Noname manuscript No.
(will be inserted by the editor)

Solving nonsmooth convex optimization with complexity $O(\varepsilon^{-1/2})$

Masoud Ahookhosh · Arnold Neumaier

the date of receipt and acceptance should be inserted later

Abstract This paper describes an algorithm for solving structured nonsmooth convex optimization problems using the optimal subgradient algorithm (OSGA), which is a first-order method with the complexity $O(\varepsilon^{-2})$ for Lipschitz continuous nonsmooth problems and $O(\varepsilon^{-1/2})$ for smooth problems with Lipschitz continuous gradient. If the nonsmoothness of the problem is manifested in a structured way, we reformulate the problem in a form that can be solved efficiently by OSGA with the complexity $O(\varepsilon^{-1/2})$. To solve the reformulated problem, we equip OSGA by an appropriate prox-function for which the OSGA subproblem can be solved either in a closed form or by a simple iterative scheme, which decreases the computational cost of applying the algorithm, especially for large-scale problems. We show that applying the new scheme is feasible for many problems arising in applications. Some numerical results are reported.

Keywords Structured nonsmooth convex optimization · Subgradient methods · Proximity operator · Optimal complexity · First-order black-box information · High-dimensional data

Mathematics Subject Classification (2000) 90C25 · 90C60 · 49M37 · 65K05 · 68Q25

1 Introduction

Subgradient methods are a class of first-order methods that have been developed to solve convex nonsmooth optimization problems, dating back to 1960, see, for example, [53, 58]. In general, they only need function values and subgradients, and not only inherit the basic features of general first-order methods such as low memory requirement and simple structure but also able to deal with every convex optimization problem. Thus they are suitable for solving convex problems involving large number of variables, say several millions. Although these features make them very attractive for large problems in applications involving high-dimensional data, they suffer from low convergence rate, which finally limits the attainable accuracy. In 1983, Nemirovski & Yudin in [40] derived the worst-case complexity bound for several classes of problems by first-order methods to achieve an ε-solution, where it is $O(\varepsilon^{-2})$ for Lipschitz continuous nonsmooth problems and $O(\varepsilon^{-1/2})$ for smooth problems with Lipschitz continuous gradient. The low convergence speed of subgradient methods suggests that they often reach an ε-solution in the number of iterations closing to the worst-case complexity bound on iterations.

In [40] it was proved that the subgradient, subgradient projection, and mirror descent methods attain the optimal complexity of first-order methods for solving Lipschitz continuous nonsmooth problems; here the mirror decent method is a generalization of the subgradient projection method, cf. [9, 11]. Nesterov in [45, 46] proposed some primal-dual subgradient schemes, which attain the the complexity $O(\varepsilon^{-2})$ for Lipschitz continuous nonsmooth problems. Juditsky & Nesterov in [30] proposed a primal-dual subgradient scheme for uniformly convex functions with an unknown convexity parameter, which attains the complexity close to the optimal bound. Nesterov in [42] and later in [41] proposed some gradient methods for solving smooth problems with Lipschitz continuous gradients attaining the complexity $O(\varepsilon^{-1/2})$. He also in [43, 44] proposed some smoothing methods for structured nonsmooth problems getting the

Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
E-mail: masoud.ahookhosh@univie.ac.at
E-mail: Arnold.Neumaier@univie.ac.at

complexity $O(\varepsilon^{-1/2})$. Smoothing methods also have been studied by many authors, see, for example, Beck & Teboulle in [10], Bot & Hendrich in [12, 13], and Devolder et al. in [24].

In many fields of applied sciences and engineering such as signal and image processing, geophysics, economic, machine learning, and statistics, there exist many applications that can be modeled as a convex optimization problem, in which the objective function is a composite function of a smooth function with Lipschitz continuous gradients and a nonsmooth function, see Ahookhosh [1] and references therein. Hence studying this class of problems using first-order methods has dominated the convex optimization literature in the recent years. Nesterov in [47, 48] proposed some gradient methods for solving composite problems obtaining the complexity $O(\varepsilon^{-1/2})$. For this class of problems, some more first-order methods with the complexity $O(\varepsilon^{-1/2})$ have been developed by Auslander & Teboulle [7], Beck & Teboulle [10], Chen [16, 17, 18, 19], Devolder et al. [23], Gonzaga et al. [28, 29], Lan [35], Lan et al. [36], and Tseng [60]. In particular, Neumaier in [49] proposed an optimal subgradient algorithm (OSGA) attaining the complexity $O(\varepsilon^{-2})$ for Lipschitz continuous nonsmooth problems and the complexity $O(\varepsilon^{-1/2})$ for smooth problems with Lipschitz continuous gradients at the same time. OSGA is a black-box method and does not need to know about global information of the objective function such as Lipschitz constants.

Content. In this paper we consider a class of structured nonsmooth convex constrained optimization problems that is a generalization of the composite problems discussed above, which is frequently found in applications. OSGA behaves well for composite problems in applications, see Ahookhosh [1] and Ahookhosh & Neumaier [5, 6], however, it does not have the complexity $O(\varepsilon^{-1/2})$ for this class of problems. Hence we first reformulate the problem considered in a way that only the smooth part remains in the objective, in the cost of adding a functional constraint to our feasible domain. Afterward, we propose a suitable prox-function and show that solving the OSGA auxiliary subproblem for the reformulated problem is equivalent to solving a proximal-like problem. It is shown that this proximal-like subproblem can be solved efficiently for many cases appearing in applications either in a closed form or by a simple iterative scheme. Due to this reformulation, the problem can be solved by OSGA with the complexity $O(\varepsilon^{-1/2})$. Finally, some numerical results are reported suggesting a good behavior of OSGA.

The remainder of this paper is organized as follows. In the next section we give some preliminary results needed later in the paper. In Section 3 we briefly review the main idea of OSGA. In Section 4 we give a reformulation for the basic problem considered and show that solving the OSGA subproblem is equivalent to solving a proximal-like problem. Section 5 points out how the proximal-like subproblem can be solved in many interesting cases. Some numerical results are reported in Section 6, and conclusions are given in Section 7.

2 Preliminaries and notation

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

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

If $\mathcal{V} = \mathbb{R}^n$, then, for $1 \leq p \leq \infty$,

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

where $x = (x_{g_1}, \cdots, x_{g_m}) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$ in which $n_1 + \cdots + n_m = n$. We set $(x)_+ = \max(x, 0)$. For a function $f : \mathcal{V} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$,

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

denotes its effective domain, and f is called proper if $\mathrm{dom} f \neq \emptyset$ and $f(x) > \infty$ for all $x \in \mathcal{V}$. Let C be a subset of $\mathcal{V}$. In particular, if C is a box, we denote it by $\mathbf{x} = [\underline{x}, \overline{x}]$, where in which $\underline{x}$ and $\overline{x}$ are the

vectors of lower and upper bounds on the components of x, respectively. The vector $g \in \mathcal{V}^*$ is called a subgradient of f at x if $f(x) \in \mathbb{R}$ and

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

The set of all subgradients is called is called the subdifferential of f at x and is denoted by $\partial f(x)$. If $f : \mathcal{V} \to \mathbb{R}$ is nonsmooth and convex, then Fermat's optimality condition for the nonsmooth convex optimization problem

$$\begin{array}{l}\min\ f(x)\\ \text{s.t. } x \in C\end{array}$$

is given by

$$0 \in \partial f(x) + N_C(x), \tag{1}$$

where $N_C(x)$ is the normal cone of C at x defined by

$$N_C(x) = \{p \in \mathcal{V} \mid \langle p, x - z\rangle \geq 0 \ \ \forall z \in C\}. \tag{2}$$

The proximal-like operator $\text{prox}^C_{\lambda f}(y)$ is the unique optimizer of the optimization problem

$$\text{prox}^C_{\lambda f}(y) := \underset{x\in C}{\text{argmin}}\, \frac{1}{2}\|x - y\|_2^2 + \lambda f(x), \tag{3}$$

where $\lambda > 0$. From (1), the first-order optimality condition for the problem (3) is given by

$$0 \in x - y + \lambda\partial f(x) + N_C(x). \tag{4}$$

If $C = \mathcal{V}$, then (4) is simplified to

$$0 \in x - y + \lambda\partial f(x), \tag{5}$$

giving the classical proximity operator. A function f is called strongly convex with the convexity parameter σ if and only if

$$f(z) \geq f(x) + \langle g, z - x\rangle + \frac{\sigma}{2}\|z - x\|_2^2, \tag{6}$$

for all $x, z \in \mathcal{V}$ where g denotes any subgradient of f at x, i.e., $g \in \partial f(x)$. The conjugate function of f is

$$f^* : \mathcal{V} \to \overline{\mathbb{R}},\ f^*(g) = \sup_{x\in\mathcal{V}}\{\langle g, x\rangle - f(x)\}.$$

If f is proper and convex, then f^* is also proper and convex. The next result, proved in [8], will be used in Proposition 23 to derive the subdifferential of some desired functions.

Proposition 21 (Fenchel-Young inequality) *Let $f : \mathcal{V} \to \overline{\mathbb{R}}$ be proper, and let $x \in \mathcal{V}$ and $g \in \mathcal{V}^*$. Then*

$$f(x) + f^*(g) \geq \langle g, x\rangle$$

Moreover, $g \in \partial f(x)$ if and only if

$$f(x) + f^*(g) = \langle g, x\rangle. \tag{7}$$

Lemma. 22 *Let $\phi : \mathcal{V} \to \mathbb{R},\ \phi(x) = \|Wx\|$, where $W : \mathcal{U} \to \mathcal{V}$ is a linear continuous invertible operator and $\|\cdot\|$ is any norm in the vector space $\mathcal{V}$. Then we have*

$$\phi^*(g) = \begin{cases} 0 & \text{if } \|\left(W^{-1}\right)^* g\|_* \leq 1,\\ +\infty & \text{otherwise.}\end{cases}$$

Proof By setting $y = Wx$ and the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\phi^*(g) &= \sup_{x\in\mathcal{V}}\{\langle g, x\rangle - \|Wx\|\} = \sup_{y\in\mathcal{U}}\left\{\langle g, W^{-1}y\rangle - \|y\|\right\} = \sup_{y\in\mathcal{U}}\left\{\langle \left(W^{-1}\right)^* g, y\rangle - \|y\|\right\}\\ &\leq \sup_{y\in\mathcal{U}}\left\{\|\left(W^{-1}\right)^* g\|_*\|y\| - \|y\|\right\} = \sup_{y\in\mathcal{U}}\left\{\left(\|\left(W^{-1}\right)^* g\|_* - 1\right)\|y\|\right\}.\end{aligned}$$

If $\|\left(W^{-1}\right)^* g\|_* \leq 1$, then $\phi^*(g) = 0$. If $\|\left(W^{-1}\right)^* g\|_* > 1$, we have $\|\left(W^{-1}\right)^* g\|_* = \sup_{\|y\|\leq 1}\langle \left(W^{-1}\right)^* g, y\rangle > 1$. Thus there exists $\widetilde{y} \in \mathcal{U}$ such that $\|\widetilde{y}\| \leq 1$ and $\langle \left(W^{-1}\right)^* g, \widetilde{y}\rangle > 1$ leading to

$$\begin{aligned}\phi^*(g) &= \sup_{x\in\mathcal{V}}\{\langle g, x\rangle - \|Wx\|\} = \sup_{y\in\mathcal{U}}\left\{\langle g, W^{-1}y\rangle - \|y\|\right\} = \sup_{y\in\mathcal{U}}\left\{\langle \left(W^{-1}\right)^* g, y\rangle - \|y\|\right\}\\ &\geq \sup_{t>0}\left\{\langle \left(W^{-1}\right)^* g, t\widetilde{y}\rangle - \|t\widetilde{y}\|\right\} = \sup_{t>0}\left\{t\left(\langle \left(W^{-1}\right)^* g, \widetilde{y}\rangle - \|\widetilde{y}\|\right)\right\} = +\infty,\end{aligned}$$

giving the result. □

We use Lemma 22 to derive the subdifferential of $\phi(x) = \|Wx\|$ for an arbitrary norm $\|\cdot\|$ in the vector space $\mathcal{V}$ and a linear continuous invertible operator W.

Proposition 23 *Let $\phi : \mathcal{V} \to \mathbb{R}$, $\phi(x) = \|Wx\|$, where $W : \mathcal{U} \to \mathcal{V}$ is a linear continuous invertible operator and $\|\cdot\|$ is any norm in the vector space $\mathcal{V}$. Then*

$$\partial\phi(x) = \begin{cases} \{g \in \mathcal{V}^* \mid \| \left(W^{-1}\right)^* g\|_* \leq 1\} & \text{if } x = 0, \\ \{g \in \mathcal{V}^* \mid \| \left(W^{-1}\right)^* g\|_* = 1,\ \langle g, x\rangle = \|Wx\|\} & \text{if } x \neq 0. \end{cases}$$

In particular, if $\|\cdot\|$ is self-dual ($\|\cdot\| = \|\cdot\|_*$), we have

$$\partial\phi(x) = \begin{cases} \{g \in \mathcal{V}^* \mid \| \left(W^{-1}\right)^* g\|_* \leq 1\} & \text{if } x = 0, \\ W^* \frac{Wx}{\|Wx\|} & \text{if } x \neq 0. \end{cases}$$

Proof If $x = 0$, the Fenchel-Young equality (7) and Lemma 22 imply

$$\phi(0) + \phi^*(g) = \phi^*(g) = \langle g, 0\rangle = 0,$$

leading to

$$\partial\phi(0) = \left\{g \in \mathcal{V}^* \mid \| \left(W^{-1}\right)^* g\|_* \leq 1\right\}.$$

If $x \neq 0$, the Fenchel-Young equality (7) implies

$$\phi(x) + \phi^*(g) = \|Wx\| + \phi^*(g) = \langle g, x\rangle,$$

leading to

$$\phi^*(g) = 0, \quad \langle g, x\rangle = \|Wx\|.$$

By setting $y = Wx$ and using the Cauchy-Schwarz inequality, we get

$$\|y\| = \|Wx\| = \langle g, x\rangle = \langle g, W^{-1}y\rangle = \langle \left(W^{-1}\right)^* g, y\rangle \leq \| \left(W^{-1}\right)^* g\|_* \|y\| \leq \|y\| \tag{8}$$

implying

$$\| \left(W^{-1}\right)^* g\|_* = 1,$$

leading to

$$\partial\phi(x) = \left\{g \in \mathcal{V}^* \mid \| \left(W^{-1}\right)^* g\|_* = 1,\ \langle g, x\rangle = \|Wx\|\right\}, \quad \text{for } x \neq 0$$

If $\|\cdot\|$ is self-dual then (8) implies

$$\langle g, x\rangle = \langle g, W^{-1}y\rangle = \langle \left(W^{-1}\right)^* g, y\rangle = \| \left(W^{-1}\right)^* g\|_* \|y\|,$$

hence $\left(W^{-1}\right)^* g = \alpha y$ for $\alpha > 0$. Since $\|\cdot\|$ is self-dual and $\| \left(W^{-1}\right)^* g\|_* = \| \left(W^{-1}\right)^* g\| = 1$, we obtain

$$1 = \| \left(W^{-1}\right)^* g\| = \alpha\|y\|,$$

leading to

$$g = W^* \frac{1}{\|y\|} y = W^* \frac{Wx}{\|Wx\|},$$

giving the result. □

In the following example we show how Proposition 23 is applied for $\phi = \|\cdot\|_\infty$, which will be needed in Section 5. The subdifferential of other norms of $\mathcal{V}$ can be computed with Proposition 23 in the same way.

