Numer Algor (2017) 76:1071–1097
DOI 10.1007/s11075-017-0297-x

ORIGINAL PAPER

Optimal subgradient algorithms for large-scale convex optimization in simple domains

Masoud Ahookhosh[1] · Arnold Neumaier[1]

Received: 25 September 2016 / Accepted: 19 February 2017 / Published online: 14 March 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract This paper describes two optimal subgradient algorithms for solving structured large-scale convex constrained optimization. More specifically, the first algorithm is optimal for smooth problems with Lipschitz continuous gradients and for Lipschitz continuous nonsmooth problems, and the second algorithm is optimal for Lipschitz continuous nonsmooth problems. In addition, we consider two classes of problems: (i) a convex objective with a simple closed convex domain, where the orthogonal projection onto this feasible domain is efficiently available; and (ii) a convex objective with a simple convex functional constraint. If we equip our algorithms with an appropriate prox-function, then the associated subproblem can be solved either in a closed form or by a simple iterative scheme, which is especially important for large-scale problems. We report numerical results for some applications to show the efficiency of the proposed schemes.

Keywords Structured convex optimization · Nonsmooth optimization · Optimal complexity · First-order black-box information · Subgradient method · High-dimensional data

Mathematics Subject Classification (2010) 90C25 · 90C60 · 49M37 · 65K05 · 68Q25

✉ Masoud Ahookhosh
masoud.ahookhosh@univie.ac.at

Arnold Neumaier
Arnold.Neumaier@univie.ac.at

[1] Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

1 Introduction

Convex optimization has been shown to provide efficient algorithms for computing reliable solutions in a broad range of applications. Many applications arising in applied sciences and engineering such as signal and image processing, machine learning, statistics, and general inverse problems can be addressed by a convex optimization problem involving high-dimensional data. In practice, solving a nonsmooth convex problem is usually more difficult and costly than a smooth one. More precisely, for a prescribed accuracy parameter $\varepsilon > 0$, the optimal complexity to achieve an ε-solution of nonsmooth Lipschitz continuous problems is $\mathcal{O}(\varepsilon^{-2})$, the superior complexity $\mathcal{O}(\varepsilon^{-1/2})$ for smooth problems with Lipschitz continuous gradients, cf. [41, 42].

Thanks to the low memory requirement and simple structure, first-order methods have received much attention during the past few decades. Indeed, they deal successfully with large-scale problems. In general, convex optimization problems can be solved by gradient-type algorithms [3, 16, 17], conjugate gradient methods [12, 18], for smooth objectives and by subgradient-type methods [22, 40], proximal gradient methods [26, 51], smoothing techniques [13, 19, 28, 45], bundle-type algorithms [37, 38], and primal-dual first-order methods [20, 21, 23] for nonsmooth objectives. Moreover, both classes can be addressed by (zero-order) coordinate descent methods and derivative-free methods. The current paper only addresses first-order methods and assumes that first-order black-box information—function values and subgradients— of the objective function are available.

Historically, subgradient methods were the first numerical schemes proposed to solve optimization problems with nonsmooth convex objective functions. In practice, subgradient methods were too slow, especially for badly scaled problems, although they attain the optimal worst-case complexity $\mathcal{O}(\varepsilon^{-2})$. In 1983, Nemirovski & Yudin in [41] derived optimal worst-case complexity bounds of first-order methods to achieve an ε-solution for several classes of problems such as Lipschitz continuous nonsmooth problems and smooth problems with Lipschitz continuous gradients. If an algorithm attains the optimal worst-case complexity bound for a class of problems, it is called optimal. The pioneering optimal first-order method dated back to Nesterov [43] in 1983. This optimal first-order method is interesting both theoretically and computationally, attracting many researchers to work in the development of such schemes, see Auslander & Teboulle [9], Beck &Teboulle [14], Devolder et al. [27], Gonzaga et al. [29, 30], Lan [38], Lan et al. [39], Nesterov [45–47], and Tseng [53]. Computational comparisons for composite functions show that optimal Nesterov-type first-order methods are substantially superior to the gradient descent and subgradient methods, see, e.g., Ahookhosh [1, 2] and Becker et al. [15].

Content In this paper, we first review the OSGA algorithm from Neumaier [48] for solving such problems and develop a new, simplified version of it. Unlike the original OSGA, the new algorithm, called OSGA-V, only needs a single solution of the OSGA subproblem, but it is still optimal for Lipschitz continuous nonsmooth problems.

We then consider structured convex constrained optimization problems frequently observed in applications. We show how to solve the OSGA subproblem for two classes of convex domains, namely (i) simple convex domains such that the

orthogonal projection is cheap to compute and (ii) sublevel sets of a convex function referred as a functional domain. For problems with a simple domain, we first introduce an appropriate prox-function and then show that the solution of the associated subproblem is obtained by a projection onto the domain followed by solving a one-dimensional nonlinear equation. It is shown that if an explicit formula for projection is available, the nonlinear equation can be solved in a closed form in many interesting cases. We also establish the optimality conditions for a functional domain and find a closed form solution for some such constraints.

Recently, Nesterov [44] proposed a fast gradient method with optimal complexity for smooth problems with Lipschitz continuous gradients, nonsmooth problems with bounded variation subgradients, and weakly smooth problems with Hölder continuous gradients. Similar to OSGA-V and OSGA, this method needs no global parameters such as Lipschitz or Hölder constants at the cost of applying a backtracking line search. Typically, this line search requires several solutions the related subproblem increasing the total computational cost, whereas OSGA-V and OSGA do not need such a line search. In addition, the subproblem of Nesterov's universal gradient method can be solved efficiently only when the nonsmooth part of the objective is sufficiently simple. For many examples such as isotropic or anisotropic total variation (cf. Section 5.1), the subproblem of Nesterov's universal gradient method can only be solved approximately, at a significant cost; on the other hand, OSGA-V and OSGA only require first-order information to handle such problems.

Finally, we report numerical results for applications to show the efficiency of the proposed schemes compared with some state-of-the-art algorithms on both nonsmooth and smooth constrained problems. Remarkably, the new simplified algorithm slightly outperforms OSGA (and all other algorithms) on smooth problems with Lipschitz continuous gradients for which the latter have provable optimal complexity $\mathcal{O}(\varepsilon^{-1/2})$. Thus suggests that a similar complexity bound can perhaps be proved for OSGA-V, but we haven't been able to do so.

The remainder of this paper is organized as follows. In Section 2, we give the basic idea of the optimal subgradient framework resulting to two algorithms. Sections 3 and 4 describes how to applying these algorithms to the convex problems with simple domains. We report numerical results and conclusions in Sections 5 and 6, respectively.

Notation and preliminaries Let $\mathcal{V}$ be a finite-dimensional vector space endowed with the norm $\|\cdot\|$, and let $\mathcal{V}^*$ denotes its dual space, formed by all linear functional on $\mathcal{V}$ where the bilinear pairing $\langle g, x\rangle$ denotes the value of the functional $g \in \mathcal{V}^*$ at $x \in \mathcal{V}$. The associated dual norm of $\|\cdot\|$ is defined by

$$\|g\|_* = \sup_{z\in V}\{\langle g, z\rangle : \|z\| \le 1\}.$$

We define $x_+ = \max\{x, 0\}$ and $x_- = \max\{-x, 0\}$. If $\mathcal{V} = \mathbb{R}^n$, then, for $1 \le p \le \infty$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \ \|x\|_{1,p} = \sum_{i=1}^{m} \|x_{g_i}\|_p,$$

where $x = (x_{g_1}, \ldots, x_{g_m}) \in \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m}$ in which $n_1 + \ldots + n_m = n$. For a function $f : \mathcal{V} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, we denote by

$$\mathrm{dom} f := \{x \in \mathcal{V} | f(x) < +\infty\}$$

its effective domain and call f proper if $\mathrm{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathcal{V}$. The vector $g \in \mathcal{V}^*$ is called a subgradient of f at x if $f(x) \in \mathbb{R}$ and

$$f(y) \geq f(x) + \langle g, y - x\rangle \text{ for all } y \in \mathcal{V}.$$

The set $\partial f(x)$ of all subgradients is called the subdifferential of f at x.

Let us consider a nonempty, closed, and convex subset C of $\mathcal{V}$. The normal cone $N_C(x)$ of C at x is defined by

$$N_C(x) = \{p \in \mathcal{V} \mid \langle p, x - z\rangle \geq 0 \forall z \in C\}. \tag{1}$$

We call C a simple convex domain if the orthogonal projection

$$P_C(y) := \operatorname*{argmin}_{x \in C} \frac{1}{2}\|x - y\|^2 \tag{2}$$

of y onto C can be found efficiently for every $y \in \mathcal{V}$. Computing the orthogonal projection is a well-studied topic in convex optimization, and the projection operator is available for many domains C either in a closed form or by a simple iterative scheme (cf. Table 5.1 in [2]).

2 Optimal subgradient algorithms

In this section, we review the main idea of the optimal subgradient framework proposed by Neumaier in [48] for solving the convex constrained minimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in C, \end{array} \tag{3}$$

where $f : C \to \mathbb{R}$ is a proper and convex function defined on a nonempty, closed, and convex subset C of $\mathcal{V}$. More specifically, we give two subgradient algorithms for problem (3), where the first one requires double solutions of the subproblem (8) (see Algorithm 1 originally proposed in [48]) and the second one needs a single solving of the subproblem (8) (see Algorithm 2). Both algorithms use first-order information, i.e., function values and subgradients, to construct a sequence of iterations $\{x_k\}_{k\geq 0} \subseteq C$ whose function values $\{f_k\}_{k\geq 0}$ converge to the minimum $\widehat{f} := f(\widehat{x})$ with the optimal complexity. The algorithms require no information regarding global parameters such as Lipschitz constants of function values and gradients.

