Soft Computing
https://doi.org/10.1007/s00500-021-05695-4

FOCUS

On the convergence of a Block-Coordinate Incremental Gradient method

Laura Palagi[1] · Ruggiero Seccia[1]

Accepted: 13 February 2021

Abstract

In this paper, we study the convergence of a block-coordinate incremental gradient method. Under some specific assumptions on the objective function, we prove that the block-coordinate incremental gradient method can be seen as a gradient method with errors and convergence can be proved by showing the error at each iteration satisfies some standard conditions. Thus, we can prove convergence towards stationary points when the block incremental gradient method is coupled with a diminishing stepsize and towards an ϵ-approximate solution when a bounded away from zero stepsize is employed.

Keywords Incremental gradient · Block-coordinate decomposition · Online optimization

1 Introduction

In this paper, we consider a block-coordinate incremental gradient algorithm, hereafter called BIG, for minimizing a finite-sum function

$$\underset{\mathbf{w}\in\mathbf{R}^n}{\text{minimize}} \quad f(\mathbf{w}) = \sum_{h=1}^{H} f_h(\mathbf{w}). \tag{1}$$

and study its convergence when $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is a continuously differentiable function.

Problem (1) is a well-known optimization problem that arises in many practical applications including the regularized empirical risk minimization (ERM) where f_h represents the loss function of the $h-$th data block and constitutes a standard approach when training machine learning models (see e.g. Bertsekas 2011; Bottou et al. 2018; Goodfellow et al. 2016 and reference therein). We focus on the case where both the number of components H and the dimension of the space n are very large, which arises in machine learning training problems when tackling Big Data applications by means of over-parametrized models such as Deep Networks. Indeed, one of the main issues when solving problem (1) through standard batch methods, namely methods that use all the terms f_h at each iteration, is related to the high computational effort needed for each objective function and gradient evaluations. The per-iteration cost depends on the size of H and n, so that when both of them are large there is an incentive to use less expensive per-iteration methods that exploit the structure of the objective function to reduce the computational burden and avoid slowing down the optimization process.

In order to overcome this computational burden, problem (1) has been mainly tackled by means of online algorithms, namely methods that at each iteration k use one or more terms f_h in the objective function to compute an update of the current solution $\mathbf{w}^k$. The reason for the great success of online methods lies mainly in the different balance of per-iteration costs and expected per-iteration improvement in the objective function, particularly in the Big Data setting when the size of H becomes very large [see e.g. comments in Bottou et al. (2018)]. Online methods can be roughly distinguished in two kinds: incremental gradient (IG) methods where the order in which the elements f_h are considered is fixed *a priori* and not changed over the iterations; stochastic gradient (SG) methods where elements f_h are chosen according to some probability distribution. IG methods can be applied

Communicated by Marcello Sanguineti.

The authors were partially supported by the project *Advanced Optimization-based machine learning models for analytics on clinical data* of Sapienza n. RM11916B7FCB6893.

✉ Ruggiero Seccia
ruggiero.seccia@uniroma1.it

[1] Department of Computer, Control and Management Engineering A. Ruberti, Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy

Published online: 04 March 2021

only to finite-sum functions, while SG methods also apply to functions with infinite terms f_h (e.g. function representing expected values). Concerning the convergence theory, while incremental methods can be considered and analysed as deterministic methods, stochastic frameworks are usually analysed recurring to probabilistic analysis. The former method and its convergence have been deeply investigated in, e.g. Bertsekas (1996), Bertsekas (2015), Bertsekas and Tsitsiklis (2000) and Solodov (1998), while the latter in, e.g. Bertsekas and Tsitsiklis (2000), Bottou (2010) and Robbins and Monro (1951).

As pointed out in Palagi and Seccia (2020), even though online methods can effectively tackle optimization problems where the dimension of H is very large, they still suffer when the search space n becomes large as well, namely when the number of variables increases. It is often the case when dealing with applications where deep learning models are employed (e.g. image recognition applications) that the number of parameters to be estimated can go above hundreds of millions (Simonyan and Zisserman 2014). An efficient solution to tackle optimization problems with a large number of variables n is represented by *Block-Coordinate Descent* (BCD) methods, which update at each iteration only a subset of the whole variables, keeping the other fixed to the current value. By exploiting the structure of the objective function (e.g. fixing some variables makes the subproblem convex or allows for parallel updates) and thanks to the lower cost of calculating the block component of the gradient, these methods lend themselves well to efficient implementations and can greatly improve optimization performance and reduce the computational effort (see e.g. Bertsekas and Tsitsiklis 1989; Buzzi et al. 2001; Grippo et al. 2016; Palagi and Seccia 2020). Their convergence has already been analysed in many works with different assumptions on both the block selection rule and the properties of the update (Beck and Tetruashvili 2013; Bertsekas and Tsitsiklis 1989; Grippo and Sciandrone 1999; Lu and Xiao 2015; Nesterov 2012; Wright 2015).

In order to leverage both the sample decomposition with respect to the elements f_h composing the objective function, typical of online methods, and the block-wise decomposition with respect to the variables, typical of BCD frameworks, an effective solution is represented by block-coordinate online methods. Block-coordinate online methods aim to reduce the per-iteration costs by operating along a twofold line: updating only on a subset of the variables $\mathbf{w}$ as in BCD methods, and on a subset (i.e. mini-batch) of the components f_h as in online methods. The behaviour of block-coordinate online methods has already been investigated in Wang and Banerjee (2014) where the strongly convex case is considered and a geometric rate of convergence in expectation has been established. Moreover, in Zhao et al. (2014) and Chauhan et al. (2017) the effectiveness of this approach has already been tested in strongly convex sparse problems such as LASSO or sparse logistic regression, respectively. In Bravi and Sciandrone (2014), a two-block decomposition method is applied for training a neural network where the objective function is assumed to be convex with respect to one of the block components (the output weights) so that exact optimization can be used for the convex block update while the other block (hidden weights) is updated using an incremental gradient update. In Palagi and Seccia (2020), the layered structure of a deep neural network has been explored to define a block layer incremental gradient (BLInG) algorithm which uses an incremental approach for updating the weights over each single layer. Numerical effectiveness of embedding block-coordinate modifications in online frameworks has already been tested and turned out to be a promising approach (Chauhan et al. 2017; Palagi and Seccia 2020; Wang and Banerjee 2014; Zhao et al. 2014).

In this paper, we present a block-coordinate incremental gradient method (BIG), which generalizes the BLInG algorithm presented in Palagi and Seccia (2020) for the deep networks training problem, and we focus on its convergence analysis. BIG can be seen as a deterministic gradient method with errors, since the selection of both the elements f_h and the blocks of variables is fixed *a priori* and not changed over the iterations so that the algorithm can be analysed as a gradient method with deterministic errors. Thus, taking steps from the deterministic convergence results for gradient methods with errors reported in Bertsekas and Tsitsiklis (2000) and Solodov (1998), we prove convergence of BIG towards stationary points and to an ϵ-approximate solution, respectively, when a diminishing and a bounded away from zero stepsizes are employed. We do not report numerical results that can be found in Palagi and Seccia (2020) where the optimization problem arising in training deep neural networks is considered. Overall, the numerical results in Palagi and Seccia (2020) suggest the effectiveness of BIG in exploiting the finite-sum objective function and the inherent block layer structure of deep neural networks.

The paper is organized as follows: in Sect. 2, preliminary results on the convergence theory of gradient methods with errors are recalled and the convergence analysis of IG is provided following standard analysis of gradient methods with errors from Bertsekas and Tsitsiklis (2000) and Solodov (1998). In Sect. 3, we show how BIG can be regarded as a gradient method with errors and prove its convergence properties. In Sect. 4, we discuss numerical performance of BIG when compared with its non-decomposed counterpart IG. Finally, in Sect. 5 conclusions are drawn and in the "Appendix" supporting material is provided.