Example. 24 *We use Proposition 23 to derive the subdifferential of $\phi = \|\cdot\|_\infty$ at arbitrary point x. We first recall that the dual norm $\|\cdot\|_\infty$ is $\|\cdot\|_1$. If $x = 0$, Proposition 23 implies*

$$\partial\phi(0) = \{g \in \mathcal{V}^* \mid \|g\|_1 \leq 1\} = \left\{ g \in \mathcal{V}^* \mid g = \sum_{i=1}^n \beta_i e_i,\ \beta \in [-1,1],\ \sum_{i=1}^n |\beta_i| \leq 1 \right\},$$

where .Then we have

$$\partial\phi(x) = \left\{ g \in \mathcal{V}^* \mid \|g\|_1 = 1, \langle g, x\rangle = \|x\|_\infty = \max_{1\leq i\leq n} |x_i| \right\} = \left\{ g \in \mathcal{V}^* \mid \sum_{j=1}^n |g_j| = 1, \sum_{j=1}^n g_j x_j = \|x\|_\infty \right\}$$

If $x \neq 0$, we set

$$\mathcal{I} := \{i \in \{1, \cdots, n\} \mid \|x\|_\infty = |x_i|\}$$

and we have $\|x\|_\infty = \sum_{i\in\mathcal{I}} \beta_i |x_i|$ and $\sum_{i\in\mathcal{I}} \beta_i = 1$ leading to

$$\partial\phi(x) = \left\{ g \in \mathcal{V}^* \mid g = \sum_{i\in\mathcal{I}} \beta_i \ \mathrm{sign}(x_i) e_i, \quad \sum_{i\in\mathcal{I}} \beta_i = 1 \right\}.$$

3 A review of OSGA

In this section we briefly review the main idea of optimal subgradient algorithm proposed by Neumaier in [49]. To this end, we first consider the convex constrained minimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in C, \end{array} \tag{9}$$

where $f : C \to \overline{\mathbb{R}}$ is a proper and convex function defined on a nonempty, closed, and convex subset C of $\mathcal{V}$. The aim is to derive an approximating a solution $\widehat{x} \in C$ using the first-order black-box information, function values and subgradients. OSGA (see Algorithm 1) is an optimal subgradient algorithm for the problem (9) that constructs a sequence of iterates whose related function values converge to the minimum with the optimal complexity. The primary objective is to monotonically reduce bounds on the error term $f(x_b) - \widehat{f}$ of the function values, where $\widehat{f}$ denotes the minimum and x_b is the best known point.

In details, OSGA considers a linear relaxation of f at z defined by

$$f(x) \geq \gamma + \langle h, x\rangle \quad \text{for all } x \in C, \tag{10}$$

where $\gamma \in \mathbb{R}$ and $h \in \mathcal{V}^*$ and a continuously differentiable prox-function $Q : C \to \mathbb{R}$ satisfying (6) and

$$Q_0 := \inf_{x\in C} Q(x) > 0 \tag{11}$$

Afterwards, OSGA requires an efficient routine for finding a maximizer $\widehat{u} = U(\gamma, h)$ and the optimal objective value $E(\gamma, h)$ of an auxiliary problem of the form

$$\begin{array}{ll} \sup & E_{\gamma,h}(x) \\ \text{s.t.} & x \in C, \end{array} \tag{12}$$

where it is known that the supremum is positive. The function $E_{\gamma,h} : C \to \mathbb{R}$ is defined by

$$E_{\gamma,h}(x) := -\frac{\gamma + \langle h, x\rangle}{Q(x)}. \tag{13}$$

with $\gamma \in \mathbb{R}$, $h \in \mathcal{V}^*$. It is assumed that $e = E(\gamma, h)$ and and $u = U(\gamma, h)$ are readily computable.

In [49] it is shown that OSGA attains the following bound on function values

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

Hence, by 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 some target tolerance $\varepsilon > 0$. In [49], it is shown that the number of iterations to achieve this optimizer is $O(\varepsilon^{-1/2})$ for smooth f with Lipschitz continuous gradients and $O(\varepsilon^{-2})$ for Lipschitz continuous nonsmooth f, which are optimal in both cases, cf. [40]. The algorithm does not needs to know about the global Lipschitz parameters and has the low memory requirement. Hence if the subproblem (12) can be solved efficiently, it is appropriate for solving large-scale problems. Numerical results reported by Ahookhosh [1] and Ahookhosh & Neumaier [3, 4], for unconstrained problems, and by Ahookhosh & Neumaier [5, 6], for simple constrained problems, demonstrate that OSGA is well-behaved for problems in applications. In the next section we show that OSGA can solve some structured problems with the complexity $O(\varepsilon^{-1/2})$. Moreover, it is shown that by selecting a suitable prox-function Q, the subproblem (12) can be solved efficiently for this class of problems.

Algorithm 1: OSGA (optimal subgradient algorithm)

Input: global parameters: $\delta, \alpha_{\max} \in]0,1[$, $0 < \kappa' \leq \kappa$; local parameters: x_0, $\mu \geq 0$, f_{target};
Output: x_b, f_{x_b};
begin
  choose an initial best point x_b;
  compute f_{x_b} and g_{x_b};
  if $f_{x_b} \leq f_{\text{target}}$ **then**
    stop;
  else
    $h = g_{x_b} - \mu g_Q(x_b)$; $\gamma = f_{x_b} - \mu Q(x_b) - \langle h, x_b \rangle$;
    $\gamma_b = \gamma - f_{x_b}$; $u = U(\gamma_b, h)$; $\eta = E(\gamma_b, h) - \mu$;
  end
  $\alpha \leftarrow \alpha_{\max}$;
  while *stopping criteria do not hold* **do**
    $x = x_b + \alpha(u - x_b)$; compute f_x and g_x;
    $g = g_x - \mu g_Q(x)$; $\overline{h} = h + \alpha(g - h)$;
    $\overline{\gamma} = \gamma + \alpha(f_x - \mu Q(x) - \langle g, x \rangle - \gamma)$;
    $x'_b = \operatorname{argmin}_{z \in \{x_b, x\}} f(z, v_z)$; $f_{x'_b} = \min\{f_{x_b}, f_x\}$;
    $\gamma'_b = \overline{\gamma} - f_{x'_b}$; $u' = U(\gamma'_b, \overline{h})$;
    $x' = x_b + \alpha(u' - x_b)$; compute $f_{x'}$;
    choose $\overline{x}_b$ in such a way that $f_{\overline{x}_b} \leq \min\{f_{x'_b}, f_{x'}\}$;
    $\overline{\gamma}_b = \overline{\gamma} - f_{\overline{x}_b}$; $\overline{u} = U(\overline{\gamma}_b, \overline{h})$; $\overline{\eta} = E(\overline{\gamma}_b, \overline{h}) - \mu$; $x_b = \overline{x}_b$; $f_{x_b} = f_{\overline{x}_b}$;
    if $f_{x_b} \leq f_{\text{target}}$ **then**
      stop;
    else
      update the parameters α, h, γ, η and u using PUS;
    end
  end
end

If the best function value f_{x_b} is stored and updated, than each iteration of OSGA requires the computation of two function values f_x and $f_{x'}$ and one subgradient g_x, i.e., for problem of the form (9), two times applying the forward linear operator $\mathcal{A}$ and one applying its adjoint $\mathcal{A}^*$ are needed. Therefore, OSGA needs a routine for computing function values and subgradients and a routine for applying forward and adjoint operators.

As discussed in [49], in order to updating the given parameters α, h, γ, η and u, OSGA uses the following scheme:

Algorithm 2: PUS (parameters updating scheme)

Input: $\delta,\ \alpha_{\max} \in]0,1[,\ 0 < \kappa' \leq \kappa,\ \alpha,\ \eta,\ \bar{h},\ \bar{\gamma},\ \bar{\eta},\ \bar{u};$
Output: $\alpha,\ h,\ \gamma,\ \eta,\ u;$
begin
  $R \leftarrow (\eta - \overline{\eta})/(\delta\alpha\eta);$
  if $R < 1$ **then**
    $h \leftarrow \overline{h};$
  else
    $\overline{\alpha} \leftarrow \min(\alpha e^{\kappa'(R-1)}, \alpha_{\max});$
  end
  $\alpha \leftarrow \overline{\alpha};$
  if $\overline{\eta} < \eta$ **then**
    $h \leftarrow \overline{h};\ \gamma \leftarrow \overline{\gamma};\ \eta \leftarrow \overline{\eta};\ u \leftarrow \overline{u};$
  end
end

4 Structured convex optimization problems

In this paper we consider the convex constrained problem

$$\begin{array}{ll} \min & f(\mathcal{A}x, \phi(x)) \\ \text{s.t.} & x \in C, \end{array} \tag{14}$$

where $f : \mathcal{U} \times \mathbb{R} \to \mathbb{R}$ is a proper and convex function that is smooth with Lipschitz continuous gradients with respect to both arguments and monotone increasing with respect to the second argument, $\mathcal{A} : \mathcal{V} \to \mathcal{U}$ is a linear operator, $C \subseteq \mathcal{V}$ is a simple convex domain, and $\phi : \mathcal{V} \to \mathbb{R}$ is a simple nonsmooth, real-valued, and convex loss function. This class of convex problems generalizes the composite problem considered in [47, 48]. As discussed in Section 2, OSGA attains the complexity $O(\varepsilon^{-2})$ for this class of problems. Hence we aim to reformulate the problem (14) in such a way that OSGA attains the complexity $O(\varepsilon^{-1/2})$. We here reformulate the problem (14) in the form

$$\begin{array}{ll} \min & \widehat{f}(x, \xi) \\ \text{s.t.} & x \in \widehat{C}, \end{array} \tag{15}$$

where

$$\widehat{f}(x, \xi) := f(\mathcal{A}x, \xi), \tag{16}$$

$$\widehat{C} := \{(x, \xi) \in \mathcal{V} \times \mathbb{R} \mid x \in C,\ \ \phi(x) \leq \xi\}. \tag{17}$$

By the assumptions about f, the reformulated function $\widehat{f}$ is smooth and has Lipschitz continuous gradients. OSGA can handle the problems of the form (15) with the complexity $O(\varepsilon^{-1/2})$ in the price of adding a functional constraint to the feasible domain C. In the next subsection we will show that how OSGA can effectively handle (15) with the feasible domain $\widehat{C}$.

Problems of the form (14) appears in many applications in the fields of signal and image processing, machine learning, statistics, economic, geophysics, and inverse problems. In the remainder of the paper we deal with such applications, however, we here mention the following example.

Example. 41 (COMPOSITE MINIMIZATION) *We consider the unconstrained minimization problem*

$$\begin{array}{ll} \min & f(\mathcal{A}x) + \phi(x) \\ \text{s.t.} & x \in C, \end{array} \tag{18}$$

where $f : \mathcal{U} \to \overline{\mathbb{R}}$ is a smooth, proper, and convex function, $\mathcal{A} : \mathcal{V} \to \mathcal{U}$ is a linear operator, and $\phi : \mathcal{V} \to \mathbb{R}$ is a simple but nonsmooth, real-valued, and convex loss function. In this case we reformulate (18) in the form (15) by defining the problem

$$\begin{array}{ll} \min & \widetilde{f}(\mathcal{A}x, \xi) \\ \text{s.t.} & \phi(x) \leq \xi, \end{array} \tag{19}$$

where $\widetilde{f}:\widetilde{C}\to\mathbb{R}$, $\widetilde{f}(\mathcal{A}x,\xi):=f(\mathcal{A}x)+\xi$ *with the feasible set* $\widetilde{C}$ *is defined by*

$$\widetilde{C}:=\{(x,\xi)\in C\times\mathbb{R}\ \mid\ \phi(x)\leq\xi\}.$$

Consider the linear inverse problem

$$y=\mathcal{A}x+\nu, \tag{20}$$

where $x\in\mathbb{R}^n$ *is the original object,* $y\in\mathbb{R}^m$ *is an observation, and* $\nu\in\mathbb{R}^m$ *is additive or impulsive noise. The objective is to recover* x *from* y *by solving (20). In practice, this problem is typically underdetermined and ill-conditioned, and* ν *is unknown. Hence* x *typically is approximated by one of the minimization problems*

$$\begin{aligned}&\min && \frac{1}{2}\|y-Ax\|_2^2+\frac{1}{2}\lambda\|x\|_2^2\\ &\text{s.t.} && x\in\mathbb{R}^n,\end{aligned} \tag{21}$$

$$\begin{aligned}&\min && \frac{1}{2}\|y-Ax\|_2^2+\lambda\|x\|_1\\ &\text{s.t.} && x\in\mathbb{R}^n,\end{aligned} \tag{22}$$

or

$$\begin{aligned}&\min && \frac{1}{2}\|y-Ax\|_2^2+\frac{1}{2}\lambda_1\|x\|_2^2+\lambda_2\|x\|_1\\ &\text{s.t.} && x\in\mathbb{R}^n.\end{aligned} \tag{23}$$

These problems can be reformulated in the form (18) by setting

$$f(x,\xi):=\frac{1}{2}\|y-Ax\|_2^2+\xi,\quad \phi(x):=\frac{1}{2}\lambda\|x\|_2^2, \tag{24}$$

$$f(x,\xi):=\frac{1}{2}\|y-Ax\|_2^2+\xi,\quad \phi(x):=\lambda\|x\|_1, \tag{25}$$

or

$$f(x,\xi):=\frac{1}{2}\|y-Ax\|_2^2+\xi,\quad \phi(x):=\frac{1}{2}\lambda_1\|x\|_2^2+\lambda_2\|x\|_1, \tag{26}$$

respectively.

4.1 **Description of OSGA's new setup**

This section devotes to solving the OSGA subproblem (12) for a problem of the form (15). To this end, we introduce some prox-function and employ it to derive an inexpensive solution of the subproblem. We generally assume that the domain C is simple enough such that $E(\eta,y)$ and $U(\eta,y)$ can be computed cheaply, in $O(n\log n)$ operations, say.

Lemma. 42 *Let* $Q:\mathcal{V}\times\mathbb{R}\to\mathbb{R}$ *be a function defined by*

$$Q(x,x_0):=Q_0+\frac{1}{2}\left(\|x\|_2^2+x_0^2\right), \tag{27}$$

where $Q_0>0$. *Then* Q *is strongly convex, and* $Q(x,x_0)>0$.

Proof Since $g_Q(x)=(x\ x_0)^T$, we obtain

$$\begin{aligned}
&Q(z,z_0)+\langle g_Q(z,z_0),(x-z,x_0-z_0)\rangle+\frac{1}{2}\|(x-z,x_0-z_0)^T\|_2^2\\
&\quad=Q_0+\frac{1}{2}\left\langle(z,z_0)^T,(z,z_0)^T\right\rangle+\left\langle(z,z_0)^T,(x-z,x_0-z_0)^T\right\rangle\\
&\qquad+\frac{1}{2}\left\langle(x-z,x_0-z_0)^T,(x-z,x_0-z_0)^T\right\rangle\\
&\quad=Q_0+\frac{1}{2}\left\langle(z,z_0)^T,(x,x_0)^T\right\rangle+\frac{1}{2}\left\langle(x,x_0)^T,(x-z,x_0-z_0)^T\right\rangle\\
&\quad=Q_0+\frac{1}{2}\left\langle(x,x_0)^T,(x,x_0)^T\right\rangle=Q_0+\frac{1}{2}\left\|(x,x_0)^T\right\|_2^2\\
&\quad=Q(x,x_0).
\end{aligned}$$

This means that Q is a strongly convex function with the convexity parameter 1, and since $Q_0>0$, we get $Q(x,x_0)>0$. □

Lemma 42 shows that the quadratic function Q defined by (27) is a prox-function. We now replace the linear relaxation (10) by

$$f(x, x_0) \geq \gamma + \langle h, x\rangle + h_0 x_0 \quad \text{for all } x \in \widehat{C}. \tag{28}$$

By using this linear relaxation and the prox-function (27), the subproblem (12) is rewritten in the form

$$\begin{array}{l} \sup\ E_{\gamma,h,h_0}(x) \\ \text{s.t.}\ \ (x, x_0) \in C \times \mathbb{R},\ \phi(x) \leq x_0, \end{array} \tag{29}$$

where $E_{\gamma,h,h_0} : \mathcal{V} \times \mathbb{R} \to \mathbb{R}$ and

$$E_{\gamma,h,h_0}(x, x_0) := \frac{\gamma + \langle h, x\rangle + h_0 x_0}{Q(x, x_0)}, \tag{30}$$

which is a differentiable function. The next result gives a bound on the error $f(x_b) - \widehat{f}$, which is important for the complexity analysis of the the new setup of OSGA.