We fix a continuously differentiable prox-function $Q : C \to \mathbb{R}$ satisfying

$$Q_0 := \min_{z \in C} Q(z) > 0 \tag{4}$$

and

$$Q(z) \geq Q(x) + \langle g_Q(x), z - x\rangle + \frac{\sigma}{2}\|z - x\|^2 \text{ for all } x, z \in C, \tag{5}$$

where $\sigma = 1$, $g_Q(x)$ denotes the gradient of Q at $x \in C$ and $\|\cdot\|$ is a norm defined on $\mathcal{V}$. At each iteration, the sequence of function values satisfies the bound

$$0 \leq f(x_b) - \widehat{f} \leq \eta Q(\widehat{x}) \tag{6}$$

on the currently best function value $f(x_b)$ with a monotonically decreasing error factor η that is guaranteed to be convergent to zero by an appropriate step-size selection strategy (see Procedure 1). Note that $\widehat{x}$ is not known; thus, the error bound is not fully constructive, but enough to guarantee the convergence of $f(x_b)$ to $\widehat{f}$ with a predictable worst case complexity. To maintain (6), we consider linear relaxations of f at z,

$$f(z) \geq \gamma + \langle h, z\rangle \text{ for all } z \in C, \tag{7}$$

where $\gamma \in \mathbb{R}$ and $h \in \mathcal{V}^*$, updated using linear underestimators available from the subgradients evaluated (see Algorithm 2). Associated to such a linear relaxation is a maximization problem of the form

$$E(\gamma_b, h) := \sup_{x \in C} E_{\gamma_b,h}(x), \tag{8}$$

where

$$\gamma_b := \gamma - f(x_b),\ E_{\gamma_b,h}(x) := -\frac{\gamma_b + \langle h, x\rangle}{Q(x)}.$$

The subproblem (8) implies that

$$\gamma_b + \langle h, z\rangle \geq -E(\gamma_b, h)Q(z) \text{ for all} z \in C.$$

Since $Q(z) \geq Q_0$ by construction, (7) implies for $z = \widehat{x}$ inequality (6) with the computable value

$$\eta := E(\gamma_b, h).$$

If x_b is not optimal, then (6) implies $\eta > 0$ and we see that the supremum is positive. It is easy to see that the function $E_{\gamma_b,h}$ is continuously differentiable, quasi-concave on $C' := \{x \in C \mid E_{\gamma_b,h}(x) > 0\}$, and has there compact level sets. Therefore, the supremum is attained, and the set of solutions of (8) is convex. For use in our algorithms, we assume that some solution $u := U(\gamma_b, h) \in C$ of (8) is efficiently computable.

By sufficiently decreasing the error factor η, the convergence to an ε-minimizer x_b is guaranteed by

$$0 \leq f(x_b) - \widehat{f} \leq \varepsilon,$$

for any accuracy tolerance $\varepsilon > 0$.

The following result provides the optimality conditions for the subproblem (8).

Proposition 1 [48, Proposition 2.2] *Let $\eta = E(\gamma, h) > 0$ and $u = U(\gamma, h)$. Then*

$$\gamma + \langle h, u\rangle = -\eta Q(u), \tag{9}$$

$$\langle \eta g_Q(u) + h, z - u\rangle \geq 0 \textit{ for all } z \in C, \tag{10}$$

$$\gamma + \langle h, z\rangle \geq \eta\Big(\frac{1}{2}\|z - u\|^2 - Q(z)\Big) \textit{ for all } z \in C. \tag{11}$$

If f is a strongly convex function, then we may know $\mu > 0$ such that $f - \mu Q$ is convex. In this case, the definition of the subgradient at x_b implies

$$f(z) - \mu Q(z) \geq f(x_b) - \mu Q(x_b) + \langle g_{x_b} - \mu g_Q(x_b), z - x_b \rangle \text{for all} z \in C, \quad (12)$$

where $g_Q(x_b)$ denotes the gradient of Q at x_b. This leads to strongly convex relaxations of the form

$$f(z) \geq \gamma + \langle h, z \rangle + \mu Q(z) \text{ for all } z \in C \quad (13)$$

with

$$\gamma = f(x_b) - \mu Q(x_b) - \langle g_{x_b}, x_b \rangle, h = g_{x_b} - \mu g_Q(x_b). \quad (14)$$

In this case, a more general linear relaxation with accumulated information is defined.

Proposition 2 [48, Proposition 3.2] *Let $x \in C$, $\alpha \in [0, 1]$, and*

$$\overline{\gamma} := \gamma + \alpha(f(x) - \mu Q(x) - \langle g_x, x \rangle - \gamma), \overline{h} := h + \alpha(g - h), g = g_x - \mu g_Q(x).$$

If (7) holds and $f - \mu Q$ is convex, then we have

$$f(z) \geq \overline{\gamma} + \langle \overline{h}, z \rangle + \mu Q(z) \textit{ for all } z \in C. \quad (15)$$

We update the parameters α, h, γ, η, and u using the following procedure until a stopping criterion holds.

Procedure PUS(parameter updating scheme)

Input: δ, $\alpha_{\max} \in]0, 1[$, $0 < \kappa' \leq \kappa$, α, η, $\overline{h}$, $\overline{\gamma}$, $\overline{\eta}$, $\overline{u}$;
Output: α, h, γ, η, u;

```
begin
    R = (η − η̄)/(δαη);
    if R < 1 then
        ᾱ = αe^{−κ};
    else
        ᾱ = min(αe^{κ'(R−1)}, α_max);
    end
    α = ᾱ;
    if η̄ < η then
        h = h̄;  γ = γ̄;  η = η̄;  u = ū;
    end
end
```

For later comparison with the new algorithm OSGA-V, we now display the original OSGA algorithm proposed by Neumaier [48].

Algorithm 1 **OSGA** (optimal subgradient algorithm)

Input: $\delta, \alpha_{\max} \in]0,1[$, $0 < \kappa' \le \kappa$; local parameters: x_0, $\mu \ge 0$;
Output: x_b, f_{x_b};

1 **begin**
2 $\quad x_b = x_0$; $h = g_{x_b} - \mu g_Q(x_b)$; $\gamma = f_{x_b} - \mu Q(x_b) - \langle h, x_b \rangle$;
3 $\quad \gamma_b = \gamma - f_{x_b}$; $u = U(\gamma_b, h)$; $\eta = E(\gamma_b, h) - \mu$; $\alpha \leftarrow \alpha_{\max}$;
4 $\quad$ **while** *stopping criteria do not hold* **do**
5 $\quad\quad x = x_b + \alpha(u - x_b)$; $g = g_x - \mu g_Q(x)$; $\overline{h} = h + \alpha(g - h)$;
6 $\quad\quad \overline{\gamma} = \gamma + \alpha(f_x - \mu Q(x) - \langle g, x \rangle - \gamma)$;
7 $\quad\quad x'_b = \operatorname{argmin}_{z \in \{x_b, x\}} f(z)$; $f_{x'_b} = \min\{f_{x_b}, f_x\}$;
8 $\quad\quad \gamma'_b = \overline{\gamma} - f_{x'_b}$; $u' = U(\gamma'_b, \overline{h})$; $x' = x_b + \alpha(u' - x_b)$;
9 $\quad\quad$ choose $\overline{x}_b$ in such a way that $f_{\overline{x}_b} \le \min\{f_{x'_b}, f_{x'}\}$;
10 $\quad\quad \overline{\gamma}_b = \overline{\gamma} - f_{\overline{x}_b}$; $\overline{u} = U(\overline{\gamma}_b, \overline{h})$; $\overline{\eta} = E(\overline{\gamma}_b, \overline{h}) - \mu$; $x_b = \overline{x}_b$; $f_{x_b} = f_{\overline{x}_b}$;
11 $\quad\quad$ update the parameters α, h, γ, η and u using PUS;
12 $\quad$ **end**
13 **end**

Since subgradient methods do not generally guarantee the monotonicity of the sequence of function values $\{f_k\}_{k\ge0}$, OSGA determines the sequence of iterations $\{x_k\}_{k\ge0}$ in the way that guarantees the monotonicity of the sequence $\{f_k\}_{k\ge0}$ (see Lines 7 and 9 of Algorithm 1) in which x_k is given by the best current point x_b (see Line 10 for the definition of x_b).

Proposition 3 *Suppose that the sequence $\{x_k\}_{k\ge0}$ is generated by OSGA.*

(i) [48, Theorem 5.2] *Suppose also that the dual norm of the subgradient g_x encountered during the iteration remains bounded by the constant c_0. Define*

$$c_1 := \frac{c_0^2}{2Q_0}, c_2 := \max\left(\frac{e^\kappa c_1}{(1-\lambda)(1-\alpha_{\max})}, \frac{\eta_0(\eta_0+\mu)}{\alpha_0}\right), c_3 := \frac{c_2}{2\lambda}.$$

The algorithm stops after at most

$$K_\mu(\alpha,\eta) := 1 + \kappa^{-1}\log\frac{c_2\alpha}{\varepsilon(\varepsilon+\mu)} + \frac{c_3}{\varepsilon(\varepsilon+\mu)} + \frac{c_3}{\eta(\eta+\mu)}$$

further iterations.

(ii) [48, Theorem 5.3] *Suppose also that f has Lipschitz continuous gradients with the constant L, and set*

$$c_4 := \max\left(\frac{\eta_0+\mu}{\alpha_0^2}, \frac{e^{2\kappa}L}{1-\alpha_{\max}}\right), c_5 := \frac{4c_4}{\lambda^2}, c_6 := \sqrt{\frac{c_4}{\lambda}}, c_7 := \frac{c_6}{\lambda}.$$

If $\mu = 0$, the algorithm stops after at most

$$K_\mu(\alpha,\eta) := 1 + \kappa^{-1}\log\left(\alpha\sqrt{\frac{c_4}{\varepsilon}}\right) + \sqrt{\frac{c_5}{\varepsilon}} - \sqrt{\frac{c_4}{\eta}}$$

and if $\mu > 0$, then

$$K_\mu(\alpha,\eta) := 1 + \frac{\log(c_6\alpha)}{\kappa} + c_7\log\frac{\eta}{\varepsilon}\sqrt{\frac{c_5}{\varepsilon}} - \sqrt{\frac{c_4}{\eta}}.$$

Theorem 4 [48, Theorem 5.1] *Suppose that $f - \mu Q$ is convex, and write $N(\varepsilon)$ for the total number of iterations needed to reach a point with $f(x) \leq f(\widehat{x}) + \varepsilon$.*

(i) *Nonsmooth complexity bound*
If the points generated by OSGA stay in a bounded region of the interior of C or if f is Lipschitz continuous in C, then $N(\varepsilon) = \mathcal{O}((\varepsilon^2 + \mu\varepsilon)^{-1})$. Thus the asymptotic worst case complexity is $\mathcal{O}(\varepsilon^{-2})$ when $\mu = 0$ and $\mathcal{O}(\varepsilon^{-1})$ when $\mu > 0$.

(ii) *Smooth complexity bound*
If the points are generated by OSGA and f has Lipschitz continuous gradients with Lipschitz constant L, then $N(\varepsilon) = \mathcal{O}(\varepsilon^{-1/2})$ if $\mu = 0$ and $N(\varepsilon) = \mathcal{O}(|\log \varepsilon|\sqrt{L/\mu})$ if $\mu > 0$.

