Bull. Korean Math. Soc. **53** (2016), No. 2, pp. 589–600
http://dx.doi.org/10.4134/BKMS.2016.53.2.589

A MEMORY EFFICIENT INCREMENTAL GRADIENT METHOD FOR REGULARIZED MINIMIZATION

Sangwoon Yun

Abstract. In this paper, we propose a new incremental gradient method for solving a regularized minimization problem whose objective is the sum of m smooth functions and a (possibly nonsmooth) convex function. This method uses an adaptive stepsize. Recently proposed incremental gradient methods for a regularized minimization problem need $O(mn)$ storage, where n is the number of variables. This is the drawback of them. But, the proposed new incremental gradient method requires only $O(n)$ storage.

1. Introduction

In this paper, we consider the regularized minimization problem whose form is

$$\min_{x\in\Re^n} \quad F_\lambda(x) := f(x) + \lambda P(x), \tag{1}$$

where $\lambda > 0$, $P : \Re^n \to (-\infty, \infty]$ is a proper, convex, lower semicontinuous (lsc) function [20], and

$$f(x) := \sum_{i=1}^{m} f_i(x), \tag{2}$$

where each function f_i is real-valued and smooth (i.e., continuously differentiable) on an open subset of $\Re^n$ containing $\mathrm{dom}P = \{x \mid P(x) < \infty\}$.

The minimization problem (1) we consider arises in many applications such as (supervised) learning [7, 12, 27], regression [17, 23], neural network training [11, 21, 29], and data mining/classification [5, 15, 22, 28]. For the ℓ_1-regularized linear least squares problem [6, 23],

$$f_i(x) = \frac{1}{2}(a_i^T x - b_i)^2, \quad P(x) = \lambda\|x\|_1,$$

Received April 9, 2015; Revised October 8, 2015.
2010 *Mathematics Subject Classification.* Primary 49M27, 49M37, 65K05, 90C25, 90C30.
Key words and phrases. incremental gradient method, nonsmooth, regularization, running average.
This work was financially supported by Basic Science Research Program through NRF funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A2056038).

where $a_i \in \Re^n$, $b_i \in \Re$, and $\lambda > 0$. In this problem, f can be interpreted as a linear model under Gaussian errors on b. For the ℓ_1-regularized logistic regression problem [15]:

$$f_i(x) = \frac{1}{m}\log(1+\exp(-(a_i^T x_{1:n-1} + b_i x_n)), \quad P(x) = \lambda\|x_{1:n-1}\|_1, \tag{3}$$

where $x_{1:n-1} = (x_1, \ldots, x_{n-1})^T$, $a_i = b_i z_i$ with $(z_i, b_i) \in \Re^{n-1} \times \{-1, 1\}$, and $\lambda > 0$. As is done in compressed sensing, lasso, and group lasso, a nonsmooth regularization term $P(x)$, such as the 1-norm, is added to avoid over-fitting and/or induces a sparse representation; see [6, 8, 9, 23, 25, 31] and references therein. Another important problem of the form (1) is L2-loss support vector regression [12]:

$$f_i(x) = \max\left(|a_i^T x - b_i| - \epsilon, 0\right), \quad P(x) = \frac{\lambda}{2}\|x\|_2^2,$$

where $a_i \in \Re^n$, $b_i \in \Re$, and $\epsilon, \lambda > 0$. Note that a_i and b_i are a given set of (observed or training) data.

In many applications, the number of functions m is large, say, more than 10^4. In this case, traditional gradient based algorithms would be inefficient since they require evaluating $\nabla f_i(x)$ for all i before x is updated. In contrast, incremental gradient methods update x after $\nabla f_i(x)$ is evaluated for only one or a few i. In the unconstrained case, i.e., $P \equiv 0$, the classical incremental gradient method has the following basic form

$$x^{k+1} = x^k + \alpha_k \nabla f_{i_k}(x^k), \quad k = 0, 1, \ldots, \tag{4}$$

where i_k is chosen to cycle through $1, \ldots, m$ (i.e., $i_0 = 1, i_1 = 2, \ldots, i_{m-1} = m, i_m = 1, \ldots$) and $\alpha_k > 0$. To guarantee global convergence of them, the stepsize requires to diminish to zero. This can lead to slow convergence; see [1, 10, 16, 18, 29, 30]. Moreover, its extension to the nonsmooth regularized minimization problem is hard.