Notation. We use boldface for denoting vectors, e.g. $\mathbf{w} = (w_1, \dots, w_n)$ and $\| \cdot \|$ to denote the euclidean norm of a vector. Given a set of indexes $\ell \subseteq \{1, \dots, n\}$, we denote by $\mathbf{w}_\ell$ the subvector of $\mathbf{w}$ made up of the components $i \in \ell$, namely $\mathbf{w}_\ell = (w_i)_{i\in\ell} \in \mathbf{R}^{|\ell|}$. The gradient of the function is denoted by $\nabla f(\mathbf{w}) \in \mathbf{R}^n$ and, given a subvector $\mathbf{w}_\ell$ of $\mathbf{w}$, we use the short notation $\nabla_\ell f(\mathbf{w}) \in \mathbf{R}^{|\ell|}$ to denote the partial derivative with respect to the block w_ℓ, i.e. $\nabla_{\mathbf{w}_\ell} f(\mathbf{w})$.

Given a partition $\mathcal{L} = \{\ell_1, \dots, \ell_L\}$ of the indexes $\{1, \dots, n\}$, namely $\cup_{i=1}^L \ell_i = \{1, \dots, n\}$ and $\ell_i \cap \ell_j = \emptyset$ for all $i \neq j$, a vector $\mathbf{w}$ can be written, by reordering its components, as $\mathbf{w} = (\mathbf{w}_{\ell_1}, \dots, \mathbf{w}_{\ell_L})$ and correspondingly $\nabla f(\mathbf{w}) = (\nabla f(\mathbf{w})_{\ell_1} \dots, \nabla f(\mathbf{w})_{\ell_L})$. Further, we use the notation $[\cdot]_\ell$ to define a vector in $\mathbf{R}^n$ where all the components are set to zero except those corresponding to the block ℓ, namely given a vector $\mathbf{w} \in \mathbf{R}^n$ the vector $[\mathbf{w}]_\ell$ is defined component-wise as

$$([\mathbf{w}]_\ell)_k = \begin{cases} w_k & \text{if } k \in \ell \\ 0 & \text{otherwise.} \end{cases}$$

Thanks to this notation, we have $\mathbf{w} = \sum_{i=1}^L [\mathbf{w}]_{\ell_i}$ and $\nabla f(\mathbf{w}) = \sum_{i=1}^L [\nabla f(\mathbf{w})]_{\ell_i}$. Moreover note that $[\mathbf{w}]_{\ell_i} \in \mathbf{R}^n$ while $\mathbf{w}_{\ell_i} \in \mathbf{R}^{|\ell_i|}$.

2 Background on Gradient method with errors

In this section, we report two main results concerning convergence of gradient methods with errors which will be useful for proving the convergence properties of BIG in Sect. 3. In particular, we consider two results, one concerning the adoption of a diminishing stepsize and the other considering a bounded away from zero stepsize, respectively, from Bertsekas and Tsitsiklis (2000) and Solodov (1998). To the best of author knowledge, these two results are among the most significant, with the former being among the ones with the less restrictive assumptions when a diminishing stepsize is employed (as highlighted by the authors neither convexity of the function nor boundedness conditions on the function or the sequence generated $\{\mathbf{w}^k\}$ are required to prove convergence), and the latter being the first convergence result for incremental methods with bounded away from zero stepsizes. After having introduced and briefly discussed these two results, in the remainder of this section we recall their implications for the standard incremental gradient method.

In both the next two propositions, it is assumed that the function f is continuously differentiable with M-Lipschitz continuous gradient, that is

$$\|\nabla f(\mathbf{u}) - \nabla f(\mathbf{v})\| \leq M\|\mathbf{u} - \mathbf{v}\| \quad \forall \mathbf{u}, \mathbf{v} \in \mathbf{R}^n. \tag{2}$$

We start by considering the work done by Bertsekas and Tsitsiklis (2000) where a diminishing stepsize is considered. The main idea is that a gradient method with errors where the error is proportional to the stepsize converges to a stationary point provided that the stepsize goes to zero but not too fast.

Proposition 1 (Proposition 1 in Bertsekas and Tsitsiklis 2000) *Let f be continuously differentiable over $\mathbf{R}^n$ satisfying (2). Let $\{\mathbf{w}^k\}$ be a sequence generated by the method*

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha^k(\mathbf{d}^k + \mathbf{e}^k)$$

where $\mathbf{d}^k$ is a descent direction satisfying for some positive scalars c_1 and c_2 and all k,

$$c_1\|\nabla f(\mathbf{w}^k)\|^2 \leq -\nabla f(\mathbf{w}^k)^T \mathbf{d}^k \qquad \|\mathbf{d}^k\| \leq c_2(1 + \|\nabla f(\mathbf{w}^k)\|)$$

and $\mathbf{e}^k$ is an error vector satisfying for all k,

$$\|\mathbf{e}^k\| \leq \alpha^k(p + q\|\nabla f(\mathbf{w}^k)\|) \tag{3}$$

where p and q are positive scalars. Assume that the stepsize α^k is chosen according to a diminishing stepsize condition, that is

$$\sum_{k=0}^{\infty} \alpha^k = \infty \qquad \sum_{k=0}^{\infty} (\alpha^k)^2 < \infty. \tag{4}$$

Then either $lim_{k\to\infty} f(\mathbf{w}^k) = -\infty$ or $\{f(\mathbf{w}^k)\}$ converges to a finite value and $\lim_{k\to\infty} \nabla f(\mathbf{w}^k) = 0$. Furthermore every accumulation point of $\mathbf{w}^k$ is a stationary point of f.

On the other hand, when it comes to the case of stepsizes bounded away from zero, Solodov proves in Solodov (1998) that a gradient method with errors has at least an accumulation point $\bar{\mathbf{w}}$ that is an ϵ-approximate solution, with the tolerance value ϵ at least linearly depending on the limiting value of the stepsize $\bar{\alpha} > 0$.

Proposition 2 (Proposition 2.2 in Solodov 1998) *Let f be continuously differentiable over a bounded set D and let f satisfying condition (2). Let $\{\mathbf{w}^k\} \subset D$ be a sequence generated by*

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha^k(\nabla f(\mathbf{w}^k) - \mathbf{e}^k).$$

Assume $\lim_{k\to\infty} \alpha^k = \bar{\alpha} > 0$ with $\alpha^k \in (\theta, 2/M - \theta)$ with $\theta \in (0, 1/M]$, and

$$\|\mathbf{e}^k\| \leq \alpha^k \bar{B} \tag{5}$$

with $\bar{B} > 0$.

Then there exist a constant $C > 0$ (independent of $\bar{\alpha}$) and an accumulation point $\bar{\mathbf{w}}$ of the sequence $\left\{\mathbf{w}^k\right\}$ such that

$$\|\nabla f(\bar{\mathbf{w}})\| \leq C\bar{\alpha} \tag{6}$$

Furthermore, if the sequence $\left\{f\left(\mathbf{w}^k\right)\right\}$ *converges then every accumulation point* $\bar{\mathbf{w}}$ *of the sequence* $\left\{\mathbf{w}^k\right\}$ *satisfies* (6).

Comparing the hypothesis in Propositions 1 and 2, we have that the former result considers a *gradient related* direction while the latter is stated only with respect to the antigradient. Moreover, the former does not require the sequence $\{w^k\}$ to stay within a bounded set, thing that instead is needed by the latter Proposition. Finally, Proposition 2 makes a stronger assumption on the error term compared to Proposition 1, which, however, can be relaxed so to consider the same bound as in (3) (see Solodov 1998 and the discussion in the following Sect. 3.2).