Proposition 43 *Let $\gamma_b := \gamma - f(x_b)$, $u := U(\gamma_b, h, h_0)$, and $\eta := E(\gamma_b, h, h_0)$. Then we have*

$$0 \leq f(x_b) - \widehat{f} \leq \eta Q(\widehat{x}, \widehat{x}_0). \tag{31}$$

In particular, if x_b is not yet optimal then the choice $u = U(\gamma_b, h, h_0)$ implies $E(\gamma_b, h, h_0) > 0$.

Proof Using (28), (29), and (30), this follows similar to Proposition 2.1 in [49]. □

Proposition 44 *Let $e := E(\gamma, h, h_0) > 0$ and $u = U(\gamma, h, h_0)$. Then*

$$\gamma + \langle h, u\rangle + h_0 u_0 = -eQ(u, u_0), \tag{32}$$

$$\langle e g_Q(u, u_0) + h, x - u\rangle + (eu_0 + h_0)(x_0 - u_0) \geq 0 \quad \text{for all } (x, x_0) \in C \times \mathbb{R},\ \phi(x) \leq x_0. \tag{33}$$

Proof The problem (29) and the definition (30) imply that the function $\zeta : C \times \mathbb{R} \to \mathbb{R}$ defined by

$$\zeta(x, x_0) := \gamma + \langle h, x\rangle + h_0 x_0 + eQ(x, x_0)$$

is nonnegative and vanishes for $(x, x_0) = (u, u_0) := U(\gamma, h, h_0)$, i.e., the identity (32) holds. Since $\zeta(x, x_0)$ is continuously differentiable with gradient $g_\zeta(x, x_0) = (h + \eta g_Q(x), eu_0 + h_0)^T$, the first order optimality condition holds, i.e.,

$$\langle g_\zeta(x, x_0), x - u\rangle + (eu_0 + h_0)(x_0 - u_0)) \geq 0 \tag{34}$$

for all $(x, x_0) \in C \times \mathbb{R},\ \phi(x) \leq x_0$, giving the results. □

The next result gives a systematic way for solving OSGA's subproblem (29) for problems of the form (15).

Theorem. 45 *Let $(u, u_0) \in \mathcal{V} \times \mathbb{R}$ be a minimizer of (29) and $e = E_{\gamma,h,h_0}(u, u_0)$. Then*

$$u := u(e, \lambda), \quad u_0 := \phi(u),$$

where $y := -e^{-1}h$, $\lambda := u_0 + e^{-1}h_0$, and

$$\widehat{u} := u(e, \lambda) := \operatorname*{argmin}_{x \in C} \frac{1}{2}\|x - y\|_2^2 + \lambda\phi(x). \tag{35}$$

Furthermore, e and λ can be computed by solving the two-dimensional system of equations

$$\begin{cases} \phi(\widehat{u}) + e^{-1}h_0 - \lambda = 0, \\ e\left(\dfrac{1}{2}(\|\widehat{u}\|_2^2 + \phi(\widehat{u})^2) + Q_0\right) + \gamma + \langle h, \widehat{u}\rangle + h_0\phi(\widehat{u}) = 0. \end{cases} \tag{36}$$

Proof From Proposition 44, at the minimizer (u, u_0), we obtain

$$e\left(\frac{1}{2}(\|u\|_2^2+u_0^2)+Q_0\right)=-\gamma-\langle h,u\rangle-h_0u_0 \tag{37}$$

and

$$\langle eu+h,x-u\rangle+(eu_0+h_0)(x_0-u_0)\geq 0 \quad \forall\ (x,x_0)\in C\times\mathbb{R},\ \phi(x)\leq x_0. \tag{38}$$

We conclude the proof in the next two steps:

Step 1. We first show that this inequality is equivalent to the following two inequalities

$$\begin{cases} eu_0+h_0\geq 0, \\ \\ \langle eg_Q(u,u_0)+h,x-u\rangle+(eu_0+h_0)(\phi(x)-u_0)\geq 0 \quad \forall (x,x_0)\in C\times\mathbb{R}. \end{cases} \tag{39}$$

Assuming that these two inequalities hold, we prove (38). From $\phi(x)\leq x_0$ and $eu_0+h_0\geq 0$, we obtain

$$\begin{aligned}\langle eg_Q(u,u_0)+h,x-u\rangle+(eu_0+h_0)(x_0-u_0)& \\ \geq \langle eg_Q(u,u_0)+h,x-u\rangle+(eu_0+h_0)(\phi(x)-u_0)&\geq 0.\end{aligned}$$

We now assume (38) and prove (39). The inequality $eu_0+h_0\geq 0$ holds; Otherwise, by selecting x_0 big enough, we get

$$\langle eg_Q(u,u_0)+h,x-u\rangle+(eu_0+h_0)(x_0-u_0)<0,$$

which is a contradiction with (38). Since $\phi(x)\leq x_0$, the second inequality in (39) holds.

Step 2. By setting $x=u$ and $u_0=\phi(u)$, we see that u is a solution of the minimization problem

$$\inf_{x\in C}\ \langle eg_Q(u,u_0)+h,x-u\rangle+(eu_0+h_0)(\phi(x)-u_0).$$

The first-order optimality condition (1) for this problem leads to

$$0\in u+e^{-1}h+(u_0+e^{-1}h_0)\ \partial\phi(u)+N_C(u). \tag{40}$$

By writing the first-order optimality condition (4) for the problem

$$\begin{aligned}&\min\ \frac{1}{2}\|x-y\|_2^2+\lambda\phi(x)\\ &\text{s.t.}\ \ x\in C,\end{aligned}$$

we get

$$0\in\widehat{u}-y+\lambda\ \partial\phi(\widehat{u})+N_C(\widehat{u}). \tag{41}$$

By comparing (40) and (41) and setting $y=-e^{-1}h$, $\lambda=u_0+e^{-1}h_0$, we conclude that both problems have the same minimizer $u=\widehat{u}$. Since $u_0=\phi(\widehat{u})$, we obtain

$$\lambda=u_0+e^{-1}h_0=\phi(\widehat{u})+e^{-1}h_0.$$

Using this and substituting $u_0=\phi(\widehat{u})$ in (37), e and λ are found by solving the system of nonlinear equations (36). This complets the proof. □

In Theorem 45, if $C=\mathcal{V}$, the problem (35) is reduced to the classical proximity operator $\widehat{u}=\text{prox}_{\lambda\phi}(y)$ defined in (3). Hence the problem (35) is called ***proximal-like***. Therefore, the world "simple" in the definition of C means that the problem (35) can be solved efficiently either in a closed form or by an inexpensive iterative scheme. To have a clear view of Theorem 45, we give the following example.

Example. 46 *Consider the ℓ_1-regularized least squares problem (22). Then the problem can be reformulated as*

$$\begin{aligned}&\min\ \frac{1}{2}\|y-Ax\|_2^2+\xi\\ &\text{s.t.}\ \ \|x\|_1\leq\xi.\end{aligned}$$

Since $\phi = \|\cdot\|_1$, the solution of (35) is $\widehat{u} = \mathrm{sign}(y_i)(|y_i|-\lambda,0)_+$ with $y = -e^{-1}h$ (see Table 1). Substituting this into (36) gives

$$\begin{cases} \displaystyle\sum_{i=1}^{n}(|y_i|-\lambda)_+ + e^{-1}h_0 - \lambda = 0, \\ \displaystyle e\left(\frac{1}{2}(\sum_{i=1}^{n}(|y_i|-\lambda)_+^2 + (\sum_{i=1}^{n}(|y_i|-\lambda)_+)^2) + Q_0\right) + \gamma + \sum_{i=1}^{n}(h_i+h_0)(|y_i|-\lambda)_+ = 0. \end{cases}$$

This is a two-dimensional system of nonsmooth equations that can be reformulated as a nonlinear least squares problem, see, for example, [51].

Theorem 45 leads to the two-dimensional nonlinear system

$$F(e,\lambda) := (f_1(e,\lambda), f_2(e,\lambda))^T = 0, \tag{42}$$

where

$$\begin{aligned} f_1(e,\lambda) &:= \phi(\widehat{u}) + e^{-1}h_0 - \lambda, \\ f_2(e,\lambda) &:= e\left(\frac{1}{2}(\|\widehat{u}\|_2^2 + \phi(\widehat{u})^2) + Q_0\right) + \gamma + \langle h, \widehat{u}\rangle + h_0\phi(\widehat{u}), \end{aligned}$$

in which $\widehat{u} = u(e,\lambda)$ and $e, \lambda > 0$. For an instance, in Example 46, we have

$$\begin{aligned} f_1(e,\lambda) &= \sum_{i=1}^{n}(|y_i|-\lambda)_+ + e^{-1}h_0 - \lambda, \\ f_2(e,\lambda) &= e\left(\frac{1}{2}(\sum_{i=1}^{n}(|y_i|-\lambda)_+^2 + (\sum_{i=1}^{n}(|y_i|-\lambda)_+)^2) + Q_0\right) + \gamma + \sum_{i=1}^{n}(h_i+h_0)(|y_i|-\lambda)_+. \end{aligned}$$

The system of nonsmooth equations (42) can be handled by the bound-constrained least-squares problem

$$\begin{aligned} \min \quad & \frac{1}{2}\|F(e,\lambda)\|_2^2 \\ \text{s.t.} \quad & e, \lambda > 0 \end{aligned} \tag{43}$$

if $f_1(e,\lambda)$ and $f_2(e,\lambda)$ are smooth and by replacing the vector (e,λ) with $(|e|,|\lambda|)$ and solving

$$\begin{aligned} \min \quad & \frac{1}{2}\|F(|e|,|\lambda|)\|_2^2 \\ \text{s.t.} \quad & e, \lambda \in \mathbb{R} \end{aligned} \tag{44}$$

if $f_1(e,\lambda)$ and $f_2(e,\lambda)$ are nonsmooth. The problem (43) can be handled by various bound-constrained nonlinear optimization schemes such as Newton and quasi-Newton methods [15, 38], Levenberg–Marquardt methods [31], and trust-region methods [21, 26]. The problems (42)and (44), such as Example 46, can be solved by the semismooth Newton method or the smoothing Newton method [56], the quasi-Newton methods [59, 37], the secant method [54], and trust-region methods [2, 55].

In view of Theorem 45, we can propose a systematic way a for solving OSGA's subproblem (29), which is summarized in next scheme.

Algorithm 3: OSS (OSGA's subproblem solver)

Input: Q_0, γ, h;
Output: u, e;
begin
  solve the system of nonlinear equation (42) approximately by a nonlinear solver to find e and λ;
  set $u = \widehat{u}(e,\lambda)$.
end

To implement Algorithm 3 (OSS), we need a reliable nonlinear solver to solve the system of nonlinear equation (42) and a routine giving the solution of the proximal-like problem (35) effectively. In Section 5 we investigate solving the proximal-like problem (35) for some practically important loss function ϕ.

4.2 **Convergence analysis**

In this section we establish the complexity bounds of OSGA for Lipschitz continuous nonsmooth problems and smooth problems with Lipschitz continuous gradients. We also show that if f is strictly convex, the sequence generated by OSGA is convergent to $\widehat{x}$.

To guarantee the existence of a minimizer for OSGA, we assume that the following conditions :

(H1) The objective function f is proper and convex;

(H2) 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 V is a finite-dimensional vector space, (H2) implies that the upper level set $N_f(x_0)$ is convex and compact. It follows from the continuity and properness of the objective function f that it attains its global minimizer on the upper level set $N_f(x_0)$. Therefore, there is at least one minimizer $\widehat{x}$, and its corresponding minimum is denoted by $\widehat{f}$.

Since the underlying problem (15) is a special case of the problem (9) considered by Neumaier in [49], the complexity results for OSGA remains valid.

Theorem. 47 *Suppose that $f - \mu Q$ is convex and $\mu \geq 0$. Then we have*

(i) (Nonsmooth complexity bound) *If the points generated by Algorithm 1 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 $O((\varepsilon^2 + \mu\varepsilon)^{-1})$. Thus the asymptotic worst case complexity is $O(\varepsilon^{-2})$ when $\mu = 0$ and $O(\varepsilon^{-1})$ when $\mu > 0$.*

(ii) (Smooth complexity bound) *If f has Lipschitz continuous gradients with Lipschitz constant L, the total number of iterations needed by Algorithm 1 to reach a point with $f(x) \leq f(\widehat{x}) + \varepsilon$ is at most $O(\varepsilon^{-1/2})$ if $\mu = 0$, and at most $O(|\log\varepsilon|\sqrt{L/\mu})$ if $\mu > 0$.*

Proof Since all assumptions of Theorem 4.1 and 4.2, Propositions 5.2 and 5.3, and Theorem 5.1 in [49] are satisfied, the results remains valid. □

Indeed, if a nonsmooth problem can be reformulates as (15) with a nonsmooth loss function ϕ, then OSGA can solve the reformulated problem with the complexity $O(\varepsilon^{-1/2})$ for an arbitrary accuracy parameter ε. The next result shows that the the sequence $\{x_k\}$ generated by OSGA is convergent to x^* if the objective f is strictly convex and $x^* \in \text{int } C$, where int C denotes the interior of C.

Proposition 48 *Suppose that f is strictly convex, then the sequence $\{x_k\}$ is generated by OSGA is convergent to x^* if $x^* \in \text{int } C$.*

Proof Since f is strictly convex, the minimizer x^* is unique. By $x^* \in \text{int } C$, there exists a small $\delta > 0$ such that the neighborhood

$$N(x^*) := \{x \in C \mid \|x - x^*\| \leq \delta\},$$

is included in C, which is a convex and compact set. Let x_δ be a minimizer of the problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \partial N(x^*), \end{array} \tag{45}$$

where $\partial N(x^*)$ denotes the boundary of $N(x^*)$. Set $\varepsilon_\delta := f(x_\delta) - f^*$ and consider the upper level set

$$N_f(x_\delta) := \{x \in C \mid f(x) \leq f(x_\delta) = f^* + \varepsilon_\delta\}.$$

Now Theorem 47 implies that the algorithm attains an ε_δ-solution of (15) in a finite number κ of iterations. Hence after κ iterations the best point x_b attained by OSGA satisfies $f(x_b) \leq f^* + \varepsilon_\delta$, i.e., $x_b \in N_f(x_\delta)$. We now show that $N_f(x_\delta) \subseteq N(x^*)$. To prove this statement by contradiction, we suppose that there exists $x \in N_f(x_\delta) \setminus N(x^*)$. Since $x \notin N(x^*)$, we have $\|x - x^*\| > \delta$. Therefore, there exists λ_0 such that

$$\|\lambda_0 x + (1 - \lambda_0)x^*\| = \delta.$$

From (45), $f(x) \leq f(x_\delta)$, and the strictly convex property of f, we obtain

$$f(x_\delta) \leq f(\lambda_0 x + (1 - \lambda_0)x^*) < \lambda_0 f(x) + (1 - \lambda_0)f(x^*) \leq \lambda_0 f(x_\delta) + (1 - \lambda_0)f(x_\delta) = f(x_\delta),$$

which is a contradiction, i.e., $N_f(x_\delta) \subseteq N(x^*)$ implying $x \in N(x^*)$ giving the results. □

5 Solving proximal-like subproblem

In this section we show that the proximal-like problem (35) can be solved in a closed form for many special cases appearing in applications. To this end, we first consider unconstrained problems ($C = \mathcal{V}$) and study some problems with simple constrained domains ($C \neq \mathcal{V}$). We give some available proximal-like operators in Table 1.

