Coordinate Descent Methods for DC Minimization: Optimality Conditions and Global Convergence

Ganzhao Yuan

Peng Cheng Laboratory, China
yuangzh@pcl.ac.cn

Abstract

Difference-of-Convex (DC) minimization, referring to the problem of minimizing the difference of two convex functions, has been found rich applications in statistical learning and studied extensively for decades. However, existing methods are primarily based on multi-stage convex relaxation, only leading to weak optimality of critical points. This paper proposes a coordinate descent method for minimizing a class of DC functions based on sequential nonconvex approximation. Our approach iteratively solves a nonconvex one-dimensional subproblem globally, and it is guaranteed to converge to a coordinate-wise stationary point. We prove that this new optimality condition is always stronger than the standard critical point condition and directional point condition under a mild *locally bounded nonconvexity assumption*. For comparisons, we also include a naive variant of coordinate descent methods based on sequential convex approximation in our study. When the objective function satisfies a *globally bounded nonconvexity assumption* and *Luo-Tseng error bound assumption*, coordinate descent methods achieve *Q-linear* convergence rate. Also, for many applications of interest, we show that the nonconvex one-dimensional subproblem can be computed exactly and efficiently using a breakpoint searching method. Finally, we have conducted extensive experiments on several statistical learning tasks to show the superiority of our approach.

1 Introduction

This paper mainly focuses on the following DC minimization problem ('$\triangleq$' means define):

$$\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^n} \; F(\mathbf{x}) \triangleq f(\mathbf{x}) + h(\mathbf{x}) - g(\mathbf{x}). \qquad (1)$$

Throughout this paper, we make the following assumptions on Problem (1). ***(i)*** $f(\cdot)$ is convex and continuously differentiable, and its gradient is coordinate-wise Lipschitz continuous with constant $\mathbf{c}_i \geq 0$ that (Nesterov 2012; Beck and Tetruashvili 2013):

$$f(\mathbf{x}+\eta e_i) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \eta e_i\rangle + \frac{\mathbf{c}_i}{2}\|\eta e_i\|_2^2 \qquad (2)$$

$\forall \mathbf{x}, \eta, i = 1, ..., n$. Here $\mathbf{c} \in \mathbb{R}^n$, and $e_i \in \mathbb{R}^n$ is an indicator vector with one on the i-th entry and zero everywhere else. ***(ii)*** $h(\cdot)$ is convex and coordinate-wise separable with $h(\mathbf{x}) = \sum_{i=1}^n h_i(\mathbf{x}_i)$. Typical examples of $h(\mathbf{x})$ include the bound constrained function and the ℓ_1 norm function. ***(iii)*** $g(\cdot)$ is convex and its associated proximal operator:

$$\min_{\eta\in\mathbb{R}} p(\eta) \triangleq \frac{a}{2}\eta^2 + b\eta + h_i(\mathbf{x}+\eta e_i) - g(\mathbf{x}+\eta e_i), \qquad (3)$$

can be computed exactly and efficiently for given $a \in \mathbb{R}_+$, $b \in \mathbb{R}$ and $i \in \{1, ..., n\}$. We remark that $g(\cdot)$ is neither necessarily differentiable nor coordinate-wise separable, and typical examples of $g(\mathbf{x})$ are the ℓ_p norm function $g(\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_p$ with $p = \{1, 2, \infty\}$, the RELU function $g(\mathbf{x}) = \|\max(0, \mathbf{A}\mathbf{x})\|_1$, and the top-$s$ norm function $g(\mathbf{x}) = \sum_{i=1}^s |\mathbf{x}_{[i]}|$. Here $\mathbf{A} \in \mathbb{R}^{m\times n}$ is an arbitrary given matrix and $\mathbf{x}_{[i]}$ denotes the ith largest component of $\mathbf{x}$ in magnitude. ***(iv)*** $F(\mathbf{x})$ only takes finite values.

DC programming. DC Programming/minimization is an extension of convex maximization over a convex set (Tao and An 1997; Thi and Dinh 2018). It is closely related to the concave-convex procedure and alternating minimization in the literature. The class of DC functions is very broad, and it includes many important classes of nonconvex functions, such as twice continuously differentiable function on compact convex set and multivariate polynomial functions (Ahmadi and Hall 2018). DC programs have been mainly considered in global optimization and some algorithms have been proposed to find global solutions to such problem (Horst and Thoai 1999; Horst and Tuy 2013). Recent developments on DC programming primarily focus on designing local solution methods for some specific DC programming problems. For example, proximal bundle DC methods (Joki et al. 2017), double bundle DC methods (Joki et al. 2018), inertial proximal methods (Maingé and Moudafi 2008), and enhanced proximal methods (Lu and Zhou 2019) have been proposed. DC programming has been applied to solve a variety of statistical learning tasks, such as sparse PCA (Sriperumbudur, Torres, and Lanckriet 2007; Beck and Teboulle 2021), variable selection (Gotoh, Takeda, and Tono 2018; Gong et al. 2013), single source localization (Beck and Hallak 2020), positive-unlabeled learning (Kiryo et al. 2017; Xu et al. 2019), and deep Boltzmann machines (Nitanda and Suzuki 2017).