2.1 Incremental Gradient as Gradient method with error

The incremental gradient framework updates the point $\mathbf{w}^k$ by moving along the gradient direction of one or few terms f_h, which are used in a fixed order. Once all the elements H composing the function in problem (1) have been considered, the outer iteration counter k is increased and the stepsize α^k is updated. The inner iteration starts with the current iterate $\mathbf{w}^k$, and it loops over the indexes $h = 1\dots, H$ using a fixed stepsize α^k; that is

$$\begin{aligned} &\mathbf{y}_0^k = \mathbf{w}^k \\ &\mathbf{y}_h^k = \mathbf{y}_{h-1}^k - \alpha^k \nabla f_h(\mathbf{y}_{h-1}^k) \qquad h = 1, \dots, H \\ &\mathbf{w}^{k+1} = \mathbf{y}_H^k \end{aligned} \tag{7}$$

Thus, an iteration of the IG method can be written as

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha^k \mathbf{d}^k, \tag{8}$$

with the direction $\mathbf{d}_k$ defined through the intermediate updates $\mathbf{y}_h^k$ defined as in (7), i.e.

$$\mathbf{d}^k = -\sum_{h=1}^{H} \nabla f_h(\mathbf{y}_{h-1}^k). \tag{9}$$

For the sake of notation, in the following we do not report explicitly the number of terms f_h used to determine the direction, but without loss of generality we assume that the index h can represent either one index or a set of indexes. In both of the two cases, the same arguments directly apply. A general scheme of an IG method is reported in Algorithm 1.

Convergence of IG has been proved both in the case a diminishing stepsize and a bounded away from zero stepsize is employed, respectively, in Bertsekas and Tsitsiklis (2000) and Solodov (1998), by showing that it satisfies the assumptions of Propositions 1 and 2. Since we follow a similar approach to prove convergence of BIG in the next section, below we report the main convergence result for the IG method when a diminishing stepsize is employed.

Algorithm 1 Incremental Gradient (IG)

```
Choose $\mathbf{w}^0 \in \mathbf{R}^n$, $\alpha^0 > 0$ and set $k = 0$;
while (stopping criterion not met) do
    Set $\mathbf{y}_0^k = \mathbf{w}^k$;
    for h=1,…,H do
        $\mathbf{y}_h^k = \mathbf{y}_{h-1}^k - \alpha^k \nabla f_h(\mathbf{y}_{h-1}^k)$
    end for
    $\mathbf{w}^{k+1} = \mathbf{y}_H^k$
    Update $\alpha^k$
    $k = k + 1$
end while
```

Proposition 3 (Proposition 2 in Bertsekas and Tsitsiklis 2000) *Let* $\{\mathbf{w}^k\}$ *be a sequence generated by* (7), (8) *and* (9). *Assume that for all* $h = 1, \dots, H$ *there exist positive constants* M, a, b *such that*

$$\|\nabla f_h(\mathbf{u}) - \nabla f_h(\mathbf{w})\| \le M\|\mathbf{u} - \mathbf{w}\| \qquad \forall\, \mathbf{u}, \mathbf{w} \in \mathbf{R}^n \tag{10}$$

$$\|\nabla f_h(\mathbf{w})\| \le a + b\|\nabla f(\mathbf{w})\| \qquad \forall\, \mathbf{w} \in \mathbf{R}^n. \tag{11}$$

Then the direction defined in (9) *can be written as*

$$\mathbf{d}^k = -\nabla f(\mathbf{w}^k) + \mathbf{e}^k$$

with $\mathbf{e}^k$ *satisfying for some positive constants* p, q

$$\|\mathbf{e}^k\| \le \alpha^k(p + q\|\nabla f(\mathbf{w}^k)\|).$$

Proposition 3 states that an IG iteration satisfies the hypothesis of Proposition 1 and convergence follows. Namely, it shows that IG can be viewed as a batch gradient method where the gradient is perturbed by an error term that is proportional to the stepsize. Thus, roughly speaking, driving the stepsize to zero will drive the error to zero as well, allowing to prove convergence. The proof of Proposition 3 provided in Bertsekas and Tsitsiklis (2000) is only shown for the case of $H = 2$, for reasons of simplicity. For the sake of completeness and to help the reader with the following convergence result, we provide in "Appendix A" the proof of Proposition 3 in the more generic case of any number of elements H. Finally, we note that condition (10) could be stated using a different Lipschitz constant for each h. However, for the sake of simplicity, we omit this detail.

We do not report the convergence result of IG in case a bounded away from zero stepsize is applied since it is not useful for proving convergence of BIG in case of a bounded away from zero stepsize. (Its convergence can be proved by following a similar reasoning to the one applied in the following Proposition 4). However, we remark that in Solodov (1998) to ensure that the error term in IG satisfies assumption (5) it is assumed that each ∇f_h is Lipschitz continuous and that the norm of each ∇f_h is bounded above by some positive constant [cfr Proposition 2.1 in Solodov (1998)].

3 The block-coordinate incremental gradient method

As already discussed in the introduction, incremental methods rely on the reduction of the complexity of a single iteration by exploiting the sum in the objective function. However, they still suffer when the dimension of the space n is large. On the other hand, BCD methods resort to simpler optimization problems by working only on a subset of variables.

Given a partition $\mathcal{L} = \{\ell_1, \dots, \ell_L\}$ of the indexes $\{1, \dots, n\}$ with $w_{\ell_i} \in \mathbf{R}^{N_i}$ and $\sum_{i=1}^{L} N_i = n$, a standard BCD method selects at a generic iteration one block ℓ_i (we omit the possible dependence on k) and updates only the block $\mathbf{w}^k_{\ell_i}$ while keeping all the other blocks fixed at the current iteration, i.e. $\mathbf{w}^{k+1}_{\ell_j} = \mathbf{w}^k_{\ell_j}$ for $j \neq i$. By fixing some variables, the obtained subproblem, besides being smaller, might have a special structure in the remaining variables that can be conveniently exploited. Further, these methods might allow a distributed/parallel implementation that can speed up the overall process (Bertsekas and Tsitsiklis 1989; Grippo and Sciandrone 1999; Wright 2015).

In order to leverage the structure of the objective function and mitigate the influence of both the number of variables n and the number of terms H, a solution is to embed the online framework into a block-coordinate decomposition scheme. Following this idea, the block-coordinate incremental gradient (BIG) method proposed here consists in updating each block of variables $\mathbf{w}_{\ell_j}$ using only one or a few terms f_h of the objective function. As done for the presentation of the IG methods, for the sake of notation, we do not report explicitly the number of terms f_h used in the updating rule, and without loss of generality we assume that the index h represents either one single term or a batch of terms. All the next arguments apply with only slight changes in the notation.

More formally, given a partition $\mathcal{L} = \{\ell_1, \dots, \ell_L\}$ of the indexes $\{1, \dots, n\}$, the BIG method selects a term $h \in \{1, \dots, H\}$ and updates all the blocks $\mathbf{w}_{\ell_j}$ sequentially with $j = 1, \dots, L$ by moving with a fixed stepsize along the gradient of f_h evaluated in successive points. Once all the elements H have been selected, the outer iteration counter k is increased and the stepsize α^k is updated. Similarly to the IG method, the BIG iteration from $\mathbf{w}^k$ to $\mathbf{w}^{k+1}$ can be described by using vectors $\mathbf{y}^k_{h,j}$ obtained in the inner iterations by sequentially using in a fixed order both the terms h and blocks j. For the sake of simplicity, we omit in the description below the dependence on k. For any fixed value of h, the inner iteration on the blocks ℓ_j is defined as

$$\begin{aligned} \mathbf{y}_{h,0} &= \mathbf{y}_{h-1,L} \\ \mathbf{y}_{h,j} &= \mathbf{y}_{h,j-1} - \alpha[\nabla f_h(\mathbf{y}_{h,j-1})]_{\ell_j} \quad \text{for } j = 1, \dots, L \end{aligned}$$

where $\mathbf{y}_{1,0} = \mathbf{w}^k$. Applying iteratively, we get for any h

$$\mathbf{y}_{h,j} = \mathbf{y}_{h-1,L} - \alpha \sum_{i=1}^{j} [\nabla f_h(\mathbf{y}_{h,i-1})]_{\ell_i} \quad \text{for } j = 1, \dots, L.$$

Developing now iteratively on h, we get

$$\begin{aligned} \mathbf{y}_{h,j} &= \mathbf{y}_{h-2,L} - \alpha \sum_{i=1}^{L} [\nabla f_{h-1}(\mathbf{y}_{h-1,i-1})]_{\ell_i} \\ &\quad -\alpha \textstyle\sum_{i=1}^{j} [\nabla f_h(\mathbf{y}_{h,i-1})]_{\ell_i} \\ &= \mathbf{w}^k - \alpha \displaystyle\sum_{t=1}^{h-1} \sum_{i=1}^{L} [\nabla f_t(\mathbf{y}_{t,i-1})]_{\ell_i} \\ &\quad -\alpha \textstyle\sum_{i=1}^{j} [\nabla f_h(\mathbf{y}_{h,i-1})]_{\ell_i} \end{aligned}$$

and we finally set $\mathbf{w}^{k+1} = \mathbf{y}_{H,L}$.