5.1 Unconstrained examples ($C = \mathcal{V}$)

We here consider several interesting unconstrained proximal problems appearing in applications and explain how the associated OSGA subproblem (35) can be solved.

In recent years the interest of applying regularizations with weighted norms is increased by emerging many applications, see, for example, [22, 57]. Let d be a vector in $\mathbb{R}^n$ such that $d_i \neq 0$ for $i = 1, \cdots, n$. Then we define the weight matrix $D := \mathrm{diag}(d)$, which is a diagonal matrix with $D_{i,i} = d_i$ for $i = 1, \cdots, n$. It is clear that D is a invertible matrix. The next result show how to compute a solution of the problem (35) for special cases of ϕ arising frequently in applications.

Proposition 51 *Let $D := \mathrm{diag}(d)$, where $d \in \mathbb{R}^n$ with $d_i \neq 0$, for $i = 1, \cdots, n$. If $\phi(x) = \|Dx\|_1$, then the proximity operator (35) is given by*

$$\big(\mathrm{prox}_{\lambda\phi}(y)\big)_i = \mathrm{sign}(y_i)(|y_i| - \lambda|d_i|)_+, \tag{46}$$

for $i = 1, \cdots, n$.

Proof The optimaliy condition (5) implies that $u = \mathrm{prox}_{\lambda\phi}(y)$ if and only if

$$0 \in u - y + \lambda\, \partial\|Du\|_1. \tag{47}$$

We consider two cases: (i) $\|D^{-1}y\|_\infty \leq \lambda$; (ii) $\|D^{-1}y\|_\infty > \lambda$.

Case (i). Let $\|D^{-1}y\|_\infty \leq \lambda$. Then we show $u = 0$ satisfies (47). If $u = 0$, Proposition 23 implies $\partial\phi(0) = \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_\infty \leq 1\}$. By substituting this into (47), we get that $u = 0$ is satisfied (47) if $y \in \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_\infty \leq \lambda\}$ leading to $\mathrm{prox}_{\lambda\phi}(y) = 0$. Since the right hand side of (46) is also zero, (46) holds.

Case (ii). Let $\|D^{-1}y\|_\infty > \lambda$. Then Case (i) implies $u \neq 0$. Since $\|\cdot\|_* = \|\cdot\|_\infty$ and D is invertible, Proposition 23 implies that

$$\partial\phi(u) = \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_\infty = 1, \quad \langle g, u\rangle = \|Du\|_1\}$$

leading to

$$\sum_{i=1}^{n}(g_i u_i - |d_i||u_i|) = 0.$$

By induction on nonzero elements of u, we get $g_i u_i = |d_i||u_i|$ for $i = 1, \cdots, n$. This implies $g_i = |d_i|\ \mathrm{sign}(u_i)$. The optimality condition (5) implies

$$0 \in u_i - y_i + \lambda|d_i|\partial|u_i|$$

for $i = 1, \cdots, n$. If $u_i > 0$, then $u_i = -\lambda|d_i| + y_i > 0$. Hence if $\lambda|d_i| < y_i$, we set $u_i = -\lambda|d_i| + y_i$. If $u_i < 0$, then $u_i = \lambda|d_i| + y_i < 0$. Hence if $\lambda|d_i| > y_i$, we set $u_i = \lambda|d_i| + y_i$. Otherwise, we have $u_i = 0$. Therefore, we obtain

$$\big(\mathrm{prox}_{\lambda\phi}(y)\big)_i = \begin{cases} y_i - \lambda|d_i| & \text{if } y_i > \lambda|d_i|, \\ y_i + \lambda|d_i| & \text{if } y_i < -\lambda|d_i|, \\ 0 & \text{otherwise,} \end{cases} \tag{48}$$

giving the result. □

Table 1: List of available proximal-like operators for several ϕ and C, where $D := \text{diag}(d)$, where $d \in \mathbb{R}^n$ with $d_i \neq 0$, for $i = 1, \cdots, n$ and Q is $n \times n$ orthogonal matrix ($Q^TQ = I$).

$\phi(x)$	C	**proximity operator** $u = \text{prox}_\phi(\text{y})$	**Reference**
$\lambda\|Dx\|_1$	$\mathcal{V}$	$u_i = \text{sign}(y_i)(\|y_i\| - \lambda\|d_i\|)_+$	Proposition 51
$\lambda\|Qx\|_1$	$\mathcal{V}$	$u = Q^T\left(\text{sign}(y)\cdot(\|y\| - \lambda)_+\right)$	Prop. 11 in [20]
$\lambda\|Dx\|_2$	$\mathcal{V}$	$u_i = \begin{cases} 0 & \text{if } \|D^{-1}y\|_2 \leq \lambda, \\ (\tau y_i)/(\tau + \lambda d_i^2), & \text{if } \|D^{-1}y\|_2 > \lambda. \end{cases}$ τ is the solution of $\sum_{i=1}^n (d_i^2 y_i^2)/(\tau + \lambda d_i^2)^2 - 1 = 0$.	Proposition 52
$\lambda\|x\|_2$	$\mathcal{V}$	$u = (1 - \lambda/\|y\|_2)_+ y$	[52]
$\frac{1}{2}\lambda\|x\|_2^2$	$\mathcal{V}$	$u = 1/(1+\lambda)y$	[52]
$\frac{1}{2}\lambda_1\|x\|_2^2 + \lambda_2\|x\|_1$	$\mathcal{V}$	$u = 1/(1+\lambda_1)\text{prox}_{\lambda_2\|\cdot\|_1}(y)$	[52]
$\lambda\|x\|_\infty$	$\mathcal{V}$	$u_i = \begin{cases} 0 & \text{if } \|y\|_1 \leq \lambda, \\ \text{sign}(y_i)u_\infty & \text{if } \|y\|_1 > \lambda, i \in \mathcal{I}, \\ y_i & \text{if } \|y\|_1 > \lambda, i \notin \mathcal{I}. \end{cases}$ $\mathcal{I} = \{l_1, \cdots, l_{\widehat{k}}\}$, $u_\infty = \frac{1}{\widehat{k}}\left(\sum_{i\in\mathcal{I}}\|y_i\| - \lambda\right)$ $\widehat{k}$ is the smallest $k \in 1, \cdots, n-1$ such that $\frac{1}{k}\left(\sum_{i\in\mathcal{I}}\|v_i\| - \lambda\right) \geq v_{k+1}$, $v_i = \|y_{l_i}\|$, $l_1, \cdots, l_n$ a permutation of $1, \cdots, n$ such that $v_1 \geq v_2 \geq \cdots \geq v_n$ and otherwise $\widehat{k} = n$	Proposition 53
$\lambda\|x\|_{1,2}$	$\mathcal{V}$	$u_{g_i} = (1 - \lambda/\|y_{g_i}\|_2)_+ y_{g_j}$	Proposition 54
$\lambda\|x\|_{1,\infty}$	$\mathcal{V}$	$u_{g_i} = \begin{cases} 0_{g_i} & \text{if } \|y_{g_i}\|_1 \leq \lambda, \\ \text{sign}(y_{g_i})u_\infty^i & \text{if } \|y_{g_i}\|_1 > \lambda,\ i \in \mathcal{I}_{g_i}, \\ y_{g_i} & \text{if } \|y_{g_i}\|_1 > \lambda,\ i \notin \mathcal{I}_{g_i}. \end{cases}$ $\mathcal{I} = \{l_{g_i}^1, \cdots, l_{g_i}^{\widehat{k}}\}$, $u_\infty^i = \frac{1}{\widehat{k}_i}\left(\sum_{j\in\mathcal{I}_{g_i}}\|y_{g_i}^j\| - \lambda\right)$, $\widehat{k}_i$ is the smallest $k \in 1, \cdots, n_i - 1$ such that $\frac{1}{k}\left(\sum_{j\in\mathcal{I}_{g_i}}\|v_{g_i}^j\| - \lambda\right) \geq v_{g_i}^{k+1}$ $v_{g_i}^j = \|y_{g_i}^j\|$, $l_{g_i}^1, \cdots, l_{g_i}^n$ a permutation of $1, \cdots, n$ such that $v_{g_i}^1 \geq v_{g_i}^2 \geq \cdots \geq v_{g_i}^{n_i}$ and otherwise $\widehat{k}_i = n_i$	Proposition 55
$\lambda\|Dx\|_1$	$x \geq 0$	$u_i = \begin{cases} y_i - \lambda\|d_i\| & \text{if } \mathcal{J} \neq \emptyset,\ y_i > \lambda\|d_i\|, \\ y_i + \lambda\|d_i\| & \text{if } \mathcal{J} \neq \emptyset,\ y_i < -\lambda\|d_i\|, \\ 0 & \text{otherwise.} \end{cases}$ $\mathcal{J} := \{j \in 1, \cdots, n \mid y_j > \lambda\|d_j\|\}$	Proposition 56
$\frac{1}{2}\lambda\|x\|_2^2$	$x \geq 0$	$u_i = \begin{cases} \underline{x}_i & \text{if } y_i \leq 0, \\ y_i/(1+\lambda) & \text{if } y_i > 0, \end{cases}$	Proposition 57
$\frac{1}{2}\lambda_1\|x\|_2^2 + \lambda_2\|Dx\|_1$	$x \geq 0$	$u_i = \begin{cases} 1/(1+\lambda)(y_i - \lambda\|d_i\|) & \text{if } \mathcal{J} \neq \emptyset,\ y_i > \lambda\|d_i\|, \\ 1/(1+\lambda)(y_i + \lambda\|d_i\|) & \text{if } \mathcal{J} \neq \emptyset,\ y_i < -\lambda\|d_i\|, \\ 0 & \text{otherwise.} \end{cases}$ $\mathcal{J} := \{j \in 1, \cdots, n \mid y_j > \lambda\|d_j\|\}$	Proposition 58
$\lambda\|Dx\|_1$	$[\underline{x}, \overline{x}]$	$u_i = \begin{cases} \underline{x}_i & \text{if } \omega > 0,\ \underline{x}_i - y_i + \lambda\|d_i\|\ \text{sign}(\underline{x}_i) \geq 0, \\ \overline{x}_i & \text{if } \omega > 0,\ \overline{x}_i - y_i + \lambda\|d_i\|\ \text{sign}(\overline{x}_i) \leq 0, \\ y_i - \lambda\|d_i\| & \text{if } \omega > 0,\ y_i > \lambda\|d_i\|, \\ y_i + \lambda\|d_i\| & \text{if } \omega > 0,\ y_i < -\lambda\|d_i\|, \\ 0 & \text{otherwise.} \end{cases}$ $\omega = \sum_{y_i+\lambda\|d_i\|<0}(y_i + \lambda\|d_i\|)\underline{x} + \sum_{y_i+\lambda\|d_i\|>0}(y_i + \lambda\|d_i\|)\overline{x}$	Proposition 56
$\frac{1}{2}\lambda\|x\|_2^2$	$[\underline{x}, \overline{x}]$	$u_i = \begin{cases} \underline{x}_i & \text{if } (1+\lambda)\underline{x}_i \geq y_i, \\ \overline{x}_i & \text{if } (1+\lambda)\overline{x}_i \leq y_i, \\ y_i/(1+\lambda) & \text{if } \underline{x}_i < y_i/(1+\lambda) < \overline{x}_i, \end{cases}$	Proposition 57
$\frac{1}{2}\lambda_1\|x\|_2^2 + \lambda_2\|Dx\|_1$	$[\underline{x}, \overline{x}]$	$u_i = \begin{cases} \underline{x}_i & \text{if } \omega > 0,\ \underline{x}_i - y_i + \lambda\|d_i\|\ \text{sign}(\underline{x}_i) \geq 0, \\ \overline{x}_i & \text{if } \omega > 0,\ \overline{x}_i - y_i + \lambda\|d_i\|\ \text{sign}(\overline{x}_i) \leq 0, \\ 1/(1+\lambda)(y_i - \lambda\|d_i\|) & \text{if } \omega > 0,\ y_i > \lambda\|d_i\|, \\ 1/(1+\lambda)(y_i + \lambda\|d_i\|) & \text{if } \omega > 0,\ y_i < -\lambda\|d_i\|, \\ 0 & \text{otherwise.} \end{cases}$ $\omega = \sum_{y_i+\lambda\|d_i\|<0}(y_i + \lambda\|d_i\|)\underline{x} + \sum_{y_i+\lambda\|d_i\|>0}(y_i + \lambda\|d_i\|)\overline{x}$	Proposition 58

Proposition 52 *Let $D := \mathrm{diag}(d)$, where $d \in \mathbb{R}^n$ and $d_i \neq 0$, for $i = 1, \cdots, n$. If $\phi(x) = \|Dx\|_2$, then the proximity operator (35) is given by $\mathrm{prox}_{\lambda\phi}(y) = 0$ if $\|D^{-1}y\|_2 \leq \lambda$ and otherwise*

$$\left(\mathrm{prox}_{\lambda\phi}(y)\right)_i = \frac{\tau y_i}{\tau + \lambda d_i^2},$$

for $i = 1, \cdots, n$, where τ is given by solving the one-dimensional nonlinear equation

$$\sum_{i=1}^{n} \frac{d_i^2 y_i^2}{(\tau + \lambda d_i^2)^2} - 1 = 0,$$

which has a unique solution.

Proof The optimaliy condition (5) shows that $u = \mathrm{prox}_{\lambda\phi}(y)$ if and only if

$$0 \in u - y + \lambda\, \partial\|D^{-1}u\|_2. \tag{49}$$

We consider two cases: (i) $\|D^{-1}y\|_2 \leq \lambda$; (ii) $\|D^{-1}y\|_2 > \lambda$.

Case (i). Let $\|D^{-1}y\|_2 \leq \lambda$. Then we show $u = 0$ satisfies (49). If $u = 0$, Proposition 23 implies $\partial\phi(0) = \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_2 \leq 1\}$. By using this and (49), we get that $u = 0$ is satisfied (49) if $y \in \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_2 \leq \lambda\}$ leading to $\mathrm{prox}_{\lambda\phi}(y) = 0$.