Coordinate descent methods. Coordinate Descent (CD) is a popular method for solving large-scale optimization problems. Advantages of this method are that compared with the full gradient descent method, it enjoys faster convergence

(Tseng and Yun 2009; Xu and Yin 2013), avoids tricky parameters tuning, and allows for easy parallelization (Liu et al. 2015). It has been well studied for convex optimization such as Lasso (Tseng and Yun 2009), support vector machines (Hsieh et al. 2008), nonnegative matrix factorization (Hsieh and Dhillon 2011), and the PageRank problem (Nesterov 2012). Its convergence and worst-case complexity are well investigated for different coordinate selection rules such as cyclic rule (Beck and Tetruashvili 2013), greedy rule (Hsieh and Dhillon 2011), and random rule (Lu and Xiao 2015; Richtárik and Takávc 2014). It has been extended to solve many nonconvex problems such as penalized regression (Breheny and Huang 2011; Deng and Lan 2020), eigenvalue complementarity problem (Patrascu and Necoara 2015), ℓ_0 norm minimization (Beck and Eldar 2013; Yuan, Shen, and Zheng 2020), resource allocation problem (Necoara 2013), leading eigenvector computation (Li, Lu, and Wang 2019), and sparse phase retrieval (Shechtman, Beck, and Eldar 2014).

Iterative majorization minimization. Iterative majorization / upper-bound minimization is becoming a standard principle in developing nonlinear optimization algorithms. Many surrogate functions such as Lipschitz gradient surrogate, proximal gradient surrogate, DC programming surrogate, variational surrogate, saddle point surrogate, Jensen surrogate, quadratic surrogate, cubic surrogate have been considered, see (Mairal 2013; Razaviyayn, Hong, and Luo 2013). Recent work extends this principle to the coordinate update, incremental update, and stochastic update settings. However, all the previous methods are mainly based on multiple-stage convex relaxation, only leading to weak optimality of critical points. In contrast, our method makes good use of sequential nonconvex approximation to find stronger stationary points. Thanks to the coordinate update strategy, we can solve the one-dimensional nonconvex subproblem *globally* by using a novel exhaustive breakpoint searching method even when $g(\cdot)$ is *non-separable* and *non-differentiable*.

Theory for nonconvex optimization. We pay specific attention to two contrasting approaches on the theory for nonconvex optimization. ***(i)*** Strong optimality. The first approach is to achieve stronger optimality guarantees for nonconvex problems. For smooth optimization, canonical gradient methods only converge to a first-order stationary point, recent works aim at finding a second-order stationary point (Jin et al. 2017). For cardinality minimization, the work of (Beck and Eldar 2013; Yuan, Shen, and Zheng 2020) introduces a new optimality condition of (block) coordinate stationary point which is stronger than that of the Lipschitz stationary point (Yuan, Li, and Zhang 2017). ***(ii)*** Strong convergence. The second approach is to provide convergence analysis for nonconvex problems. The work of (Jin et al. 2017) establishes a global convergence rate for nonconvex matrix factorization using a regularity condition. The work of (Attouch et al. 2010) establishes the convergence rate for general nonsmooth problems by imposing Kurdyka-Łojasiewicz inequality assumption of the objective function. The work of (Dong and Tao 2021; Yue, Zhou, and So 2019) establish linear convergence rates under the *Luo-Tseng error bound assumption*. Inspired by these works, we prove that the proposed CD method has strong optimality guarantees and convergence guarantees.

Contributions. The contributions of this paper are as follows: ***(i)*** We propose a new CD method for minimizing DC functions based on sequential nonconvex approximation (See Section 4). ***(ii)*** We prove that our method converge to a coordinate-wise stationary point, which is always stronger than the optimality of standard critical points and directional points when the objective function satisfies a *locally bounded nonconvexity assumption*. When the objective function satisfies a *globally bounded nonconvexity assumption* and *Luo-Tseng error bound assumption*, CD methods achieve *Q-linear* convergence rate (See Section 5). ***(iii)*** We show that, for many applications of interest, the one-dimensional subproblem can be computed exactly and efficiently using a breakpoint searching method (See Section 6). ***(iv)*** We have conducted extensive experiments on some statistical learning tasks to show the superiority of our approach (See Section 7).

