A TEMPLATE FOR THE *arxiv* STYLE

A PREPRINT

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

1 Introduction

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, η_0
for $k = 1, 2, ...$ **do**
  D_k - update by using information of $f(w)$
  η_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

Some assumptions for our problem:

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

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

1.1 Citations

Kingma and Ba [2014], Jahani et al. [2021], Sadiev et al. [2022], Beznosikov et al. [2022], Stich [2019]

References

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

Majid Jahani, Sergey Rusakov, Zheng Shi, Peter Richtárik, Michael W Mahoney, and Martin Takáč. Doubly adaptive scaled algorithm for machine learning using second-order information. *arXiv preprint arXiv:2109.05198*, 2021.

Abdurakhmon Sadiev, Aleksandr Beznosikov, Abdulla Jasem Almansoori, Dmitry Kamzolov, Rachael Tappenden, and Martin Takáč. Stochastic gradient methods with preconditioned updates. *arXiv preprint arXiv:2206.00285*, 2022.

Aleksandr Beznosikov, Aibek Alanov, Dmitry Kovalev, Martin Takáč, and Alexander Gasnikov. On scaled methods for saddle point problems. *arXiv preprint arXiv:2206.08303*, 2022.

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.