It is clear that OSGA needs, two function evaluations (Lines 7 and 9), a subgradient (Line 5), and two solutions of the subproblem (8) (Lines 8 and 10). We now construct a variant OSGA-V of this algorithm that requires only a single solution of the subproblem (8). The hope is that these modifications result in a gain in efficiency. Inspection of the proof of the complexity of OSGA in [48] reveals that part of the argumentation still applies without change, while another part needs to be adapted to work with the modifications. In the following, we only give the part of the proof that had to be changed.

Suppose that the solution u of the subproblem (8) is given. We generate the new point x by a convex combination of x_b and u, i.e.,

$$x := x_b + \alpha(u - x_b), \tag{16}$$

where $\alpha \in [0, 1]$ is a step-size (see Procedure 1). Then, we update the linear relaxation given in Proposition 2. Since our linear relaxations (15) (and relatively the function $E_{\gamma,h}$) are constructed based on the subgradient information, we keep track the best point so far leading to

$$x_b' := \operatorname*{argmin}_{z \in \{x_b, x\}} f(z).$$

Afterwards, we update the linear relaxation information given in Proposition 2 based on the new point x_b', solve the subproblem (8) to attain the new trial step u', and produce the point x' as a convex combination of x_b' and u', i.e.,

$$x' := x_b' + \alpha(u - x_b'),$$

for $\alpha \in [0, 1]$. The new x_b is produced in such a way guaranteeing $f_{x_b} \leq \min\{f_{x_b'}, f_{x'}\}$.

The results of our discussion is the following single-projection optimal subgradient algorithm.

Algorithm 2 **OSGA-V** (single-projection optimal subgradient algorithm)

Input: $\delta, \alpha_{\max} \in]0,1[$, $0 < \kappa' \leq \kappa$; local parameters: x_0, $\mu \geq 0$;
Output: x_b, f_{x_b};

1 **begin**
2 $\quad x_b = x_0$; $h = g_{x_b} - \mu g_Q(x_b)$; $\gamma = f_{x_b} - \mu Q(x_b) - \langle h, x_b \rangle$;
3 $\quad \gamma_b = \gamma - f_{x_b}$; $u = U(\gamma_b, h)$; $\eta = E(\gamma_b, h) - \mu$; $\alpha = \alpha_{\max}$;
4 $\quad$ **while** *stopping criteria do not hold* **do**
5 $\quad\quad x = x_b + \alpha(u - x_b)$; $g = g_x - \mu g_Q(x)$; $\overline{h} = h + \alpha(g - h)$;
6 $\quad\quad \overline{\gamma} = \gamma + \alpha(f_x - \mu Q(x) - \langle g, x \rangle - \gamma)$; $x'_b = \operatorname{argmin}_{z \in \{x_b, x\}} \{f(z)\}$;
7 $\quad\quad \gamma'_b = \overline{\gamma} - f_{x'_b}$; $u' = U(\gamma'_b, \overline{h})$; $\eta = E(\gamma'_b, \overline{h}) - \mu$; $x' = x'_b + \alpha(u' - x'_b)$;
8 $\quad\quad$ choose x_b in such a way that $f_{x_b} \leq \min\{f_{x'_b}, f_{x'}\}$;
9 $\quad\quad$ update the parameters α, h, γ, η and u using PUS;
10 $\quad$ **end**
11 **end**

Note that OSGA-V needs in each iteration a single solution of the subproblem (8) (Line 7) and a subgradient (Line 5) and two function values (Lines 6 and 8).

To guarantee the existence of a minimizer for OSGA-V and OSGA, we assume that the upper level set $N_f(x_0) := \{x \in C \mid f(x) \leq f(x_0)\}$ is bounded, for the starting point x_0. Since f is convex, the upper level set $N_f(x_0)$ is closed, and $\mathcal{V}$ is a finite-dimensional vector space, $N_f(x_0)$ is convex and compact. It follows from the continuity and properness of the objective function f that it attains its global minimizer on $N_f(x_0)$. Therefore, there is at least a minimizer $\widehat{x}$, and the corresponding minimum is denoted by $\widehat{f}$.

Proposition 5 *In Algorithm 2, the error factors are related by*

$$\overline{\eta} - (1-\alpha)\eta \leq \frac{\alpha^2 \|g(x)\|_*^2}{2(1-\alpha)(\eta + \mu) Q_0}. \tag{17}$$

Proof We first establish some inequalities needed for the later estimation. By convexity of Q and the definition of $\overline{h}$,

$$\begin{aligned}\alpha\mu\Big(Q(u') - Q(x) + \langle g_Q(x), x\rangle\Big) &\geq \alpha\mu\langle g_Q(x), u'\rangle = \langle h - \overline{h} + \alpha(g(x) - h), u'\rangle \\ &= (1-\alpha)\langle h, u'\rangle + \langle \alpha g(x) - \overline{h}, u'\rangle.\end{aligned}$$

By the definition of x, we have

$$(1-\alpha)(x_b - x) = -\alpha(u - x).$$

Hence, (12) (with $\mu = 0$) implies

$$(1-\alpha)(f(x_b) - f(x)) \geq (1-\alpha)\langle g(x), x_b - x\rangle = -\alpha\langle g(x), u - x\rangle.$$

By the definition of $\overline{\gamma}$, we conclude from these two inequalities that

$$\begin{aligned}\overline{\gamma} - f(x) + \alpha\mu Q(u') &= (1-\alpha)(\gamma - f(x)) - \alpha\langle g(x), x\rangle \\ &\quad + \alpha\mu\Big(Q(u') - Q(x) + \langle g_Q(x), x\rangle\Big) \\ &\geq (1-\alpha)\Big(\gamma - f(x) + \langle h, u'\rangle\Big) + \alpha\langle g(x), u' - x\rangle - \langle \overline{h}, u'\rangle \\ &\geq (1-\alpha)\Big(\gamma - f(x_b) + \langle h, u'\rangle\Big) + \alpha\langle g(x), u' - u\rangle - \langle \overline{h}, u'\rangle.\end{aligned}$$

Then, this (9) (with $\overline{\gamma}_b = \overline{\gamma} - f(x_b')$ in place of γ and $\overline{h}$ in place of h) and $E(\overline{\gamma}_b, \overline{h}) = \overline{\eta} + \mu$ give

$$\begin{aligned}(\overline{\eta} + \mu - \alpha\mu)Q(u') &= f(x_b') - \overline{\gamma} - \langle \overline{h}, u' \rangle - \alpha\mu Q(u') \\ &\leq f(x_b') - f(x) - \alpha\langle g(x), u' - u \rangle \\ &\quad -(1-\alpha)\Big(\gamma - f(x_b) + \langle h, u' \rangle\Big).\end{aligned} \tag{18}$$

Since $E(\gamma_b, h) > 0$ by Proposition 1, we may use (11) with $\gamma_b = \gamma - f(x_b)$ in place of γ and $\eta + \mu = E(\gamma_b, h)$, and find

$$(\eta + \mu)Q(u') \geq f(x_b) - \gamma - \langle h, u' \rangle + \frac{\eta + \mu}{2}\|u' - u\|^2. \tag{19}$$

Now (18) and (19) imply

$$\begin{aligned}(\overline{\eta} - (1-\alpha)\eta)Q(u') &= (\overline{\eta} + \mu - \alpha\mu)Q(u') - (1-\alpha)(\eta + \mu)Q(u') \\ &\leq f(x_b') - f(x) - (1-\alpha)\Big(\gamma - f(x_b) + \langle h, u' \rangle\Big) \\ &\quad -\alpha\langle g(x), u' - u \rangle \\ &\quad -(1-\alpha)\Big(f(x_b) - \gamma - \langle h, u' \rangle + \frac{\eta + \mu}{2}\|u' - u\|^2\Big) \\ &= f(x_b') - f(x) + \overline{S},\end{aligned}$$

where

$$\begin{aligned}\overline{S} &:= -\alpha\langle g(x), u' - u \rangle - \frac{(1-\alpha)(\eta + \mu)}{2}\|u' - u\|^2 \\ &\leq \alpha\|g(x)\|_*\|u' - u\| - \frac{(1-\alpha)(\eta + \mu)}{2}\|u' - u\|^2 \\ &= \frac{\alpha^2\|g(x)\|_*^2 - (\alpha\|g(x)\|_* + (1-\alpha)(\eta + \mu)\|u' - u\|)^2}{2(1-\alpha)(\eta + \mu)} \\ &\leq \frac{\alpha^2\|g(x)\|_*^2}{2(1-\alpha)(\eta + \mu)}.\end{aligned} \tag{20}$$

If $\overline{\eta} \leq (1-\alpha)\eta$ then (17) holds trivially. Now let $\overline{\eta} > (1-\alpha)\eta$. Then

$$(\overline{\eta} - (1-\alpha)\eta)Q_0 \leq (\overline{\eta} - (1-\alpha)\eta)Q(u') \leq f(x_b') - f(x) + \overline{S}. \tag{21}$$

Since $f(x_b') \leq f(x)$, we conclude that (17) holds. Thus, (17) holds generally. □

The remainder of the arguments in [48] apply unchanged for OSGA-V in place of OSGA when no smoothness is assumed. In particular, part (i) of Proposition 3 above remains valid for OSGA-V. As a consequence, we find as in [48] the following complexity result for OSGA-V.

Theorem 6 *Suppose that $f - \mu Q$ is convex, then if the points generated by OSGA-V stay in a bounded region of the interior of C, or if f is Lipschitz continuous in C, the total number of iterations needed to reach a point with $f(x) \leq f(\widehat{x}) + \varepsilon$ is at most $\mathcal{O}((\varepsilon^2 + \mu\varepsilon)^{-1})$. Thus the asymptotic worst case complexity is $\mathcal{O}(\varepsilon^{-2})$ when $\mu = 0$ and $\mathcal{O}(\varepsilon^{-1})$ when $\mu > 0$.*

Thus, OSGA-V attains like OSGA the optimal complexity $\mathcal{O}\left(\varepsilon^{-2}\right)$ for Lipschitz continuous nonsmooth f (cf. Nemirovsky & Yudin [41] and Nesterov [42]).

Note that Theorem 4 also asserts that OSGA attains the optimal complexity $\mathcal{O}\left(\varepsilon^{-1/2}\right)$ for smooth f with Lipschitz continuous gradients. We were not able to prove this for OSGA-V. But in our numerical experiments, we nevertheless observed that OSGA-V converged on smooth problems with essentially the same speed as OSGA.

If the subproblem (8) can be solved efficiently, OSGA-V and OSGA are appropriate for solving large-scale problems. Numerical results reported by Ahookhosh [1] and Ahookhosh & Neumaier [4, 5], for unconstrained problems, and Ahookhosh & Neumaier [6, 7], for constrained problems, show the promising behavior of OSGA for practical problems. In general, solving the subproblem (8) is the key factor for applying OSGA-V and OSGA, especially for large-scale problems, which is not trivial. Therefore, in the next section, we show that by selecting a suitable prox-function, the subproblem (8) can be solved efficiently for problems with simple domains.