Notations. Vectors are denoted by boldface lowercase letters, and matrices by boldface uppercase letters. The Euclidean inner product between $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$ or $\mathbf{x}^T\mathbf{y}$. We denote $\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. $\mathbf{x}_i$ denotes the i-th element of the vector $\mathbf{x}$. $\mathbb{E}[\cdot]$ represents the expectation of a random variable. $\odot$ and $\div$ denote the element-wise multiplication and division between two vectors, respectively. For any extended real-valued function $h : \mathbb{R}^n \to (-\infty, +\infty]$, the set of all subgradients of h at $\mathbf{x}$ is defined as $\partial h(\mathbf{x}) = \{\mathbf{g} \in \mathbb{R}^n : h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$, the conjugate of $h(\mathbf{x})$ is defined as $h^*(\mathbf{x}) \triangleq \max_{\mathbf{y}}\{\langle \mathbf{y}, \mathbf{x} \rangle - h(\mathbf{y})\}$, and $(\partial h(\mathbf{x}))_i$ denotes the subgradient of $h(\mathbf{x})$ at $\mathbf{x}$ for the i-th componnet. $\text{diag}(\mathbf{c})$ is a diagonal matrix with $\mathbf{c}$ as the main diagonal entries. We define $\|\mathbf{d}\|_{\mathbf{c}}^2 = \sum_i \mathbf{c}_i \mathbf{d}_i^2$. $\text{sign}(\cdot)$ is the signum function. $\mathbf{I}$ is the identity matrix of suitable size. The directional derivative of $F(\cdot)$ at a point $\mathbf{x}$ in its domain along a direction $\mathbf{d}$ is defined as: $F'(\mathbf{x}; \mathbf{d}) \triangleq \lim_{t \downarrow 0} \frac{1}{t}(F(\mathbf{x} + t\mathbf{d}) - F(\mathbf{x}))$. $\text{dist}(\Omega, \Omega') \triangleq \inf_{\mathbf{v} \in \Omega, \mathbf{v}' \in \Omega'} \|\mathbf{v} - \mathbf{v}'\|$ denotes the distance between two sets.

2 Motivating Applications

A number of statistical learning models can be formulated as Problem (1), which we present some instances below.

• **Application I: ℓ_p Norm Generalized Eigenvalue Problem**. Given arbitrary data matrices $\mathbf{G} \in \mathbb{R}^{m \times n}$ and $\mathbf{Q} \in \mathbb{R}^{n \times n}$ with $\mathbf{Q} \succ \mathbf{0}$, it aims at solving the following problem:

$$\bar{\mathbf{v}} \in \arg\max_{\mathbf{v}} \|\mathbf{G}\mathbf{v}\|_p, \ s.t. \ \mathbf{v}^T\mathbf{Q}\mathbf{v} = 1. \quad (4)$$

with $p \geq 1$. Using the Lagrangian dual, we have the following equivalent unconstrained problem:

$$\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}} \ \tfrac{\alpha}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \|\mathbf{G}\mathbf{x}\|_p, \quad (5)$$

for any given $\alpha > 0$. The optimal solution to Problem (4) can be recovered as $\bar{\mathbf{v}} = \pm\bar{\mathbf{x}} \cdot (\bar{\mathbf{x}}^T\mathbf{Q}\bar{\mathbf{x}})^{-\frac{1}{2}}$. Refer to the appendix for a detailed discussion.

• **Application II: Approximate Sparse/Binary Optimization**. Given a channel matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$, a structured signal $\mathbf{x}$ is transmitted through a communication channel, and received as $\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{v}$, where $\mathbf{v}$ is the Gaussian noise. If $\mathbf{x}$

has s-sparse or binary structure, one can recover $\mathbf{x}$ by solving the following optimization problem (Gotoh, Takeda, and Tono 2018; Jr. 1972):

$$\min_{\mathbf{x}}\ \tfrac{1}{2}\|\mathbf{G}\mathbf{x}-\mathbf{y}\|_2^2,\ s.t.\ \|\mathbf{x}\|_0 \le s,$$
$$\text{or}\quad \min_{\mathbf{x}}\ \tfrac{1}{2}\|\mathbf{G}\mathbf{x}-\mathbf{y}\|_2^2,\ s.t.\ \mathbf{x}\in\{-1+1\}^n.$$