Case (ii). Let $\|D^{-1}y\|_2 > \lambda$. Then Case (i) implies $u \neq 0$. Proposition 23 implies $\partial\phi(u) = D^T Du/\|Du\|_2$, and the optimality conditions (5) yields

$$u - y + \lambda\, D^T \frac{Du}{\|Du\|_2} = 0.$$

By using this and setting $\tau = \|Du\|_2$, we get

$$\left(1 + \frac{\lambda d_i^2}{\tau}\right) u_i - y_i = 0,$$

leading to

$$u_i = \frac{\tau y_i}{\tau + \lambda d_i^2},$$

for $i = 1, \cdots, n$. Substituting this into $\tau = \|Du\|_2$ implies

$$\sum_{i=1}^{n} \frac{d_i^2 y_i^2}{(\tau + \lambda d_i^2)^2} = 1.$$

We define the function $\psi :]0, +\infty[\to \mathbb{R}$ by

$$\psi(\tau) := \sum_{i=1}^{n} \frac{d_i^2 y_i^2}{(\tau + \lambda d_i^2)^2} - 1,$$

where it is clear that ψ is decreasing

$$\lim_{\tau\to 0} \psi(\tau) = \frac{1}{\lambda^2} \sum_{i=1}^{n} \frac{y_i^2}{d_i^2} - 1 = \frac{1}{\lambda^2}\left(\|D^{-1}y\|_2^2 - \lambda^2\right), \qquad \lim_{\tau\to+\infty} \psi(\tau) = -1.$$

By $\|D^{-1}y\|_2 > \lambda$ and the mean value theorem, we get that there exists $\widehat{\tau} \in]0, +\infty[$ such that $\psi(\widehat{\tau}) = 0$, giving the results. □

W here emphasize that if $D = I$ (I denotes the identity matrix) then the proximity operator for $\phi(\cdot) = \|\cdot\|_2$ is given by

$$\mathrm{prox}_{\lambda\phi}(y) = (1 - \lambda/\|y\|_2)_+ y,$$

see, for example, [52]. If one solves the equation $\psi(\tau) = 0$ approximately, and an initial interval $[a, b]$ is available such that $\psi(a)\psi(b) < 0$, then a solution can be computed to ε-accuracy using the bisection scheme in $O(\log_2((b-a)/\varepsilon))$ iterations, see, for example, [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 one can also use MATLAB's `fzero` function combining the bisection scheme, the inverse quadratic interpolation, and the secant method.

Proposition 53 *Let $\phi(\cdot)=\|\cdot\|_\infty$. Then the proximity operator (35) is given by* $\text{prox}_{\lambda\phi}(y)=0$ *if* $\|y\|_1\leq\lambda$ *and otherwise*

$$(\text{prox}_{\lambda\phi}(y))_i=\begin{cases}0 & \text{if } \|y\|_1\leq\lambda,\\ \text{sign}(y_i)u_\infty & \text{if } \|y\|_1>\lambda,\ i\in\mathcal{I},\\ y_i & \text{if } \|y\|_1>\lambda,\ i\notin\mathcal{I},\end{cases}\tag{50}$$

for $i=1,\cdots,n$, *where*

$$u_\infty:=\frac{1}{\widehat{k}}\left(\sum_{i\in\mathcal{I}}|y_i|-\lambda\right)\tag{51}$$

with

$$\mathcal{I}:=\{l_1,\cdots,l_{\widehat{k}}\}\tag{52}$$

in which $\widehat{k}$ *is the smallest* $k\in\{1,\cdots,n-1\}$ *such that*

$$\frac{1}{\widehat{k}}\left(\sum_{i=1}^{\widehat{k}}v_i-\lambda\right)\geq v_{\widehat{k}+1}\tag{53}$$

where $v_i:=|y_{l_i}|$ *and* $l_1,\cdots,l_n$ *is a permutation of* $1,\cdots,n$ *such that* $v_1\geq v_2\geq\cdots\geq v_n$. *If (53) is not satisfied for* $k\in\{1,\cdots,n-1\}$, *then* $\widehat{k}=n$.

Proof The optimaliy condition (5) shows that $u=\text{prox}_{\lambda\phi}(y)$ if and only if

$$0\in u-y+\lambda\,\partial\|u\|_\infty.\tag{54}$$

We consider two cases: (i) $\|y\|_1\leq\lambda$; (ii) $\|y\|_1>\lambda$.

Case (i). Let $\|y\|_1\leq\lambda$. Then we show $u=0$ satisfies (54). If $u=0$, the subdifferential of ϕ derived in Example 24 is $\partial\phi(0)=\{g\in\mathcal{V}^*\mid\|g\|_1\leq1\}$. By substituting this into (54), we get that $u=0$ is satisfied (54) if $y\in\{g\in\mathcal{V}^*\mid\|g\|_1\leq1\}$ leading to $\text{prox}_{\lambda\phi}(y)=0$. Since the right hand side of (46) is also zero, (46) holds.

Case (ii). Let $\|y\|_1>\lambda$. From Case (i), we have $u\neq0$. We show that, for $i=1,\cdots,n$,

$$u_i=\begin{cases}\text{sign}(y_i)u_\infty & \text{if } i\in\mathcal{I},\\ y_i & \text{otherwise},\end{cases}\tag{55}$$

with $\mathcal{I}$ defined in (52), satisfies (54). Hence, using the subdifferential of ϕ derived in Example 24, there exist coefficients β_j, for $j\in\mathcal{I}$, such that

$$u-y+\lambda\sum_{j\in\mathcal{I}}\beta_j\,\text{sign}(u_j)e_j=0,\tag{56}$$

where

$$\beta_j\geq0\quad j\in\mathcal{I},\qquad\sum_{j\in\mathcal{I}}\beta_j=1.\tag{57}$$

Let u be the vector defined in (55). Let us define

$$\beta_j:=\frac{|y_j|-u_\infty}{\lambda},\tag{58}$$

for $j\in\mathcal{I}=\{l_1,\cdots,l_{\widehat{k}}\}$ with u_∞ defined in (51). We show that the choice (58) satisfies (56) and (57). We first show $u_\infty>0$. It follows from (51) and (53) if $\widehat{k}<n$ and from $\|y\|_1>\lambda$ if $\widehat{k}=n$. By (55) and (56), we have

$$\begin{aligned}u_i-y_i+\lambda\beta_i\,\text{sign}(u_i)&=\text{sign}(y_i)u_\infty-y_i+(|y_i|-u_\infty)\,\text{sign}(\text{sign}(y_i)u_\infty)\\&=\text{sign}(y_i)u_\infty-y_i+(|y_i|-u_\infty)\,\text{sign}(y_i)=0,\end{aligned}$$

for $i \in \mathcal{I}$. For $i \notin \mathcal{I}$, we have $u_i - y_i = 0$. Hence (56) is satisfied componentwise. It remains to show that (57) holds. From (53), we have that $|y_i| \geq u_\infty$, for $i \in \mathcal{I}$. This and (58) imply $\beta_i \geq 0$ for $i \in \mathcal{I}$. From (51), we obtain

$$\sum_{i=1}^{\widehat{k}} \beta_i = \frac{1}{\lambda} \sum_{i=1}^{\widehat{k}} |y_i| - \frac{\widehat{k}}{\lambda} u_\infty = \frac{1}{\lambda} \sum_{i=1}^{\widehat{k}} |y_i| - \frac{1}{\lambda} \left(\sum_{i=1}^{\widehat{k}} |y_i| - \lambda \right) = 1,$$

giving the results. □

Grouped variables typically appear in high-dimensional statistical learning problems. For example, in data mining applications, categorical features are encoded by a set of dummy variables forming a group. Another interesting example is learning sparse additive models in statistical inference, where each component function can be represented using basis expansions and thus can be treated as a group. For such problems, see [39] and references therein, it is more natural to select groups of variables instead of individual ones when a sparse model is preferred.

In the following two results we show how the proximity operator $\text{prox}_{\lambda\phi}(\cdot)$ can be computed for the mixed-norms $\phi(\cdot) = \|\cdot\|_{1,2}$ and $\phi(\cdot) = \|\cdot\|_{1,\infty}$, which are especially important in the context of sparse optimization and sparse recovery with grouped variables.

Proposition 54 *Let $\phi(\cdot) = \|\cdot\|_{1,2}$. Then the proximity operator (35) is given by*

$$(\text{prox}_{\lambda\phi}(y))_{g_i} = \left(1 - \frac{\lambda}{\|y_{g_i}\|_2}\right)_+ y_{g_i}. \tag{59}$$

for $i = 1, \cdots, m$.

Proof Since $u = (u_{g_1}, \cdots, u_{g_m}) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$ and ϕ is separable with respect to the grouped variables, we fix the index $i \in \{1, \cdots, m\}$. The optimaliy condition (5) shows that $u_{g_i} = \text{prox}_{\lambda\phi}(y_{g_i})$ if and only if

$$0 \in u_{g_i} - y_{g_i} + \lambda\, \partial \|u_{g_i}\|_2, \tag{60}$$

for $i = 1, \cdots, m$. We now consider two cases: (i) $\|y_{g_i}\|_2 \leq \lambda$; (ii) $\|y_{g_i}\|_2 > \lambda$.

Case (i). Let $\|y_{g_i}\|_2 \leq \lambda$. Then we show $u_{g_i} = 0$ satisfies (60). If $u_{g_i} = 0$, Proposition 23 implies $\partial\phi(0_{g_i}) = \{g \in \mathbb{R}^{n_i} \mid \|g_{g_i}\|_2 \leq 1\}$. By substituting this into (60), we get that $u_{g_i} = 0$ is satisfied in (60) if $y_{g_i} \in \{g \in \mathbb{R}^{n_i} \mid \|g_{g_i}\|_2 \leq \lambda\}$ leading to $\text{prox}_{\lambda\phi}(y_{g_i}) = 0_{g_i}$. Since the right hand side of (59) is also zero, (59) holds.

Case (ii). Let $\|y_{g_i}\|_2 > \lambda$. Then Case (i) implies that $u_{g_i} \neq 0$. From Proposition 23, we obtain

$$\partial_\phi(u_{g_i}) = \left\{ \frac{u_{g_i}}{\|u_{g_i}\|_2} \right\}, \tag{61}$$

where $i = 1, \cdots, m$ and $\|y_{g_i}\|_2 > \lambda$. Then (60) and (61) imply

$$u_{g_i} - y_{g_i} + \lambda \frac{u_{g_i}}{\|u_{g_i}\|_2} = 0,$$

leading to

$$\left(1 + \frac{\lambda}{\|u_{g_i}\|_2}\right) u_{g_i} = y_{g_i}$$

implying $u_{g_i} = \mu_i y_{g_i}$. By substituting this into the previous identity and solving with respect to μ_i, we get

$$u_{g_i} = \left(1 - \frac{\lambda}{\|y_{g_i}\|_2}\right)_+ y_{g_i},$$

implying the result is valid. □

Proposition 55 *Let $\phi(\cdot)=\|\cdot\|_{1,\infty}$. Then $\mathrm{prox}_{\lambda\phi}(y_{g_i})=0_{g_i}$ if $\|y_{g_i}\|_1\leq\lambda$, for $i=1,\cdots,m$ and otherwise*

$$(\mathrm{prox}_{\lambda\phi}(y_{g_i}))^j_{g_i}=\begin{cases}0 & \text{if } \|y_{g_i}\|_1\leq\lambda,\\ \mathrm{sign}(y^j_{g_i})u^i_\infty & \text{if } \|y_{g_i}\|_1>\lambda,\ j\in\mathcal{I}_{g_i},\\ y^j_{g_i} & \text{if } \|y_{g_i}\|_1>\lambda,\ j\notin\mathcal{I}_{g_i},\end{cases}\tag{62}$$

for $i=1,\cdots,m$, where

$$u^i_\infty:=\frac{1}{\widehat{k}_i}\left(\sum_{j\in\mathcal{I}_{g_i}}|y^j_{g_i}|-\lambda\right)\tag{63}$$

with

$$\mathcal{I}_{g_i}:=\{l^1_{g_i},\cdots,l^{\widehat{k}_i}_{g_i}\}\tag{64}$$

in which $\widehat{k}_i$ is the smallest $k\in\{1,\cdots,n_i-1\}$ such that

$$\frac{1}{\widehat{k}_i}\left(\sum_{j=1}^{\widehat{k}_i}v^j_{g_i}-\lambda\right)\geq v^{\widehat{k}_i+1}_{g_i}\tag{65}$$

where $v^j_{g_i}:=|y_{l^j_{g_i}}|$ and $l^1_{g_i},\cdots,l^{n_i}_{g_i}$ is a permutation of $1,\cdots,n_i$ such that $v^1_{g_i}\geq v^2_{g_i}\geq\cdots\geq v^{n_i}_{g_i}$. If (53) is not satisfied for $k\in\{1,\cdots,n_i-1\}$, then $\widehat{k}_i=n_i$, for $i=1,\cdots,m$.

Proof Since $u=(u_{g_1},\cdots,u_{g_m})\in\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_m}$ and ϕ is separable with respect to the grouped variables, we fix the index $i\in\{1,\cdots,m\}$. The optimaliy condition (5) shows that $u_{g_i}=\mathrm{prox}_{\lambda\phi}(y_{g_i})$ if and only if

$$0\in u_{g_i}-y_{g_i}+\lambda\,\partial\|u_{g_i}\|_\infty.\tag{66}$$

We now consider two cases: (i) $\|y_{g_i}\|_1\leq\lambda$; (ii) $\|y_{g_i}\|_1>\lambda$.

Case (i). Let $\|y_{g_i}\|_1\leq\lambda$. Then we show $u_{g_i}=0$ satisfies (66). If $u_{g_i}=0$, the subdifferential of ϕ derived in Example 24 is $\partial\phi(0_{g_i})=\{g\in\mathbb{R}^{n_i}\mid\|g\|_1\leq1\}$. By substituting this into (66), we get that $u_{g_i}=0$ is satisfied (66) if $y_{g_i}\in\{g\in\mathbb{R}^{n_i}\mid\|g\|_1\leq1\}$ leading to $\mathrm{prox}_{\lambda\phi}(y_{g_i})=0_{g_i}$.