3 Structured convex constrained problems in simple domains

In this section, we consider the convex constrained optimization problem (3), where C is a simple convex domain, i.e., we call it a simple domain problem. This problem appears in many applications such as signal and image processing, machine learning, statistics, and inverse problem (see Sections 5.1 and 5.2).

We here consider the quadratic prox function

$$Q(z) := \frac{1}{2}\|z\|_2^2 + Q_0; \tag{22}$$

see, e.g., [1]. We show that the solution of the subproblem (8) can be found either in a closed form or by a simple iterative scheme. In particular, we address some convex domains that a closed form solution for the subproblem (8) can be found.

The next result shows that the solution of the subproblem (8) is given by the orthogonal projection (2) of $y := \eta^{-1}h$ on the domain C followed by solving a one-dimensional nonlinear equation to determine $\eta := E(\gamma_b, h)$.

Theorem 7 *Let u be a maximizer of* (8) *and also let $\eta = E_{\gamma,h}(u) > 0$. Then*

$$u = \widehat{u}(\eta) := P_C(y),\ y := -\eta^{-1}h,$$

where, η is a solution of the univariate equation

$$\varphi(\eta) = 0$$

with

$$\varphi(\eta) := \eta\left(\frac{1}{2}\|\widehat{u}(\eta)\|_2^2 + Q_0\right) + \gamma + \langle h, \widehat{u}(\eta)\rangle. \tag{23}$$

Proof From Proposition 5.1 in [48], at the maximizer u, we obtain

$$\eta Q(u) = -\gamma - \langle h, u\rangle \tag{24}$$

and

$$\langle \eta u + h, z - u \rangle \geq 0 \text{ for all } z \in C. \tag{25}$$

By setting $z = u$ in this variational inequality, it follows that u is a solution of the minimization problem

$$\inf_{z \in C} \langle \eta u + h, z - u \rangle.$$

The first-order optimality condition for this problem is

$$0 \in \eta u + h + N_C(u). \tag{26}$$

Since $\eta > 0$, u satisfies

$$u = \operatorname*{argmin}_{z \in C} \frac{1}{2} \|\eta z + h\|_2^2 = \operatorname*{argmin}_{z \in C} \frac{1}{2} \|z - y\|_2^2 = P_C(y) = \widehat{u}(\eta),$$

where $y = -\eta^{-1} h$ giving the result. □

Theorem 7 gives a way to compute a solution of the subproblem (8) involving a projection onto the domain C and solving the one-dimensional nonlinear equation. This equation can be solved exactly for some projection operators, see Table 1. However, one can solve this nonlinear equation approximately using zero finding schemes, see, e.g., Chapter 5 of [50]. We apply the results of Theorem 7 in the next scheme to solve (8):

Algorithm 3 **SUS** (subproblem solver algorithm)

Input: Q_0, γ, h. a program for evaluating $\varphi(\eta)$ defined in (23);
Output: u, η;
begin
  solve the nonlinear equation $\varphi(\eta) = 0$ either in a closed form or approximately by a root finding solver;
  set $u = \widehat{u}(\eta)$.
end

To implement Algorithm 3 (SUS), we first need to solve the projection problem (2) effectively, see Table 5.1 of [2]. If one solves the equation $\varphi(\eta) = 0$ approximately, and an initial interval $[a, b]$ is available such that $\varphi(a)\varphi(b) < 0$, then a solution can be computed to an ε-accuracy using the bisection scheme in $\mathcal{O}(\log_2((b-a)/\varepsilon))$ iterations, see, e.g., [50]. However, it is preferable to use a more sophisticated zero finder like the secant bisection scheme (Algorithm 5.2.6, [50]). If an interval $[a, b]$ with sign change is available[1], one can also use MATLAB's `fzero` function combining the bisection scheme, the inverse quadratic interpolation, and the secant method.

In the following, we investigate special domains C, where the nonlinear equation $\varphi(\eta) = 0$ can be solved explicitly, see Table 1.

The next result shows how to the solution of (8) is given for the simple domain $C = \{x \in \mathcal{V} \mid Ax = b\}$.

[1] Without a sign change, `fzero` is unreliable; it fails on the simple quadratic $x^2 - 0.0001 = 0$ with starting point 0.2.

Table 1 List of domains C where $\varphi(\eta) = 0$ can be solved explicitly

Defining constraint $c(x)$	Solution
$Ax = b$	Proposition 8
$\langle a, x\rangle = b$	Corollary 9
$\langle a, x\rangle \leq b$	Proposition 10
$x \geq 0$	Proposition 11
$\|x\|_2 \leq \xi, \xi > 0$	Proposition 12

Proposition 8 *If $C = \{x \in \mathcal{V} \mid Ax = b\}$ is an affine set, then the subproblem (8) is solved by $u = P_C(-\eta^{-1}h)$, where*

$$P_C(y) = y - A^{\dagger}(Ay - b). \tag{27}$$

and

$$\eta = \frac{-\beta_2 + \sqrt{\beta_2^2 - 4\beta_1\beta_3}}{2\beta_1} = \frac{-2\beta_3}{\beta_2 + \sqrt{\beta_2^2 - 4\beta_1\beta_3}}, \tag{28}$$

with

$$\beta_1 := \frac{1}{2}\left\|A^{\dagger}b\right\|_2^2 + Q_0, \beta_2 := \langle A^{\dagger}(Ah), A^{\dagger}b\rangle + \gamma, \beta_3 := \frac{1}{2}\left\|A^{\dagger}(Ah)\right\|_2^2 + \frac{1}{2}\|h\|_2^2. \tag{29}$$

Proof The projection operator on C is given by (27). This and $y = -\eta^{-1}h$ give

$$P_C(-\eta^{-1}h) = -\eta^{-1}(A^{\dagger}(Ah + \eta b) - h).$$

This, together with (24), yields

$$\begin{aligned}
\eta Q(u) + \gamma + \langle h, u\rangle &= \eta\left(\tfrac{1}{2}\left(\|P_C\left(-\eta^{-1}h\right)\|_2^2\right) + Q_0\right) + \gamma + \langle h, P_C(-\eta^{-1}h)\rangle \\
&= \tfrac{1}{2}\|A^{\dagger}(Ah + \eta b)\|_2^2 + \tfrac{1}{2}\|h\|_2^2 - \langle A^{\dagger}(Ah + \eta b), h\rangle + Q_0\eta^2 \\
&\quad + \gamma\eta + \langle A^{\dagger}(Ah + \eta b) - h, h\rangle \\
&= \left(\tfrac{1}{2}\|A^{\dagger}b\|_2^2 + Q_0\right)\eta^2 + (\langle A^{\dagger}(Ah), A^{\dagger}b\rangle + \gamma)\eta \\
&\quad + \tfrac{1}{2}\|A^{\dagger}(Ah)\|_2^2 + \tfrac{1}{2}\|h\|_2^2 \\
&= \beta_1\eta^2 + \beta_2\eta + \beta_3 = 0,
\end{aligned}$$

where β_1, β_2, and β_3 are defined in (29). Since the subproblem (8) is the maximization, the bigger root of this equation is selected, which is given by (28). □

Note that in (28), the first form of η is numerically stable when $\beta_2 \leq 0$ and the second form when $\beta_2 > 0$. In the following, the same holds whenever two formulas for η are given.

The following result shows how to the solution of (8) is given for the simple domain $C = \{x \in \mathcal{V} \mid a^T x = b\}$.

Corollary 9 *If $C = \{x \in \mathcal{V} \mid a^T x = b\}$ is a hyperplane, then the subproblem (8) is solved by $u = P_C(-\eta^{-1}h)$, where*

$$P_C(y) = y - \left(\frac{\langle a, y\rangle - b}{\|a\|_2^2}\right) a, \tag{30}$$

and η is given by (28) with

$$\beta_1 := \frac{b}{2\|a\|_2^2} + Q_0, \beta_2 := \frac{b\langle a, h\rangle}{\|a\|_2^2} + \gamma, \beta_3 := \frac{1}{2}\frac{\langle a, h\rangle^2}{\|a\|_2^2} - \frac{1}{2}\|h\|_2^2. \tag{31}$$

Proof Since the hyperplane $C = \{x \in \mathcal{V} \mid a^T x = b\}$ is an affine set, this is a special case of Proposition 8. □

The subsequent result shows how to the solution of (8) is given for the simple domain $C = \{x \in \mathcal{V} \mid \langle a, x\rangle \leq b\}$.

Proposition 10 *If $C = \{x \in \mathcal{V} \mid \langle a, x\rangle \leq b\}$ is a halfspace, then the subproblem (8) is solved by $u = P_C(-\eta^{-1}h)$, where*

$$P_C(y) = y - \frac{(\langle a, y\rangle - b)_+}{\|a\|_2^2} a \tag{32}$$

and $\eta \in \{\eta_1, \eta_2\}$, where

$$\eta_1 = \frac{-\gamma + \sqrt{\gamma^2 + 4Q_0\beta}}{2Q_0} = \frac{2\beta}{\gamma + \sqrt{\gamma^2 + 4Q_0\beta}}, \tag{33}$$

with $\beta := \frac{1}{2}\|h\|_2^2 \geq 0$ and η_2 is given by (28) with β_1, β_2, and β_3 given in (31). More specifically, if $\langle a, h\rangle \geq \eta_1^{-1}b$ and $\langle a, h\rangle \geq \eta_2^{-1}b$, then $\eta = \eta_1$. If $\langle a, h\rangle \leq \eta_1^{-1}b$ and $\langle a, h\rangle < \eta_2^{-1}b$, then $\eta = \eta_2$. If $\langle a, h\rangle \geq \eta_1^{-1}b$ and $\langle a, h\rangle < \eta_2^{-1}b$, then $\eta = \max\{\eta_1, \eta_2\}$.