Hence, an iteration of BIG method can be written as

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha^k \mathbf{d}^k \tag{12}$$

where the direction $\mathbf{d}_k$ is defined through the intermediate updates $\mathbf{y}^k_{h,j} \in \mathbf{R}^n$ as

$$\mathbf{d}^k = -\sum_{h=1}^{H} \sum_{j=1}^{L} [\nabla f_h(\mathbf{y}^k_{h,j-1})]_{\ell_j} \tag{13}$$

with

$$\begin{aligned} &\mathbf{y}^k_{1,0} = \mathbf{w}^k, \; \mathbf{y}^k_{h,0} = \mathbf{y}^k_{h-1,L} \\ &\mathbf{y}^k_{h,j} = \mathbf{w}^k - \alpha^k \\ &\left(\textstyle\sum_{t=1}^{h-1} \sum_{i=1}^{L} [\nabla f_t(\mathbf{y}^k_{t,i-1})]_{\ell_i} + \sum_{i=1}^{j} [\nabla f_h(\mathbf{y}^k_{h,i-1})]_{\ell_i}\right) \\ &\mathbf{w}^{k+1} = \mathbf{y}^k_{H,L}. \end{aligned} \tag{14}$$

The scheme of BIG is reported in Algorithm 2.

3.1 Convergence of BIG with diminishing stepsize

Convergence of BIG can be proved under suitable assumptions by looking at the iteration generated by the algorithm as a gradient method with errors. Below we report the main convergence result in case a diminishing stepsize is employed, namely when α^k is updated according to (4).

Proposition 4 (Convergence of BIG - Diminishing stepsize) *Let* $\{\mathbf{w}^k\}$ *be a sequence generated by* (12), (13) *and* (14).

Algorithm 2 Block-coordinate incremental gradient (BIG)

```
Given $\mathcal{L} = \{\ell_1, \dots, \ell_L\}$
Choose $\mathbf{w}^0 \in \mathbf{R}^n$, $\alpha^0 > 0$, and set $k = 0$;
while (stopping criterion not met) do
    Set $\mathbf{y}_{1,0}^k = \mathbf{w}^k$;
    for $h = 1, \dots, H$ do
        for $j = 1, \dots, L$ do
            $\mathbf{y}_{h,j}^k = \mathbf{y}_{h,j-1}^k - \alpha^k [\nabla f_h(\mathbf{y}_{h,j-1}^k)]_{\ell_j}$
        end for
        $\mathbf{y}_{h+1,0}^k = \mathbf{y}_{h,L}^k$
    end for
    $\mathbf{w}^{k+1} = \mathbf{y}_{H,L}^k$
    Update $\alpha^k$
    $k = k + 1$
end while
```

Assume that (10) *and* (11) *hold for each $h = 1, \dots, H$, i.e. there exist positive constants M, a, b such that*

$$\|\nabla f_h(\mathbf{u}) - \nabla f_h(\mathbf{w})\| \le M\|\mathbf{u} - \mathbf{w}\| \qquad \forall\, \mathbf{u}, \mathbf{w} \in \mathbf{R}^n$$
$$\|\nabla f_h(\mathbf{w})\| \le a + b\|\nabla f(\mathbf{w})\| \qquad \forall\, \mathbf{w} \in \mathbf{R}^n.$$

Further assume that the stepsize α^k satisfies (4)*, i.e.*

$$\sum_{k=0}^{\infty} \alpha^k = \infty \qquad \sum_{k=0}^{\infty} (\alpha^k)^2 < \infty.$$

Then we have that either $lim_{k\to\infty} f(\mathbf{w}^k) = -\infty$ or $f(\mathbf{w}^k)$ converges to a finite value and $\lim_{k\to\infty} \nabla f(\mathbf{w}^k) = 0$. Furthermore every accumulation point of $\mathbf{w}^k$ is a stationary point of f.

Proof We show that the assumptions of Proposition 1 are satisfied.
First of all, note that by the definition of norm we have $\|[\mathbf{w}]_{\ell_j}\| \le \|\mathbf{w}\|$ for all $\mathbf{w} \in \mathbf{R}^n$ and $j = \{1, \dots, L\}$. In turn, this yields for each $h = 1, \dots, H$ and $j = 1, \dots, L$

$$\|[\nabla f_h(\mathbf{u}) - \nabla f_h(\mathbf{v})]_{\ell_j}\| \le \|\nabla f_h(\mathbf{u}) - \nabla f_h(\mathbf{v})\| \le M\|\mathbf{u} - \mathbf{v}\| \quad \forall \mathbf{u}, \mathbf{v} \in \mathbf{R}^n$$

and

$$\|[\nabla f_h(\mathbf{u})]_{\ell_j}\| \le \|\nabla f_h(\mathbf{u})\| \le a + b\|\nabla f(\mathbf{u})\| \quad \forall \mathbf{u} \in \mathbf{R}^n.$$

We start by remarking that (10) implies (2). Indeed, $\forall \mathbf{u}, \mathbf{v} \in \mathbf{R}^n$ we have that

$$\begin{aligned}\|\nabla f(\mathbf{u}) - \nabla f(\mathbf{v})\| &= \left\| \sum_{h=1}^{H} \left(\nabla f_h(\mathbf{u}) - \nabla f_h(\mathbf{v}) \right) \right\| \\ &\le \sum_{h=1}^{H} \|\nabla f_h(\mathbf{u}) - \nabla f_h(\mathbf{v})\| \\ &\le M \sum_{h=1}^{H} \|\mathbf{u} - \mathbf{v}\| \le \widetilde{M}\|\mathbf{u} - \mathbf{v}\|.\end{aligned}$$

For the sake of simplicity, we report the proof for the case $H = 2$, $L = 2$. The proof in the case of generic values H, L is reported in “Appendix B”. The BIG iteration can be written as

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha^k \left(\sum_{h=1}^{2} \sum_{j=1}^{2} \left[\nabla f_h(\mathbf{y}_{h,j-1}^k) \right]_{\ell_j} \right)$$

which can be seen as

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha^k \left(\nabla f(\mathbf{w}^k) + \mathbf{e}^k \right)$$

with the error

$$\mathbf{e}^k = \sum_{h=1}^{2} \sum_{j=1}^{2} \left[\nabla f_h(\mathbf{y}_{h,j-1}^k) - \nabla f_h(\mathbf{w}^k) \right]_{\ell_j}.$$

