Journal of Global Optimization (2023) 87:533–550
https://doi.org/10.1007/s10898-022-01173-9

An indefinite proximal subgradient-based algorithm for nonsmooth composite optimization

Rui Liu[1] · **Deren Han**[1] · **Yong Xia**[1]

Received: 21 August 2021 / Accepted: 19 April 2022 / Published online: 16 September 2022

Abstract
We propose an indefinite proximal subgradient-based algorithm (IPSB) for solving nonsmooth composite optimization problems. IPSB is a generalization of the Nesterov's dual algorithm, where an indefinite proximal term is added to the subproblems, which can make the subproblem easier and the algorithm efficient when an appropriate proximal operator is judiciously setting down. Under mild assumptions, we establish sublinear convergence of IPSB to a region of the optimal value. We also report some numerical results, demonstrating the efficiency of IPSB in comparing with the classical dual averaging-type algorithms.

Keywords Nonsmooth optimization · Composite convex optimization · Nesterov's dual averaging · Subgradient

1 Introduction

Consider the nonsmooth composite convex optimization problem

$$\min_{x\in Q} \{F(x) := f(x) + h(x)\}, \tag{1.1}$$

where $Q \subseteq \mathbb{R}^n$ is a simple closed convex set, $f, h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are convex (not necessarily smooth) and $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is nonsmooth. Moreover, h is assumed to be the summation of a quadratic convex and a convex function (SQCC). Problem (1.1) has received much attention due to its broad applications in several different areas such as signal processing, system identification, machine learning and image processing; see, for instance, [6, 7, 10] and references therein.

Among the numerical algorithms for solving nonsmooth optimization problems (1.1) such as splitting algorithms [9], cutting plane methods [21], ellipsoid methods [11], bun-

✉ Deren Han
handr@buaa.edu.cn

Rui Liu
by1909010@buaa.edu.cn

Yong Xia
yxia@buaa.edu.cn

[1] LMIB, School of Mathematical Sciences, Beihang University, Beijing 100191, China

dle methods [17], gradient sampling methods [4] and smoothing methods [19], subgradient methods [25] are fundamental, which have been extensively studied due to their applicability to a wide variety of problems and low requirement on memory [3, 8, 22, 23]. The iteration complexity for applying a subgradient method to solve the general nonsmooth convex minimization problem is $O(1/\epsilon^2)$, i.e., after $O(1/\epsilon^2)$ iterations, the difference between the objective function value and the optimum is about ϵ; see [21]. For problems equipped with additional structure, various approaches are proposed such as smoothing schemes [19], fast iterative shrinkage-thresholding algorithm [1], bundle method [17], to improve the iteration complexity to $O(1/\epsilon)$.

Note that for the nonsmooth optimization problems, it is usually not the case that the subgradient vanishes at the solution point, and as a consequence, the stepsize in the subgradient-based method should be approaching zero. Such a vanishing property of the stepsize slows down the convergence rate of the subgradient method [20]. To deal with this undesirable phenomenon, Nesterov proposed a dual averaging (DA) scheme [20]. Each iteration of DA scheme takes the form

$$x_{k+1} = \operatorname*{argmin}_{x\in Q} \left\{ \sum_{i=0}^{k} \langle \lambda_i D_i, x - x_0\rangle + \beta_{k+1} r(x) \right\}, \ D_k \in \partial F(x_k), \ \forall k \geq 0, \tag{1.2}$$

where λ_k, $\forall k \geq 0$ are stepsizes, $\{\beta_k\}_{k=0}^{\infty}$ is a positive nondecreasing sequence and $r(\cdot)$ is an auxiliary strongly convex function. Following the DA scheme, Xiao [26] proposed a regularized dual averaging (RDA) scheme, which generates the iterate by minimizing a problem that involves all the past subgradients of f and the whole function h,

$$x_{k+1} = \arg\min_{x\in Q} \left\{ \sum_{i=0}^{k} \langle g_i, x - x_0\rangle + (k+1)h(x) + \beta_{k+1} r(x) \right\}, \ g_i \in \partial f(x_i), \ \forall k \geq 0, \tag{1.3}$$

where x_0 is the minimizer of h over Q. Setting the auxiliary function $r(\cdot)$ as $\frac{1}{2}\|\cdot - x_0\|^2$ in the above RDA scheme (1.3) becomes the so-called proximal subgradient-based (PSB) method

$$x_{k+1} = \arg\min_{x\in Q} \left\{ \sum_{i=0}^{k} \langle g_i, x - x_0\rangle + (k+1)h(x) + \frac{\beta_{k+1}}{2}\|x - x_0\|^2 \right\}, \ g_i \in \partial f(x_i), \ \forall k \geq 0. \tag{1.4}$$

The regularization function $r(\cdot)$ is crucial in RDA and PSB, which plays a similar role as the proximal term in the classical proximal point algorithm (PPA) [5, 18, 24]. On one hand, it ensures the existence and uniqueness of the solution of the subproblems, and makes the subproblems stable. On the other hand, it also influences the efficiency of the algorithms. Recently, much attention was paid on relaxing the strong convexity requirements on the proximal term in PPA [13] and related algorithms such as augmented Lagrangian method [12] and alternating direction method of multipliers [14, 16], and such a strategy achieves great success in numerical experiments. In this paper, we relax $r(\cdot)$ in (1.3) to an indefinite one, yielding the following dynamic regularized dual averaging (DRDA) scheme

$$x_{k+1} = \arg\min_{x\in Q} \left\{ \sum_{i=0}^{k} \langle g_i, x - x_0\rangle + (k+1)h(x) + \beta_{k+1} r_k(x) \right\}, \ g_i \in \partial f(x_i), \ \forall k \geq 0, \tag{1.5}$$

where $(k+1)h(\cdot) + \beta_{k+1} r_k(\cdot)$ is strongly convex for each k. Note that under this requirement, even if the function $h(\cdot)$ is convex, $r_k(\cdot)$ could be carefully chosen to be nonconvex. Specially, we introduce an appropriate indefinite item in (1.4) and then propose the indefinite

proximal subgradient-based (IPSB) algorithm. Convergence rate is established under mild assumptions. We do numerical experiments on the regularized least squares problem and elastic net regression. Numerical results demonstrate the efficiency of IPSB in comparing with the existing algorithms SDA and PSB.

The rest of this paper is organized as follows. In the following subsection, we introduce some notations and preliminaries. Section 2 reviews the simple dual averaging algorithm, the proximal subgradient-based algorithm and gives our new extensions. Section 3 presents the convergence analysis. Numerical experiments are performed in Sect. 4. We make conclusions in Sect. 5.

1.1 Notations and preliminaries

In this subsection, we present some definitions and preliminary results that will be used in our analysis later. Let Q be a closed convex set in $\mathbb{R}^n$. We use $\langle s, x\rangle$ and $s^T x$ to denote the inner product of s and x, two real vectors with the same dimension. Let $\mathbb{S}^n$ denote the set of symmetric matrices of order n, and I denote the identity matrix whose dimension is clear from the context. The Euclidean norm defined by $\sqrt{\langle\cdot,\cdot\rangle}$ is denoted by $\|\cdot\|$. Let $[m]$ denote the set $\{1, 2, \ldots, m\}$. The ball with center x and radius r reads as

$$B_r(x) = \{y \in \mathbb{R}^n : \|y - x\| \le r\}.$$

The subdifferential of a convex function f at point $x \in \mathrm{dom} f$ is given by

$$\partial f(x) := \{g \in \mathbb{R}^n : f(y) \ge f(x) + \langle g, y - x\rangle, \forall y \in \mathbb{R}^n\},$$

and any element in $\partial f(x)$ is called a subgradient of f at x, where $\mathrm{dom} f$ is the domain of f, i.e., the set of $x \in \mathbb{R}^n$ such that $f(x)$ is finite.