Proof The projection operator on C is given by (32). This gives

$$P_C(-\eta^{-1}h) = -\eta^{-1}\left(h + \frac{(\langle a, h\rangle + \eta b)_-}{\|a\|_2^2} a\right). \tag{34}$$

If $\langle a, h\rangle \geq -\eta b$, we obtain

$$P_C(-\eta^{-1}h) = -\eta^{-1}h,$$

leading to

$$\begin{aligned}\eta Q(P_C(-\eta^{-1}h)) + \gamma + \langle h, P_C(-\eta^{-1}h)\rangle &= \tfrac{1}{2}\eta^{-1}\|h\|_2^2 + Q_0\eta + \gamma - \eta^{-1}\|h\|_2^2 \\ &= Q_0\eta^2 + \gamma\eta - \tfrac{1}{2}\|h\|_2^2 = Q_0\eta^2 + \gamma\eta - \beta = 0.\end{aligned}$$

This identity leads to a solution, say η_1. If $\langle a, h\rangle < -\eta b$, (30) is valid and η is computed by (28) where β_1, β_2, and β_3 is defined in (31), say η_2. After computing

η_1 and η_2, we check whether the inequalities $\langle a, h\rangle \geq -\eta_1 b$ and $\langle a, h\rangle < -\eta_2 b$ are satisfied. Since the subproblem (8) has a solution, at least one of the conditions has to be satisfied. If one of them is satisfied, the corresponding η and (34) give the solution. If both of them hold, the solution with the bigger η is considered. □

The next result shows how to the solution of (8) is given for the simple domain $C = \{x \in \mathbb{R}^n \mid x_i \geq 0 i = 1, \ldots, n\}$.

Proposition 11 *If $C = \{x \in \mathbb{R}^n \mid x_i \geq 0 i = 1, \ldots, n\}$ is the nonnegative orthant, then the subproblem (8) is solved by $u = P_C(-\eta^{-1}h)$, where*

$$P_C(y) = y_+ \tag{35}$$

and η is given by

$$\eta = \frac{-\gamma + \sqrt{\gamma^2 + 4Q_0\beta}}{2Q_0} = \frac{2\beta}{\gamma + \sqrt{\gamma^2 + 4Q_0\beta}},$$

with $\beta := \frac{1}{2}\|h_-\|_2^2 \geq 0$.

Proof The projection operator on C is given by (35) leading to

$$P_C(-\eta^{-1}h) = -\eta^{-1}h_-.$$

This and (24) imply

$$\begin{aligned}\eta Q\left(P_C\left(-\eta^{-1}h\right)\right)+\gamma+\langle h, P_C\left(-\eta^{-1}h\right)\rangle &= \tfrac{1}{2}\eta^{-1}\|h_-\|_2^2 + Q_0\eta + \gamma - \eta^{-1}\langle h, h_-\rangle \\ &= Q_0\eta^2 + \gamma\eta + \tfrac{1}{2}\|h_-\|_2^2 - \langle h, h_-\rangle \\ &= Q_0\eta^2 + \gamma\eta - \beta = 0,\end{aligned}$$

giving the result. □

The following result shows how to the solution of (8) is given for the simple domain $C = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq \xi\}$, for $\xi > 0$.

Proposition 12 *Let $C = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq \xi\}$ be the Euclidean ball. Then*

$$P_C(y) = \begin{cases} \xi y/\|y\|_2 & \|y\|_2 > \xi, \\ y & \|y\|_2 \leq \xi, \end{cases} \tag{36}$$

If $\|\eta^{-1}h\|_2 \leq \xi$ where η is given by

$$\eta = \frac{-\gamma + \sqrt{\gamma^2 + 4Q_0\beta}}{2Q_0} = \frac{2\beta}{\gamma + \sqrt{\gamma^2 + 4Q_0\beta}},$$

with $\beta := \frac{1}{2}\|h\|_2^2 \geq 0$, then $u = -\eta^{-1}h$; otherwise, the solution of the subproblem (8) is given by

$$u = -\frac{\xi}{\|h\|_2}h, \eta = -\frac{2(\gamma + \xi\|h\|_2)}{\xi^2 + 2Q_0}.$$

Proof The projection operator on C is given by (36), leading to

$$P_C(-\eta^{-1}h) = \begin{cases} -\xi h/\|h\|_2 & \|h\|_2 > \eta\xi, \\ -\eta^{-1}h & \|h\|_2 \le \eta\xi. \end{cases}$$

We first assume that $\|h\|_2 \le \eta\xi$ implying $P_C(-\eta^{-1}h) = -\eta^{-1}h$. Substituting this into (24) yields

$$\begin{aligned} \eta Q(P_C(-\eta^{-1}h)) + \gamma + \langle h, P_C(-\eta^{-1}h)\rangle &= \tfrac{1}{2}\eta^{-1}\|h\|_2^2 + Q_0\eta + \gamma - \eta^{-1}\|h\|_2^2 \\ &= Q_0\eta^2 + \gamma\eta - \tfrac{1}{2}\|h\|_2^2 = Q_0\eta^2 + \gamma\eta - \beta = 0, \end{aligned}$$

giving the result. If this η satisfies $\|h\|_2 \le \eta\xi$, then $u = -\eta^{-1}h$; otherwise, we assume that $\|h\|_2 > \eta\xi$. Substituting $P_C(-\eta^{-1}h) = -\xi h/\|h\|_2$ into (24) yields

$$\eta\left(\frac{1}{2}\xi^2 + Q_0\right) + \gamma - \xi\|h\|_2 = 0,$$

implying

$$\eta = -\frac{2(\gamma + \xi\|h\|_2)}{\xi^2 + 2Q_0}$$

and $u = -\xi h/\|h\|_2$. This completes the proof. □

To solve bound-constrained problems with OSGA-V and OSGA, we developed an algorithm that can find the global solution of the subproblem (8) by solving a sequence of one-dimensional rational optimization problems, see Algorithm 3 in [6]. Notice that the constraint $C := \{x \in \mathcal{V} \mid \|x\|_\infty \le \xi\}$ is a special case of bound-constrained problems with $\underline{x} = -\xi\mathbf{1}$ and $\overline{x} = \xi\mathbf{1}$, where $\mathbf{1}$ is a n-dimensional vector with all elements equal to unity.

4 Solving structured problems with a functional constraint

In this section, we consider the structured convex constrained problem (3) with the domain

$$C := \{x \in \mathcal{V} | \phi(x) \le \xi\}. \tag{37}$$

The aim of this section is to find a solution of the subproblem (8) directly by using the KKT optimality conditions, especially when no efficient method for finding the projection onto C is known.

In the reminder of this section, we assume that the functional constraint satisfies the Cottle constraint qualification, see [10], i.e.,
(H1) for all $x \in C$, either $\phi(x) < 0$ or $0 \notin \partial_\phi(x)$.

The next result gives the optimality conditions for solving the problem (8) with the domain (37).

Theorem 13 *Let (H1) hold for the problem* (3)*, with C satisfying* (37)*. Then, for a real constant ξ, the supremum*

$$\sup_{\phi(x)\le\xi} \frac{-\gamma - \langle h, x\rangle}{Q(x)} \tag{38}$$

is attained, and every maximizer u satisfies either

$$u = -\eta^{-1}h, \phi(u) < \xi \tag{39}$$

or

$$\frac{1}{\mu}\frac{-\eta u - h}{Q(u)} \in \partial\phi(u), \mu > 0, \phi(u) = \xi, \tag{40}$$

where $\eta := E_{\gamma_b,h}(u)$.

Proof Let us define the function

$$E_{\gamma,h} : C \to \mathbb{R}, E_{\gamma,h}(x) := -\frac{\gamma + \langle h, x\rangle}{Q(x)}.$$

Since this function is differentiable, by differentiating both sides of the equality $E_{\gamma,h}(x)Q(x) = -\gamma - \langle h, x\rangle$ with respect to x, we obtain

$$\nabla E_{\gamma,h}(x) = \frac{-E_{\gamma,h}(x)x - h}{Q(x)}. \tag{41}$$

In view of the KKT optimality conditions for inequality constrained nonsmooth problems, see [10], we have the optimality conditions

$$\begin{cases} 0 \in \nabla E_{\gamma,h}(u) + \mu\partial\phi(u), \\ \phi(u) \le \xi, \\ \mu \ge 0, \\ \mu(\phi(u) - \xi) = 0, \end{cases} \tag{42}$$

for (38). Now, by substituting (41) into (42), setting $\eta := -(\gamma + \langle h, u\rangle)/Q(u)$, and distinguishing between $\mu = 0$ and $\mu > 0$, we obtain either (39) or (40). □

Theorem 13 gives the optimality conditions for a general function ϕ, however, in view of Theorem 7, it is especially useful when the projection onto $C = \{x \mid \phi(x) \le \xi\}$ is not efficiently available.

We here need the following result, where the proof is given in Proposition 2.1.17 of [2].

Proposition 14 *(see, e.g.,* [5]*) Let* $\phi : \mathcal{V} \to \mathbb{R}, \phi(x) = \|x\|$. *Then the subdifferential of* ϕ *is*

$$\partial\phi(x) = \begin{cases} \{g \mid \|g\|_* \le 1\} & \text{if} x = 0, \\ \{g \mid \|g\|_* = 1, \langle g, x\rangle = \|x\|\} & \text{if} x \ne 0. \end{cases}$$

Moreover, if $\|\cdot\|$ *is self-dual, then*

$$\partial\phi(x) = \begin{cases} \{g \mid \|g\|_* \le 1\} & \text{if} x = 0, \\ x/\|x\| & \text{if} x \ne 0. \end{cases}$$

To conclude this section, we derive the solution of the subproblem (8) for some ϕ such as $\|\cdot\|_{1,2}$ that appear in many applications. In 2004, Yuan and Lin in [54] proposed an interesting regularizer called grouped LASSO for the linear regression. Later Kim et al. in [33] proposed a constrained ridge regression model

$$\begin{array}{ll} \min & \frac{1}{2}\|y - \mathcal{A}x\|_2^2 \\ \text{s.t.} & \|x\|_{1,2} \le \xi, \end{array} \tag{43}$$

in which ξ is a nonnegative real constant and $\|x\|_{1,2}$ is a so-called the $l_{1,2}$ group norm. We consider this constraint in the next result.