Here, $\|\cdot\|_0$ is the number of non-zero components. Using the equivalent variational reformulation of the ℓ_0 (pseudo) norm $\|\mathbf{x}\|_0 \le s \Leftrightarrow \|\mathbf{x}\|_1 = \sum_{i=1}^{s} |\mathbf{x}_{[i]}|$ and the binary constraint $\{-1,+1\}^n \Leftrightarrow \{\mathbf{x}| -\mathbf{1} \le \mathbf{x} \le \mathbf{1}, \|\mathbf{x}\|_2^2 = n\}$, one can solve the following approximate sparse/binary optimization problem (Gotoh, Takeda, and Tono 2018; Yuan and Ghanem 2017, 2016):

$$\min_{\mathbf{x}}\ \tfrac{1}{2}\|\mathbf{G}\mathbf{x}-\mathbf{y}\|_2^2 + \rho(\|\mathbf{x}\|_1 - \textstyle\sum_{i=1}^{s}|\mathbf{x}_{[i]}|) \tag{6}$$
$$\min_{\|\mathbf{x}\|_\infty \le 1}\ \tfrac{1}{2}\|\mathbf{G}\mathbf{x}-\mathbf{y}\|_2^2 + \rho(\sqrt{n} - \|\mathbf{x}\|). \tag{7}$$

• **Application III: Generalized Linear Regression**. Given a sensing matrix $\mathbf{G}\in\mathbb{R}^{m\times n}$ and measurements $\mathbf{y}\in\mathbb{R}^m$, it deals with the problem of recovering a signal $\mathbf{x}$ by solving $\bar{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\sigma(\mathbf{G}\mathbf{x}) - \mathbf{y}\|_2^2$. When $\sigma(\mathbf{z}) = \max(0,\mathbf{z})$ or $\sigma(\mathbf{z}) = |\mathbf{z}|$, this problem reduces to the one-hidden-layer ReLU networks (Zhang et al. 2019) or the amplitude-base phase retrieval problem (Candès, Li, and Soltanolkotabi 2015). When $\mathbf{y}\ge\mathbf{0}$, we have the following equivalent DC program:

$$\min_{\mathbf{x}\in\mathbb{R}^n}\ \tfrac{1}{2}\|\sigma(\mathbf{G}\mathbf{x})\|_2^2 - \langle \mathbf{1}, \sigma(\mathrm{diag}(\mathbf{y})\mathbf{G})\mathbf{x})\rangle + \tfrac{1}{2}\|\mathbf{y}\|_2^2. \tag{8}$$

3 Related Work

We now present some related DC minimization algorithms.

(i) Multi-Stage Convex Relaxation (MSCR)(Zhang 2010; Bi, Liu, and Pan 2014). It solves Problem (1) by generating a sequence $\{\mathbf{x}^t\}$ as:

$$\mathbf{x}^{t+1} \in \arg\min_{\mathbf{x}}\ f(\mathbf{x}) + h(\mathbf{x}) - \langle \mathbf{x}-\mathbf{x}^t,\ \mathbf{g}^t\rangle \tag{9}$$

where $\mathbf{g}^t \in \partial g(\mathbf{x}^t)$. Note that Problem (9) is convex and can be solved via standard proximal gradient method. The computational cost of MSCR could be expensive for large-scale problems, since it is K times that of solving Problem (9) with K being the number of outer iterations.

(ii) Proximal DC algorithm (PDCA) (Gotoh, Takeda, and Tono 2018). To alleviate the computational issue of solving Problem (9), PDCA exploits the structure of $f(\cdot)$ and solves Problem (1) by generating a sequence $\{\mathbf{x}^t\}$ as: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \mathcal{Q}(\mathbf{x},\mathbf{x}^t) + h(\mathbf{x}) - \langle \mathbf{x}-\mathbf{x}^t,\ \mathbf{g}^t\rangle$, where $\mathcal{Q}(\mathbf{x},\mathbf{x}^t) \triangleq f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t),\ \mathbf{x}-\mathbf{x}^t\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{x}^t\|_2^2$, and L is the Lipschitz constant of $\nabla f(\cdot)$.

(iii) Toland's duality method (Toland 1979; Beck and Teboulle 2021). Assuming $g(\mathbf{x})$ has the following structure $g(\mathbf{x}) = \bar{g}(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}}\{\langle \mathbf{A}\mathbf{x},\mathbf{y}\rangle - \bar{g}^*(\mathbf{y})\}$. This approach rewrites Problem (1) as the following equivalent problem using the conjugate of $g(\mathbf{x})$: $\min_{\mathbf{x}}\min_{\mathbf{y}}\ f(\mathbf{x}) + h(\mathbf{x}) - \langle \mathbf{y},\mathbf{A}\mathbf{x}\rangle + \bar{g}^*(\mathbf{y})$. Exchanging the order of minimization yields the equivalent problem: $\min_{\mathbf{y}}\min_{\mathbf{x}} f(\mathbf{x}) + h(\mathbf{x}) - \langle \mathbf{y},\mathbf{A}\mathbf{x}\rangle + \bar{g}^*(\mathbf{y})$. The set of minimizers of the inner problem with respect to $\mathbf{x}$ is $\partial h^*(\mathbf{A}^T\mathbf{y}) + \nabla f^*(\mathbf{A}^T\mathbf{y})$, and the minimal value is $-f^*(\mathbf{A}^T\mathbf{y}) - h^*(\mathbf{A}^T\mathbf{y}) + \bar{g}^*(\mathbf{y})$. We have the Toland-dual problem which is also a DC program:

$$\min_{\mathbf{y}}\ \bar{g}^*(\mathbf{y}) - f^*(\mathbf{A}^T\mathbf{y}) - h^*(\mathbf{A}^T\mathbf{y}) \tag{10}$$

This method is only applicable when the minimization problem with respect to $\mathbf{x}$ is simple so that it has an analytical solution. Toland's duality method could be useful if one of the subproblems is easier to solve than the other.

(iv) Subgradient descent method (SubGrad). It uses the iteration $\mathbf{x}^{t+1} = \mathcal{P}(\mathbf{x}^t - \eta^t\mathbf{g}^t)$, where $\mathbf{g}^t \in \partial F(\mathbf{x}^t)$, η^t is the step size, and $\mathcal{P}$ is the projection operation on some convex set. This method has received much attention recently due to its simplicity (Zhang et al. 2019; Davis et al. 2018; Davis and Grimmer 2019; Li et al. 2021).

4 Coordinate Descent Methods for DC Minimization

This section presents a new Coordinate Descent (CD) method for solving Problem (1), which is based on Sequential NonConvex Approximation (SNCA). For comparisons, we also include a naive variant of CD methods based on Sequential Convex Approximation (SCA) in our study. These two methods are denoted as ***CD-SNCA*** and ***CD-SCA***, respectively.

Coordinate descent is an iterative algorithm that sequentially minimizes the objective function along coordinate directions. In the t-th iteration, we minimize $F(\cdot)$ with respect to the i^t variable while keeping the remaining $(n-1)$ variables $\{\mathbf{x}_j^t\}_{j\neq i^t}$ fixed. This is equivalent to performing the following one-dimensional search along the i^t-th coordinate: $\bar{\eta}^t \in \arg\min_{\eta\in\mathbb{R}}\ f(\mathbf{x}^t + \eta e_{i^t}) + h(\mathbf{x}^t + \eta e_{i^t}) - g(\mathbf{x}^t + \eta e_{i^t})$. Then $\mathbf{x}^t$ is updated via: $\mathbf{x}^{t+1} = \mathbf{x}^t + \bar{\eta}^t e_{i^t}$. However, the one-dimensional problem above could be still hard to solve when $f(\cdot)$ and/or $g(\cdot)$ is complicated. One can consider replacing $f(\cdot)$ and $g(\cdot)$ with their majorization function:

$$f(\mathbf{x}^t + \eta e_{i^t}) \le \mathcal{S}_{i^t}(\mathbf{x}^t,\eta)$$
$$\text{with } \mathcal{S}_i(\mathbf{x},\eta) \triangleq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \eta e_i\rangle + \frac{\mathbf{c}_i}{2}\eta^2, \tag{11}$$
$$-\,g(\mathbf{x}^t + \eta e_{i^t}) \le \mathcal{G}_{i^t}(\mathbf{x}^t,\eta)$$
$$\text{with } \mathcal{G}_i(\mathbf{x},\eta) \triangleq -g(\mathbf{x}) - \langle \partial g(\mathbf{x}),\ (\mathbf{x}+\eta e_i) - \mathbf{x}\rangle. \tag{12}$$

► Choosing the Majorization Function

1. **Sequential NonConvex Approximation Strategy**. If we replace $f(\mathbf{x}^t + \eta e_{i^t})$ with its upper bound $\mathcal{S}_{i^t}(\mathbf{x}^t,\eta)$ as in (11) while keep the remaining two terms unchanged, we have the resulting subproblem as in (13), which is a nonconvex problem. It reduces to the proximal operator computation as in (3) with $a = \mathbf{c}_{i^t} + \theta$ and $b = \nabla_{i^t} f(\mathbf{x}^t)$. Setting the subgradient with respect to η of the objective function in (13) to zero, we have the following *necessary but not sufficient* optimality condition for (13):

   $$0 \in [\nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1}) - \partial g(\mathbf{x}^{t+1})]_{i^t} + (\mathbf{c}_{i^t} + \theta)\bar{\eta}^t.$$