Recently, Tseng and Yun [26] proposed incremental gradient methods to solve the problem (1), i.e., the (nonsmooth) regularized minimization problem. The method proposed in [26] has the following form

$$g^k = \sum_{i=1}^m \nabla f_i(x^{\tau_i^k}), \tag{5}$$

$$d^k = \arg\min_{d \in \Re^n} \left\{ \langle g^k, d\rangle + \frac{1}{2}\langle d, H^k d\rangle + \lambda P(x^k + d) \right\}, \tag{6}$$

$$x^{k+1} = x^k + \alpha_k d^k, \tag{7}$$

where $\tau_i^k \le k$ for $i = 1, \ldots, m$, $H^k \succ 0_n$, $\alpha_k \in (0, 1]$, and $x^0, x^{-1}, \ldots$ in $\mathrm{dom} P$ are given. In particular, when there is no regularization term, i.e., $P \equiv 0$,

(6)–(7) with $H^k = I$ and $\alpha_k = \alpha$, and

$$\tau_i^k = \begin{cases} k & \text{if } i = (k \bmod m) + 1; \\ \tau_i^{k-1} & \text{otherwise,} \end{cases} \quad 1 \le i \le m,\ k \ge m, \tag{8}$$

reduces to the incremental gradient method proposed by Blatt et al. [4] which has the form

$$g^k = g^{k-1} + \nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^{k-m}),$$
$$x^{k+1} = x^k - \alpha_k g^k.$$

The algorithm (5)–(7) is a more general method and the gradient components can be partially asynchronously updated [3, 24]. Hence the incremental gradient method in [26] has several advantages over the classical incremental gradient method (4). But, this incremental gradient method requires $O(mn)$ storage, which is expensive when m is large. This is the main drawback of it.

In this paper, we propose a new incremental gradient (IG) method that, instead of storing a past gradient of f_i for each i, uses a running average of *all* past gradients. This has the advantage of using only $O(n)$ storage. Specifically, the proposed IG method replaces (5) by

$$g^k = \frac{k}{k+1} g^{k-1} + \frac{m}{k+1} \nabla f_{i_k}(x^k) \quad \text{with} \quad i_k = (k \bmod m) + 1, \tag{9}$$

with $g^{-1} = 0$. This gradient update is also used in a subgradient averaging method of Nesterov [19, Section 6] and its extension by Xiao [31] (also see [13]) for convex stochastic optimization of the form (1) with f being the expectation of convex functions parametrized by a random variable. In contrast to the IG method proposed in [26], the proposed IG method can choose i_k randomly.

In our notation, $\Re^n$ denotes the space of n-dimensional real column vectors, ^{T} denotes transpose. For any $x \in \Re^n$, x_j denotes the jth component of x, and $\|x\|_p = \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}$ for $1 \le p < \infty$ and $\|x\|_\infty = \max_j |x_j|$. For simplicity, we write $\|x\| = \|x\|_2$. For any $x, y \in \Re^n$, $\langle x, y\rangle = x^T y$ (so $\|x\| = \sqrt{\langle x, x\rangle}$. For $n \times n$ real symmetric matrices A, B, we write $A \succeq B$ (respectively, $A \succ B$) to mean that $A - B$ is positive semidefinite (respectively, positive definite). We denote by I the identity matrix and by 0_n the $n \times n$ matrix of zero entries. Unless otherwise specified, $\{x^k\}$ denotes the sequence $x^0, x^1, \ldots$.

2. Memory efficient incremental gradient method

In this section we describe the proposed IG method in which g^k is updated by a weighted average of past component gradients (9) and give some lemmas that will be used in our convergence analysis, i.e., Theorem 3.1.

We make the following standard assumptions about functions $f_1, \ldots, f_m$:

Assumption 1.

$$\|\nabla f_i(y) - \nabla f_i(z)\| \le L_i \|y - z\| \quad \forall y,\ z \in \mathrm{dom} P, \tag{10}$$

for some $L_i \geq 0$, $i = 1, \ldots, m$. Let $L = \sum_{i=1}^m L_i$.

The following assumptions are required for global convergence analysis of the proposed method.

Assumption 2. $\underline{\sigma} I \preceq H^k$ for all k, where $0 < \underline{\sigma}$.

Convergence of the proposed method requires $g^k - \nabla f(x^k) \to 0$. To ensure this, the stepsize α_k needs to be chosen carefully.

Assumption 3. (a) $\displaystyle\sum_{k=0}^{\infty} \alpha_k = \infty$.