Then we have

$$\begin{aligned}\|\mathbf{e}^k\| &= \left\| \sum_{h=1}^{2} \sum_{j=1}^{2} \left[\nabla f_h(\mathbf{y}_{h,j-1}^k) - \nabla f_h(\mathbf{w}^k) \right]_{\ell_j} \right\| \\ &\le \sum_{h=1}^{2} \sum_{j=1}^{2} \left\| \nabla f_h(\mathbf{y}_{h,j-1}^k) - \nabla f_h(\mathbf{w}^k) \right\| \\ &\le M \sum_{h=1}^{2} \sum_{j=1}^{2} \left\| \mathbf{y}_{h,j-1}^k - \mathbf{w}^k \right\| \\ &= M \left(\left\| \mathbf{y}_{1,0}^k - \mathbf{w}^k \right\| + \left\| \mathbf{y}_{1,1}^k - \mathbf{w}^k \right\| \right. \\ &\quad \left. + \left\| \mathbf{y}_{2,0}^k - \mathbf{w}^k \right\| + \left\| \mathbf{y}_{2,1}^k - \mathbf{w}^k \right\| \right).\end{aligned}$$

Let focus on the terms in the last inequality one by one taking into account the inner iterations which are written as

$$\begin{aligned}&\mathbf{y}_{1,0}^k = \mathbf{w}^k \\ &\mathbf{y}_{1,1}^k = \mathbf{y}_{1,0}^k - \alpha^k \left[\nabla f_1(\mathbf{y}_{1,0}^k) \right]_{\ell_1} \\ &\mathbf{y}_{1,2}^k = \mathbf{y}_{1,1}^k - \alpha^k \left[\nabla f_1(\mathbf{y}_{1,1}^k) \right]_{\ell_2} \\ &\mathbf{y}_{2,0}^k = \mathbf{y}_{1,2}^k \\ &\mathbf{y}_{2,1}^k = \mathbf{y}_{2,0}^k - \alpha^k \left[\nabla f_2(\mathbf{y}_{2,0}^k) \right]_{\ell_1} \\ &\mathbf{y}_{2,2}^k = \mathbf{y}_{2,1}^k - \alpha^k \left[\nabla f_2(\mathbf{y}_{2,1}^k) \right]_{\ell_2} \\ &\mathbf{w}^{k+1} = \mathbf{y}_{2,2}^k.\end{aligned}$$

Hence we have

$$\begin{aligned}
\left\|\mathbf{y}_{1,0}^k - \mathbf{w}^k\right\| &= 0\\
\left\|\mathbf{y}_{1,1}^k - \mathbf{w}^k\right\| &\le \alpha^k \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_1}\right\| \le \alpha^k\left(a + b\left\|\nabla f(\mathbf{w}^k)\right\|\right)\\
\left\|\mathbf{y}_{2,0}^k - \mathbf{w}^k\right\| &\le \left\|\mathbf{y}_{2,0}^k - \mathbf{y}_{1,1}^k\right\| + \left\|\mathbf{y}_{1,1}^k - \mathbf{w}^k\right\|\\
&= \alpha^k\left(\left\|\left[\nabla f_1(\mathbf{y}_{1,1}^k)\right]_{\ell_2}\right\| + \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_1}\right\|\right)\\
&\le \alpha^k\left(\left\|\left[\nabla f_1(\mathbf{y}_{1,1}^k) - \nabla f_1(\mathbf{w}^k)\right]_{\ell_2}\right\|\right.\\
&\quad \left. + \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_2}\right\| + \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_1}\right\|\right)\\
&\le \alpha^k\left(M\left\|\mathbf{y}_{1,1}^k - \mathbf{w}^k\right\| + \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_2}\right\|\right.\\
&\quad \left. + \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_1}\right\|\right)\\
&\le \alpha^k\left(M\left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_1}\right\| + \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_2}\right\|\right.\\
&\quad \left. + \left\|\left[\nabla f_1(\mathbf{w}^k)\right]_{\ell_1}\right\|\right)\\
&\le \alpha^k\,(M+2)\left(a + b\left\|\nabla f(\mathbf{w}^k)\right\|\right)\\
&\le \alpha^k\left(a + b\left\|\nabla f(\mathbf{w}^k)\right\|\right)
\end{aligned}$$

where without loss of generality we have redefined $a(M+2)$ and $b(M+2)$ as a and b.

$$\begin{aligned}
\left\|\mathbf{y}_{2,1}^k - \mathbf{w}^k\right\| &\le \left\|\mathbf{y}_{2,1}^k - \mathbf{y}_{2,0}^k\right\| + \left\|\mathbf{y}_{2,0}^k - \mathbf{w}^k\right\|\\
&= \alpha^k\left\|\left[\nabla f_2(\mathbf{y}_{2,0}^k)\right]_{\ell_1}\right\| + \left\|\mathbf{y}_{2,0}^k - \mathbf{w}^k\right\|\\
&\le \alpha^k\left\|\left[\nabla f_2(\mathbf{y}_{2,0}^k) - \nabla f_2(\mathbf{w}^k)\right]_{\ell_1}\right\|\\
&\quad + \alpha^k\left\|\left[\nabla f_2(\mathbf{w}^k)\right]_{\ell_1}\right\| + \left\|\mathbf{y}_{2,0}^k - \mathbf{w}^k\right\|\\
&\le \alpha^k M\left\|\mathbf{y}_{2,0}^k - \mathbf{w}^k\right\| + \alpha^k\left\|\left[\nabla f_2(\mathbf{w}^k)\right]_{\ell_1}\right\|\\
&\quad + \left\|\mathbf{y}_{2,0}^k - \mathbf{w}^k\right\|\\
&\le \left(\alpha^k M + 1\right)\left\|\mathbf{y}_{2,0}^k - \mathbf{w}^k\right\|\\
&\quad + \alpha^k\left\|\left[\nabla f_2(\mathbf{w}^k)\right]_{\ell_1}\right\|\\
&\le \left(\alpha^k M + 1\right)\alpha^k\left(a + b\left\|\nabla f(\mathbf{w}^k)\right\|\right)\\
&\quad + \alpha^k\left(a + b\left\|\nabla f(\mathbf{w}^k)\right\|\right)\\
&\le \alpha^k\left(\widehat{a} + \widehat{b}\left\|\nabla f(\mathbf{w}^k)\right\|\right).
\end{aligned}$$

This implies there exist positive constants A, B such that

$$\|\mathbf{e}^k\| \le \alpha^k\left(A + B\left\|\nabla f(\mathbf{w}^k)\right\|\right).$$

Then all the hypothesis of Proposition 1 hold and the thesis follows. □

The assumptions done in Proposition 4 are the same of those done when proving convergence of IG in Proposition 3. Overall, the Lipschitz condition (10) is quite natural when studying convergence analysis of finite-sum problems and directly implies that the whole objective function has a Lipschitz gradient. On the other hand, condition (11) is a stronger and less usual assumption, requiring the gradient of each term to be linearly bounded by the real gradient. As observed in Bertsekas and Tsitsiklis (2000), this assumption is guaranteed to hold when the functions f_h are quadratic convex as in the case of linear least squares.

3.2 Convergence of BIG with bounded away from zero stepsize

So far we have shown that a block incremental gradient method converges to a stationary point as long as a diminishing stepsize is employed. However, the diminishing stepsize rule could be cumbersome to implement leading to slow convergence in case it is not properly tuned. As a consequence, a more practical updating rule commonly used when dealing with incremental gradient methods is to keep the stepsize fixed for a certain number of iterations and then reduce it by a small factor. This updating rule is straightforward to be implemented and can be controlled more easily than the diminishing one.

We have already seen that BIG can be written as a gradient method with error, i.e.

$$\mathbf{w}^{k+1} = \mathbf{w}^{\mathbf{k}} - \alpha^k(\nabla f(\mathbf{w}^{\mathbf{k}}) - \mathbf{e}^k).$$

Hence, in order to apply the results of Proposition 2 to the BIG method, we need to show that under some standard assumptions the error term in BIG satisfies condition (5). This is the aim of the following proposition.