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called strongly convex if there exists a constant $\kappa > 0$ such that

$$f(x) \ge f(y) + \langle g, x - y\rangle + \frac{\kappa}{2}\|x - y\|^2,\ \forall x, y \in \mathbb{R}^n,\ \forall g \in \partial f(y),$$

where the constant κ is called the strong convexity parameter.

For $M \in \mathbb{R}^{n\times n}$, we use the notation $\|x\|_M^2$ to denote $x^T Mx$ even if M is not positive semidefinite. Denote by $\mathrm{tr}\,(M)$ the trace of the matrix M.

Definition 1.1 (*SQCC*) A function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called the summation of quadratic convex and convex functions (SQCC) if there exists a (nonlinear) quadratic convex function $q : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and a convex function $\tilde{h} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that

$$h(x) = q(x) + \tilde{h}(x),\ \forall x \in \mathbb{R}^n.$$

Since h is SQCC, there exists a non-zero positive semidefinite matrix $\Sigma_h \in \mathbb{S}^n$ such that for all $x, y \in \mathbb{R}^n$,

$$h(y) \ge h(x) + \langle u, y - x\rangle + \frac{1}{2}\|y - x\|_{\Sigma_h}^2,\ \forall u \in \partial h(x), \tag{1.6}$$

or equivalently,

$$\langle x - y, u - v\rangle \ge \|x - y\|_{\Sigma_h}^2,\ \forall u \in \partial h(x), v \in \partial h(y).$$

2 A new proximal subgradient algorithm

In the first subsection, we briefly review two existing algorithms SDA and PSB. Then in the second subsection, we describe the indefinite proximal subgradient-based (IPSB) algorithm.

2.1 SDA and PSB

We start from the classical subgradient algorithm [3] for minimizing the problem (1.1)

$$x_{k+1} = P_Q(x_k - \lambda_k d_k),\ k \in \mathbb{N}, \tag{2.1}$$

where P_Q denotes the projection onto Q, d_k is either a subgradient $D_k \in \partial F(x_k)$ or the normalized subgradient $D_k/\|D_k\|$, and the sequence of the stepsizes $\{\lambda_k\}_{k=0}^{\infty}$ satisfies the divergent-series rule:

$$\lambda_k > 0,\ \lambda_k \to 0,\ \sum_{i=0}^{\infty} \lambda_k = \infty.$$

In order to avoid taking decreasing stepsizes (i.e., $\lambda_k \to 0$) as in the classical subgradient algorithm, Nesterov [20] proposed the SDA algorithm,

$$x_{k+1} = \operatorname*{argmin}_{x\in Q} \left\{ \sum_{i=0}^{k} \langle \lambda_i D_i, x - x_0\rangle + \frac{\beta_{k+1}}{2}\|x - x_0\|^2 \right\},\ D_k \in \partial F(x_k),\ \forall k \geq 0, \tag{2.2}$$

where $\{\beta_{k+1}\}_{k=0}^{\infty}$ is a positive nondecreasing sequence and x_0 denotes the initial point. There are two simple strategies for choosing $\{\lambda_i\}_{i=0}^{\infty}$, either $\lambda_i \equiv 1$ or $\lambda_i = 1/\|d_i\|$. SDA can solve the generalized nonsmooth convex optimization problem and it has been proved to be optimal from the view point of worst-case black-box lower complexity bounds [20]. By considering problems with additive structure as in (1.1), Xiao [26] proposed the RDA scheme. A detailed algorithm under the RDA scheme is PSB, which is as follows

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x\in Q} \left\{ \sum_{i=0}^{k} (\langle g_i, x - x_0\rangle + h(x)) + \frac{\beta_{k+1}}{2}\|x - x_0\|^2 \right\} \\ &= \operatorname*{argmin}_{x\in Q} \left\{ \sum_{i=0}^{k} \langle g_i, x - x_0\rangle + (k+1)h(x) + \frac{\beta_{k+1}}{2}\|x - x_0\|^2 \right\}, \end{aligned} \tag{2.3}$$

where $g_i \in \partial f(x_i)$, $\forall i \geq 0$, the stepsize $\lambda_i \equiv 1$, $\{\beta_{k+1}\}_{k=0}^{\infty}$ is nondecreasing, and $x_0 \in \operatorname{argmin}_{x\in Q} h(x)$. The above iteration (2.3) reduces to (2.2) when $h \equiv 0$.

2.2 Algorithm IPSB

Motivated by indefinite approaches, we extend RDA to the following dynamic regularized dual averaging (DRDA) scheme

$$x_{k+1} = \operatorname*{arg\,min}_{x\in Q} \left\{ \sum_{i=0}^{k} \langle g_i, x - x_0\rangle + (k+1)h(x) + \beta_{k+1} r_k(x) \right\}. \tag{2.4}$$

We only assume that the sum $(k+1)h(x)+\beta_{k+1}r_k(x)$ is strongly convex. A simple choice of $r_k(x)$ is

$$r_k(x)=\frac{1}{2}\|x-x_0\|^2_{G_{k+1}},$$

where $G_{k+1}=I-(k+1)\Sigma_h/\beta_{k+1}$. Algorithm 1 describes the algorithm in detail.

Algorithm 1: Indefinite proximal subgradient-based (IPSB) algorithm

Initialization: Set $s_0=0$, $\gamma>0$ and $\{\tilde{\beta}_{k+1}\}_{k=0}^{\infty}$. Initialize $k=0$ and choose

$$x_0\in\arg\min_{x\in Q}h(x);$$

Output: $\hat{x}_{k+1}=\frac{1}{k+1}\sum_{i=0}^{k}x_i$;

1: Stop if a termination criterion is met. Otherwise, compute $g_k\in\partial f(x_k)$;

2: Set $s_{k+1}=s_k+g_k$;

3: Choose $\beta_{k+1}=\gamma\tilde{\beta}_{k+1}$, and set $G_{k+1}:=I-\frac{k+1}{\beta_{k+1}}\Sigma_h$;

4: Solve

$$x_{k+1}=\arg\min_{x\in Q}\left\{\langle s_{k+1},x-x_0\rangle+(k+1)h(x)+\frac{\beta_{k+1}}{2}\|x-x_0\|^2_{G_{k+1}}\right\};$$

5: Let $k=k+1$ and go to step 1.

In Algorithm 1, the choice of the indefinite matrix G_k in step 3 guarantees the strong convexity of the subproblem minimized in step 4. In some specially structured problems, the introduction of G_k can make the subproblem in step 4 much easier to solve.

Remark 2.1 Note that as the progressing of the iteration, the influence of the initial point x_0 should be vanishing. In other words, the auxiliary quadratic term should be as small as possible. By comparing the auxiliary functions in the k-th step of the algorithms PSB and IPSB, we can obtain

$$\begin{aligned}\frac{1}{2}\|x-x_0\|^2_{G_{k+1}}&=\frac{1}{2}\|x-x_0\|^2_{I-(k+1)\Sigma_h/\beta_{k+1}}\\&=\frac{1}{2}\|x-x_0\|^2-\frac{k+1}{2}\|x-x_0\|^2_{\Sigma_h/\beta_{k+1}}\\&<\frac{1}{2}\|x-x_0\|^2,\end{aligned}$$

which indicates that the indefinite term can reduce the impact of x_0 on the k-th subproblem as k increases.