Proposition 15 *Let $\mathcal{V}$ be a real finite-dimensional vector space with the induced norm $\phi(\cdot) = \|\cdot\|_{1,2}$. Then the subproblem (8) is solved by*

$$u_{g_i} = -\eta^{-1} h_{g_i} \text{for all} i = 1, \ldots, m, \tag{44}$$

and

$$\eta = \frac{-\gamma + \sqrt{\gamma^2 + 4Q_0\beta}}{2Q_0} = \frac{-2\beta}{\gamma + \sqrt{\gamma^2 - 4Q_0\beta}}, \mu = 0,$$

where $\beta := \sum_{i=1}^m \|h_{g_i}\|_2^2 - \frac{1}{2}\|h\|_2^2$, if $\phi(u) < \xi$; otherwise, it is solved by

$$u_i = \rho_i h_{g_i}, \rho_i = \frac{\|h_{g_i}\|_2 - \mu\left(\frac{1}{2}\xi^2 + Q_0\right)}{\eta\|h_{g_i}\|_2} \text{for all} i = 1, \ldots, m,$$

and

$$\eta = -\frac{\gamma + \langle h, u\rangle}{\frac{1}{2}\xi^2 + Q_0} = -\frac{2(\gamma + \sum_{i=1}^n \tau_i^2 \|h_{g_i}\|_2^2)}{\sum_{i=1}^n \tau_i^2 \|h_{g_i}\|_2^2 + 2Q_0}, \mu = \frac{2(\sum_{i=1}^m \|h_{g_i}\|_2 + \eta\xi)}{m(\sum_{i=1}^m \tau_i^2 \|h_{g_i}\|_2^2 + 2Q_0)}.$$

Proof Since $\eta > 0$ and $\xi > 0$, $u = 0$ if $h_{g_i} = 0$, for $i = 1, \ldots, m$, satisfying (44). Let us consider $u \neq 0$. In view of Proposition 14, we get

$$\partial\phi(u_{g_i}) = \left\{\frac{u_{g_i}}{\|u_{g_i}\|_2}\right\} \text{for all } i = 1, \ldots, m,$$

leading to

$$\partial\phi(u) = \left\{\left(\frac{u_{g_1}}{\|u_{g_1}\|_2}, \ldots, \frac{u_{g_m}}{\|u_{g_m}\|_2}\right)\right\}.$$

We apply Theorem 13 leading to two cases: (i) (39) holds; (ii) (40) holds.

Case (i). The condition (39) holds. Then, we have $u_{g_i} = -\eta^{-1} h_{g_i}$ for $i = 1, \ldots, n$. By substituting $u = (u_{g_1}, \ldots, u_{g_n})$ into the identity $E_{\gamma,h}(u) = \eta$, we get

$$\eta = \frac{-\gamma + \sum_{i=1}^m \|h_{g_i}\|_2^2 \eta^{-1}}{\frac{1}{2}\|h\|_2^2 \eta^{-2} + Q_0},$$

implying

$$Q_0\eta^2 + \gamma\eta + \frac{1}{2}\|h\|_2^2 - \sum_{i=1}^m \|h_{g_i}\|_2^2 = 0.$$

By using the bigger root of this equation, we get

$$\eta = \frac{-\gamma + \sqrt{\gamma^2 + 4Q_0\beta}}{2Q_0},$$

where $\beta := \sum_{i=1}^m \|h_{g_i}\|_2^2 - \frac{1}{2}\|h\|_2^2$.

Case (ii). The condition (40) holds. Then, we have

$$\frac{-\eta u_{g_i} - h_{g_i}}{\frac{1}{2}\|u\|_2^2 + Q_0} = -\mu \frac{u_{g_i}}{\|u_{g_i}\|_2} \text{for all } i = 1, \ldots, m.$$

Since $\phi(u) = \|u\| = \xi$, we equivalently get

$$\left(\frac{1}{2}\|u\|_2^2 + Q_0\right)\left(-\frac{\eta}{\frac{1}{2}\|u\|_2^2 + Q_0} + \frac{\mu}{\|u_{g_i}\|_2}\right) u_{g_i} = h_{g_i},$$

implying $u_{g_i} = \tau_i h_{g_i}$. If $h_{g_i} = 0$, then $u_{g_i} = 0$. Now let $h_{g_i} \neq 0$. Substituting $u_{g_i} = \tau_i h_{g_i}$ into the previous identity, it follows that

$$\left(\frac{1}{2}\sum_{i=1}^{m} \tau_i^2 \|h_{g_i}\|_2^2 + Q_0\right)\left(-\frac{\eta}{\frac{1}{2}\sum_{i=1}^{m} \tau_i^2 \|h_{g_i}\|_2^2 + Q_0} + \frac{\mu}{\tau_i \|h_{g_i}\|_2}\right) \tau_i h_{g_i} = h_{g_i},$$

giving

$$-\eta \tau_i \|h_{g_i}\|_2 + \mu\left(\frac{1}{2}\sum_{i=1}^{m} \tau_i^2 \|h_{g_i}\|_2^2 + Q_0\right) = \|h_{g_i}\|_2 \text{for all } i = 1, \ldots, m.$$

Applying a summation from both sides, together with $\sum_{i=1}^{m} \tau_i \|h_{g_i}\|_2 = \xi$, yields

$$-\eta\xi + m\mu\left(\frac{1}{2}\sum_{i=1}^{m} \tau_i^2 \|h_{g_i}\|_2^2 + Q_0\right) = \sum_{i=1}^{m} \|h_{g_i}\|_2, \tag{45}$$

implying

$$\mu = \frac{2(\sum_{i=1}^{m} \|h_{g_i}\|_2 + \eta\xi)}{m(\sum_{i=1}^{m} \tau_i^2 \|h_{g_i}\|_2^2 + 2Q_0)}.$$

By substituting this into (45), we have

$$\tau_i = -\frac{1}{m\eta\|h_{g_i}\|_2}\left(m\|h_{g_i}\|_2 - \sum_{i=1}^{m} \|h_{g_i}\|_2 - \eta\xi\right),$$

leading to

$$u = (\tau_i h_{g_1}, \ldots, \tau_m h_{g_m}).$$

By substituting this into $E_{\gamma,h}(u) = \eta$, we get

$$\eta = -\frac{\gamma + \langle h, u\rangle}{\frac{1}{2}\xi^2 + Q_0} = -\frac{2(\gamma + \sum_{i=1}^{n} \tau_i^2 \|h_{g_i}\|_2^2)}{\sum_{i=1}^{n} \tau_i^2 \|h_{g_i}\|_2^2 + 2Q_0},$$

giving the result.

□

5 Numerical experiments and applications

This section discusses numerical results and comparisons of OSGA and OSGA-V with some state-of-the-art first-order solvers on some image deblurring and ridge regression problems. For image deblurring problem, we consider a ℓ_1 data fidelity regularized with isotropic total variation that both part of the objective are nonsmooth. Therefore, we compare with DRPD-1, DRPD-2 (Douglas-Rachford primal-dual schemes proposed by Boţ & Hendrich [20]), and ADMM (alternating direction method of multipliers [25]). In this case, we used the algorithms provided by the respective authors. Since the ridge regression has a smooth convex objective, we compared with PGA (gradient projection algorithm), SPG-A (the spectral gradient projection [18] with the Amini et al. nonmonotone term [8]), NESCO (Nesterov's composite optimal algorithm [47]), and NESUN (Nesterov's universal gradient algorithm [44]). We used the default parameter values reported in the corresponding papers or packages. In our comparisons, we did not consider the popular forward-backward solver FISTA since it is not designed to handle constrained problems of the form (3).

The codes of OSGA and OSGA-V are written in MATLAB, where they use the parameters

$$\delta = 0.9, \alpha_{max} = 0.7, \kappa = \kappa' = 0.5.$$

and the prox-function (22) with $Q_0 = \frac{1}{2}\|x_0\|_2 + \epsilon$, where ϵ is the machine precision. All numerical experiments were executed on a PC Intel Core i7-3770 CPU 3.40GHz 8 GB RAM.

5.1 Image deblurring with nonnegativity constraints

Inverse problems are appearing in many fields of applied sciences and engineering. This is particularly happen when researchers use digital images to record and analyze results from experiments in many fields such as astronomy, medical sciences, biology, geophysics, and physics. In these cases, observing blurred and noisy images is a common phenomenon happening frequently because of environmental effects and imperfections in the imaging system.

The process of reconstructing or estimating a true image from a degraded observation is known as the image restoration, also called deblurring or deconvolution. Image restoration is addressed by considering a constraint satisfaction problem of the form

$$\mathcal{A}x = b, x \in C,$$

where C is a convex domain C that is commonly a box or the nonnegative orthant. This is an ill-posed problem, cf. Neumaier [49], and can be handled by solving the regularized ℓ_1 problem

$$\begin{array}{ll} \min & \|\mathcal{A}x - b\|_1 + \lambda\|x\|_{ITV} \\ \text{s.t.} & x \in C, \end{array} \tag{46}$$

where $\|\cdot\|_{ITV}$ is called isotropic total variation (cf. [24]) given by

$$\begin{aligned} \|x\|_{ITV} = & \textstyle\sum_i^{m-1}\sum_j^{n-1}\sqrt{(x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2} \\ & + \textstyle\sum_i^{m-1}|x_{i+1,n} - x_{i,n}| + \sum_j^{n-1}|x_{m,j+1} - x_{m,j}|, \end{aligned}$$

for $x \in \mathbb{R}^{m\times n}$. Since both $\|\mathcal{A}x - b\|_1$ and $\|x\|_{ITV}$ are nonsmooth the Nesterov-type optimal methods like NESCO cannot be applied. In addition, the associated subproblem of the universal gradient method NESUN can be only solved approximately with an iterative scheme which is costly. Therefore, we will not consider them in our comparison.

In many applications, the variable x describes physical quantities, which is meaningful if each component of x is restricted to be nonnegative. This constraint is referred as the nonnegativity constraint; it is especially useful for restoring blurred and noisy images, see [11, 31, 32].

We also consider the restoration of the 1024×1024 blurred/noisy Titan image using (46). The true image is available in http://photojournal.jpl.nasa.gov/Help/ImageGallery.html.

The blurred/noisy image is constructed from the 7×7 Gaussian kernel with standard deviation 5 and salt-and-pepper impulsive noise with the level 50%. To recover the image, we use DRPD-1, DRPD-2, ADMM, OSGA-V, and OSGA. The algorithms are stopped after 100 iterations, and three different regularization parameters are considered. The results of implementation are reported in Table 2 and Fig. 1.

The comparison concerning the quality of the recovered image is made via the so-called peak signal-to-noise ratio (PSNR) defined by

$$\mathrm{PSNR} = 20\log_{10}\left(\frac{\sqrt{mn}}{\|x - x_c\|_F}\right) \tag{47}$$

and the improvement in signal-to-noise ratio (ISNR) defined by

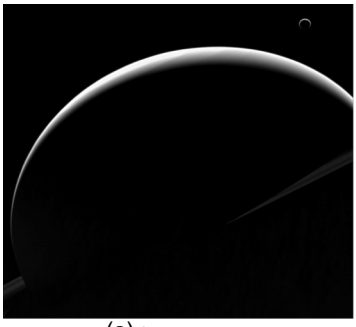
$$\mathrm{ISNR} = 20\log_{10}\left(\frac{\|y - x_c\|_F}{\|x - x_c\|_F}\right), \tag{48}$$

where $\|\cdot\|_F$ is the Frobenius norm, x_c denotes the $m \times n$ clean image, y is the observed image, and pixel values are in [0, 1].