Proposition 5 *Let* $\{\mathbf{w}^k\}$ *be a sequence generated by* (12), (13) *and* (14). *Assume that for each* $h \in \{1, \dots, H\}$ *condition* (10) *is satisfied, namely*

$$\|\nabla f_h(\mathbf{u}) - \nabla f_h(\mathbf{w})\| \le M\|\mathbf{u} - \mathbf{w}\| \qquad \forall\, \mathbf{u}, \mathbf{w} \in \mathbf{R}^n$$

and there exist a positive constant $\bar{B}$ *such that it holds*

$$\|\nabla f_h(\mathbf{w}^k)\| \le \bar{B} \quad \forall\, h \in \{1, \dots, H\}. \tag{15}$$

Further assume that the stepsize α^k satisfies

$$\lim_{k\to\infty} \alpha^k = \bar{\alpha} > 0.$$

Then the error term $\mathbf{e}^k$ satisfies condition (5).

Proof Similarly to what done in Proposition 4 (cfr the extended proof in “Appendix B”) the general BIG iteration can be written as

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha^k(\nabla f(\mathbf{w}^k) - \mathbf{e}^k)$$

where

$$\mathbf{e}^k = \sum_{h=1}^{H}\sum_{j=1}^{L}\left([\nabla f_h(\mathbf{y}^k_{h,j-1})]_{\ell_j} - [\nabla f_h(\mathbf{w}^k)]_{\ell_j}\right).$$

As done in “Appendix B” to prove (19), thanks to the sample Lipschitz condition (10), we obtain the bound

$$\left\|\mathbf{e}^k\right\| \le M\sum_{h=1}^{H}\sum_{j=0}^{L-1}\left\|\mathbf{y}^k_{h,j} - \mathbf{w}^k\right\|.$$

Now reasoning in a similar way to what done in Eq. (21) and thanks to the hypothesis (15), we can get the following bound on two iterates

$$\left\|\mathbf{y}^k_{h,j} - \mathbf{w}^k\right\| \le (\alpha^k M + 1)\left\|\mathbf{y}^k_{p(h,j-1)} - \mathbf{w}^k\right\| + \alpha^k\bar{B}$$

where $p(h, j-1)$ is used to denote the estimate before considering the term $\mathbf{y}^k_{h,j}$, as described in (20). Thus, by iteratively applying this bound, we obtain

$$\|\mathbf{e}^k\| \le \alpha^k\bar{B}$$

for some positive constant $\bar{B}$. □

Then the following convergence result for BIG with a bounded away from zero stepsize directly applies by considering the results from Propositions 2 and 5.

Proposition 6 (Convergence of BIG - Bounded away from zero stepsize) *Let $\{\mathbf{w}^k\}$ be a sequence generated by* (12), (13) *and* (14). *Assume that all the iterates $\mathbf{w}^k$ and $\mathbf{y}^k_{h,j}$ belong to some bounded set D.*
Assume that for each $h \in \{1,\ldots,H\}$ conditions (10) *and* (15) *hold and that the stepsize α^k satisfies*

$$\lim_{k\to\infty} \alpha^k = \bar{\alpha} > 0,$$

where $\alpha^k \in (\theta, 2/L - \theta)$ with $\theta \in (0, 1/L]$. Then there exist a constant $C > 0$ (independent of $\bar{\alpha}$) and an accumulation point $\bar{\mathbf{w}}$ of the sequence $\{\mathbf{w}^k\}$ such that

$$\|\nabla f(\mathbf{N})\| \le C\bar{\alpha} \tag{16}$$

Furthermore, if the sequence $\{f(\mathbf{w}^k)\}$ converges, then every accumulation point $\bar{\mathbf{w}}$ of the sequence $\{\mathbf{w}^k\}$ satisfies (16).

Thus Proposition 6 implies that BIG with a bounded away from zero stepsize can only achieve a neighbourhood of a stationary point. Overall, it was a predictable result. Indeed, since the error term satisfies (5), if we consider the scalar product $\nabla f(\mathbf{w}^k)^T\mathbf{d}^k$ and assume that $\alpha^k = \bar{\alpha} > 0$, it yields

$$\begin{aligned}\nabla f(\mathbf{w}^k)^T\mathbf{d}^k &= \nabla f(\mathbf{w}^k)^T\left(-\nabla f(\mathbf{w}^k) + \mathbf{e}^k\right)\\ &\le -\|\nabla f(\mathbf{w}^k)\|^2 + \|\nabla f(\mathbf{w}^k)\|\|\mathbf{e}^k\|\\ &\le -\|\nabla f(\mathbf{w}^k)\|^2 + \|\nabla f(\mathbf{w}^k)\|\bar{\alpha}\bar{B}.\end{aligned}$$

This shows how within the region

$$\left\{\mathbf{w} \in \mathbf{R}^n \ : \ \|\nabla f(\mathbf{w})\| > \bar{\alpha}\bar{B}\right\}$$

BIG computes directions which are actually descent directions, while in the complementary region the behaviour is unpredictable. Moreover the size of this region linearly depends on the constant stepsize $\bar{\alpha}$ employed.

It is interesting to note that, by fixing $\theta = 0$ we get the stepsize $\alpha \in (0, \frac{2}{L}]$, which is the same stepsize needed to prove convergence towards exact stationary points for the standard gradient descent method. That is, if BIG has a cost per iteration much cheaper than the batch gradient descent, up to H times, the price to pay is that it does not converge towards stationary points, but lends in a ϵ-accurate solution.

As underlined in Solodov (1998), the assumptions on the norm of the error term in Proposition 2 (namely condition (5)) could be relaxed so to consider the more general case $\|\mathbf{e}^k\| \le \alpha^k(a+b\nabla f(\mathbf{w}^k))$ for some positive constants a, b. However, this would lead to a third-degree inequality to determine the allowed interval for the stepsize α^k which is not trivial to solve.

Moreover, note that the boundedness assumption on the iterates is not very restrictive. Indeed, it is satisfied as long as the level set $\{\mathbf{w} \mid f(\mathbf{w}) \le \rho_1\}$ is bounded for some $\rho_1 > f(\mathbf{w}^0)$ and the iterates stay within that region, as is usually the case in the optimization problem behind training a Deep Neural Network (Solodov 1998; Zhi-Quan and Paul 1994). Note that also the Lipschitz and boundedness conditions on the gradient of the objective function are satisfied whenever each term f_h is twice continuously differentiable and the iterates stay within a compact set.

Finally, we remark that as a further example of an optimization problem where conditions (10) and (15) are satisfied (and consequently (11) holds as well), we can consider the LogitBoost algorithm (Collins et al. 2002). Indeed, given a classification problem, in the nonlinearly separable case and when the features are linearly independent on the training set, then the objective function has a sample Lipschitz gradient and each gradient can be bounded above [see Remark 3 on Blatt et al. (2007) for a deeper discussion on the properties of the LogitBoost algorithm].

4 Discussion on numerical performance

As a block-coordinate descent method BIG can lead to improvements in performance in all those cases where the structure of the objective function can be leveraged to define problems easier to solve (e.g. subproblems might be less computationally expensive to treat or might become separable). On the other hand, as an incremental method, BIG owns good properties when dealing with large-scale finite-sum problems, namely in all those cases where the function can be expressed as a large sum of similar terms so that each gradient and objective function computations might require an excessive computational effort. Thus, BIG might be employed in all those cases where the objective function has both some block structure that can be exploited and a finite-sum structure.