Case (ii). Let $\|y_{g_i}\|_1>\lambda$. From Case (i), we have $u_{g_i}\neq0$. We show that

$$u^j_{g_i}=\begin{cases}\mathrm{sign}(y^j_{g_i})u^i_\infty & \text{if } i\in\mathcal{I}_{g_i},\\ y^j_{g_i} & \text{otherwise,}\end{cases}\tag{67}$$

with $\mathcal{I}_{g_i}$ defined in (64), satisfies (66). Hence, using the subdifferential of ϕ derived in Example 24, there exist coefficients $\beta^j_{g_i}$, for $j\in\mathcal{I}_{g_i}$, such that

$$u_{g_i}-y_{g_i}+\lambda\sum_{j\in\mathcal{I}}\beta^j_{g_i}\,\mathrm{sign}(u^j_{g_i})e_j=0,\tag{68}$$

where

$$\beta^j_{g_i}\geq0\ \ j\in\mathcal{I}_{g_i},\qquad\sum_{j\in\mathcal{I}_{g_i}}\beta^j_{g_i}=1.\tag{69}$$

Let u_{g_i} be the vector defined in (67). Let us define

$$\beta^j_{g_i}:=\frac{|y^j_{g_i}|-u^i_\infty}{\lambda},\tag{70}$$

for $j\in\mathcal{I}_{g_i}=\{l^i_1,\cdots,l^i_{\widehat{k}_i}\}$ with u^i_∞ defined in (63). We show that the choice (70) satisfies (68). We first show $u^i_\infty>0$. It follows from (63) and (65) if $\widehat{k}_i<n$ and from $\|y_{g_i}\|_1>\lambda$ if $\widehat{k}_i=n$. By (67) and (68), we have

$$\begin{aligned}u^j_{g_i}-y^j_{g_i}+\lambda\beta^j_{g_i}\,\mathrm{sign}(u^j_{g_i})&=\mathrm{sign}(y^j_{g_i})u^i_\infty-y^j_{g_i}+\big(|y^j_{g_i}|-u^i_\infty\big)\,\mathrm{sign}(\mathrm{sign}(y^j_{g_i})u^i_\infty)\\&=\mathrm{sign}(y^j_{g_i})u^i_\infty-y^j_{g_i}+\big(|y^j_{g_i}|-u^i_\infty\big)\,\mathrm{sign}(y^j_{g_i})=0,\end{aligned}$$

for $j \in \mathcal{I}_{g_i}$. For $j \notin \mathcal{I}^j_{g_i}$, we have $u^j_{g_i} - y^j_{g_i} = 0$. Hence (68) is satisfied componentwise. It remains to show that (69) holds. From (65), we have that $|y^j_{g_i}| \geq u^i_\infty$, for $j \in \mathcal{I}_{g_i}$. This and (70) imply $\beta^j_{g_i} \geq 0$ for $j \in \mathcal{I}_{g_i}$. From (63), we obtain

$$\sum_{j=1}^{\widehat{k}_i} \beta^j_{g_i} = \frac{1}{\lambda} \sum_{j=1}^{\widehat{k}_i} |y^j_{g_i}| - \frac{\widehat{k}^i}{\lambda} u^i_\infty = \frac{1}{\lambda} \sum_{j=1}^{\widehat{k}_i} |y^j_{g_i}| - \frac{1}{\lambda} \left(\sum_{j=1}^{\widehat{k}_i} |y^j_{g_i}| - \lambda \right) = 1,$$

giving the results. □

5.2 **Constrained examples ($C \neq \mathcal{V}$)**

In this section we consider the subproblem (35) and show how it can be solved for some ϕ and C. More precisely, we solve the minimization problem

$$\begin{array}{ll} \min & \frac{1}{2}\|x - y\|_2^2 + \lambda\varphi(x) \\ \text{s.t.} & x \in C, \end{array}$$

where $\varphi(x)$ is a simple convex function and C is a simple domain. We consider a few examples of this form.

Proposition 56 *Let $\varphi(x) = \|Dx\|_1$ and $C = [\underline{x}, \overline{x}]$. Then the global minimizer of the subproblem (35) is given by*

$$(\mathrm{prox}^C_{\lambda\phi}(y))_i = \begin{cases} \underline{x}_i & \text{if } \omega(y,\lambda) > 0,\ \underline{x}_i - y_i + \lambda|d_i|\ \mathrm{sign}(\underline{x}_i) \geq 0, \\ \overline{x}_i & \text{if } \omega(y,\lambda) > 0,\ \overline{x}_i - y_i + \lambda|d_i|\ \mathrm{sign}(\overline{x}_i) \leq 0, \\ y_i - \lambda|d_i| & \text{if } \omega(y,\lambda) > 0,\ y_i > \lambda|d_i|, \\ y_i + \lambda|d_i| & \text{if } \omega(y,\lambda) > 0,\ y_i < -\lambda|d_i|, \\ 0 & \text{otherwise.} \end{cases} \tag{71}$$

for $i = 1, \cdots, n$, where

$$\omega(y,\lambda) := \sum_{y_i + \lambda|d_i| < 0} (y_i + \lambda|d_i|)\underline{x} + \sum_{y_i + \lambda|d_i| > 0} (y_i + \lambda|d_i|)\overline{x}. \tag{72}$$

Proof The optimaliy condition (4) shows that $u = \mathrm{prox}^C_{\lambda\phi}(y)$ if and only if

$$0 \in u - y + \lambda\ \partial\|Du\|_1 + N_C(u), \tag{73}$$

where $N_C(u)$ is the normal cone of C at u defined in (2). Let us define $\omega(y,\lambda)$ by (72). We now consider two cases: (i) $\omega(y,\lambda) \leq 0$; (ii) $\omega(y,\lambda) > 0$.

Case (i). Let $\omega(y,\lambda) \leq 0$. Then we show that $u = 0$ satisfies (73). If $u = 0$, Proposition 23 implies $\partial\phi(0) = \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_\infty \leq 1\}$. This and (73) leads to

$$y - \lambda\ \partial\phi(0) \in N_C(0), \tag{74}$$

where

$$N_C(0) = \{p \in \mathcal{V} \mid \forall z \in [\underline{x}, \overline{x}], \langle p, z\rangle \leq 0\} = \left\{ p \in \mathcal{V} \mid \sum_{p_i < 0} p_i\underline{x} + \sum_{p_i > 0} p_i\overline{x} \leq 0 \right\}.$$

This and (74) leads to

$$\begin{array}{ll} \max & \sum_{p_i<0} p_i\underline{x} + \sum_{p_i>0} p_i\overline{x} \\ \text{s.t.} & p \in \{y - g \mid \|D^{-1}g\|_\infty \leq \lambda,\ g \in \mathbb{R}^n\}. \end{array}$$

It is clear that $p = y - \lambda|D\mathbf{1}|$ is the solution of this problem, where $\mathbf{1}$ is the vector of all ones. Then if $\omega(y,\lambda) \leq 0$, then $u = 0$ is satisfied (73) leading to $\mathrm{prox}_{\lambda\phi}(y) = 0$.

Case (ii). Let $\omega(y,\lambda) > \lambda$. Then Case (i) implies $u \neq 0$. Proposition 23 yields

$$\partial\phi(u) = \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_\infty = 1,\ \ \langle g, u\rangle = \|Du\|_1\}$$

leading to

$$\sum_{i=1}^{n}(g_iu_i - |d_iu_i|) = 0.$$

By induction on nonzero elements of u, we get $g_iu_i = |d_iu_i|$, for $i = 1, \cdots, n$. This implies that $g_i = |d_i|\ \text{sign}(u_i)$ if $u_i \neq 0$. This and the definition of $N_C(u)$ imply

$$u_i - y_i + \lambda(\partial\|Du\|_1)_i \begin{cases} \geq 0 & \text{if } u_i = \underline{x}_i, \\ \leq 0 & \text{if } u_i = \overline{x}_i, \\ = 0 & \text{if } \underline{x}_i < u_i < \overline{x}_i, \end{cases}$$

for $i = 1, \cdots, n$, and equivalently for $u \neq 0$, we get

$$u_i - y_i + \lambda|d_i|\ \text{sign}(u_i) \begin{cases} \geq 0 & \text{if } u_i = \underline{x}_i, \\ \leq 0 & \text{if } u_i = \overline{x}_i, \\ = 0 & \text{if } \underline{x}_i < u_i < \overline{x}_i, \end{cases} \tag{75}$$

for $i = 1, \cdots, n$. If $u_i = \underline{x}_i$, substituting $u_i = \underline{x}_i$ in (75) implies $\underline{x}_i - y_i + \lambda|d_i|\ \text{sign}(\underline{x}_i) \geq 0$. If $u_i = \overline{x}_i$, substituting $u_i = \overline{x}_i$ in (75) implies $\overline{x}_i - y_i + \lambda|d_i|\ \text{sign}(\overline{x}_i) \leq 0$. If $\underline{x}_i < u_i < \overline{x}_i$, there are three possibilities: (a) $u_i > 0$; (b) $u_i < 0$; (c) $u_i = 0$. In Case (a), the fact $\text{sign}(u_i) = 1$ and (75) imply $u_i = y_i - \lambda|d_i| > 0$. In Case (b), the condition $\text{sign}(u_i) = -1$ and (75) imply $u_i = y_i + \lambda|d_i| < 0$. In Case (c), we get $u_i = 0$. This completes the proof. □

Proposition 57 *Let $\varphi(x) = \frac{1}{2}\|x\|_2^2$ and $C = [\underline{x}, \overline{x}]$. Then the global minimizer of the subproblem (35) is given by*

$$(\text{prox}_{\lambda\phi}^{C}(y))_i = \begin{cases} \underline{x}_i & \text{if } (1+\lambda)\underline{x}_i \geq y_i, \\ \overline{x}_i & \text{if } (1+\lambda)\overline{x}_i \leq y_i, \\ y_i/(1+\lambda) & \text{if } \underline{x}_i < y_i/(1+\lambda) < \overline{x}_i, \end{cases} \tag{76}$$

for $i = 1, \cdots, n$.

Proof The function $\varphi(x) = \frac{1}{2}\|x\|_2^2$ is differentiable, i.e.,

$$\partial\varphi(x) = \{x\}.$$

This and the definition of $N_C(u)$ imply

$$u_i - y_i + \lambda u_i \begin{cases} \geq 0 & \text{if } u_i = \underline{x}_i, \\ \leq 0 & \text{if } u_i = \overline{x}_i, \\ = 0 & \text{if } \underline{x}_i < u_i < \overline{x}_i, \end{cases} \tag{77}$$

for $i = 1, \cdots, n$. If $u_i = \underline{x}_i$, substituting $u_i = \underline{x}_i$ in (77) implies $(1+\lambda)\underline{x}_i \geq y_i$. If $u_i = \overline{x}_i$, substituting $u_i = \overline{x}_i$ in (77) implies $(1+\lambda)\overline{x}_i \leq y_i$. If $\underline{x}_i < u_i < \overline{x}_i$, then $u_i = y_i/(1+\lambda)$. This gives the result. □

Proposition 58 *Let $\varphi(x) = \frac{1}{2}\lambda_1\|x\|_2^2 + \lambda_2\|Dx\|_1$ and $C = [\underline{x}, \overline{x}]$. Then the global minimizer of the subproblem (35) is determined by*

$$(\text{prox}_{\lambda\phi}^{C}(y))_i = \begin{cases} \underline{x}_i & \text{if } \omega(y,\lambda) > 0,\ (1+\lambda_1)\underline{x}_i - y_i + \lambda_2|d_i|\ \text{sign}(\underline{x}_i) \geq 0, \\ \overline{x}_i & \text{if } \omega(y,\lambda) > 0,\ (1+\lambda_1)\overline{x}_i - y_i + \lambda_2|d_i|\ \text{sign}(\overline{x}_i) \leq 0, \\ 1/(1+\lambda_1)(y_i - \lambda_2|d_i|) & \text{if } \omega(y,\lambda) > 0,\ y_i > \lambda_2|d_i|, \\ 1/(1+\lambda_1)(y_i + \lambda_2|d_i|) & \text{if } \omega(y,\lambda) > 0,\ y_i < -\lambda_2|d_i|, \\ 0 & \text{otherwise}, \end{cases} \tag{78}$$

for $i = 1, \cdots, n$, where $\omega(y,\lambda)$ is defined by (72).

Proof Since $\mathcal{V}$ is finite-dimensional and $\text{dom}\left(\frac{1}{2}\lambda_1\|x\|_2^2\right) \cap \text{dom}\lambda_2\|Dx\|_1 \neq \emptyset$, we get

$$\partial\left(\frac{1}{2}\lambda_1\|x\|_2^2 + \lambda_2\|Dx\|_1\right) = \lambda_1\partial\left(\frac{1}{2}\|x\|_2^2\right) + \lambda_2\partial\left(\|Dx\|_1\right). \tag{79}$$

The optimaliy condition (4) shows that $u = \text{prox}_{\lambda\phi}^{C}(y)$ if and only if

$$0 \in u - y + \lambda_1 u + \lambda_2\ \partial\|Du\|_1 + N_C(u), \tag{80}$$

where $N_C(u)$ is the normal cone of C at u defined in (2). Let us consider $\omega(y,\lambda)$ defined by (72). We now consider two cases: (i) $\omega(y,\lambda) \le 0$; (ii) $\omega(y,\lambda) > 0$.

Case (i). Let $\omega(y,\lambda) \le 0$. Then we show that $u = 0$ satisfies (80). If $u = 0$, Proposition 23 implies $\partial\phi(0) = \{g \in \mathcal{V}^* \mid \|D^{-1}g\|_\infty \le 1\}$. This and (80) leads to

$$y - \lambda\, \partial\phi(0) \in N_C(0), \tag{81}$$

where

$$N_C(0) = \{p \in \mathcal{V} \mid \forall z \in [\underline{x}, \overline{x}], \langle p, z\rangle \le 0\} = \left\{ p \in \mathcal{V} \mid \sum_{p_i<0} p_i\underline{x} + \sum_{p_i>0} p_i\overline{x} \le 0 \right\}.$$

This and (81) leads to

$$\begin{array}{ll} \max & \displaystyle\sum_{p_i<0} p_i\underline{x} + \sum_{p_i>0} p_i\overline{x} \\ \text{s.t.} & p \in \{y - g \mid \|D^{-1}g\|_\infty \le \lambda,\ g \in \mathbb{R}^n\}. \end{array}$$

It is clear that $p = y - \lambda|D\mathbf{1}|$ is the solution of this problem, where $\mathbf{1}$ is the vector of all one's. Then if $\omega(y,\lambda) \le 0$, then $u = 0$ is satisfied (80) leading to $\mathrm{prox}_{\lambda\phi}(y) = 0$.

Case (ii). Let $\omega(y,\lambda) > \lambda$. Then Case (i) implies $u \ne 0$. From (79) and the definition of $N_C(u)$, we obtain

$$u_i - y_i + \lambda_1 u_i + \lambda_2 \partial|d_i u_i| \begin{cases} \ge 0 & \text{if } u_i = \underline{x}_i, \\ \le 0 & \text{if } u_i = \overline{x}_i, \\ = 0 & \text{if } \underline{x}_i < u_i < \overline{x}_i, \end{cases}$$

for $i = 1, \cdots, n$. This leads to

$$(1+\lambda_1)u_i - y_i + \lambda_2|d_i|\ \mathrm{sign}(u_i) \begin{cases} \ge 0 & \text{if } u_i = \underline{x}_i, \\ \le 0 & \text{if } u_i = \overline{x}_i, \\ = 0 & \text{if } \underline{x}_i < u_i < \overline{x}_i, \end{cases} \tag{82}$$

for $i = 1, \cdots, n$. If $u_i = \underline{x}_i$, substituting $u_i = \underline{x}_i$ in (75) implies $(1+\lambda_1)\underline{x}_i - y_i + \lambda_2|d_i|\ \mathrm{sign}(\underline{x}_i) \ge 0$. If $u_i = \overline{x}_i$, substituting $u_i = \overline{x}_i$ in (75) implies $(1+\lambda_1)\overline{x}_i - y_i + \lambda_2|d_i|\ \mathrm{sign}(\overline{x}_i) \le 0$. If $\overline{x}_i < u_i < \overline{x}_i$, there are three possibilities: (i) $u_i > 0$; (ii) $u_i < 0$; (iii) $u_i = 0$. In Case (i), the fact $\mathrm{sign}(u_i) = 1$ and (75) imply $u_i = 1/(1+\lambda_1)(y_i - \lambda_2|d_i|) > 0$. In Case (ii), the condition $\mathrm{sign}(u_i) = -1$ and (75) imply $u_i = 1/(1+\lambda_1)(y_i + \lambda_2|d_i|) < 0$. In Case (iii), we get $u_i = 0$ giving the result. □

Let $x \ge 0$ be nonnegativity constraints. These constraints are important in many applications, especially if x describes physical quantities, see, for example, [25, 32, 33]. Since nonnegativity constraints can be regarded as especial case of bound-constrained domain, Propositions 56, 57, and 58 can be used to derive the results for nonnegativity constraints (see Table 1).