Remark 2.2 The following choice of the sequence $\{\tilde{\beta}_{k+1}\}_{k=0}^{\infty}$ initialized in Algorithm 1 is due to Nesterov [20]:

$$\tilde{\beta}_1=\hat{\lambda},\ \tilde{\beta}_{k+1}=\tilde{\beta}_k+\frac{1}{\tilde{\beta}_k},\ k\in\mathbb{N},\tag{2.5}$$

where $\hat{\lambda}>0$ is an initial parameter.

For the sequence $\{\tilde{\beta}_{k+1}\}_{k=0}^{\infty}$, we have the following estimation, which corrected the previous estimation in [20, Lemma 3].

Lemma 2.1 *Based on (2.5), we have*

$$\sqrt{\hat{\lambda}^2+2k-2} \le \tilde{\beta}_k \le \hat{\lambda}+\frac{1}{\hat{\lambda}}+\sqrt{\hat{\lambda}^2+2k-4},\ \forall k \ge 1. \tag{2.6}$$

Proof According to (2.5), we can obtain $\tilde{\beta}_1=\hat{\lambda}$ and $\tilde{\beta}_k^2=\tilde{\beta}_{k-1}^2+\tilde{\beta}_{k-1}^{-2}+2$. Consequently,

$$\tilde{\beta}_k^2 \ge \tilde{\beta}_{k-1}^2+2 \ge \tilde{\beta}_1^2+2(k-1)=\hat{\lambda}^2+2(k-1),\ \forall k \ge 2,$$

which implies the left-hand side of estimation (2.6). Conversely, we can derive that

$$\begin{aligned}\tilde{\beta}_k &= \tilde{\beta}_{k-1}+\frac{1}{\tilde{\beta}_{k-1}} \le \tilde{\beta}_{k-1}+\frac{1}{\sqrt{\hat{\lambda}^2+2(k-2)}} \le \tilde{\beta}_1+\sum_{t=0}^{k-2}\frac{1}{\sqrt{\hat{\lambda}^2+2t}}\\ &= \hat{\lambda}+\sum_{t=0}^{k-2}\frac{1}{\sqrt{\hat{\lambda}^2+2t}}.\end{aligned} \tag{2.7}$$

From

$$\frac{1}{\sqrt{\hat{\lambda}^2+2t}} \le \frac{2}{\sqrt{\hat{\lambda}^2+2t}+\sqrt{\hat{\lambda}^2+2(t-1)}}=\sqrt{\hat{\lambda}^2+2t}-\sqrt{\hat{\lambda}^2+2(t-1)},$$

we have

$$\sum_{t=0}^{k-2}\frac{1}{\sqrt{\hat{\lambda}^2+2t}}=\frac{1}{\hat{\lambda}}+\sum_{t=1}^{k-2}\frac{1}{\sqrt{\hat{\lambda}^2+2t}} \le \frac{1}{\hat{\lambda}}+\sqrt{\hat{\lambda}^2+2(k-2)}-\hat{\lambda}. \tag{2.8}$$

Finally, the right-hand side of the estimation (2.6) follows from substituting (2.8) into (2.7). □

3 Convergence analysis

Similar to Nesterov's analysis [20], the convergence of the algorithm IPSB is established. First let us define two auxiliary functions as follows

$$U_k(s) := \max_{x\in\mathcal{F}_D}\{\langle s, x-x_0\rangle - kh(x)\}, \tag{3.1}$$

$$V_k(s) := \max_{x\in Q}\left\{\langle s, x-x_0\rangle - kh(x)-\frac{\beta_k}{2}\|x-x_0\|_{G_k}^2\right\}, \tag{3.2}$$

where $\mathcal{F}_D=\{x\in Q:\frac{1}{2}\|x-x_0\|^2\le D\}$, $D>0$ and $G_k=I-k\Sigma_h/\beta_k,\ \forall k\ge 1$. Let $x_0\in\arg\min_{x\in Q}h(x)$. Since $s_0=0$, we have

$$V_1(-s_0)=\max_{x\in Q}\left\{-h(x)-\frac{\beta_1}{2}\|x-x_0\|_{G_1}^2\right\}=\max_{x\in Q}\left\{-h(x)-\frac{1}{2}\|x-x_0\|_{\beta_1 I-\Sigma_h}^2\right\}, \tag{3.3}$$

Notice that (3.3) is a concave maximization problem and then it has a unique optimal solution in the closed convex set Q. According to Danskin's theorem [2, Proposition B.25], we obtain that both $V_1(-s_0)$ and $\nabla V_1(-s_0)$ are well defined. Let

$$T := V_1(-s_0)+\langle -g_0, \nabla V_1(-s_0)\rangle + h(x_1). \tag{3.4}$$

In the following, the first lemma studies the relation between $U_k(s)$ and $V_k(s)$, and the second lemma studies the smoothness of function $V_k(s)$.

Lemma 3.1 *For any $s \in \mathbb{R}^n$ and $k \in \mathbb{N}$, we have*

$$U_k(s) \leq \beta_k D + V_k(s) \tag{3.5}$$

Proof According to the definitions (3.1), (3.2) and $\mathcal{F}_D = \{x \in Q : \frac{1}{2}\|x - x_0\|^2 \leq D\}$, we have

$$\begin{aligned}
U_k(s) &= \max_{x \in \mathcal{F}_D} \{\langle s, x - x_0\rangle - kh(x)\} \\
&\leq \min_{\beta \geq 0} \max_{x \in Q} \left\{\langle s, x - x_0\rangle - kh(x) - \beta(\frac{1}{2}\|x - x_0\|^2 - D)\right\} \\
&\leq \max_{x \in Q} \left\{\langle s, x - x_0\rangle - kh(x) - \beta_k(\frac{1}{2}\|x - x_0\|^2 - D)\right\} \\
&\leq \max_{x \in Q} \left\{\langle s, x - x_0\rangle - kh(x) - \frac{\beta_k}{2}\|x - x_0\|^2 + \beta_k D + \frac{k}{2}\|x - x_0\|^2_{\Sigma_h}\right\} \\
&= \beta_k D + V_k(s),
\end{aligned}$$

where the first inequality corresponds to the partial Lagrangian relaxation, and the last inequality holds as $\Sigma_h \succeq 0$. □

Lemma 3.2 *The well-defined function $V_k(s)$ is convex and continuously differentiable. Then we have*

$$\nabla V_k(s) = x_k(s) - x_0, \ \forall k \geq 1, \tag{3.6}$$

where $x_k(s)$ is the minimizer of the function $V_k(s)$. In addition, $\nabla V_k(s)$ is $1/\beta_k$-Lipschitz continuous, i.e., there exists a constant $1/\beta_k > 0$ such that

$$\|\nabla V_k(s) - \nabla V_k(t)\| \leq \frac{1}{\beta_k}\|s - t\|, \ \forall s, t \in \mathbb{R}^n.$$

Proof Since the objective function of problem (3.2) is β_k-strongly concave with respect to x, $x_k(s)$ is the unique maximizer of $V_k(s)$. Then (3.6) follows from Danskin's theorem [2, Proposition B.25].