With the aim to provide the reader with an application of BIG to a real problem, we can consider an estimation problem where given some data $\{\mathbf{x}_h, y_h\}_{h=1}^{H}$, where $\mathbf{x}_h \in \mathbf{R}^d$ represents the input features and $y_h \in \mathbf{R}$ is the output we want to estimate, we want to determine the relation between the input and the output by means of a nonlinear least-square function of the form

$$\underset{\mathbf{w},\mathbf{v}}{\text{minimize}} \quad \sum_{h=1}^{H} \left(\phi(\mathbf{w}; \mathbf{x}_h)^T \mathbf{v} - y_h \right)^2 + \rho \|\mathbf{w}\|^2 + \rho \|\mathbf{v}\|^2 \tag{17}$$

where $\phi(\mathbf{w}; \mathbf{x}_h)$ represents a nonlinear transformation of the input $\mathbf{x_h}$. This formulation is quite general and includes several applications such as some kernel methods (Shawe-Taylor and Cristianini 2004) and neural networks (Goodfellow et al. 2016). Problem (17) presents a finite-sum with two-block structure that perfectly fit with the advantages led by BIG. Indeed, when considering the general term f_h only with respect to the block $\mathbf{v}$ it is strictly convex while when considering the block $\mathbf{w}$ the problem is still nonlinear but has a smaller size and other interesting properties might come out according to the type of nonlinear transformation ϕ.

As a particular instance of problem (17), in Palagi and Seccia (2020) extensive numerical results have been reported when dealing with the mean squared error optimization problem behind the training phase of a deep neural network. In this class of problems, indeed, a natural block decomposition with respect to the weights of each layer appears. Then the performance of BIG has been analysed when the layered structure of the model is exploited. In particular, the standard IG method is compared to the application of BIG when each layer of the model defines a block of variables $\mathbf{w}_\ell$ and several numerical results are discussed. We do not report numerical results here which can be found in Palagi and Seccia (2020). However, we remark how numerical results in Palagi and Seccia (2020) suggest that BIG outperform IG, especially when considering deeper and wider models, namely neural networks with a larger number of layers or neurons per layer. Moreover, from a machine learning perspective, it is interesting to underline how BIG seems to lead to better performance compared to IG also on the generalization error, namely the error on new samples never seen before by the estimation model.

5 Conclusion

In this paper, we have extended the convergence theory of incremental methods by providing convergence results of a block-coordinate incremental gradient method under two different stepsize updating rules. The analysis has shown how the BIG algorithm can be seen as a gradient method with errors; thus, its convergence can be proved by recalling known convergence results (Bertsekas and Tsitsiklis 2000; Solodov 1998).

Author Contributions Authors contributed equally at the conceptualization, formal analysis, writing reviewing and editing of the manuscript.

Funding Open access funding provided by Université degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Data availability No data have been used in this work.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

A Proof of Proposition 3

Here we report a proof of Proposition 3 for a general number of terms H.

Proof First we recall that condition (10) implies the gradient satisfies the Lipschitz condition (2) with constant at most equal to HM as shown in Sect. 3.1.

Let consider now the general kth iteration of IG

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha^k \sum_{h=1}^{H} \nabla f_h(\mathbf{y}_{h-1}^k) = \mathbf{w}^k + \alpha^k \left(-\nabla f(\mathbf{w}^k) + \mathbf{e}^k\right),$$

where

$$\begin{aligned}\mathbf{e}^k &= \sum_{h=1}^{H}\left(\nabla f_h(\mathbf{w}^k) - \nabla f_h(\mathbf{y}_{h-1}^k)\right)\\ &= \sum_{h=2}^{H}\left(\nabla f_h(\mathbf{w}^k) - \nabla f_h(\mathbf{y}_{h-1}^k)\right).\end{aligned}$$

Taking the norm of the error and recalling (10), we obtain

$$\begin{aligned}\|\mathbf{e}^k\| &= \left\|\sum_{h=2}^{H}\left(\nabla f_h(\mathbf{w}^k) - \nabla f_h(\mathbf{y}_{h-1}^k)\right)\right\|\\ &\le \sum_{h=2}^{H} M \left\|\mathbf{w}^k - \mathbf{y}_{h-1}^k\right\| = M \sum_{h=1}^{H-1}\left\|\mathbf{w}^k - \mathbf{y}_h^k\right\|.\end{aligned}$$

Now, for each of the element in the last sum it holds

$$\begin{aligned}\|\mathbf{w}^k - \mathbf{y}_h^k\| &= \left\|\mathbf{w}^k + \sum_{j=1}^{h-1}(-\mathbf{y}_{h-j}^k + \mathbf{y}_{h-j}^k) - \mathbf{y}_h^k\right\|\\ &= \left\|(\mathbf{y}_0^k - \mathbf{y}_1^k) + (\mathbf{y}_1^k - \mathbf{y}_2^k) + \cdots + (\mathbf{y}_{h-1}^k - \mathbf{y}_h^k)\right\|\\ &= \left\|\alpha^k \sum_{j=1}^{h} \nabla f_j(\mathbf{y}_{j-1}^k)\right\| \le \alpha^k \sum_{j=1}^{h}\left\|\nabla f_j(\mathbf{y}_{j-1}^k)\right\|\\ &= \alpha^k \left(\sum_{j=1}^{h}\left\|\nabla f_j(\mathbf{y}_{j-1}^k) - \nabla f_j(\mathbf{w}^k) + \nabla f_j(\mathbf{w}^k)\right\|\right)\\ &\le \alpha^k \left(\sum_{j=1}^{h}\left\|\nabla f_j(\mathbf{w}^k)\right\| + \sum_{j=1}^{h}\left\|\nabla f_j(\mathbf{w}^k) - \nabla f_j(\mathbf{y}_{j-1}^k)\right\|\right)\\ &\le \alpha^k \left(\sum_{j=1}^{h}\left\|\nabla f_j(\mathbf{w}^k)\right\| + M\sum_{j=1}^{h}\left\|\mathbf{w}^k - \mathbf{y}_{j-1}^k\right\|\right)\\ &\le \alpha^k \left(ha + hb\left\|\nabla f(\mathbf{w}^k)\right\| + M\sum_{j=1}^{h}\left\|\mathbf{w}^k - \mathbf{y}_{j-1}^k\right\|\right).\end{aligned}$$

Recalling that $\mathbf{w}^k = \mathbf{y}_0^k$, we get

$$\begin{aligned}\left\|\mathbf{w}^k - \mathbf{y}_h^k\right\| \le \alpha^k \Bigg(& ha + hb\left\|\nabla f(\mathbf{w}^k)\right\|\\ & + M\sum_{j=2}^{h}\left\|\mathbf{w}^k - \mathbf{y}_{j-1}^k\right\|\Bigg),\end{aligned} \tag{18}$$

with

$$\|\mathbf{w}^k - \mathbf{y}_1^k\| = \alpha^k \|\nabla f_1(\mathbf{w}^k)\| \le \alpha^k (a + b\|\nabla f(\mathbf{w}^k)\|).$$

So

$$\begin{aligned}\|\mathbf{e}^k\| \le M\alpha^k \Bigg(& \sum_{h=1}^{H-1}\left(ha + hb\left\|\nabla f(\mathbf{w}^k)\right\|\right)\\ & + \sum_{h=1}^{H-1}\sum_{j=2}^{h}\left\|\mathbf{w}^k - \mathbf{y}_{j-1}^k\right\|\Bigg).\end{aligned}$$

By recursively applying (18) to the elements of the last sum, and recalling that α^k is bounded above, we get a bound like (3). Then the hypothesis of Proposition 1 hold and the thesis follows. □

B Proof of Proposition 4

Proof A general iteration of BIG method is defined as

$$\begin{aligned}\mathbf{y}_{1,0}^k &= \mathbf{w}^k\\ \mathbf{y}_{1,1}^k &= \mathbf{y}_{1,0}^k - \alpha^k \left[\nabla f_1(\mathbf{y}_{1,0}^k)\right]_{\ell_1}\\ &\vdots\\ \mathbf{y}_{1,L}^k &= \mathbf{y}_{1,L-1}^k - \alpha^k \left[\nabla f_1(\mathbf{y}_{1,L-1}^k)\right]_{\ell_L}\\ \mathbf{y}_{2,0}^k &= \mathbf{y}_{1,L}^k\\ \mathbf{y}_{2,1}^k &= \mathbf{y}_{2,0}^k - \alpha^k \left[\nabla f_2(\mathbf{y}_{2,0}^k)\right]_{\ell_1}\\ &\vdots\\ \mathbf{y}_{H,L}^k &= \mathbf{y}_{H,L-1}^k - \alpha^k \left[\nabla f_H(\mathbf{y}_{H,L-1}^k)\right]_{\ell_{L-1}}\\ \mathbf{w}^{k+1} &= \mathbf{y}_{H,L}^k\end{aligned}$$