(b) $\displaystyle\lim_{\ell\to\infty} \sum_{j=0}^{\ell} \frac{j+1}{\ell+1}\delta_j = 0$, where $\displaystyle\delta_j := \max_{i=0,1,\ldots,m} \|x^{k+i} - x^{k+m}\|\Big|_{k=jm-1}$ $(x^{-1} = x^0)$.

Assumption 3(b) is satisfied by, for example, the adaptive stepsize rule

$$\alpha_{k+i} = \min\left\{1, \frac{\phi(j+1)}{(j+1)\|d^{k+i}\|}\right\}\Big|_{k=jm-1}, \quad 0 \leq i < m,\ j = 0, 1, \ldots, \tag{11}$$

where $\phi : [1, \infty) \to (0, \infty)$ is continuous, decreasing, and $\lim_{t\to\infty} \phi(t) = 0$. This is because, by (11), $\delta_j \leq m\frac{\phi(j+1)}{j+1}$. Also, for any $\epsilon > 0$, there exists $\bar{j}$ such that $\phi(j+1) \leq \epsilon$ for all $j > \bar{j}$. Thus

$$\sum_{j=0}^{\ell} \frac{j+1}{\ell+1}\delta_j \leq m\sum_{j=0}^{\ell} \frac{\phi(j+1)}{\ell+1} \leq m\left(\sum_{j=0}^{\bar{j}} \frac{\phi(j+1)}{\ell+1} + \frac{\ell-\bar{j}}{\ell+1}\epsilon\right) \to m\epsilon \ \text{ as } \ell \to \infty.$$

If in addition $\int_1^\infty \frac{\phi(t)}{t} dt = \infty$ and $\{d^k\}$ is bounded (such as when dom$P$ is bounded), then Assumption 3(a) holds as well. Examples of such ϕ include $\phi(t) = \frac{1}{\ln t}$ and $\phi(t) = \frac{1}{\ln t \ln(\ln t)}$.

Now, we formally describe our proposed method below.

Algorithm 1 Memory Efficient Incremental Gradient Method

Choose $x^0 \in$ dom$P$ and set $g^{-1} = 0$. For $k = 0, 1, 2, \ldots$, generate x^{k+1} from x^k according to the following iteration:

Step 1: Choose $H^k \succ 0_n$.

Step 2: Solve

$$d^k = \arg\min_{d\in\Re^n} \left\{\langle g^k, d\rangle + \frac{1}{2}\langle d, H^k d\rangle + \lambda P(x^k + d)\right\}$$

with $g^k = \frac{k}{k+1}g^{k-1} + \frac{m}{k+1}\nabla f_{i_k}(x^k)$ with $i_k = (k \bmod m) + 1$.

Step 3: Set $x^{k+1} = x^k + \alpha_k d^k$ with $\alpha_k \in (0, 1]$.

In what follows, for any $x \in \mathrm{dom}P$, $g \in \Re^n$, and $H \succ 0_n$, we denote

$$d_H^g(x) \;:=\; \arg\min_d \left\{ \langle g, d\rangle + \frac{1}{2}\langle d, Hd\rangle + \lambda P(x+d) \right\}.$$

Thus $d^k = d_{H^k}^{g^k}(x^k)$. For H diagonal and P separable piecewise-linear/quadratic, $d_H^g(x)$ is computable in closed form. Some examples are given below.

1. For $P(x) = \|x\|_1$, $d_H^g(x)_j = -\mathrm{mid}\{\frac{g_j-\lambda}{H_{jj}}, x_j, \frac{g_j+\lambda}{H_{jj}}\}$.
2. For $P(x) = \|x\|_1 + \frac{\omega}{2}\|x\|^2$ $(\omega > 0)$ [9],
$$d_H^g(x)_j = -\mathrm{mid}\left\{ \frac{g_j - \lambda + \lambda\omega}{H_{jj}+\lambda\omega}, x_j, \frac{g_j+\lambda+\lambda\omega}{H_{jj}+\lambda\omega} \right\}.$$
3. For $P(x) = \|x\|_1 + \iota_B(x)$, where $\iota_B(x)$ is the indicator function of $B = \{x \mid \ell \le x \le u\}$ with $\ell \le u$ (possibly with $-\infty$ or ∞ components),
$$d_H^g(x)_j = -\mathrm{mid}\left\{ \ell_j - x_j, -\mathrm{mid}\{\frac{g_j-\lambda}{H_{jj}}, x_j, \frac{g_j+\lambda}{H_{jj}}\}, u_j - x_j \right\},$$

where $\mathrm{mid}\{a,b,c\}$ denotes the median (mid-point) of a, b, c.

