, also we will compare it with the modified AdamW algorithm where the regularizer is put under the hessian, also we will compare it under the condition that now the minimal eigenvalue of hessian is not epsilon, but decoupled weight hyperparameter.We will conduct the same experiments with SGD and L2 regularization and SGD with decoupled weitght at the same parameters. All experiments will be performed on batches of size 128, on a ResNet18 grid of 200 epochs. All this will be implemented using the PyTorch library on dataset CIFAR10.

4 Theory

To proof this algorigthm we use simplified form of AdamW algorithm

Algorithm 2 AdamW

```
Require: $r, \varepsilon, f$
  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:

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

□

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:

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

□

References

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.