2. **Sequential Convex Approximation Strategy**. If we replace $f(\mathbf{x}^t + \eta e_{i^t})$ and $-g(\mathbf{x}^t + \eta e_{i^t})$ with their respective upper bounds $\mathcal{S}_{i^t}(\mathbf{x}^t,\eta)$ and $\mathcal{G}_{i^t}(\mathbf{x}^t,\eta)$ as in (11) and (12), while keep the term $h(\mathbf{x}^t + \eta e_{i^t})$ unchanged, we have

Algorithm 1: **Coordinate Descent Methods for Minimizing DC functions using *SNCA* or *SCA* strategy.**

```
Input: an initial feasible solution $\mathbf{x}^0$, $\theta > 0$. Set $t = 0$.
while not converge do
    (S1) Use some strategy to find a coordinate $i^t \in \{1, ..., n\}$ for the $t$-th iteration.
    (S2) Solve the following nonconvex or convex subproblem globally and exactly.
    • Option I: Sequential NonConvex Approximation (SNCA) strategy.
        $\bar{\eta}^t \in \bar{\mathcal{M}}_{i^t}(\mathbf{x}^t) \triangleq \arg\min_{\eta} \ \mathcal{M}_{i^t}(\mathbf{x}^t, \eta)$    (13)
        with $\mathcal{M}_i(\mathbf{x}, \eta) \triangleq \mathcal{S}_i(\mathbf{x}, \eta) + h_i(\mathbf{x} + \eta e_i)$
            $-g(\mathbf{x} + \eta e_i) + \frac{\theta}{2}\|(\mathbf{x} + \eta e_i) - \mathbf{x}\|_2^2$
    • Option II: Sequential Convex Approximation (SCA) strategy.
        $\bar{\eta}^t \in \bar{\mathcal{P}}_{i^t}(\mathbf{x}^t) \triangleq \arg\min_{\eta} \ \mathcal{P}_{i^t}(\mathbf{x}^t, \eta)$    (14)
        $\mathcal{P}_i(\mathbf{x}, \eta) \triangleq \mathcal{S}_i(\mathbf{x}, \eta) + h_i(\mathbf{x} + \eta e_i)$
            $+\mathcal{G}_i(\mathbf{x}, \eta) + \frac{\theta}{2}\|(\mathbf{x} + \eta e_i) - \mathbf{x}\|_2^2$
    (S3) $\mathbf{x}^{t+1} = \mathbf{x}^t + \bar{\eta}^t \cdot e_{i^t} \quad (\Leftrightarrow \mathbf{x}_{i^t}^{t+1} = \mathbf{x}_{i^t}^t + \bar{\eta}^t)$
    (S4) Increment $t$ by 1.
end while
```

the resulting subproblem as in (14), which is a convex problem. We have the following *necessary and sufficient* optimality condition for (14):

$$0 \in [\nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1}) - \partial g(\mathbf{x}^t)]_{i^t} + (\mathbf{c}_{i^t} + \theta)\bar{\eta}^t.$$

►Selecting the Coordinate to Update

There are several fashions to decide which coordinate to update in the literature (Tseng and Yun 2009). **(i)** ***Random rule***. i^t is randomly selected from $\{1, ..., n\}$ with equal probability. **(ii)** ***Cyclic rule***. i^t takes all coordinates in cyclic order $1 \to 2 \to ... \to n \to 1$. **(iii)** ***Greedy rule***. Assume that $\nabla f(\mathbf{x})$ is Lipschitz continuous with constant L. The index i^t is chosen as $i^t = \arg\max_j |\mathbf{d}_j^t|$ where $\mathbf{d}^t = \arg\min_{\mathbf{d}} \ h(\mathbf{x}^t + \mathbf{d}) + \frac{L}{2}\|\mathbf{d}\|_2^2 + \langle \nabla f(\mathbf{x}^t) - \partial g(\mathbf{x}^t)), \mathbf{d}\rangle$. Note that $\mathbf{d}^t = \mathbf{0}$ implies that $\mathbf{x}^t$ is a critical point.

We summarize ***CD-SNCA*** and ***CD-SCA*** in Algorithm 1.