For any $l(x) \in \partial h(x)$, $s_1, s_2 \in \mathbb{R}^n$, according to the first-order optimality conditions, we have

$$\begin{aligned}
\langle -s_1 + kl(x_k(s_1)) + \beta_k G_k(x_k(s_1) - x_0), x_k(s_2) - x_k(s_1)\rangle \geq 0, \\
\langle -s_2 + kl(x_k(s_2)) + \beta_k G_k(x_k(s_2) - x_0), x_k(s_1) - x_k(s_2)\rangle \geq 0.
\end{aligned}$$

Adding these two inequalities together, we can get

$$\begin{aligned}
\langle s_2 - s_1, x_k(s_1) - x_k(s_2)\rangle &\leq k\langle l(x_k(s_2)) - l(x_k(s_1)), x_k(s_1) - x_k(s_2)\rangle \\
&\quad + \langle \beta_k G_k(x_k(s_2) - x_k(s_1)), x_k(s_1) - x_k(s_2)\rangle \\
&\leq -k\|x_k(s_1) - x_k(s_2)\|^2_{\Sigma_h} - \beta_k\|x_k(s_1) - x_k(s_2)\|^2_{G_k} \\
&\leq -\beta_k\|x_k(s_1) - x_k(s_2)\|^2, \ \forall k \geq 1,
\end{aligned}$$

where last inequality follows from $G_k = I - \frac{k}{\beta_k}\Sigma_h$. Thus, we have

$$\begin{aligned}\|x_k(s_1) - x_k(s_2)\|^2 &\leq -\frac{1}{\beta_k}\langle s_2 - s_1, x_k(s_1) - x_k(s_2)\rangle \\ &\leq \frac{1}{\beta_k}\|s_2 - s_1\|\|x_k(s_1) - x_k(s_2)\|, \forall k \geq 1,\end{aligned}$$

which is equivalent to

$$\|\nabla V_k(s_1) - \nabla V_k(s_2)\| \leq \frac{1}{\beta_k}\|s_2 - s_1\|, \ \forall k \geq 1.$$

□

Let $F_D^* = \min_{x\in\mathcal{F}_D} F(x)$. According to the convexity of the objective function, we have

$$\begin{aligned}F(\hat{x}_{k+1}) - F_D^* &\leq \frac{1}{k+1}\sum_{i=0}^{k}[f(x_i) + h(x_i)] - \min_{x\in\mathcal{F}_D}[f(x) + h(x)] \\ &= \frac{1}{k+1}\max_{x\in\mathcal{F}_D}\sum_{i=0}^{k}[f(x_i) - f(x) + h(x_i) - h(x)] \\ &\leq \frac{1}{k+1}\max_{x\in\mathcal{F}_D}\sum_{i=0}^{k}[\langle g_i, x_i - x\rangle + h(x_i) - h(x)]. \qquad (3.7)\end{aligned}$$

Consequently, we define the gap function as

$$\delta_{k+1} := \max_{x\in\mathcal{F}_D}\sum_{i=0}^{k}[\langle g_i, x_i - x\rangle + h(x_i) - h(x)].$$

It follows from the inequality (3.5) that

$$\begin{aligned}\delta_{k+1} &= \sum_{i=0}^{k}[\langle g_i, x_i - x_0\rangle + h(x_i)] + U_{k+1}(-s_{k+1}) \qquad (3.8)\\ &\leq \sum_{i=0}^{k}[\langle g_i, x_i - x_0\rangle + h(x_i)] + \beta_{k+1}D + V_{k+1}(-s_{k+1}) \\ &:= \Delta_{k+1}. \qquad (3.9)\end{aligned}$$

Remark 3.1 For any fixed k, there exists a constant P that satisfies $\max_{i\in[k]}\frac{1}{2}\|x_i - x_0\|^2 \leq P$. Thus we have

$$\frac{1}{2}\sum_{i=1}^{k}\|x_{i+1} - x_0\|_{\Sigma_h}^2 \leq \lambda_{max} k P, \qquad (3.10)$$

where λ_{max} is the maximum eigenvalue of Σ_h.

Now we present the upper bounds as follows.

Theorem 3.1 *Let the sequence $\{x_i\}_{i=0}^k \subset Q$ and $\{g_i\}_{i=0}^k \subset \mathbb{R}^n$ be generated by Algorithm 1. Let sequence $\{\beta_i\}_{i=0}^k$ satisfies $\beta_k = \gamma\tilde{\beta}_k$, where $\{\tilde{\beta}_i\}_{i=1}^k$ is defined in (2.5), $\tilde{\beta}_0 = \tilde{\beta}_1$ and $\gamma > 0$. Then*

1. *For any $k \in \mathbb{N}$, we have*

$$\delta_k \le \Delta_k \le \beta_k D + T + \frac{1}{2}\sum_{i=0}^{k-1}\frac{1}{\beta_i}\|g_i\|^2 + \lambda_{max}(k-1)P. \tag{3.11}$$

2. *Assume that*

(1) *the sequence $\{g_k\}_{k\ge 0}$ is bounded, which means that*

$$\exists L > 0, \ such\ that\ \|g_k\| \le L, \ \forall k \ge 0, \tag{3.12}$$

(2) *there exists a solution x^* satisfying*

$$\langle g, x - x^*\rangle \ge 0, \ g \in \partial f(x), \ \forall x \in Q. \tag{3.13}$$

Then it holds that

$$\|x_k - x^*\|^2 \le \frac{2T + 2\lambda_{max}(k-1)P}{\beta_k} + \|x^* - x_0\|^2_{G_k} + L^2. \tag{3.14}$$

3. *Let x^* be an interior solution, i.e., there exist $r, D > 0$ satisfying $B_r(x^*) \subseteq \mathcal{F}_D$. Assume there is a $\Gamma_h > 0$ such that*

$$\max_{\substack{z\in\partial h(y)\\ y\in B_r(x^*)}} \|z\| \le \Gamma_h.$$

Then we have

$$\|\bar{s}_{k+1}\| \le \frac{1}{r(k+1)}\left[\beta_{k+1}D + T + \frac{1}{2}\sum_{i=0}^{k}\frac{1}{\beta_i}\|g_i\|^2 + \lambda_{max}kP\right] + \Gamma_h, \tag{3.15}$$

where $\bar{s}_{k+1} = \frac{1}{k+1}\sum_{i=0}^{k} g_k$.

Proof 1. According to the definitions of $V_k(s)$ and G_k, for any integer $k \ge 1$, we have

$$\begin{aligned}
V_{k-1}(-s_k) &= \max_{x\in Q}\left\{\langle -s_k, x - x_0\rangle - (k-1)h(x) - \frac{\beta_{k-1}}{2}\|x - x_0\|^2_{G_{k-1}}\right\}\\
&\ge \langle -s_k, x_k - x_0\rangle - (k-1)h(x_k) - \frac{\beta_{k-1}}{2}\|x_k - x_0\|^2_{G_{k-1}}\\
&= V_k(-s_k) + h(x_k) + \frac{\beta_k}{2}\|x_k - x_0\|^2_{G_k} - \frac{\beta_{k-1}}{2}\|x_k - x_0\|^2_{G_{k-1}}\\
&\ge V_k(-s_k) + h(x_k) + \frac{\beta_k - \beta_{k-1}}{2}\|x_k - x_0\|^2 - \frac{1}{2}\|x_k - x_0\|^2_{\Sigma_h}\\
&\ge V_k(-s_k) + h(x_k) - \frac{1}{2}\|x_k - x_0\|^2_{\Sigma_h}.
\end{aligned}$$

According to Lemma 3.2, we have

$$V_k(s + \sigma) \le V_k(s) + \langle \sigma, \nabla V_k(s)\rangle + \frac{1}{2\beta_k}\|\sigma\|^2, \ \forall s, \sigma \in \mathbb{R}^n. \tag{3.16}$$