We have the following lemma, whose proof is identical to that of [25, Eq. (8)] and is thus omitted.

Lemma 2.1. *For any $x \in \mathrm{dom}P$, $g \in \Re^n$, and $H \succ 0_n$, let $d = d_H^g(x)$. Then*

$$\langle g, d\rangle + \lambda P(x+d) - \lambda P(x) \le -\langle d, Hd\rangle.$$

We say that $x \in \Re^n$ is *a stationary point* of F_λ if $x \in \mathrm{dom}P$ and $F_\lambda{}'(x; d) \ge 0$ for all $d \in \Re^n$. The following result from [25, Lemma 2] characterizes stationarity in terms of $d_H^{\nabla f(x)}(x)$.

Lemma 2.2. *For any $H \succ 0_n$, an $x \in \mathrm{dom}P$ is a stationary point of F_λ if and only if $d_H^{\nabla f(x)}(x) = 0$.*

3. Convergence analysis

In this section, we analyze convergence properties of our proposed IG method under Assumptions 1–3. The proof uses Lemmas 2.1 and 2.2.

Theorem 3.1. *Let $\{x^k\}$, $\{d^k\}$, $\{H^k\}$, $\{\alpha_k\}$ be sequences generated by Algorithm 1 under Assumptions 1, 2 and 3. Then the following results hold.*

(a) $\{\|x^{k+1} - x^k\|\} \to 0$ *and* $\{\|\nabla f(x^k) - g^k\|\} \to 0$.
(b) $\liminf_{k\to\infty} \|d^k\| = 0$.
(c) *If $\{x^k\}$ is bounded, then there exists a cluster point of $\{x^k\}$ that is a stationary point of* (1).

Proof. (a) Assumption 3(b) implies $\{\delta_j\} \to 0$, so $\{\|x^k - x^{k+1}\|\} \to 0$. For each $j \in \{0, 1, \dots\}$, letting $k = jm - 1$, we have $i_{k+1} = 1, i_{k+2} = 2, \dots, i_{k+m} = m$, and hence

$$g^{k+m} = \frac{k+m}{k+m+1} g^{k+m-1} + \frac{m}{k+m+1} \nabla f_m(x^{k+m})$$

$$
\begin{aligned}
&= \frac{k+m}{k+m+1}\left(\frac{k+m-1}{k+m}g^{k+m-2} + \frac{m}{k+m}\nabla f_{m-1}(x^{k+m-1})\right)\\
&\quad + \frac{m}{k+m+1}\nabla f_m(x^{k+m})\\
&= \frac{k+m-1}{k+m+1}g^{k+m-2} + \frac{m}{k+m+1}\nabla f_{m-1}(x^{k+m-1})\\
&\quad + \frac{m}{k+m+1}\nabla f_m(x^{k+m})\\
&= \frac{k+1}{k+m+1}g^{k} + \frac{m}{k+m+1}\sum_{i=1}^{m}\nabla f_i(x^{k+i})\\
&= \frac{j}{j+1}g^{k} + \frac{1}{j+1}\sum_{i=1}^{m}\nabla f_i(x^{k+i}).
\end{aligned}
$$

Thus, upon letting

$$e_j := \|g^k - \nabla f(x^k)\| \,\big|_{k=jm-1}$$

and using $k = jm-1$, we have

$$
\begin{aligned}
e_{j+1} &= \left\|g^{k+m} - \nabla f(x^{k+m})\right\|\\
&= \left\|\frac{j}{j+1}(g^k - \nabla f(x^{k+m})) + \frac{1}{j+1}\sum_{i=1}^{m}(\nabla f_i(x^{k+i}) - \nabla f_i(x^{k+m}))\right\|\\
&\le \frac{j}{j+1}\left\|g^k - \nabla f(x^{k+m})\right\| + \frac{1}{j+1}\sum_{i=1}^{m}\|\nabla f_i(x^{k+i}) - \nabla f_i(x^{k+m})\|\\
&\le \frac{j}{j+1}(\|g^k - \nabla f(x^k)\| + L\|x^k - x^{k+m}\|) + \frac{1}{j+1}\sum_{i=1}^{m}L_i\|x^{k+i} - x^{k+m}\|\\
&\le \frac{j}{j+1}(e_j + L\delta_j) + \frac{1}{j+1}L\delta_j\\
&= \frac{j}{j+1}e_j + L\delta_j,
\end{aligned}
$$

where the second inequality uses (10) and Assumption 1. Propagating this recursion backwards yields

$$e_{j+1} \le L\left(\frac{1}{j+1}\delta_0 + \frac{2}{j+1}\delta_1 + \cdots + \frac{j}{j+1}\delta_{j-1} + \delta_j\right).$$

Under Assumption 3(b), the right-hand side tends to zero as $j \to \infty$. This shows that $\{e_j\} \to 0$.