Remarks. ***(i)*** We use a proximal term for the subproblems in (13) and (14) with θ being the proximal point parameter. This is to guarantee sufficient descent condition and global convergence for Algorithm 1. As can be seen in Theorem 5.10 and Theorem 5.12, the parameter θ is critical for ***CD-SNCA***. ***(ii)*** Problem (13) can be viewed as *globally* solving the following nonconvex problem which has a bilinear structure: $(\bar{\eta}^t, \mathbf{y}) = \arg\min_{\eta, \mathbf{y}} \ \mathcal{S}_{i^t}(\mathbf{x}^t, \eta) + \frac{\theta}{2}\eta^2 + h(\mathbf{x}^t + \eta e_{i^t}) - \langle \mathbf{y}, \mathbf{x}^t + \eta e_{i^t}\rangle + g^*(\mathbf{y})$. ***(iii)*** While we apply CD to the primal, one may apply to the dual as in Problem (10). ***(iv)*** The nonconvex majorization function used in ***CD-SNCA*** is always a lower bound of the convex majorization function used in ***CD-SCA***, i.e., $\mathcal{M}_i(\mathbf{x}, \eta) \leq \mathcal{P}_i(\mathbf{x}, \eta), \ \forall i, \mathbf{x}, \eta$.

5 Theoretical Analysis

This section provides a novel optimality analysis and a novel convergence analysis for Algorithm 1. Due to space limit, all proofs are placed in the appendix.

We introduce the following useful definition.

Definition 5.1. (**Globally or Locally Bounded Nonconvexity**) ***(a)*** A function $z(\mathbf{x})$ is called to be globally ρ-bounded nonconvex if: $\forall \mathbf{x}, \mathbf{y}, \ z(\mathbf{x}) \leq z(\mathbf{y}) + \langle \mathbf{x} - \mathbf{y}, \ \partial z(\mathbf{x})\rangle + \frac{\rho}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ with $\rho < +\infty$. ***(b)*** In particular, $z(\mathbf{x})$ is locally ρ-bounded nonconvex if $\mathbf{x}$ is restricted to some point $\ddot{\mathbf{x}}$ with $\mathbf{x} = \ddot{\mathbf{x}}$.

Remarks. ***(i)*** Globally ρ-bounded nonconvexity of $z(\mathbf{x})$ is equivalent to $z(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x}\|_2^2$ is convex, and this notation is also referred as *semi-convex*, *approximate convex*, or *weakly-convex* in the literature (cf. (Böhm and Wright 2021; Davis et al. 2018; Li et al. 2021)). ***(ii)*** Many nonconvex functions in the robust statistics literature are *globally* ρ-bounded nonconvex, examples of which includes the *minimax concave penalty*, the *fractional penalty*, the *smoothly clipped absolute deviation*, and the *Cauchy loss* (c.f. (Böhm and Wright 2021)). ***(iii)*** Any globally ρ-bounded nonconvex function $z(\mathbf{x})$ can be rewritten as a DC function that $z(\mathbf{x}) = \frac{\rho}{2}\|\mathbf{x}\|^2 - g(\mathbf{x})$, where $g(\mathbf{x}) = \frac{\rho}{2}\|\mathbf{x}\|^2 - z(\mathbf{x})$ is convex and $(-g(\mathbf{x}))$ is globally (2ρ)-bounded nonconvex.

Globally bounded nonconvexity could be a strong definition; one may use a weaker definition of locally bounded nonconvexity instead. The following lemma shows that some nonconvex functions are locally bounded nonconvex.

Lemma 5.2. *The function* $z(\mathbf{x}) \triangleq -\|\mathbf{x}\|_p$ *with* $p \in [1, \infty)$ *is concave and locally* ρ*-bounded nonconvex with* $\rho < +\infty$.

Remarks. By Lemma 5.2, we have that the functions $z(\mathbf{x}) = -\|\mathbf{G}\mathbf{x}\|_p$ in (5) and $z(\mathbf{x}) = -\rho\|\mathbf{x}\|$ in (7) are locally ρ-bounded nonconvex. Using similar strategies, one can conclude that the functions $z(\mathbf{x}) = -\sum_{i=1}^{s}|\mathbf{x}_{[i]}|$ and $z(\mathbf{x}) = -\langle \mathbf{1}, \sigma(\text{diag}(\mathbf{y})\mathbf{G})\mathbf{x})\rangle$ as in (6) and (8) are locally ρ-bounded nonconvex.

We assume that the random-coordinate selection rule is used. After t iterations, Algorithm 1 generates a random output $\mathbf{x}^t$, which depends on the observed realization of the random variable: $\xi^{t-1} \triangleq \{i^0, \ i^1, ..., i^{t-1}\}$.

5.1 Optimality Analysis

We now provide an optimality analysis of our method. Since the coordinate-wise optimality condition is novel in this paper, we clarify its relations with existing optimality conditions formally.

Definition 5.3. (Critical Point) A solution $\check{\mathbf{x}}$ is called a critical point if (Toland 1979): $0 \in \nabla f(\check{\mathbf{x}}) + \partial h(\check{\mathbf{x}}) - \partial g(\check{\mathbf{x}})$.