Summing up these equations, we get that the new point $\mathbf{w}^{k+1}$ is a gradient descent method plus an error, that is

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \alpha^k(\nabla f(\mathbf{w}^k) - \mathbf{e}^k)$$

with

$$\mathbf{e}^k = \sum_{h=1}^{H}\sum_{j=1}^{L}\left([\nabla f_h(\mathbf{y}_{h,j-1}^k) - \nabla f_h(\mathbf{w}^k)]_{\ell_j}\right).$$

By using the sample Lipschitz hypothesis (10), we obtain

$$\begin{aligned}\left\|\mathbf{e}^k\right\| &= \left\|\sum_{h=1}^{H}\sum_{j=1}^{L}\left([\nabla f_h(\mathbf{y}_{h,j-1}^k) - \nabla f_h(\mathbf{w}^k)]_{\ell_j}\right)\right\| \\ &\leq M\sum_{h=1}^{H}\sum_{j=1}^{L}\left\|\mathbf{y}_{h,j-1}^k - \mathbf{w}^k\right\| \\ &= M\sum_{h=1}^{H}\sum_{j=0}^{L-1}\left\|\mathbf{y}_{h,j}^k - \mathbf{w}^k\right\|. \qquad (19)\end{aligned}$$

We want to prove that each term in (19) is bounded by a quantity which satisfies (3). Given the general update $\mathbf{y}_{h,j}^k$, the previous point depends on both the values of h and j. Then, to make notation easier, we introduce the following notation $p(h, j-1)$

$$p(h, j-1) = \begin{cases}(h, j-1) & \text{if } j-1 > 0\\ (h-1, L-1) & \text{if } j-1 = 0.\end{cases} \qquad (20)$$

We can now bound the general term $\left\|\mathbf{y}_{h,j}^k - \mathbf{w}^k\right\|$ as follows

$$\begin{aligned}\left\|\mathbf{y}_{h,j}^k - \mathbf{w}^k\right\| &= \left\|\mathbf{y}_{h,j}^k - \mathbf{y}_{p(h,j-1)}^k + \mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| \\ &\leq \left\|\mathbf{y}_{h,j}^k - \mathbf{y}_{p(h,j-1)}^k\right\| + \left\|\mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| \\ &= \alpha^k\left\|[\nabla f_h(\mathbf{y}_{p(h,j-1)}^k)]_{\ell_j}\right\| + \left\|\mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| \\ &\leq \alpha^k\left\|[\nabla f_h(\mathbf{y}_{p(h,j-1)}^k)\right. \\ &\quad \left.-\nabla f_h(\mathbf{w}^k) + \nabla f_h(\mathbf{w}^k)]_{\ell_j}\right\| \\ &\quad + \left\|\mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| \qquad (21)\\ &\leq \alpha^k M\left\|\mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| + \alpha^k\left\|\nabla f_h(\mathbf{w}^k)\right\| \\ &\quad + \left\|\mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| \\ &\leq (\alpha^k M + 1)\left\|\mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| + \alpha^k\left\|\nabla f_h(\mathbf{w}^k)\right\| \\ &\leq (\alpha^k M + 1)\left\|\mathbf{y}_{p(h,j-1)}^k - \mathbf{w}^k\right\| \\ &\quad + \alpha^k(a + b\left\|\nabla f(\mathbf{w}^k)\right\|).\end{aligned}$$

By iteratively applying this bound on each term of (19), and recalling that

$$\left\|\mathbf{y}_{1,0}^k - \mathbf{w}^k\right\| = \alpha^k\left\|\left[\nabla f_1(\mathbf{y}_{1,0}^k)\right]_{\ell_1}\right\|,$$

we get there exist positive constants a and b such that

$$\|\mathbf{e}^k\| \leq \alpha^k\left(a + b\left\|\nabla f(\mathbf{w}^k)\right\|\right).$$

Then all the hypothesis of Proposition 1 hold and the thesis follows. □

References

Beck A, Tetruashvili L (2013) On the convergence of block coordinate descent type methods. SIAM J Optim 23(4):2037–2060

Bertsekas DP (1996) Incremental least squares methods and the extended Kalman filter. SIAM J Optim 6(3):807–822. https://doi.org/10.1137/S1052623494268522

Bertsekas DP (2011) Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. Optim Mach Learn 2010(1–38):3

Bertsekas DP (2015) Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. CoRR http://arxiv.org/abs/1507.01030

Bertsekas DP, Tsitsiklis JN (1989) Parallel and distributed computation: numerical methods, vol 23. Prentice Hall, Englewood Cliffs

Bertsekas DP, Tsitsiklis JN (2000) Gradient convergence in gradient methods with errors. SIAM J Optim 10(3):627–642. https://doi.org/10.1137/S1052623497331063

Blatt D, Hero AO, Gauchman H (2007) A convergent incremental gradient method with a constant step size. SIAM J Optim 18(1):29–51

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: in COMPSTAT

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. SIAM Rev 60(2):223–311

Bravi L, Sciandrone M (2014) An incremental decomposition method for unconstrained optimization. Appl Math Comput 235:80–86

Buzzi C, Grippo L, Sciandrone M (2001) Convergent decomposition techniques for training RBF neural networks. Neural Comput 13(8):1891–1920

Chauhan VK, Dahiya K, Sharma A (2017) Mini-batch block-coordinate based stochastic average adjusted gradient methods to solve big data problems. In: Proceedings of the Ninth Asian conference on machine learning, proceedings of machine learning research, vol 77, pp. 49–64. PMLR. http://proceedings.mlr.press/v77/chauhan17a.html

Collins M, Schapire RE, Singer Y (2002) Logistic regression, adaboost and bregman distances. Mach Learn 48(1–3):253–285

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge

Grippo L, Manno A, Sciandrone M (2016) Decomposition techniques for multilayer perceptron training. IEEE Trans Neural Netw Learn Syst 27(11):2146–2159. https://doi.org/10.1109/TNNLS.2015.2475621

Grippo L, Sciandrone M (1999) Globally convergent block-coordinate techniques for unconstrained optimization. Optim Methods Softw 10(4):587–637. https://doi.org/10.1080/10556789908805730

Lu Z, Xiao L (2015) On the complexity analysis of randomized block-coordinate descent methods. Math Program 152(1–2):615–642

Nesterov Y (2012) Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J Optim 22(2):341–362

Palagi L, Seccia R (2020) Block layer decomposition schemes for training deep neural networks. J Global Optim 77(1):97–124

Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22:400–407

Shawe-Taylor J, Cristianini N et al (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Solodov MV (1998) Incremental gradient algorithms with stepsizes bounded away from zero. Comput Optim Appl 11(1):23–35

Wang H, Banerjee A (2014) Randomized block coordinate descent for online and stochastic optimization. arXiv preprint arXiv:1407.0107

Wright SJ (2015) Coordinate descent algorithms. Math Program 151(1):3–34. https://doi.org/10.1007/s10107-015-0892-3

Zhao T, Yu M, Wang Y, Arora R, Liu H (2014) Accelerated mini-batch randomized block coordinate descent method. In: Advances in neural information processing systems, pp 3329–3337

Zhi-Quan L, Paul T (1994) Analysis of an approximate gradient projection method with applications to the backpropagation algorithm. Optim Methods Softw 4(2):85–101