By replacing "$k = jm-1$" in the above argument with "$k = jm-1+\nu$", where $\nu \in \{1, \dots, m-1\}$, we obtain

$$\|x^k - x^{k+m}\| \le \|x^k - x^{k+m-\nu}\| + \|x^{k+m-\nu} - x^{k+m}\| \le \delta_j + 2\delta_{j+1}$$

and similarly $\|x^{k+i} - x^{k+m}\| \le \delta_j + 2\delta_{j+1}$, so a similar argument yields

$$e_{j+1} \le \frac{j + \frac{\nu}{m}}{j + 1 + \frac{\nu}{m}} e_j + L(\delta_j + 2\delta_{j+1}) \le \frac{j+1}{j+2} e_j + L(\delta_j + 2\delta_{j+1})$$

and then $\{e_j\} \to 0$. Since the choice of ν was arbitrary, this proves that $\{\|\nabla f(x^k) - g^k\|\} \to 0$.

(b) Let

$$\Delta_k = \langle g^k, d^k \rangle + \lambda P(x^k + d^k) - \lambda P(x^k).$$

Then, for each $k \in \{0, 1, \dots\}$,

$$\begin{aligned} F_\lambda(x^{k+1}) - F_\lambda(x^k) &= F_\lambda(x^k + \alpha_k d^k) - F_\lambda(x^k) \\ &= f(x^k + \alpha d^k) - f(x^k) + \lambda P(x^k + \alpha d^k) - \lambda P(x^k) \\ &\le \alpha \langle \nabla f(x^k), d^k \rangle + \alpha^2 \frac{L}{2} \|d^k\|^2 + \alpha(\lambda P(x^k + d^k) - \lambda P(x^k)) \\ &\le \alpha_k \langle \nabla f(x^k) - g^k, d^k \rangle + \alpha_k^2 \frac{L}{2} \|d^k\|^2 + \alpha_k \Delta_k \\ &\le \alpha_k \|\nabla f(x^k) - g^k\| \|d^k\| + \alpha_k^2 \frac{L}{2} \|d^k\|^2 - \alpha_k \underline{\sigma} \|d^k\|^2 \\ &= -\alpha_k \left(\underline{\sigma} \|d^k\| - \|\nabla f(x^k) - g^k\| - \frac{L}{2} \|x^{k+1} - x^k\| \right) \|d^k\|, \end{aligned} \tag{12}$$

where the first inequality uses the convexity of P, $\alpha \in (0, 1]$, and the Lipschitz continuity of ∇f on $\mathrm{dom} P$ [2, page 667] and the third inequality uses $\Delta_k \le -\langle d^k, H^k d^k \rangle \le -\underline{\sigma} \|d^k\|^2$ (see Lemma 2.1 and Assumption 2).

We argue $\liminf_{k\to\infty} \|d^k\| = 0$ by contradiction. Suppose the contrary, so that there exists an $\epsilon > 0$ such that $\|d^k\| \ge \epsilon$ for all k. By (a), we have $\{\|\nabla f(x^k) - g^k\|\} \to 0$ and $\{\|x^{k+1} - x^k\|\} \to 0$ and hence there exists an integer $\bar{k}$ such that

$$\|\nabla f(x^k) - g^k\| \le \frac{1}{4} \underline{\sigma} \|d^k\|, \qquad \frac{L}{2} \|x^{k+1} - x^k\| \le \frac{1}{4} \underline{\sigma} \|d^k\| \quad \forall k \ge \bar{k}.$$

Then (12) yields

$$F_\lambda(x^{k+1}) - F_\lambda(x^k) \le -\alpha_k \underline{\sigma} \frac{\|d^k\|^2}{2} \le -\alpha_k \underline{\sigma} \frac{\epsilon^2}{2} \qquad \forall k \ge \bar{k},$$

so that

$$\lim_{k\to\infty} F_\lambda(x^{k+1}) \le F_\lambda(x^{\bar{k}}) - \sum_{k=\bar{k}}^{\infty} \alpha_k \underline{\sigma} \frac{\epsilon^2}{2} = -\infty,$$

where the equality is due to Assumption 3(a). This contradicts $\inf F_\lambda > -\infty$.