Substituting s_k into (3.16) yields that

$$\begin{aligned} & V_k(-s_k) + h(x_k) - \frac{1}{2}\|x_k - x_0\|^2_{\Sigma_h} \\ & \le V_{k-1}(-s_k) = V_{k-1}(-s_{k-1} - g_{k-1}) \\ & \le V_{k-1}(-s_{k-1}) + \langle -g_{k-1}, \nabla V_{k-1}(-s_{k-1})\rangle + \frac{1}{2\beta_{k-1}}\|g_{k-1}\|^2 \\ & = V_{k-1}(-s_{k-1}) + \langle -g_{k-1}, x_{k-1} - x_0\rangle + \frac{1}{2\beta_{k-1}}\|g_{k-1}\|^2, \ \forall k \ge 1, \end{aligned}$$

which further implies that

$$\begin{aligned} & \langle g_{k-1}, x_{k-1} - x_0\rangle + h(x_k) \\ & \le V_{k-1}(-s_{k-1}) - V_k(-s_k) + \frac{1}{2\beta_{k-1}}\|g_{k-1}\|^2 + \frac{1}{2}\|x_k - x_0\|^2_{\Sigma_h}, \ \forall k \ge 1. \end{aligned}$$

By summing the above inequality from 1 to k, we obtain

$$\begin{aligned} & \sum_{i=1}^{k}[\langle g_i, x_i - x_0\rangle + h(x_{i+1})] \\ & \le V_1(-s_1) - V_{k+1}(-s_{k+1}) + \tfrac{1}{2}\sum_{i=1}^{k}\left[\tfrac{1}{\beta_i}\|g_i\|^2 + \|x_{i+1} - x_0\|^2_{\Sigma_h}\right], \end{aligned}$$

which is equivalent to

$$\begin{aligned} & \sum_{i=0}^{k}[\langle g_i, x_i - x_0\rangle + h(x_i)] + V_{k+1}(-s_{k+1}) \le V_1(-s_1) + h(x_0) + h(x_1) - h(x_{k+1}) \\ & + \frac{1}{2}\sum_{i=1}^{k}\left[\frac{1}{\beta_i}\|g_i\|^2 + \|x_{i+1} - x_0\|^2_{\Sigma_h}\right]. \end{aligned} \tag{3.17}$$

By combining with (3.16) and $s_1 = s_0 + g_0$, we have

$$\begin{aligned} V_1(-s_1) = V_1(-s_0 - g_0) & \le V_1(-s_0) + \langle -g_0, \nabla V_1(-s_0)\rangle + \frac{1}{2\beta_1}\|g_0\|^2 \\ & = T - h(x_1) + \frac{1}{2\beta_0}\|g_0\|^2, \end{aligned}$$

where the equality follows from (3.4) and $\beta_0 = \beta_1$. By noting that $x_0 = \arg\min_{x\in Q} h(x)$, we can obtain that

$$h(x_0) \le h(x_{k+1}).$$

Finally, combining (3.9), (3.17) and the above inequalities, we conclude that

$$\Delta_{k+1} \le \beta_{k+1}D + T + \frac{1}{2}\sum_{i=0}^{k}\frac{1}{\beta_i}\|g_i\|^2 + \frac{1}{2}\sum_{i=1}^{k}\|x_{i+1} - x_0\|^2_{\Sigma_h}.$$

2. Notice that $x_k = \arg\min_{x\in Q}\langle s_k, x - x_0\rangle + kh(x) + \frac{\beta_k}{2}\|x - x_0\|^2_{G_k}$. By the convexity of the objective function, we have

$$\langle s_k + kl_k + \beta_k G_k(x_k - x_0), x - x_k\rangle \ge 0, \ \forall x \in Q. \tag{3.18}$$

Notice that $G_k = I - k\Sigma_h/\beta_k$. Then we can define the following β_k-strongly convex function

$$\phi_k(x) := kh(x) + \frac{\beta_k}{2}\|x - x_0\|^2_{G_k},\ k \in \mathbb{N},$$

which implies that

$$\phi_k(x) \geq \phi_k(x_k) + \langle kl_k + \beta_k G_k(x_k - x_0), x - x_k\rangle + \frac{\beta_k}{2}\|x_k - x\|^2. \tag{3.19}$$

By taking $\phi_k(x_k)$ from the right-hand side of the inequality (3.19) to the left-hand side, we can get

$$\begin{aligned}&\langle kl_k + \beta_k G_k(x_k - x_0), x - x_k\rangle + \frac{\beta_k}{2}\|x_k - x\|^2\\ &\leq k[h(x) - h(x_k)] + \frac{\beta_k}{2}\|x - x_0\|^2_{G_k} - \frac{\beta_k}{2}\|x_k - x\|^2_{G_k}.\end{aligned}$$

Combining with (3.18) yields that

$$\begin{aligned}\frac{\beta_k}{2}\|x_k - x\|^2 &\leq kh(x) - kh(x_k) + \frac{\beta_k}{2}\|x - x_0\|^2_{G_k} - \frac{\beta_k}{2}\|x_k - x_0\|^2_{G_k}\\ &\quad + \langle kl_k + \beta_k G_k(x_k - x_0), x_k - x\rangle\\ &\leq kh(x) - kh(x_k) + \frac{\beta_k}{2}\|x - x_0\|^2_{G_k} - \frac{\beta_k}{2}\|x_k - x_0\|^2_{G_k} - \langle s_k, x_k - x\rangle\\ &= V_k(s_k) + kh(x) + \frac{\beta_k}{2}\|x - x_0\|^2_{G_k} + \langle s_k, x - x_0\rangle\\ &= V_k(s_k) + \sum_{i=0}^{k-1}\langle g_i, x_i - x_0\rangle + \sum_{i=0}^{k-1} h(x_i)\\ &\quad + \frac{\beta_k}{2}\|x - x_0\|^2_{G_k} + \sum_{i=0}^{k-1}\langle g_i, x - x_i\rangle + kh(x) - \sum_{i=0}^{k-1} h(x_i).\end{aligned} \tag{3.20}$$

Furthermore, we notice that (3.17) is taken into the following form

$$V_{k+1}(-s_{k+1}) + \sum_{i=0}^{k}[\langle g_i, x_i - x_0\rangle + h(x_i)] \leq T + \frac{1}{2}\sum_{i=0}^{k}\frac{1}{\beta_i}\|g_i\|^2 + \frac{1}{2}\sum_{i=1}^{k}\|x_{i+1} - x_0\|^2_{\Sigma_h}. \tag{3.21}$$

By substituting (3.20) into (3.21), we can get

$$\begin{aligned}\frac{\beta_k}{2}\|x_k - x\|^2 \leq T\ &+\frac{1}{2}\sum_{i=0}^{k-1}\frac{1}{\beta_i}\|g_i\|^2 + \frac{1}{2}\sum_{i=1}^{k-1}\|x_{i+1} - x_0\|^2_{\Sigma_h}\\ &+\frac{\beta_k}{2}\|x - x_0\|^2_{G_k} + \left\{\sum_{i=0}^{k-1}[f(x) + h(x)] - [f(x_i) + h(x_i)]\right\}.\end{aligned}$$

Finally, we set $x = x^* := \arg\min_{x\in\mathcal{F}_D} f(x) + h(x)$. Then it holds that

$$\frac{\beta_k}{2}\|x_k - x^*\|^2 \leq T + \frac{1}{2}\sum_{i=0}^{k-1}\frac{1}{\beta_i}\|g_i\|^2 + \frac{1}{2}\sum_{i=1}^{k-1}\|x_{i+1} - x_0\|^2_{\Sigma_h} + \frac{\beta_k}{2}\|x^* - x_0\|^2_{G_k}.$$

According to the conditions (2.5) and (3.12), we obtain the inequality (3.14).