6 Numerical experiments

In this section we report some numerical results to compare the performance of the new setup of OSGA (OSGA-O) with some state-of-the-art solvers. In our comparison with OSGA, and OSGA-O, we consider PGA (proximal gradient algorithm [52]), NSDSG (nonsummable diminishing subgradient algorithm [14]), FISTA (Beck and Tebolle's fast proximal gradient algorithm [10]), NESCO (Nesterov's composite optimal algorithm [47]), NESUN (Nesterov's universal gradient algorithm [48]), NES83 (Nesterov's 1983 optimal algorithm [42]), NESCS (Nesterov's constant step optimal algorithm [41]), and NES05 (Nesterov's 2005 optimal algorithm [43]). Indeed, we adapt NES83, NESCS, and NES05 by passing a subgradient in the place of the gradient to be able to apply them to nonsmooth problems (see Ahookhosh [1]). The codes of these algorithms are written in MATLAB, where we use the parameters proposed in the associated papers.

We divide the solvers into two classes: (i) proximal-based methods (PGA, FISTA, NESCO, and NESUN) that can directly applied to nonsmooth problems; (ii) Subgradient-based methds (NSDSG,

NES83, NESCS, and NES05) in which the nonsmooth first-order oracle required, where NES83, NESCS, and NES05 are adapted to take a subgradient in the place of the gradient. We set

$$\widehat{L} := \max_{1\leq i\leq n} \|a_i\|^2,$$

where a_i, for $i = 1, 2, \cdots, n$, is the i-th column of A. In the implementation, NESCS, NES05, PGA, and FISTA use $L = 10^4\widehat{L}$, and NSDSG employs $\alpha_0 = 10^{-7}$. OSGA and OSGA-O use the parameters

$$\delta = 0.9, \quad \alpha_{max} = 0.7, \quad \kappa = \kappa' = 0.5, \quad f_{\text{target}} = -\infty,$$

and the prox-function (27) 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. To solve the nonlinear system of equations (36), we first consider the nonlinear least-squares problem (44) and solve it by the MATLAB internal function `fminsearch`[1], which is a derivative-free solver handling both smooth and nonsmooth problems.

We consider solving an underdetermined system

$$Ax = y,$$

where A is a $m \times n$ matrix ($m \leq n$) and y is a m-vector. Underdetermined system of linear equations is frequently appeared in many applications of linear inverse problem such as those in the fields signal and image processing, geophysics, economics, machine learning, and statistics. The objective is to recover x from the observed vector y, and matrix A by some optimization models. Due to the ill-conditioned feature of the problem, the most popular optimization models are (21), (22), and (23), where (21) is smooth and (22) and (23) are nonsmooth. In Subsection 6.1 we report numerical results with the ℓ_1 minimization (22) and in Subsection 7 give results regarding the Elastic Net minimization problem (23). Let us set $m = 5000$ and $n = 10000$. The data A, y, and x_0 for problem (22) is randomly generated by

$$\texttt{A} = \texttt{rand}(\texttt{m},\texttt{n}), \quad \texttt{y} = \texttt{rand}(\texttt{1},\texttt{m}), \quad \texttt{x}_0 = \texttt{rand}(\texttt{1},\texttt{n}),$$

where `rand` generates uniformly distributed random numbers between 0 and 1. In our implementation we stop the algorithms after 30 seconds of the running time.

6.1 ℓ_1 minimization

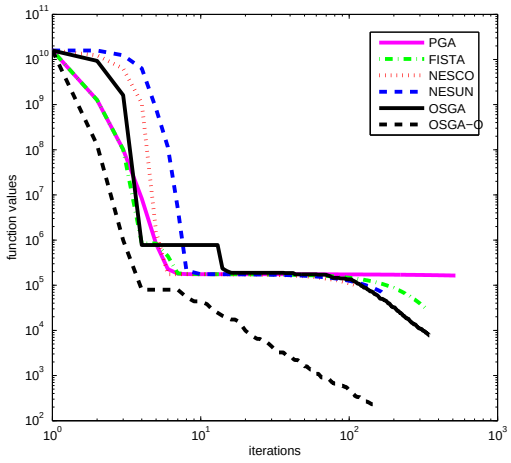
We here consider the ℓ_1 minimization problem (22), reformulate it as a minimization problem of the form (19) with the objective and the constraint described in (25), and solve the reformulated problem by OSGA-O. We give some numerical results and a comparison among OSGA-O,OSGA and some state-of-the-art solvers.

We consider 6 different regularization parameters and report numerical results for PGA, FISTA, NESCO, NESUN, OSGA, and OSGA-O in Table 2 and for NSDSG, NES83, NESCS, NES05, OSGA, and OSGA-O in Table 3. We illustrate function values versus iterations for both classes of solvers with the regularization parameters $\lambda = 1$, $\lambda = 10^{-1}$, and $\lambda = 10^{-1}$ in Figure 1.

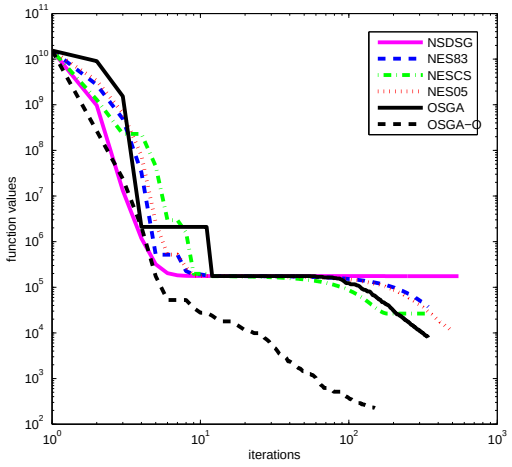
Table 2: Function values of PGA, FISTA, NESCO, NESUN, OSGA, and OSGA-O for solving the ℓ_1 minimization problem (22) with several regularization parameters.

Regularization parameter	**PGA**	**FISTA**	**NESCO**	**NESUN**	**OSGA**	**OSGA-O**
$\lambda = 1$	163813.29	29254.36	95648.33	65551.18	7705.11	224.05
$\lambda = 10^{-1}$	161882.90	28826.02	73503.70	52081.40	5654.31	209.12
$\lambda = 10^{-2}$	160173.56	14223.04	69294.26	62919.36	3668.93	1134.07
$\lambda = 10^{-3}$	153709.75	16835.48	88402.78	60112.50	6065.37	418.97
$\lambda = 10^{-4}$	158812.94	12630.09	74889.01	55774.92	76092.84	364.55
$\lambda = 10^{-5}$	155573.77	19060.71	60964.63	64549.46	8454.57	418.63

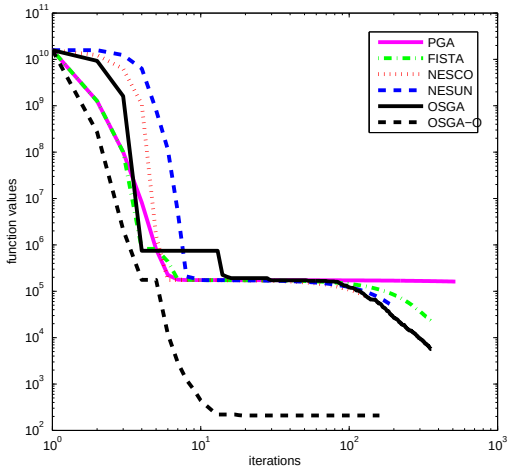
[1] The function `fminsearch` is a derivative-free solver for unconstrained optimization problems based on Nelder-Mead simplex direct search method performing well for two-dimensional problems, see, for example, [27, 34]



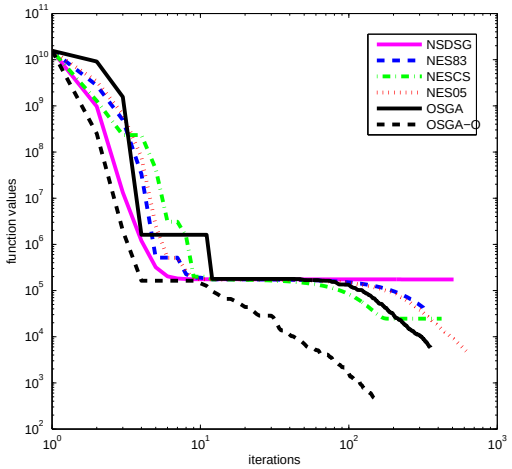
(a) $\lambda = 1$



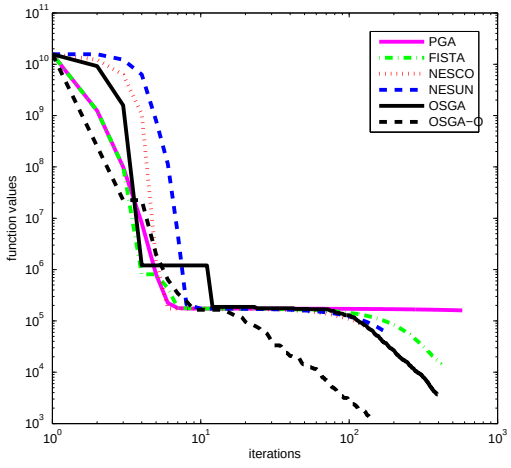
(b) $\lambda = 1$



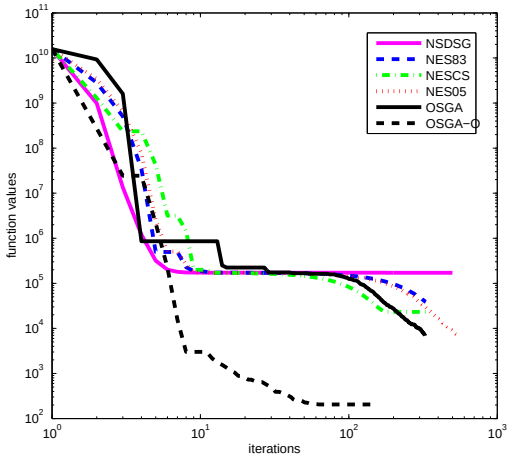
(c) $\lambda = 10^{-1}$



(d) $\lambda = 10^{-1}$



(e) $\lambda = 10^{-2}$



(f) $\lambda = 10^{-2}$

Fig. 1: A comparison among first-order methods for solving ℓ_1 minimization problem: Subfigures (a), (c), and (e) illustrate a comparison of function values versus iterations among PGA, FISTA, NESCO, NESUN, OSGA, and OSGA-O for $\lambda = 1$, $\lambda = 10^{-1}$,$\lambda = 10^{-2}$, respectively; Subfigures (b), (d), and (f) illustrate a comparison of function values versus iterations among NSDSG, NES83, NESCS, NES05, OSGA, and OSGA-O for $\lambda = 1$, $\lambda = 10^{-1}$,$\lambda = 10^{-2}$, respectively. The algorithms stopped after 30 seconds.

Table 3: Function values of NSDSG, NES83, NESCS, NES05, OSGA, and OSGA-O for solving the ℓ_1 minimization problem (22) with several regularization parameters.

Regularization parameter	**NSDSG**	**NES83**	**NESCS**	**NES05**	**OSGA**	**OSGA-O**
$\lambda = 1$	174672.58	37465.49	26510.61	10918.82	789485	224.47
$\lambda = 10^{-1}$	173411.33	34340.16	42601.60	4893.27	583.5.84	444.79
$\lambda = 10^{-2}$	170935.10	38405.66	23236.88	6730.46	6881.61	204.62
$\lambda = 10^{-3}$	170272.21	35226.85	23708.27	9457.75	6295.59	203.87
$\lambda = 10^{-4}$	172775.42	35738.34	24127.66	87641.85	69687.72	452.96
$\lambda = 10^{-5}$	171480.55	25210.72	23291.35	6074.28	6750.34	204.52

The results of Tables 2 and 3 show that OSGA-O obtains the best function values for the ℓ_1 minimization problem. From Figure 1, it can be seen that the worst results are obtained by NSDSG and PGA; FISTA, NESCO, NESUN, NES83, NESCS, NES05 and OSGA are comparable to some extent; OSGA-O is significantly superior to the other methods.

7 Elastic Net minimization

We consider the Elastic Net minimization problem (23), reformulate it as a minimization problem of the form (19) with the objective and the constraint given in (26), and solve the reformulated problem by OSGA-O. We then give some numerical results and a comparison among OSGA-O,OSGA and some state-of-the-art solvers.

We consider 6 different regularization parameters and report numerical results for PGA, FISTA, NESCO, NESUN, OSGA, and OSGA-O in in Table 4 and for NSDSG, NES83, NESCS, NES05, OSGA, and OSGA-O in Table 5. We then illustrate function values versus iterations for both classes of solvers with $\lambda = 1$, $\lambda = 10^{-1}$, and $\lambda = 10^{-1}$ in Figure 2.

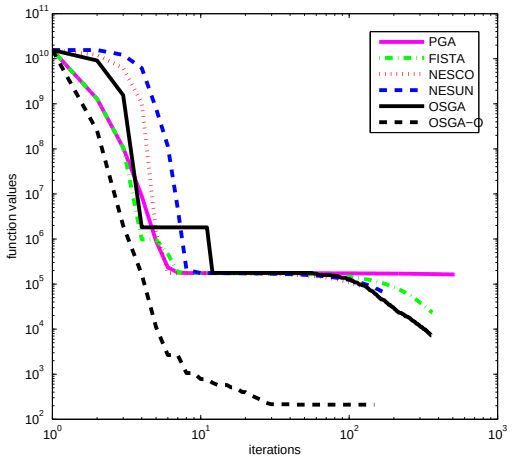
Table 4: Function values of PGA, FISTA, NESCO, NESUN, OSGA, and OSGA-O for solving Elastic Net problem with several regularization parameters.

Regularization parameter	**PGA**	**FISTA**	**NESCO**	**NESUN**	**OSGA**	**OSGA-O**
$\lambda = 1$	163170.01	23221.75	81268.32	61438.91	7242.91	209.89
$\lambda = 10^{-1}$	155821.50	18934.35	83111.27	57095.06	5439.81	213.44
$\lambda = 10^{-2}$	160336.99	23076.71	80901.20	50686.03	4891.32	1866.06
$\lambda = 10^{-3}$	159181.19	26166.87	92879.45	56724.63	8078.09	205.84
$\lambda = 10^{-4}$	163193.78	27845.31	855474.09	62852.97	7759.84	208.86
$\lambda = 10^{-5}$	163771.85	26335.87	90349.26	67991.56	7908.11	210.66

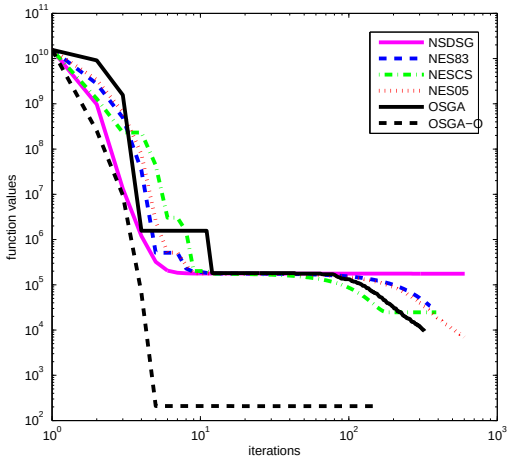
Table 5: Function values of NSDSG, NES83, NESCS, NES05, OSGA, and OSGA-O for solving Elastic Net problem with several regularization parameters.