(c) Suppose $\{x^k\}$ is bounded. By (b), $\{d^k\}_{k\in K} \to 0$ for some $K \subseteq \{0, 1, \dots\}$. Since $\{\|\nabla f(x^k) - g^k\|\} \to 0$, for every limit point $\bar{x}$ of a convergent subsequence $\{x^k\}_{k\in\tilde{K}\subseteq K}$, $\{g^k\}_{k\in\tilde{K}} \to \nabla f(\bar{x})$.

Then (6) implies that, for any $x \in \mathrm{dom} P$, we have

$$\langle g^k, d^k \rangle + \frac{1}{2} \langle d^k, H^k d^k \rangle + \lambda P(x^k + d^k) \le \langle g^k, x - x^k \rangle + \frac{1}{2} \langle x - x^k, H^k (x - x^k) \rangle + \lambda P(x)$$

for all $k \in \tilde{K}$, so the lsc property of P yields in the limit that

$$\lambda P(\bar{x}) \leq \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2}\langle x - \bar{x}, \bar{H}(x - \bar{x})\rangle + \lambda P(x) \quad \forall x \in \mathrm{dom} P,$$

where $\bar{H}$ is any cluster point of $\{H^k\}_{k\in\tilde{K}}$. Since $H^k \succeq \underline{\sigma} I$ for all $k \in \tilde{K}$, $\bar{H} \succ 0_n$. This shows that $d_{\bar{H}}(\bar{x}) = 0$ so that, by Lemma 2.2, $\bar{x}$ is a stationary point of (1). Hence every cluster point of $\{x^k\}_{k\in K}$ is a stationary point of (1). □

4. Numerical experiments

In this section we support Theorem 3.1 by numerical experiments. We apply our proposed method to solve the ℓ_1-regularized logistic regression problem (3) on randomly generated data. Note that we test our proposed method only on small size problems to show it works well.

Here, we assume that there are $j, k \in \{1, \ldots, m\}$ such that $b_j = 1$ and $b_k = -1$. If we take $x^0 = 0$, $\lambda\|x_{1:n-1}\|_1 \leq \log 2$ for all x with $x \in \mathcal{X}_0 := \{x \mid F_\lambda(x) \leq F_\lambda(x^0)\}$ since $\log(1+\exp(-(a_i^T x_{1:n-1} + b_i x_n))) > 0$ for all $x \in \Re^n$ and $i = 1, \ldots, m$. Hence $|x_i| \leq \frac{\log 2}{\lambda}$ for $i = 1, \ldots, n-1$. Also, since $\lambda\|x_{1:n-1}\|_1 \geq 0$, $\log(1 + \exp(-(a_i^T x_{1:n-1} + b_i x_n))) \leq m \log 2$ for all $x \in \mathcal{X}_0$ and $i = 1, \ldots, m$. Hence, for each i, $\exp(-(a_i^T x_{1:n-1} + b_i x_n)) \leq 2^m - 1$. This together with the assumption on the choice of b and the boundedness of $x_{1:n-1}$ with $x \in \mathcal{X}_0$ implies that

$$-m \log 2 - \frac{\log 2}{\lambda} \min_{b_j>0} \|a_i\|_1 \leq x_n \leq m \log 2 + \frac{\log 2}{\lambda} \min_{b_j<0} \|a_i\|_1. \tag{13}$$

Therefore $\mathcal{X}_0$ with $x^0 = 0$ is bounded. Optimal solutions are contained in $\mathcal{X}_0$. Hence we apply our proposed method to the following bounded ℓ_1-regularized logistic regression problem:

$$\min_x \sum_{i=1}^m \frac{1}{m} \log(1 + \exp(-(a_i^T x_{1:n-1} + b_i x_n)) + \lambda\|x_{1:n-1}\|_1 + \iota_B(x), \tag{14}$$

where $B = \left\{x \mid |x_i| \leq \frac{\log 2}{\lambda} \text{ for } i = 1, \ldots, n-1,\ x_n \text{ satisfies (13)}\right\}$. Since B is bounded, this implies that Assumption 3 with (11) and $\phi(t) = \frac{1}{\ln t}$ is satisfied.

Our proposed method is implemented as follows. We choose $H^k = I$. The stepsize α_k is chosen by the rule (11) with $\phi(t) = \frac{1}{\ln t}$.

