Paper Reading Assignment 2

HU Yang

Dual Averaging Method for
Regularized Stochastic Learning and Online Optimization

SDSC8014 - Online learning

February 10, 2022

1 Introduction

This paper combines two domains of *regularized stochastic learning* and *online optimization problems*. In this paper, the objective function consists of: 1. loss function, and 2. a simple regularization term. Based on the objective function, the paper mainly developed a novel online algorithm, the regularized dual averaging (RDA) method.

Usually, the learning task is difficult to optimize with traditional techniques, like stochastic gradient descent. Traditional techniques are abused of low accuracy when dealing with regularization terms.

The basic mechanism can be described as follows:

- learning process has many iterations.
- The samples in each iteration are different, and remain unknown until given in the beginning of each iteration. The learned parameters of the objective function are modified in each iteration.
- the learning parameters in each objective function are slightly adjusted by solving an optimization problem. Also, all past sub-gradients are incorporated in the computation.

Numerical experiments of the l1-RDA method using the MNIST dataset of handwritten digits as learning tasks demonstrated that the RDA method can be very effective for sparse online learning problems.

2 Preliminaries

2.1 Regularized stochastic learning

$$\min_{w} \phi(w) := \mathbb{E}_z f(w, z) + \psi(w) \tag{1}$$

In Eq.1, the first term on the right hand side is the loss function, while the second one is a simple regularization term (e.g., l1-norm, l2-norm).

2.2 Regularized online optimization

Definition 1. *The regret of the regularized online algorithm is defined as:*

$$R_t(w) := \sum_{\tau=1}^{t}(f_\tau(w_\tau) + \Psi(w_\tau)) - \sum_{\tau=1}^{t}(f_\tau(w) + \Psi(w)) \tag{2}$$

where $\Psi(w)$ is a strong convex regularization term.

(Well, these expressions are quite similar to the content delivered in class.)

3 Technological details

In each iteration, if it's able to find the closed-form solution for the auxiliary optimization problem, the computational complexity per iteration is $O(n)$. Moreover, the RDA method converges to the optimal solution of 1 with the optimal rate $O(1/\sqrt{t})$. If the the regularization function term is strongly convex, a better rate $O(\ln t/t)$ can be achieved.

Algorithm 1 Regularized dual averaging (RDA) method

```
Input: A strong convex function, $h(w)$, with module 1, and satisfy
    $$w_0 = \arg\min_w h(w) \in \arg\min_w \Psi(w) \tag{3}$$
    , and a predetermined non-negative and non-decreasing sequence, $\beta_t$;
Output: Final learned parameters $w_{t+1}$;
initialize $w_1 = w_0, \overline{g}_0 = 0$;
for do$t = 1, 2, 3...$
    Compute sub-gradient $g_t \in \partial f_t(w_t)$
    Update the average sub-gradient $\overline{g}_t$ :
        $$\overline{g}_t = \frac{t-1}{t}\overline{g}_{t-1} + \frac{1}{t}g_t \tag{4}$$
    Compute the next iterate $w_{t+1}$:
        $$w_{t+1} = \arg\min_w \langle \overline{g}_t, w\rangle + \Psi(w) + \frac{\beta_t}{t}h(w) \tag{5}$$
end for
```

Definition 2. *A function h is called* ***strongly convex*** *with respect to a norm if there exists a constant $\sigma > 0$ such that:*

$$h(aw + (1-a)u) \leq ah(w) + (1-a)h(u) - \frac{\sigma}{2}a(1-a)\|w-u\|^2 \tag{6}$$

In algorithm 1, $h(w)$ must satisfy definition 2. Step 1 is to compute a sub-gradient of function $f(t)$ w.r.t w_t, function $f(t)$ may not have gradient. In step 3, it's assumed that minimization problem in (6) can be solved with a closed form. If h(w) is not strong convex, or σ in Eq.6 $= 0$, just replace $\beta_t = \gamma\sqrt{t}$ (where $\gamma > 0$). If the regularization term is strong convex, any non-negative, non-decreasing sequence dominated by ln$t$ can be applied.

Theorem 1. Let the sequences $\{w_\tau\}_{\tau=1}^{t}$ and $\{g_\tau\}_{\tau=1}^{t}$ be generated by Algorithm 1. Assume there is a constant L that $\|g_t\|_* \leq L$ for all $t \geq 1$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. Then for any $t \leq 1$, we have:

$$R_t(w) \leq (\beta_0 - \beta_1)h(w_2) + \beta_t D^2 + \frac{L^2}{2}\sum_{\tau=0}^{t-1}\frac{1}{\sigma\tau + \beta_\tau} t = 1, 2, 3, ... \tag{7}$$

The authors provided the proof of Theorem 3 in Xiao(2009) [1]. By gradual enlarging and reducing, given $\sigma = 0$ and $\gamma^\star = L/D$, the regret $R_t(w)$ is bounded by $2LD\sqrt{t}$.

Theorem 2. Assume there exists an optimal solution $w^\star$ to the simple optimization problem in each iteration of algorithm 1 that satisfies $h(w*) \leq D^2$ for some $D > 0$, and there is an L¿0 such that $E\,\|g\|_*^2 \leq L^2$ for all $g \in \partial f(w, z)$. Then for any $t \geq 1$, we have:

$$E\,\phi(\frac{1}{t}\sum_{\tau=1}^{t}) - \phi(w^\star) \geq \frac{\delta_t}{t} \tag{8}$$

The $O(\text{ln}t)$ regret bound can also be induced in a similar way according to Xiao(2009).

[1] L. Xiao. Dual averaging method for regularized stochastic learning and online optimization. Technical Report MSR-TR-2009-100, Microsoft Research, 2009.