3. Based on (3.8), we can obtain

$$\begin{aligned}\delta_{k+1} &= \sum_{i=0}^{k}[\langle g_i, x_i - x^*\rangle + h(x_i)] + \max_{x\in\mathcal{F}_D}\{\langle s_{k+1}, x^* - x\rangle - (k+1)h(x)\}\\ &= \sum_{i=0}^{k}[\langle g_i, x_i - x^*\rangle + h(x_i) - h(x^*)] + \max_{x\in\mathcal{F}_D}\{\langle s_{k+1}, x^* - x\rangle + (k+1)h(x^*)\\ &\quad - (k+1)h(x)\}\\ &\geq \sum_{i=0}^{k}\{f(x_i) + h(x_i) - f(x^*) - h(x^*)\} + \max_{x\in\mathcal{F}_D}\{\langle s_{k+1}, x^* - x\rangle + (k+1)h(x^*)\\ &\quad - (k+1)h(x)\}\\ &\geq \max_{x\in B_r(x^*)}\{\langle s_{k+1}, x^* - x\rangle + (k+1)h(x^*) - (k+1)h(x)\}.\end{aligned}$$

Notice that

$$\bar{x} = \arg\max_{x\in B_r(x^*)}\langle s_{k+1}, x^* - x\rangle.$$

Then we have $\|x^* - \bar{x}\| = r$ and

$$\langle s_{k+1}, x^* - \bar{x}\rangle = \|s_{k+1}\|\|x^* - \bar{x}\| = r\|s_{k+1}\|.$$

Thus, it holds that

$$\begin{aligned}\delta_{k+1} &\geq \max_{x\in B_r(x^*)}\{\langle s_{k+1}, x^* - x\rangle + (k+1)h(x^*) - (k+1)h(x)\}\\ &\geq \langle s_{k+1}, x^* - \bar{x}\rangle + (k+1)h(x^*) - (k+1)h(\bar{x})\\ &\geq r\|s_{k+1}\| + (k+1)\langle l(\bar{x}), x^* - \bar{x}\rangle\\ &\geq r\|s_{k+1}\| - (k+1)\|l(\bar{x})\|\|x^* - \bar{x}\|\\ &= r\|s_{k+1}\| - (k+1)r\|l(\bar{x})\|,\end{aligned}$$

which implies

$$\frac{1}{k+1}\|s_{k+1}\| \leq \frac{1}{r(k+1)}\delta_{k+1} + \|l(\bar{x})\| \leq \frac{1}{r(k+1)}\delta_{k+1} + \Gamma_h.$$

Then (3.15) follows from (3.11). □

As a main result, we can now estimate the upper bound on the complexity of IPSB in the following.

Theorem 3.2 *Assume there exists a constant $L > 0$ such that $\|g_k\| \leq L$, $\forall k \geq 0$. Denote by $\{x_i\}_{i=0}^k$ the sequence generated by Algorithm 1. Let $F_D^* = \min_{x\in\mathcal{F}_D} F(x)$. Then we have*

$$F(\hat{x}_{k+1}) - F_D^* - \lambda_{max}P \leq \frac{\tilde{\beta}_{k+1}}{k+1}\left(\gamma D + \frac{L^2}{2\gamma}\right) + \frac{T - \lambda_{max}P}{k+1}. \tag{3.22}$$

Proof By combining (2.5) with the inequalities (3.7) and (3.11), we have

$$F(\hat{x}_{k+1}) - F_D^* \leq \frac{1}{k+1}\delta_{k+1}(D) \leq \frac{1}{k+1}\left[\beta_{k+1}D + T + \frac{1}{2}\sum_{i=0}^{k}\frac{1}{\beta_i}\|g_i\|^2 + \lambda_{max}kP\right]$$
$$\leq \frac{\tilde{\beta}_{k+1}}{k+1}\left(\gamma D + \frac{L^2}{2\gamma}\right) + \frac{T - \lambda_{max}P}{k+1} + \lambda_{max}P,$$

which finishes the proof of the inequality (3.22). □

Remark 3.2 According to Lemma 2.1, we know that the sequence $\{\tilde{\beta}_k\}_{k=0}^{\infty}$ can be used for balancing the terms appearing in the right-hand side of inequality (3.11). It follows from Theorem 3.2 that IPSB converges to the region of the optimal value with rate $O(1/\sqrt{k})$.

4 Numerical experiments

In this section, we perform numerical experiments to compare the algorithms IPSB, SDA and PSB on two kinds of test problems. All experiments were implemented in MTALAB 2018b and run on a laptop with a dual core (1.6 + 1.8 GHz) processor and 8 GB RAM.

4.1 Regularized least squares problem

In this subsection, we test the regularized least squares problem

$$\min_{x\in\mathbb{R}^n}\left\{\|Ax - b\|_2 + \bar{\rho}\max_{i\in[m]} f_i(x)\right\}, \tag{4.1}$$

where $A \in \mathbb{R}^{n_1\times n_2}$, $b \in \mathbb{R}^{n_1}$ and $f_i(x), i \in [m]$ are all positive and strongly convex. Notice that (4.1) is a special case of (1.1) with $f(x) = \|Ax - b\|_2$ and $h(x) = \bar{\rho}\max_{i\in[m]} f_i(x)$.

In our first test, we set $m = 2$, $f_1(x) = \|x\|_B^2$ and $f_2(x) = \|x - c\|_B^2$, where $B \in \mathbb{R}^{n_2\times n_2}$ and $c \in \mathbb{R}^{n_2}$. We set $B \neq 0$ to be positive semidefinite but singular so that the function h is SQCC. In fact, we have $\Sigma_h = 2B$. Applying Algorithm 1 to solve (4.1) reduces to

$$\begin{cases} s_{k+1} = s_k + g_k, \\ x_{k+1} = \arg\min\limits_{x}\left\{\langle s_{k+1}, x\rangle + (k+1)\max\{\|x\|_B^2, \|x - c\|_B^2\} + \dfrac{\beta_{k+1}}{2}\|x - x_0\|^2_{I - \frac{2(k+1)}{\beta_{k+1}}B}\right\}, \end{cases}$$

where $g_k \in \partial f(x_k)$ and

$$\partial f(x) = \begin{cases} \dfrac{A^T(Ax - b)}{\|Ax - b\|}, & if\ Ax - b \neq 0, \\ \{A^T x \in \mathbb{R}^n : \|x\| \leq 1\}, & if\ Ax - b = 0. \end{cases}$$

The three different algorithms in comparison for solving (4.1) are explicitly reformulated as

$$SDA: x_{k+1} = x_0 - \frac{1}{\beta_{k+1}} \sum_{i=0}^{k} (g_i + l_i),$$

$$PSB: x_{k+1} = \begin{cases} (2(k+1)B + \beta_{k+1} I)^{-1} (\beta_{k+1} x_0 - s_{k+1}), & \text{if } \|x\|_B^2 > \|x - c\|_B^2, \\ (2(k+1)B + \beta_{k+1} I)^{-1} (\beta_{k+1} x_0 - s_{k+1}), & \text{if } \|x\|_B^2 < \|x - c\|_B^2, \\ \left(2(k+1)B + \beta_{k+1} \bar{G}\right)^{-1} (\beta_{k+1} x_0 - s_{k+1} + \bar{\rho}_1 Bc), & \text{otherwise}, \end{cases}$$

$$IPSB: x_{k+1} = \begin{cases} x_0 - \frac{1}{\beta_{k+1}} (2(k+1)Bx_0 + s_{k+1}), & \text{if } \|x\|_B^2 > \|x - c\|_B^2, \\ x_0 - \frac{1}{\beta_{k+1}} (2(k+1)Bx_0 + s_{k+1} - 2(k+1)Bc), & \text{if } \|x\|_B^2 < \|x - c\|_B^2, \\ x_0 - \frac{1}{\beta_{k+1}} (2(k+1)Bx_0 + s_{k+1} + 2\bar{\rho}_2 Bc), & \text{otherwise}, \end{cases}$$

where expression for $\bar{\rho}_1$, $\bar{\rho}_2$ and $\bar{G}$ is as follows

$$\bar{\rho}_1 = \frac{(k+1)c^T (Bc + s_{k+1} - \beta_{k+1} x_0)}{c^T Bc},$$

$$\bar{\rho}_2 = \frac{1}{4\|Bc\|^2} \left(\beta_{k+1} c^T B(c - x_0) + 2c^T B s_{k+1} + 4(k+1)c^T B \cdot Bx_0 \right),$$

$$\bar{G} = I - \frac{Bc \cdot c^T}{c^T Bc}.$$

In addition, $l_k \in \partial h(x_k)$ and

$$\partial h(x) = \{ l \in \mathbb{R}^{n_2 \times n_2} : l = 2Bx + 2\alpha Bc,\ \alpha \in [0, 1] \}.$$

