Journal of Computational Mathematics
Vol.41, No.3, 2023, 525–541.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2210-m2021-0106

EXTENDED REGULARIZED DUAL AVERAGING METHODS FOR STOCHASTIC OPTIMIZATION*

Jonathan W. Siegel[1)]
Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA
Email: jus1949@psu.edu
Jinchao Xu
Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA
Email: jxx1@psu.edu

Abstract

We introduce a new algorithm, extended regularized dual averaging (XRDA), for solving regularized stochastic optimization problems, which generalizes the regularized dual averaging (RDA) method. The main novelty of the method is that it allows a flexible control of the backward step size. For instance, the backward step size used in RDA grows without bound, while for XRDA the backward step size can be kept bounded. We demonstrate experimentally that additional control over the backward step size can speed up the convergence of the algorithm while preserving desired properties of the iterates, such as sparsity. Theoretically, we show that the XRDA method achieves the same convergence rate as RDA for general convex objectives.

Mathematics subject classification: 90C25, 90C30.
Key words: Convex Optimization, Subgradient Methods, Structured Optimization, Nonsmooth Optimization.

1. Introduction

Optimizing convex objectives with a regularization term which promotes a certain structure of the minimizer, for example the ℓ^1-norm promoting sparsity or the nuclear norm promoting low rank [8], are very common and important in machine learning. Stochastic optimization of such objectives poses a unique challenge since traditional methods such as stochastic gradient descent (SGD) [14] are often not effective at producing the desired structure (i.e. sparsity or low-rank) of the iterates [24]. This motivated the introduction of the well-known RDA method and its variants, which are able to effectively produce the desired structure of the iterates [9,24]. We introduce a generalization of RDA, called extended regularized dual averaging (XRDA), which we show can significantly improve convergence while still preserving the desired structure of the iterates.

Let us begin by describing these issues in detail and giving an overview of RDA. Consider the subgradient descent and dual averaging methods [18] for minimizing a Lipschitz convex function F, given by

$$x_{n+1} = x_n - s_n g_n, \tag{1.1}$$

* Received April 14, 2021 / Revised version received July 28, 2022 / Accepted October 27, 2022 / Published online April 8, 2023 /
1) Corresponding author

where $g_n \in \partial F(x_n)$. It is well known that with a step size $s_n = n^{-1/2}$ this method attains a convergence rate of $\mathcal{O}(n^{-1/2}\log n)$. The simple dual averaging (SDA) method of Nesterov [18], which is given by

$$x_{n+1} = \arg\min_x \left(\sum_{i=1}^{n} \langle s_i g_i, x \rangle + \frac{\alpha_{n+1}}{2} \|x - x_1\|^2 \right), \tag{1.2}$$

generalizes subgradient descent (note that setting $\alpha_n = 1$ recovers the iteration (1.1)). The advantage of the more general SDA method is that by setting $s_n = 1$ and $\alpha_n = \sqrt{n}$, the logarithmic factor in the convergence rate can be removed [18]. To illustrate the relationship between this new method and original subgradient descent, we rewrite this new iteration as

$$x_{n+1} = \frac{\alpha_n}{\alpha_{n+1}} x_n + \left(1 - \frac{\alpha_n}{\alpha_{n+1}}\right) x_1 - \tilde{s}_n g_n, \tag{1.3}$$

where the effective step size is $\tilde{s}_n = s_n/\alpha_{n+1}$. Thus the difference between this new method and the subgradient descent method is an averaging with the initial iterate x_1. It is remarkable that this provides a significant improvement in the convergence rate. These results hold in more generality with the ℓ^2 distance replaced by the Bregman distance $D_\phi(x, y)$ with respect to a convex function ϕ, which we describe in more detail in Section 2, but for simplicity we stay with the current setting throughout the introduction.

Next, we consider the composite (or regularized) optimization problem

$$\arg\min_{x \in A} \big[f(x) = F(x) + G(x) \big], \tag{1.4}$$

where $F(x)$ is a convex Lipschitz function and $G(x)$ is a convex function. A standard method for solving this is the forward-backward subgradient method

$$\begin{aligned} x_{n+\frac{1}{2}} &= x_n - s_n g_n, \\ x_{n+1} - x_{n+\frac{1}{2}} &\in -s_n \partial G(x_{n+1}), \end{aligned} \tag{1.5}$$

where $g_n \in \partial F(x_n)$ [3]. Note that the second step above corresponds to backward Euler and is known as the proximal map for G [6]. With a choice of step size $s_n = \mathcal{O}(n^{-1/2})$, this method also achieves a convergence rate of $\mathcal{O}(n^{-1/2}\log n)$. As before the logarithm can be removed by introducing a similar averaging with x_1 as in the SDA method. Note also that the constants in the convergence rates only depend upon the Lipschitz constant of F and not G, which is the advantage of using the forward-backward splitting.

In many cases of practical interest, the subgradients $g_i \in \partial F(x_i)$ in the forward step are not computed exactly, but rather replaced by an unbiased sample $\tilde{g}_i$ at x_i, i.e. $\mathbb{E}(\tilde{g}_i) \in \partial F(x_i)$. Using this sample in (1.5) results in forward-backward stochastic gradient descent. Stochastic gradient descent has proven extremely useful for training a variety of machine learning models [14, 15]. However, for problems where the minimizer is expected to have a special structure, forward-backward stochastic gradient descent often has the drawback that the iterates it produces do not have the desired structure. A very common example is sparsity. For instance, consider the forward-backward stochastic gradient descent algorithm applied to an objective $F(x) + G(x)$ with $G(x) = \lambda\|x\|_1$ an l^1 regularization term

$$x_{n+\frac{1}{2}} = x_n - s_n \tilde{g}_i,$$