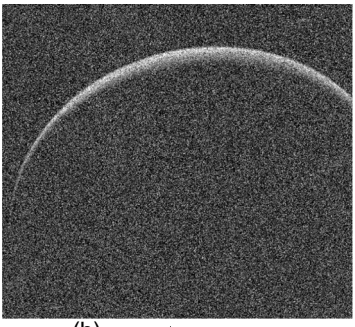
The results of Table 2 shows that OSGA-V and OSGA produce comparable or better results than the others with respect to final function values; on the other hand, they

Table 2 Results summary for image deblurring with L1ITV; each method was stopped after 100 iterations

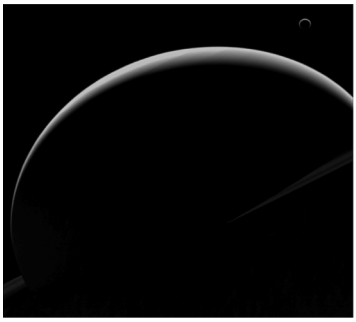
	λ	DRPD-1	DRPD-2	ADMM	OSGA	OSGA-V
f_b	3×10^{-2}	2.62581e+5	2.63318e+5	2.75685e+5	2.63044e+5	2.63024e+5
PSNR		30.88	32.88	14.20	34.97	36.44
Time		26.53	17.09	17.35	19.44	18.14
f_b	7×10^{-2}	2.62248e+5	2.64435e+5	2.67511e+5	2.62375e+5	2.623259e+5
PSNR		35.05	31.60	25.50	40.88	41.37
Time		26.14	17.17	16.96	19.02	18.03
f_b	1×10^{-1}	2.63555e+5	2.66710e+5	2.73446e+5	2.63458e+5	2.63431e+5
PSNR		39.84	31.58	39.83	41.33	41.77
Time		27.78	17.87	17.94	20.59	18.80



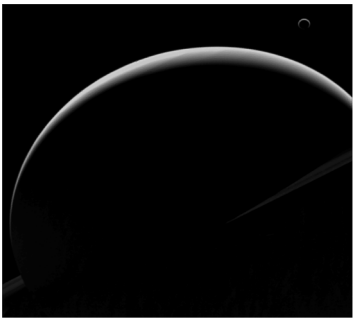
(a) Clean image



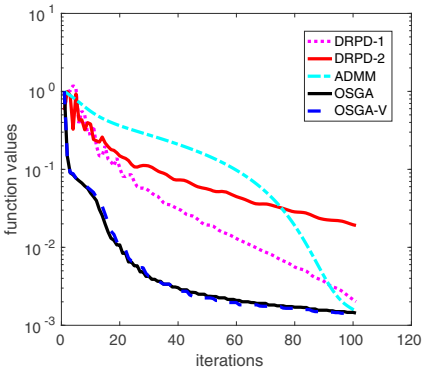
(b) Blurred/noisy image



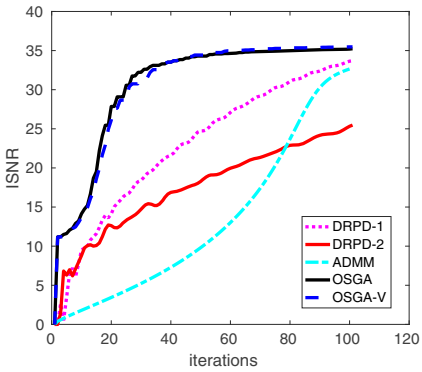
(c) OSGA: $f = 2.63458e + 5$, PSNR $= 41.33$, T $= 20.59$



(d) OSGA-V: $f = 2.63431e + 5$, PSNR $= 41.77$, T $= 18.80$



(e) Function value error δ_k versus iterations, $\lambda = 10^{-1}$



(f) ISNR versus iterations, $\lambda = 10^{-1}$

Fig. 1 A comparison among DRPD-1, DRPD-2, ADMM, OSGA-V, and OSGA for deblurring the 1024×1024 Titan image with $\lambda = 10^{-1}$. The blurred/noisy image was constructed by the 7×7 Gaussian kernel with standard deviation 5 and salt-and-pepper impulsive noise with the level 50 %. The algorithms were stopped after 100 iterations. Subfigures **a** and **b** display the clean and blurred/noisy images and recovered image by OSGA, respectively. Subfigures **c** and **d** show the recovered images by OSGA and OSGA-V, respectively. Subfigures **e** and **f** illustrate the relative error of function values δ_k (51) versus iterations and ISNR (48) versus iterations, respectively

outperform the others in the sense of PSNR. Again, OSGA-V needs less time than OSGA. Subfigures (a), (b), (c), and (f) of Fig. 1 display the clean image, blurred/noisy image, recoverd image by OSGA, and recovered imege by OSGA-V for $\lambda = 10^{-1}$,

respectively. Subfigures (e) and (f) of Fig. 1 show that OSGA-V and OSGA attain the best function values and ISNR for $\lambda = 10^{-1}$.

5.2 Ridge regression

We here consider a ℓ_2-constrained least squares of the form (50) (so-called ridge regression, cf. [36]) and report some numerical results.

Let $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ be an ill-conditioned or a singular linear operator and $y \in \mathbb{R}^m$ be a vector of observations. The linear inverse problem is the quest of finding $x \in \mathbb{R}^n$ such that

$$y = \mathcal{A}x + \nu, \tag{49}$$

with an unknown but small additive noise $\nu \in \mathbb{R}^m$. The problem is solvable if one knows additional qualitative information about x. This qualitative information is encoded in a constraint on x, under which the Euclidean norm of ν is minimized. If the qualitative information consists of a bound ξ on the Euclidean norm of x, the constrained optimization problem resulting takes the form

$$\begin{array}{ll} \min & \frac{1}{2}\|y - \mathcal{A}x\|_2^2 \\ \text{s.t.} & \|x\|_2 \le \xi, \end{array} \tag{50}$$

in which ξ is a nonnegative real constant. This problem often occurs in the fields of applied sciences and engineering, see [33, 52].

The problem is generated by

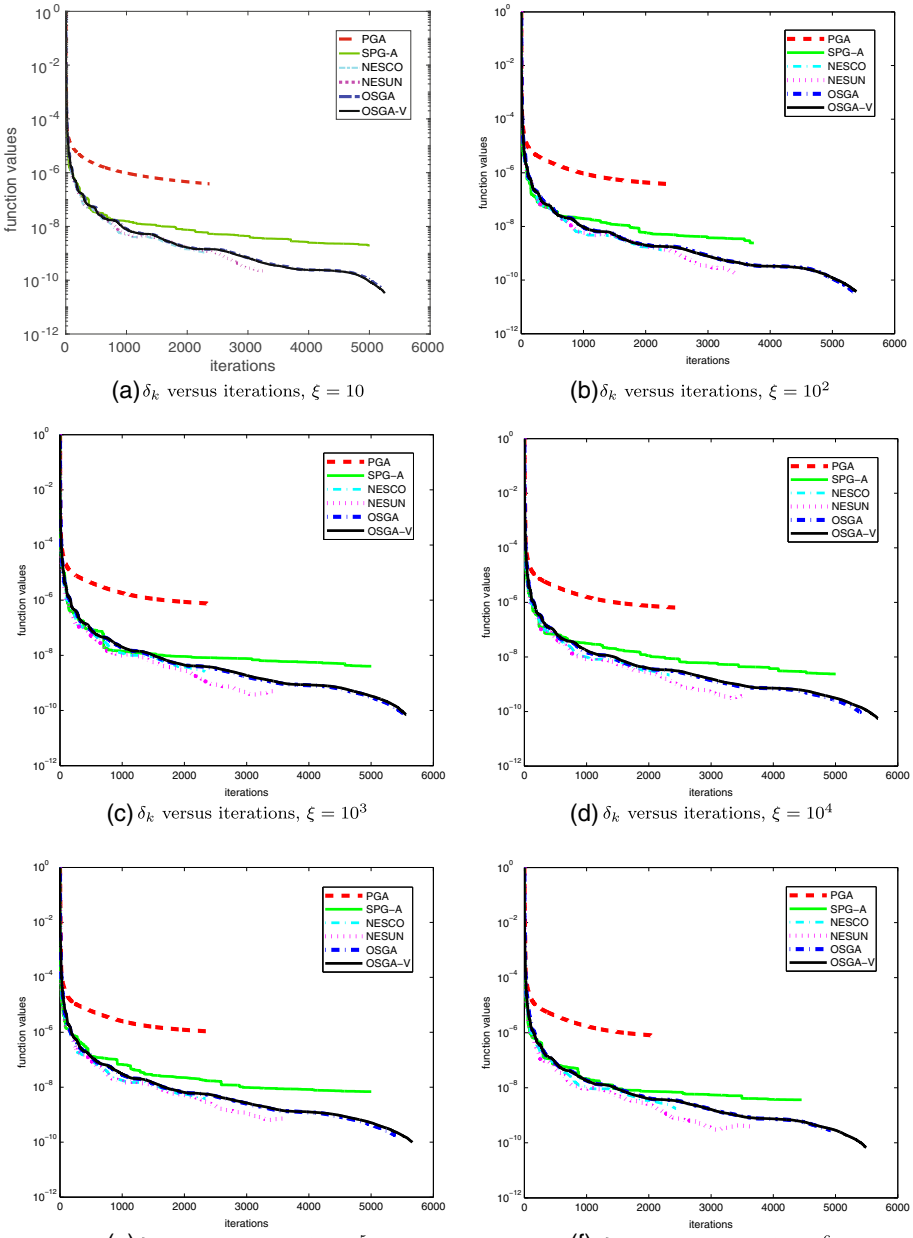
$$[\mathtt{A}, \mathtt{z}, \mathtt{x}] = \mathtt{i_laplace(n)}, \mathtt{y} = \mathtt{z} + \mathtt{0.1} * \mathtt{rand},$$

where $n = 5000$ is the problem dimension and `i_laplace.m` is an ill-posed test problem generator using the inverse Laplace transformation from Regularization Tools MATLAB package (see [35]), which is available in http://www.imm.dtu.dk/~pcha/Regutools/.