We stop the algorithm when the relative error of the iterates satisfies the following condition:

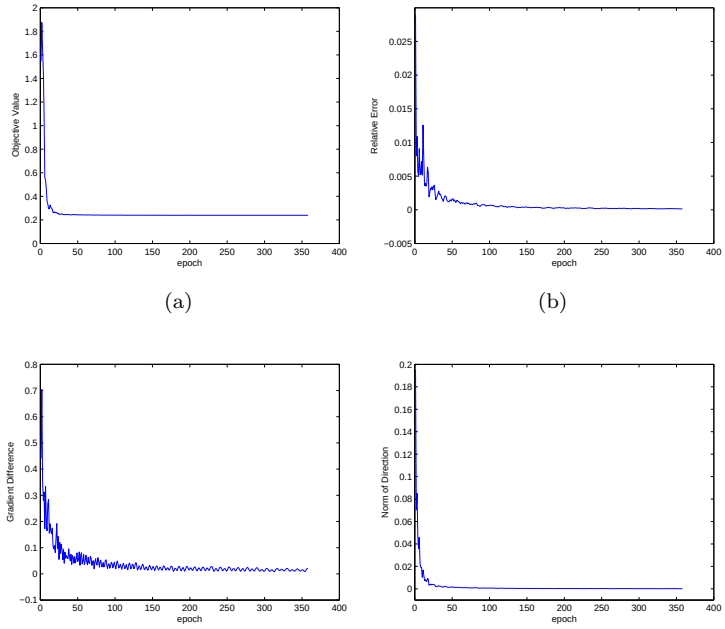
$$\frac{\|x^k - x^{k-1}\|}{\max\{1, \|x^k\|\}} \leq \mathrm{Tol}, \tag{15}$$

where Tol is a moderately small tolerance.

All runs are performed on a Desktop with Intel Core i7-3770 CPU (3.40GHz) and 8GB Memory, running 64-bit windows 8.1 and MATLAB (Version 8.3). Throughout the experiments, we choose the initial iterate to be $x^0 = 0$.

TABLE 1. Test results with ten random data sets with $m = 100$ and $n = 101$ with Tol $= 10^{-4}$.

	$\lambda = 0.0473272$	$\lambda = 0.0474351$	$\lambda = 0.0494608$	$\lambda = 0.0543597$	$\lambda = 0.0516248$
iters	35508	44336	33911	53836	28049
obj	0.2393	0.2308	0.2365	0.2453	0.2586
	$\lambda = 0.0503979$	$\lambda = 0.0503678$	$\lambda = 0.0527553$	$\lambda = 0.0529956$	$\lambda = 0.0527555$
iters	36393	35306	52595	28196	38182
obj	0.2402	0.2341	0.2369	0.2368	0.2313



(a) (b)

(c) (d)

FIGURE 1. (a) Objective values versus epoch (one epoch is m iterations). It shows that the objective values eventually decrease. (b) Relative errors versus epoch. It shows that the relative errors eventually converge to zero. (c) The differences between the gradient of f at x^k and the approximated gradient g^k versus epoch. It shows that the differences eventually converge to zero. (d) The norm of directions versus epoch. It shows that the norm of directions also eventually converge to zero. Note that these are numerical results when $\lambda = 0.0473272$.

Table 1 reports the number of iterations and the final objective value for the bounded ℓ_1-regularized logistic regression problem (14) on ten randomly generated data with size $m = 100$ and $n = 101$. As suggested in [14], each randomly generated problem has an equal number of positive and negative data points. Features of positive (negative) points are independent and identically distributed, drawn from a normal distribution $\mathcal{N}(\xi, 1)$, where ξ is in turn drawn from a uniform distribution on $[0, 1]([-1, 0])$. For each instance, we chose $\lambda = 0.1\lambda_{\max}$ where $\lambda_{\max} = \frac{1}{m}\|\frac{m_-}{m}\sum_{b_i=1} a_i + \frac{m_+}{m}\sum_{b_i=-1} a_i\|_\infty$, m_- is the number of negative points, and m_+ is the number of positive points.

Figure 1(a) shows that the objective values eventually decrease and Figure 1(b)-(d) show that the relative errors, the differences between the gradient of f at x^k and the approximated gradient g^k, the norm of directions converge to zero, respectively. Note that similar performance is observed in other randomly generated data. Hence Figure 1 supports that Theorem 3.1 works well for the bounded ℓ_1-regularized logistic regression problem (14).