Regularization parameter	**NSDSG**	**NES83**	**NESCS**	**NES05**	**OSGA**	**OSGA-O**
$\lambda = 1$	175349.77	33979.90	24608.58	69894.52	9476.02	207.28
$\lambda = 10^{-1}$	171367.80	28446.35	24270.92	5675.27	7624.94	201.33
$\lambda = 10^{-2}$	174310.41	22264.53	24207.23	9436.33	6944.39	2608.05
$\lambda = 10^{-3}$	171458.76	31253.67	24127.23	9476.60	8160.70	209.95
$\lambda = 10^{-4}$	164912.11	21584.42	23423.70	9174.28	4433.32	445.25
$\lambda = 10^{-5}$	180334.49	31765.64	24834.49	7656.03	6871.54	204.15

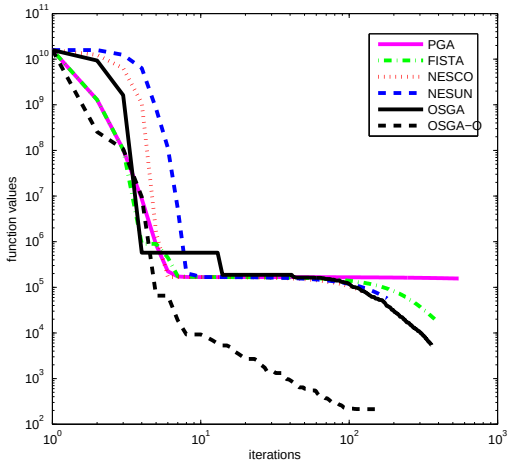
The results of Tables 4 and 5 show that the best function values are obtained by OSGA-O. In Figure 2, we can see that the worst results are obtained by NSDSG and PGA; FISTA, NESCO, NESUN, NES83, NESCS, NES05 and OSGA behave competitively; again, OSGA-O outperforms the other methods considerably.



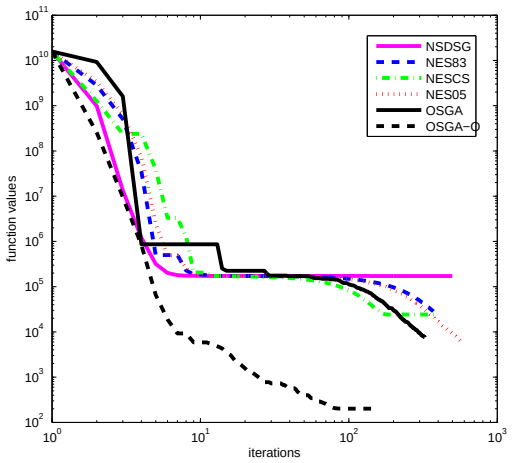
(a) $\lambda = 1$



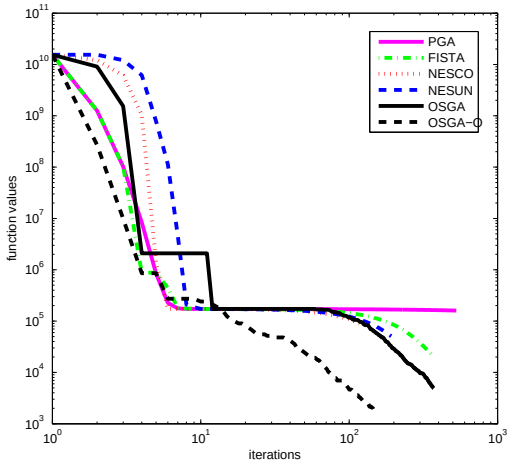
(b) $\lambda = 1$



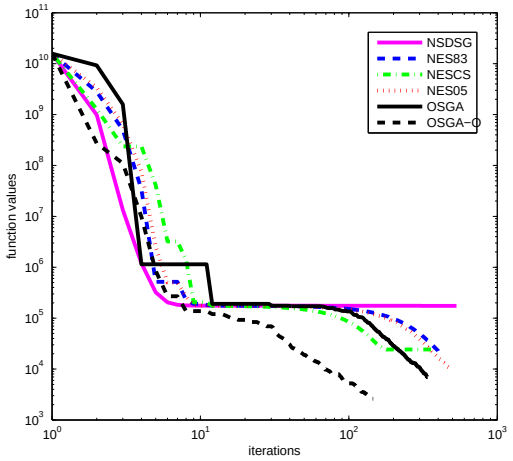
(c) $\lambda = 10^{-1}$



(d) $\lambda = 10^{-1}$



(e) $\lambda = 10^{-2}$



(f) $\lambda = 10^{-2}$

Fig. 2: A comparison among first-order methods for solving Elastic Net minimization problem: Subfigures (a), (c), and (e) illustrate a comparison of function values versus iterations among PGA, FISTA, NESCO, NESUN, OSGA, and OSGA-O for $\lambda = 1$, $\lambda = 10^{-1}$,$\lambda = 10^{-2}$, respectively; Subfigures (b), (d), and (f) illustrate a comparison of function values versus iterations among NSDSG, NES83, NESCS, NES05, OSGA, and OSGA-O for $\lambda = 1$, $\lambda = 10^{-1}$,$\lambda = 10^{-2}$, respectively. The algorithms stopped after 30 seconds.

8 Conclusions

This paper discusses the solution of structured nonsmooth convex optimization with the complexity $O(\varepsilon^{-1/2})$, which is optimal for smooth problems with Lipschitz continuous gradients. If the nonsmoothness of the problem is manifested in a structured way, we reformulate the problem in a form that the objective is smooth with Lipschitz continuous gradients in the price of adding a functional constraint to the feasible domain. Afterwards, a new setup of the optimal subgradient algorithm (OSGA-O) is developed to solve the problem with the complexity $O(\varepsilon^{-1/2})$. It is proved that OSGA-O's auxiliary subproblem is equivalent to a proximal-like problem, which is well-studied due to its appearance in Nesterov-type optimal methods for composite minimization. We either give explicit formulas or simple iterative schemes for solving several proximal-like problems appearing in the applications. We finally give numerical results indicating a good behavior of OSGA.

Acknowledgement. Thanks to Stephen M. Robinson and Defeng Sun for their comments about solving nonsmooth equations.

References

1. Ahookhosh, M.: Optimal subgradient algorithms with application to large-scale linear inverse problems, submitted (2014), `http://arxiv.org/abs/1402.7291`. [2, 6, 21]
2. Ahookhosh, M., Amini, K., Kimiaei, M.: A globally convergent trust-region method for large-scale symmetric nonlinear systems, Numerical Functional Analysis and Optimization, **36**, 830–855 (2015) [11]
3. Ahookhosh, M., Neumaier, A.: High-dimensional convex optimization via optimal affine subgradient algorithms, in ROKS workshop, 83-84 (2013) [6]
4. Ahookhosh, M., Neumaier, A.: An optimal subgradient algorithm with subspace search for costly convex optimization problems, Submitted, (2015) `http://www.optimization-online.org/DB_FILE/2015/04/4852.pdf` [6]
5. Ahookhosh, M., Neumaier, A.: An optimal subgradient algorithms for large-scale bound-constrained convex optimization, Submitted, (2015). `http://arxiv.org/abs/1501.01497` [2, 6]
6. Ahookhosh, M., Neumaier, A.: An optimal subgradient algorithms for large-scale convex optimization in simple domains, Submitted, (2015). `http://arxiv.org/abs/1501.01451` [2, 6]
7. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization, SIAM Journal on Optimization, **16**, 697–725 (2006) [2]
8. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, CMS Books Math., Springer-Verlag, New York, (2011) [3]
9. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization, Operations Research Letters, **31**(3), 167–175 (2003) [1]
10. Beck, A., Teboulle, M.: Smoothing and first order methods: A unified framework, SIAM Journal on Optimization, **22**, 557–580 (2012) [2, 21]
11. Beck, A., Ben-Tal, A., Guttmann-Beck, N., Tetruashvili, L.: The CoMirror algorithm for solving nonsmooth constrained convex problems, Operations Research Letters, **38**(6), 493–498 (2010) [1]
12. Boţ, R.I., Hendrich, C.: A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems, Computational Optimization and Applications, **54**(2), 239–262 (2013) [2]
13. Boţ, R.I., Hendrich, C.: On the acceleration of the double smoothing technique for unconstrained convex optimization problems, Optimization, **64**(2), 265–288 (2015) [2]
14. Boyd, S., Xiao, L., Mutapcic, A.: Subgradient methods, Notes for EE392o, Stanford University, (2003), `http://www.stanford.edu/class/ee392o/subgrad_method.pdf`. [21]
15. Byrd, R.H., Lu, P., Nocedal, J. and Zhu,C.: A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing, **16**, 1190–1208 (1995) [11]
16. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems, SIAM Journal on Optimization, **24**(4), 1779–1814 (2014) [2]
17. Chen, Y., Lan, G., Ouyang, Y.: Accelerated scheme for a class of variational inequalities, (2014) `http://arxiv.org/pdf/1403.4164v1.pdf` [2]
18. Chen, Y., Lan, G., Ouyang, Y.: An accelerated linearized alternating direction method of multipliers (2014) `http://arxiv.org/pdf/1401.6607v3.pdf` [2]
19. Chen, Y., Lan, G., Ouyang, Y., Zhang, W.: Fast bundle-level type methods for unconstrained and ball-constrained convex optimization, (2014) `http://arxiv.org/pdf/1412.2128v1.pdf` [2]
20. Combettes, P., Pesquet, J.C.: Proximal splitting methods in signal processing, Fixed-Point Algorithms for Inverse Problems in Science and Engineering, 185–212 (2011) [14]
21. Conn, A.R., Gould, N.I.M., Toint, Ph.L.: Global convergence of a class of trust region algorithms for optimization with simple bounds, SIAM Journal on Numerical Analysis, **25**, 433–460 (1988) [11]
22. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery, Communications on Pure and Applied Mathematics, 63(1) (2010), 1–38. [13]
23. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle, Mathematical Programming, **146**, 37–75 (2013) [2]
24. Devolder, O., Glineur, F., Nesterov, Y.: Double smoothing technique for large-scale linearly constrained convex optimization, SIAM Journal on Optimization, **22**(2), 702–727 (2012) [2]

25. Esser, E., Lou, Y., Xin, J.: A method for finding structured sparse solutions to nonnegative least squares problems with applications, SIAM Journal on Imaging Science, **6**(4), 2010–2046 (2013) [21]
26. Friedlander, A., Martínez, J.M., Santos, S.A.: A new trust region algorithm for bound constrained minimization, Applied Mathematics and Optimization, **30**, 235–266 (1994) [11]
27. Hansen, N., Auger, A., Ros, R., Finck, S., Posik, P.: Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009, in Proc. Workshop GECCO, 1689–1696 (2010) [22]
28. Gonzaga, C.C., Karas, E.W.: Fine tuning Nesterov's steepest descent algorithm for differentiable convex programming, Mathematical Programming, **138**, 141–166 (2013) [2]
29. Gonzaga, C.C., Karas, E.W., Rossetto, D.R.: An optimal algorithm for constrained differentiable convex optimization, SIAM Journal on Optimization, **23**(4), 1939–1955 (2013) [2]
30. Juditsky, A., and Nesterov, Y.: Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization, Stochastic Systems, 4(1) (2014), 44–80. [1]
31. Kanzow, C., Yamashita, N., Fukushima, M.: Levenberg-Marquardt methods with strong local convergence properties for solving equations with convex constraints, Journal of Computational and Applied Mathematics, **172**, 375–397 (2004) [11]
32. Kaufman, L., Neumaier, A.: PET regularization by envelope guided conjugate gradients, IEEE Transactions on Medical Imaging, **15**, 385–389 (1996) [21]
33. Kaufman, L., Neumaier, A.: Regularization of ill-posed problems by envelope guided conjugate gradients, Journal of Computational and Graphical Statistics, **6**(4), 451–463 (1997) [21]
34. Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: Convergence properties of the Nelder-Mead simplex method in low dimensions, SIAM Journal on Optimization, 9, 112–147 (1998) [22]
35. Lan, G.: Bundle-level type methods uniformly optimal for smooth and non-smooth convex optimization, Mathematical Programming, (2013), DOI 10.1007/s10107-013-0737-x. [2]
36. Lan, G., Lu, Z., Monteiro, R.D.C.: Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming, Mathematical Programming, **126**, 1–29 (2011) [2]
37. Li, D.H., Yamashita, N., Fukushima, M.: Nonsmooth equation based bfgs method for solving KKT systems in mathematical programming, Journal of Optimization Theory and Applications, **109** (1), 123–167 (2001) [11]
38. Lin, C.J., Moré, J.J.: Newton's method for large-scale bound constrained problems, SIAM Journal on Optimization, **9**, 1100–1127 (1999) [11]
39. H. Liu, J. Zhang, X. Jiang, and J. Liu: The group Dantzig selector, Journal of Machine Learning Research - Proceedings Track, 9, 461–468 (2010) [17]
40. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization, Wiley, New York (1983) [1, 6]
41. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, Kluwer, Dordrecht, (2004) [1, 21]
42. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Doklady AN SSSR (In Russian), 269 (1983), 543–547. English translation: Soviet Math. Dokl., **27**, 372–376 (1983) [1, 21]
43. Nesterov, Y.: Smooth minimization of non-smooth functions, Mathematical Programming, **103**, 127–152 (2005) [2, 21]
44. Nesterov, Y.: Excessive gap technique in nonsmooth convex minimization, SIAM Journal on Optimization, **16**, 235–249 (2005) [2]
45. Nesterov, Y.: Barrier subgradient method, Mathematical Programming, **127**, 31–56 (2011) [1]
46. Nesterov, Y.: Primal-dual subgradient methods for convex problems, Mathematical Programming, **120**, 221–259 (2006) [1]
47. Nesterov, Y.: Gradient methods for minimizing composite objective function, Mathematical Programming, **140**, 125–161 (2013) [2, 7, 21]
48. Nesterov, Y.: Universal gradient methods for convex optimization problems, Mathematical Programming, DOI 10.1007/s10107-014-0790-0, (2014) [2, 7, 21]
49. Neumaier, A.: OSGA: a fast subgradient algorithm with optimal complexity, Mathematical Programming, DOI 10.1007/s10107-015-0911-4, (2015) [2, 5, 6, 9, 12]
50. Neumaier, A.: Introduction to Numerical Analysis, Cambridge University Press, Cambridge, (2001). [15]
51. Pang, J.S., Qi, L.: Nonsmooth equations: motivation and algorithms, SIAM Journal on Optimization, 3, 443–465 (1993) [11]
52. Parikh, N., Boyd, S.: Proximal Algorithms, Foundations and Trends in Optimization, **1**(3), 123–231 (2013) [14, 15, 21]
53. Polyak, B.: Introduction to Optimization, Optimization Software, Inc., Publications Division, NewYork, (1987) [1]
54. Potra, F.A., Qi, L., Sun, D.: Secant methods for semismooth equations, Numerische Mathematik, **80**(2), 305–324 (1998) [11]
55. Qi, L.: Trust region algorithms for solving nonsmooth equations, SIAM Journal on Optimization, **5**, 219–230 (1995) [11]
56. Qi, L., Sun, D.: A survey of some nonsmooth equations and smoothing Newton methods, Progress in Optimization, **30**, 121–146 (1999) [11]
57. Rauhut, H., Ward, R.: Interpolation via weighted l_1-minimization, Applied and Computational Harmonic Analysis, (2015) [13]
58. Shor, N.Z.: Minimization Methods for Non-differentiable Functions, Springer Series in Computational Mathematics, Springer, (1985) [1]
59. Sun, D., Han, J.: Newton and quasi-Newton methods for a class of nonsmooth equations and related problems, SIAM Journal on Optimization, **7**(2), 463–480 (1997) [11]
60. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization, Technical report, Mathematics Department, University of Washington, (2008), `http://pages.cs.wisc.edu/~brecht/cs726docs/Tseng.APG.pdf` [2]