Since (50) is smooth and the projection onto $C = \{x \in \mathbb{R}^n \mid \|x\| \le \xi\}$ is available. We employ PGA, SPG-A, NESCO, NESUN, OSGA-V, and OSGA (see Proposition

Table 3 Result summary of function values for the ridge regression; the values improve uniformly from the left to the right

ξ	**PGA**	SPG-A	NESCO	NESUN	OSGA	OSGA-V
10	5.6234e-3	3.3007e-5	1.9145e-5	7.5211e-6	5.0749e-6	4.7653e-6
10^2	5.5816e-3	4.0555e-5	2.4331e-5	7.3372e-6	5.3821e-6	5.3808e-6
10^3	1.2031e-2	7.4762e-5	5.1284e-5	1.8230e-5	1.1402e-5	1.1219e-5
10^4	9.9266e-3	4.4910e-5	4.1212e-5	1.3725e-5	9.2309e-6	8.6053e-6
10^5	1.8227e-2	1.3232e-4	7.9612e-5	2.7936e-5	1.8227e-5	1.6849e-5
10^6	1.2533e-2	6.7878e-5	3.7444e-5	1.5768e-5	1.4334e-5	1.0897e-5



(a) δ_k versus iterations, $\xi = 10$

(b) δ_k versus iterations, $\xi = 10^2$

(c) δ_k versus iterations, $\xi = 10^3$

(d) δ_k versus iterations, $\xi = 10^4$

(e) δ_k versus iterations, $\xi = 10^5$

(f) δ_k versus iterations, $\xi = 10^6$

Fig. 2 A comparison among PGA, SPG-A, NESCO, NESUN, OSGA-V, and OSGA for solving the problem (50) based on the relative error of function values δ_k (51). Each algorithm was stopped after 60 seconds

12) to solve this minimization problem. The parameters of SPG-A are the same as those reported in the associated papers, but SPG-A uses

$$\gamma_k := \begin{cases} \gamma_0/2 & \text{if } k = 1, \\ (\gamma_{k-1} + \gamma_{k-2})/2 & \text{if } k \geq 2. \end{cases}$$

All algorithms were stopped after 60-s running time.

In Table 3, we consider $\xi = 10, 10^2, 10^3, 10^4, 10^5, 10^6$ and report the best attained function values and the running time. The results imply that OSGA-V and OSGA attain the better function values; however, OSGA-V get the best results. To see the results of implementation in more details, we demonstrate the relative error of function values

$$\delta_k := \frac{f_k - \widehat{f}}{f_0 - \widehat{f}} \tag{51}$$

in Fig. 2, where $\widehat{f}$ denotes the minimum and f_0 shows the function value on an initial point x_0. From Fig. 2, it is clear that OSGA-V and OSGA outperform the others. Remarkably OSGA-V performs best, although the proved complexity of OSGA and the others for this smooth problem is superior to that proved for OSGA-V.

6 Final remarks

In this paper, two optimal subgradient methods, OSGA and OSGA-V, were discussed for solving structured convex constrained optimization; the second being a simplified version of the first needing less work. Finding a solution of the subproblem is investigated in the presence of some convex constraints. Two types of convex constraints were considered, namely, simple convex domains, in which the orthogonal projection onto the domains is effectively available, and functional constraints, defined as the sublevel sets of simple convex functions. In each case , practically interesting examples were discussed for which the subproblem can be solved efficiently. Numerical results and comparisons with some state-of-the-art algorithms were reported showing that OSGA and OSGA-V are efficient and reliable for solving convex optimization problems in applications.

Acknowledgments Open access funding provided by University of Vienna. We would like to thank Radu Bot and Min Tao for making their codes DRPD-1, DRPD-2, and ADMM available for us. We would also like to thanks two anonymous referees whose helped to improve the paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Ahookhosh, M.: Optimal subgradient algorithms with application to large-scale linear inverse problems, Submitted (2015), arXiv:1402.7291

2. Ahookhosh, M.: High-dimensional nonsmooth convex optimization via optimal subgradient methods. PhD Thesis, University of Vienna, pp. 1–206 (2015)
3. Ahookhosh, M., Ghaderi, S.: On efficiency of nonmonotone Armijo-type line searches. Appl. Math. Model. **43**, 170–190 (2017)
4. Ahookhosh, M., Neumaier, A.: High-dimensional convex optimization via optimal affine subgradient algorithms. In: ROKS workshop, pp. 83–84 (2013)
5. Ahookhosh, M., Neumaier, A.: An optimal subgradient algorithm with subspace search for costly convex optimization problems, submitted, http://www.optimization-online.org/DB_FILE/2015/04/4852.pdf (2016)
6. Ahookhosh, M., Neumaier, A.: An optimal subgradient algorithm for large-scale bound-constrained convex optimization, submitted, http://arxiv.org/abs/1501.01497 (2015)
7. Ahookhosh, M., Neumaier, A.: Solving nonsmooth convex optimization with complexity $o(\varepsilon^{-1/2})$, submitted, http://www.optimization-online.org/DB_FILE/2015/05/4900.pdf (2016)
8. Amini, K., Ahookhosh, M., Nosratipour, H.: An inexact line search approach using modified nonmonotone strategy for unconstrained optimization. Numerical Algorithms **66**, 49–78 (2014)
9. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. **16**, 697–725 (2006)
10. Bagirov, A., Karmitsa, N., Mäkelä, M.M.: Introduction to Nonsmooth Optimization: Theory, Practice and Software, Springer International Publishing (2014)
11. Bardsley, J., Vogel, C.R.: A nonnegatively constrained convex programming method for image reconstruction. SIAM J. Sci. Comput. **25**, 1326–1343 (2003)
12. Barzilai, J., Borwein, J.M.: Two point step size gradient method. IMA J. Numer. Anal. **8**, 141–148 (1988)
13. Beck, A., Teboulle, M.: Smoothing and first order methods: a unified framework. SIAM J. Optim. **22**, 557–580 (2012)
14. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
15. Becker, S.R., Candès, E.J., Grant, M.C.: Templates for convex cone problems with applications to sparse signal recovery. Math. Program. Comput. **3**, 165–218 (2011)
16. Bertsekas, D.P. Nonlinear programming, 2nd ed. Athena Scientific, Belmont (1999)
17. Bertsekas, D.P., Tsitsiklis, J.N.: Gradient convergence in gradient methods with errors. SIAM J. Optim. **10**, 627–642 (2000)
18. Birgin, E.G., Martinez, J.M., Raydan, M.: Nonmonotone spectral projected gradient methods on convex sets. SIAM J. Optim. **10**, 1196–1211 (2000)
19. Boţ, R.I., Hendrich, C.: A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems. Comput. Optim. Appl. **54**(2), 239–262 (2013)
20. Boţ, R.I., Hendrich, C.: A Douglas-Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. SIAM J. Optim. **23**(4), 2541–2565 (2013)
21. Boţ, R.I., Csetnek, E.R., Hendrich, C.: A primal-dual splitting algorithm for finding zeros of sums of maximally monotone operators. SIAM J. Optim. **23**, 2011–2036 (2013)
22. Boyd, S., Xiao, L., Mutapcic, A.: Subgradient methods, notes for EE392o, stanford university, http://www.stanford.edu/class/ee392o/subgrad_method.pdf (2003)
23. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
24. Chambolle, A., Caselles, V., Cremers, D., Novaga, M., Pock, M.: An introduction to total variation for image analysis. In: Theoretical foundations and numerical methods for sparse recovery, vol. 9, pp. 263–340. , Radon Series Comp. Appl. Math., De Gruyter (2010)
25. Chan, R.H., Tao, M., Yuan, X.: Constrained total variation deblurring models and fast algorithms based on alternating direction method of multipliers. SIAM J. Imaging Sci. **6**(1), 680–697 (2013)
26. Combettes, P., Pesquet, J.-C.: Proximal splitting methods in signal processing (2011)
27. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**, 37–75 (2013)
28. Devolder, O., Glineur, F., Nesterov, Y.: Double smoothing technique for large-scale linearly constrained convex optimization. SIAM J. Optim. **22**(2), 702–727 (2012)
29. Gonzaga, C.C., Karas, E.W.: Fine tuning Nesterov's steepest descent algorithm for differentiable convex programming. Math. Program. **138**, 141–166 (2013)
30. Gonzaga, C.C., Karas, E.W., Rossetto, D.R.: An optimal algorithm for constrained differentiable convex optimization. SIAM J. Optim. **23**(4), 1939–1955 (2013)

31. Kaufman, L., Neumaier, A.: PET regularization by envelope guided conjugate gradients. IEEE Trans. Med. Imaging **15**, 385–389 (1996)
32. Kaufman, L., Neumaier, A.: Regularization of ill-posed problems by envelope guided conjugate gradients. J. Comput. Graph. Stat. **6**(4), 451–463 (1997)
33. Kim, Y., Kim, J., Kim, Y.: Blockwise sparse regression. Stat. Sin. **16**(2), 375–390 (2006)
34. Hager, W.W., Zhang, H.: A survey of nonlinear conjugate gradient methods. Pac. J. Optim. **2**, 35–58 (2006)
35. Hansen, P.: Regularization tools version 4.0 for matlab 7.3. Numerical Algorithms **46**, 189–194 (2007)
36. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970)
37. Lemarchal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. Math. Program. **69**(1-3), 111–147 (1995)
38. Lan, G.: Bundle-level type methods uniformly optimal for smooth and non-smooth convex optimization. Math. Program. **149**(1), 1–45 (2015)
39. Lan, G., Lu, Z., Monteiro, R.D.C.: Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. Math. Program. **126**, 1–29 (2011)
40. Nedić, A., Bertsekas, D.P.: Incremental subgradient methods for nondifferentiable optimization. SIAM J. Optim. **12**, 109–138 (2001)
41. Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization. Wiley, New York (1983)
42. Nesterov, Y.: Introductory lectures on convex optimization: a basic course. Kluwer, Dordrecht (2004)
43. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $o(1/k^2)$. Doklady AN SSSR (In Russian) **269**, 543–547 (1983). English translation: Soviet Math. Dokl., 27 (1983), 372–376
44. Nesterov, Y.: Universal gradient methods for convex optimization problems. Math. Program. **152**(1), 381–404 (2015)
45. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**, 127–152 (2005)
46. Nesterov, Y.: Excessive gap technique in nonsmooth convex minimization. SIAM J. Optim. **16**, 235–249 (2005)
47. Nesterov, Y.: Gradient methods for minimizing composite objective function. Math. Program. **140**, 125–161 (2013)
48. Neumaier, A.: OSGA: a fast subgradient algorithm with optimal complexity. Math. Program. **158**(1), 1–21 (2016)
49. Neumaier, A.: Solving ill-conditioned and singular linear systems: a tutorial on regularization. SIAM Rev. **40**(3), 636–666 (1998)
50. Neumaier, A.: Introduction to numerical analysis. Cambridge University Press, Cambridge (2001)
51. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends in Optimization **1**(3), 123–231 (2013)
52. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. **58**, 267–288 (1996)
53. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization, technical report, mathematics department, university of washington, http://pages.cs.wisc.edu/brecht/cs726docs/tseng.APG.pdf (2008)
54. Yuan, M., Lin, Y.: Model selection and estimation in regress ion with grouped variables. J. R. Stat. Soc. **68**, 49–67 (2006)