5. Conclusions and extensions

In this paper we have proposed the new incremental gradient method for minimizing the sum of smooth functions and a (possibly nonsmooth) convex function. The proposed method uses much less storage and so is a memory efficient method.

Our adaptive stepsize rule (11) is somewhat complicated. Can a diminishing stepsize which is used for classical incremental gradient methods or a constant stepsize be used? These stepsize rules are much simple to use. Can Theorem 3.1(c) be strengthened to show that every cluster point of $\{x^k\}$ is stationary? In this paper, we have tested our proposed method on small size problems only to show Theorem 3.1 works well. Hence, it will be interested in the comprehensive numerical study of the proposed method. These are some issues that need further investigation.

References

[1] D. P. Bertsekas, *A new class of incremental gradient methods for least squares problems*, SIAM J. Optim. **7** (1997), no. 4, 913–926.

[2] ———, *Nonlinear Programming*, **2**, Athena Scientific, Belmont, MA, 1999.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, 1989.

[4] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM J. Optim. **18** (2007), no. 1, 29–51.

[5] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian, *Mathematical programming for data mining: formulations and challenges*, INFORMS J. Comput. **11** (1999), no. 3, 217–238.

[6] S. Chen, D. Donoho, and M. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. **20** (1998), no. 1, 33–61.

[7] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, 2000.

[8] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math. **57** (2004), no. 11, 1413–1457.
[9] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized lienar models via coordinate descent*, Report, Department of Statistics, Stanford University, Stanford, May 2009.
[10] A. A. Gaivoronski, *Convergence properties of back-propagation for neural nets via theory of stochastic gradient methods. Part I*, Optim. Methods Softw. **4** (1994), 117–134.
[11] L. Grippo, *A class of unconstrained minimization methods for neural network training*, Optim. Methods Softw. **4** (1994), 135–150.
[12] C.-H. Ho and C.-J. Lin, *Large-scale linear support vector regression*, J. Mach. Learn. Res. **13** (2012), 3323–3348.
[13] A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro, *Stochastic approximation approach to stochastic programming*, SIAM J. Optim. **19** (2009), 1574–1609.
[14] K. Koh, S.-J. Kim, and S. Boyd, *An interior-point method for large-scale ℓ_1-regularized logistic regression*, J. Mach. Learn. Res. **8** (2007), 1519–1555.
[15] S. Lee, H. Lee, P. Abeel, and A. Ng, *Efficient ℓ_1-regularized logistic regression*, In Proceedings of the 21st National Conference on Artificial Intelligence, 2006.
[16] Z.-Q. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropagation algorithm*, Optim. Methods Softw. **4** (1994), 85–101.
[17] O. L. Mangasarian and D. R. Musicant, *Large scale kernel regression via linear programming*, Mach. Learn. **46** (2002), 255–269.
[18] O. L. Mangasarian and M. V. Solodov, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optim. Methods Softw. **4** (1994), 103–116.
[19] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program. **120** (2009), no. 1, 221–259.
[20] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.
[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation*, in Parallel Distributed Processing–Explorations in the Microstructure of Cognition, edited by Rumelhart and McClelland, 318–362, MIT press, Cambridge, 1986.
[22] S. Sardy and P. Tseng, *AMlet, RAMlet, and GAMlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets*, J. Comput. Graph. Statist. **13** (2004), no. 2, 283–309.
[23] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288.
[24] P. Tseng, *On the rate of convergence of a partially asynchronous gradient projection algorithm*, SIAM J. Optim. **1** (1991), no. 4, 603–619.
[25] P. Tseng, and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program. **117** (2009), no. 1-2, 387–423.
[26] ———, *Incrementally updated gradient methods for constrained and regularized optimization*, J. Optim. Theory Appl. **160** (2014), no. 3, 832–853.
[27] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 2000.
[28] L. Wang, *Efficient regularized solution path algorithms with applications in machine learning and data mining*, Ph.D thesis, University of Michigan, 2008.
[29] H. White, *Learning in artificial neural networks: a statistical perspective*, Neural Comput. **1** (1989), 425–464.
[30] ———, *Some asymptotic results for learning in single hidden–layer feedforward network models*, J. Amer. Statist. Assoc. **84** (1989), no. 408, 1003–1013.
[31] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, J. Mach. Learn. Res. **11** (2010), 2543–2596.

Department of Mathematics Education
Sungkyunkwan University
Seoul 03063, Korea
E-mail address: yswmathedu@skku.edu