In our experiments, we choose $\bar{\rho} = 1$ and $n_1 \times n_2 \in \{400 \times 900, 800 \times 2000, 1500 \times 3000\}$. In Algorithm 1, we set $\gamma = 20$, $\hat{\lambda} = 1e - 3$ and the termination criterion is set as either $|F(\hat{x}^k) - F(\hat{x}^{k-1})| \leq 10^{-3}$ or the number of iterations reaches 300. Starting from a fixed seed, we independently randomly generate $x^* = (10, \ldots, 10) \in \mathbb{R}^{n_2}$, $c \in \mathbb{R}^{n_2}$ from standard normal distribution $\mathcal{N}(0, 0.25)$ and then generate each elment of A from $\mathcal{N}(0, 20^2)$. We set $b \in \mathbb{R}^{n_1}$ as follows

$$b_i = \sum_{j=1}^{n_2} A_{ij} x^*,\ i \in [n_1].$$

The matrix B is constructed by randomly generating eigenvalues and eigenvectors. The first ten eigenvalues of B are random positive numbers and the rest are zero. We construct the eigenvecters by randomly generating orthogonal matrix with uniformly distributed random elements. When $n_1 = 400$, $n_2 = 900$, MATLAB code to generate the above data is as follows. The others are similar.

```
n_1=400; n2=900;
randn('seed',0)
A = 20*randn(n1,n2); x_op = randn(n2,1);
b = A*x_op;
rand('seed',0)
y = 5e2*rand(10,1); x = [y; zeros(n2-10,1)];
X = diag(x); U = orth(rand(n2,n2));
B = U' * X * U;
c = 0.5*randn(n2,1);
```

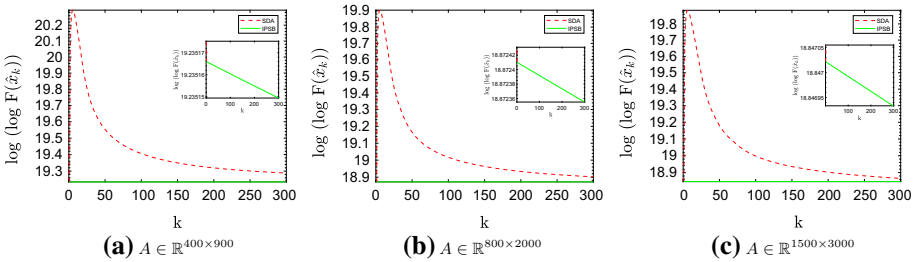

Fig. 1 Numerical comparison between algorithms SDA and IPSB

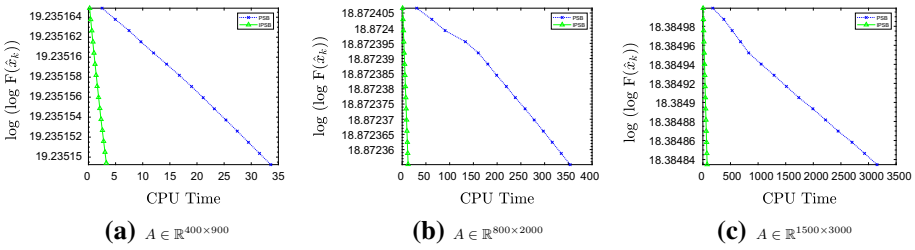

Fig. 2 Numerical comparison between algorithms PSB and IPSB

where x_op corresponds to the variable x^*.

We plot the variants of $\log(\log F(\hat{x}_k))$ versus the iterations number k and CPU runtime in Figs. 1 and 2, respectively.

As shown in Fig. 1, IPSB would have a better function value than SDA in the iteration process. By zooming in on the details of the figures, it can be seen that the value generated by IPSB is decreasing rather than constant.

PSB runs much slower than IPSB because of the heavy computation cost of matrix inversion. It is shown in Fig. 2 that IPSB is much more efficient than PSB.

4.2 Elastic net regression

The elastic net is a regularized regression model [27] by linearly combining LASSO and ridge regression. It is formulated as

$$\min_{\omega\in\mathbb{R}^n} \|y - X\omega\|_2^2 + \eta_1\|\omega\|_1 + \eta_2\|\omega\|_2^2, \tag{4.2}$$

where p is the number of samples, n is the number of features, $y \in \mathbb{R}^p$ is the response vector, $X \in \mathbb{R}^{p\times n}$ is the design matrix, and $\eta_1, \eta_2 > 0$ are regularization parameters. It corresponds to setting $f(\omega) = \eta_1\|\omega\|_1$ and $h(\omega) = \|y - X\omega\|_2^2 + \eta_2\|\omega\|_2^2$ in (1.1). The iteration schemes of three different algorithms in comparison for solving (4.2) are reformulated as

Table 1 Computational cost in each iteration

Algorithm	SDA	PSB	IPSB
Complexity per iteration	$\mathcal{O}(kpn)$	$\mathcal{O}(n^3 + pn^2 + kn)$	$\mathcal{O}(kn + pn)$

$$SDA: \omega_{k+1} = \omega_0 + \frac{1}{\beta_{k+1}} \sum_{i=0}^{k} \left(2X^T(X\omega_i - y) + \eta_1 sgn(\omega_i) + 2\eta_2\omega_i\right),$$

$$PSB: \omega_{k+1} = ((k+1)(2X^TX + 2\eta_2 I) + \beta_{k+1} I)^{-1} \left(\beta_{k+1}\omega_0 + 2X^T y - \eta_1 \sum_{i=0}^{k} sgn(\omega_i)\right),$$

$$IPSB: \omega_{k+1} = \omega_0 - \frac{1}{\beta_{k+1}} \left((k+1)(2X^T(X\omega_0 - y) + 2\eta_2\omega_0) + \eta_1 \sum_{i=0}^{k} sgn(\omega_i)\right),$$

where the initial point ω_0 is given by $(X^TX + \eta_2 I)^{-1}Xy = \arg\min_{\omega\in\mathbb{R}^n} h(\omega)$, $sgn(\cdot)$ is the sign function, and the sequence $\{\beta_k\}_{k\geq 0}$ utilizes the form (2.5). We list in Table 1 the computational complexity in each iteration. It demonstrates that in each iteration PSB has the highest computational cost when n is much larger than k, and SDA takes the highest cost when k is much larger than p and n.

We set the termination criterion as

$$\frac{|f(\bar{\omega}) - \bar{f}|}{\bar{f}} \leq \epsilon^{rel},$$

where $\bar{f}$ is an approximation of the optimal value obtained by running 500 iterations of SDA in advance, $\bar{\omega} = \sum_{i=1}^{t} \omega_i/n$ and t is the realistic number of iterations until termination.

We conduct the experiments with the following synthetic data and real data, respectively.

Synthetic data: Starting a fixed seed, we independently and randomly generate $X_{ij} \sim \mathcal{N}(0, 0.01)$, $\omega^* \sim \mathcal{N}(0, 1)$, $\epsilon_i \sim \mathcal{N}(0, 0.04)$, and then set $y_i = \sum_{j=1}^{n} X_{ij}\omega^* + \epsilon_i$, $i \in [p]$, $j \in [n]$. We choose $p \times n \in \{300 \times 1000, 500 \times 2000, 700 \times 3000, 1000 \times 5000\}$. The hyperparameters used for the synthetic data are set as

$$\gamma = 10^2,\ \beta_0 = 3 \times 10^{-2},\ \epsilon^{rel} = 10^{-1}.$$

When $p = 300$, $n = 1000$, the MATLAB code to generate the data is as follows. The others are similar.

```
p=300; n=1000;
randn('seed',0);
X=0.1.*randn(p,n);
x_op=randn(n,1);
ep=0.2.*randn(p,1);
y=X*x_op+ep;
```

where x_op and ep correspond to the variables ω^* and ϵ respectively.

MNIST data [15]: There are 70, 000 samples from the images of 10 digits in the MNIST data set, each with a 28×28 gray-scale pixel-map, for a total of 784 features. We take the digits 8 and 9. Thus we have $p = 13783$ and $n = 784$. Moveover, let $y \in \{+1, -1\}^n$ be the binary label. The hyperparameters used for MNIST data are as follows

$$\gamma = 10^3,\ \beta_0 = 10^{-3},\ \epsilon^{rel} = 10^{-2}.$$

Table 2 Numerical results for synthetic data

p	n	SDA			PSB			IPSB		
		Iter.	Time	Accu.	Iter.	Time	Accu.	Iter.	Time	Accu.
300	1000	178	0.5548	0.9007	121	10.2557	0.9009	**120**	**0.0575**	0.9009
500	2000	121	1.3441	0.9005	84	50.9602	0.9012	**83**	**0.2241**	0.9011
700	3000	98	1.9449	0.9009	71	146.9056	0.9007	**70**	**0.4141**	0.9002
1000	5000	76	3.0670	0.9001	63	629.0511	0.9004	**66**	**1.0915**	0.9001

Bold values indicate the result of running the algorithm proposed in this article, which have certain advantages in comparison

Table 3 Numerical results for MNIST data

p	n	SDA			PSB			IPSB		
		Iter.	Time	Accu.	Iter.	Time	Accu.	Iter.	Time	Accu.
13783	784	29	2.57	0.9902	31	2.14	0.9901	**32**	**0.62**	0.9901

Bold values indicate the result of running the algorithm proposed in this article, which have certain advantages in comparison

Tables 2 and 3 represent the experimental results for synthetic data and MNIST data, respectively. In both Tables, we report the results of the numbers of iterations (Iter.), running time in seconds and the accuracy (Accu.) defined as $1 - |f(\bar{\omega}) - \bar{f}|/\bar{f}$.

In synthetic data, SDA takes the largest number of iterations among the three, IPSB runs in less CPU time than the other two algorithms, and PSB is the most inefficient one. In MNIST data, the three algorithms take almost the same number of iterations so that IPSB takes the least CPU time.

5 Conclusions

Nesterov's dual averaging scheme succeeds in avoiding that stepsizes decrease as in the subgradient methods for nonsmooth convex minimizing problem. It is then extended to solve problems with an additional regularization, denoted by (RDA).

In this paper, we propose the dynamic regularized dual averaging scheme by relaxing the positive definite regularization term in RDA, which can not only reduce the impact of the initial point on the subproblems in later iterations but also make the new subproblem in each iteration easy to solve. Under this new scheme, we proposed indefinite proximal subgradient-based (IPSB) algorithm. We analyze the convergence rate of IPSB, which is $O(1/\sqrt{k})$, where k is the number of iterations. And IPSB converges to a region of the optimal value. Numerical experiments on regularized least squares problem and elastic net regression show that IPSB is more efficient than the existing algorithms SDA and PSB. Future works include more real applications of IPSB and further improvement of IPSB by, for example, relaxing the condition on the initial point.

Acknowledgements The authors thank the editor and the referees for the valuable comments/suggestions, which help us improve the paper greatly. The research of the second author was partially supported by NSFC with Nos. 12131004 and 12126603; and the research of the third author was partially supported by NSFC with No. 12171021 and by Beijing NSF with No. Z180005.

Data availability statements The authors confirm that all data generated or analysed during this study are included in the paper.

References

1. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
2. Bertsekas, D.P.: Nonlinear Programming. Taylor & Francis, Milton Park (1997)
3. Boyd, S., Xiao, L., Mutapcic, A.: Subgradient methods. Lecture Notes of EE392o, Stanford University, Autumn Quarter, 2004:2004–2005 (2003)
4. Burke, J.V., Curtis, F.E., Lewis, A.S., Overton, M.L., Simões, L.E.: Gradient sampling methods for nonsmooth optimization. In: Numerical Nonsmooth Optimization, pp. 201–225. Springer (2020)
5. Cai, X.-J., Guo, K., Jiang, F., Wang, K., Wu, Z.-M., Han, D.-R.: The developments of proximal point algorithms. J. Oper. Res. Soc. China 1–43 (2022)
6. Combettes,P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer (2011)
7. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward–backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
8. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(7), 2021–2059 (2011)
9. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1), 293–318 (1992)
10. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**(6), 721–741 (1984)
11. Grötschel, M., Lovász, L., Schrijver, A.: The ellipsoid method. In: Geometric Algorithms and Combinatorial Optimization, pp. 64–101. Springer (1993)
12. He, B., Ma, F., Yuan, X.: Optimal proximal augmented Lagrangian method and its application to full Jacobian splitting for multi-block separable convex minimization problems. IMA J. Numer. Anal. **40**(2), 1188–1216 (2020)
13. Jiang, F., Cai, X., Han, D.: The indefinite proximal point algorithms for maximal monotone operators. Optimization **70**(8), 1759–1790 (2021)
14. Jiang, F., Wu, Z., Cai, X.: Generalized ADMM with optimal indefinite proximal term for linearly constrained convex optimization. J. Ind. Manag. Optim. **16**(2), 835–856 (2020)
15. LeCun, Y., Cortes, C., Burges, C.J.C.: The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/ (2017)
16. Li, M., Sun, D., Toh, K.C.: A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. SIAM J. Optim. **26**(2), 922–950 (2016)
17. Mäkelä, M.: Survey of bundle methods for nonsmooth optimization. Optim. Methods Softw. **17**(1), 1–29 (2002)
18. Martinet, B.: Regularization d'inequations variationelles par approximations successives. Revue Francaise d'Informatique et de Recherche Opérationelle **4**, 154–159 (1970)
19. Nesterov, Y.: Smooth minimization of nonsmooth functions. Math. Program. **103**(1), 127–152 (2005)
20. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009)
21. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer, Berlin (2013)
22. Ram, S.S., Nedić, A., Veeravalli, V.V.: Incremental stochastic subgradient algorithms for convex optimization. SIAM J. Optim. **20**(2), 691–717 (2009)
23. Ram, S.S., Nedić, A., Veeravalli, V.V.: Distributed stochastic subgradient projection algorithms for convex optimization. J. Optim. Theory Appl. **147**(3), 516–545 (2010)
24. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**(5), 877–898 (1976)
25. Shor, N.Z.: Minimization Methods for Non-differentiable Functions. Springer Series in Computational Mathematics, Springer, Berlin (1985)
26. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **11**(10), 2543–2596 (2010)
27. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**, 301–320 (2005)