Remarks. ***(i)*** The expression above is equivalent to $(f(\check{\mathbf{x}}) + \partial h(\check{\mathbf{x}})) \cap \partial g(\check{\mathbf{x}}) \neq \emptyset$. The sub-differential is always non-empty on convex functions; that is why we assume that $F(\cdot)$ can be repressed as the difference of two convex functions. ***(ii)*** Existing methods such as MSCR, PDCA, and SubGrad as shown in Section (3) are only guaranteed to find critical points of Problem (1).

Definition 5.4. (Directional Point) A solution $\grave{\mathbf{x}}$ is called a directional point if (Pang, Razaviyayn, and Alvarado 2017): $F'(\grave{\mathbf{x}}; \mathbf{y} - \grave{\mathbf{x}}) \geq 0, \ \forall \mathbf{y} \in \text{dom}(F) \triangleq \{\mathbf{x} : |F(\mathbf{x})| < +\infty\}$.

Remarks. The work of (Pang, Razaviyayn, and Alvarado 2017) characterizes different types of stationary points, and proposes an enhanced DC algorithm that subsequently converges to a directional point. However, they only consider the case $g(\mathbf{x}) = \max_{i \in I} g_i(\mathbf{x})$ where each $g_i(\mathbf{x})$ is continuously differentiable and convex and I is a finite index set.

Definition 5.5. (Coordinate-Wise Stationary Point) A solution $\ddot{\mathbf{x}}$ is called a coordinate-wise stationary point if the following holds: $0 \in \arg\min_\eta \mathcal{M}_i(\ddot{\mathbf{x}}, \eta)$ for all $i = 1, ..., n$, where $\mathcal{M}_i(\mathbf{x}, \eta) \triangleq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \ \eta e_i \rangle + \frac{\mathbf{c}_i}{2}\eta^2 + h_i(\mathbf{x} + \eta e_i) - g(\mathbf{x} + \eta e_i) + \frac{\theta}{2}\eta^2$, and $\theta \geq 0$ is a constant.

Remarks. **(i)** Coordinate-wise stationary point states that if we minimize the majorization function $\mathcal{M}_i(\mathbf{x}, \eta)$, we can not improve the objective function value for $\mathcal{M}_i(\mathbf{x}, \eta)$ for all $i \in \{1, ..., n\}$. **(ii)** For any coordinate-wise stationary point $\ddot{\mathbf{x}}$, we have the following necessary but not sufficient condition: $\forall i \in \{1, ..., n\}, \ 0 \in \partial \mathcal{M}_i(\ddot{\mathbf{x}}, \eta) \triangleq (\mathbf{c}_i + \theta)\eta + [\nabla f(\ddot{\mathbf{x}}) + \partial h(\ddot{\mathbf{x}} + \eta e_i) - \partial h(\ddot{\mathbf{x}} + \eta e_i)]_i$ with $\eta = 0$, which coincides with the critical point condition. Therefore, any coordinate-wise stationary point is a critical point.

The following lemma reveals a *quadratic growth condition* for any coordinate-wise stationary point.

Lemma 5.6. *Let $\ddot{\mathbf{x}}$ be any coordinate-wise stationary point. Assume that $z(\mathbf{x}) \triangleq -g(\mathbf{x})$ is locally ρ-bounded nonconvex at the point $\ddot{\mathbf{x}}$. We have: $\forall \mathbf{d}, \ F(\ddot{\mathbf{x}}) - F(\ddot{\mathbf{x}} + \mathbf{d}) \leq \frac{1}{2}\|\mathbf{d}\|^2_{(\mathbf{c}+\theta+\rho)}$.*

Remarks. Recall that a solution $\dot{\mathbf{x}}$ is said to be a local minima if $F(\dot{\mathbf{x}}) \leq F(\dot{\mathbf{x}} + \mathbf{d})$ for a sufficiently small constant δ that $\|\mathbf{d}\| \leq \delta$. The coordinate-wise optimality condition does not have any restriction on $\mathbf{d}$ with $\|\mathbf{d}\| \leq +\infty$. Thus, neither the optimality condition of coordinate-wise stationary point nor that of the local minima is stronger than the other.

We use $\check{\mathbf{x}}$, $\grave{\mathbf{x}}$, $\ddot{\mathbf{x}}$, and $\bar{\mathbf{x}}$ to denote any critical point, directional point, coordinate-wise stationary point, and optimal point, respectively. The following theorem establishes the relations between different types of stationary points list above.

Theorem 5.7. ***(Optimality Hierarchy between the Optimality Conditions).*** *Assume that the assumption made in Lemma 5.6 holds, we have: $\{\bar{\mathbf{x}}\} \overset{(a)}{\subseteq} \{\ddot{\mathbf{x}}\} \overset{(b)}{\subseteq} \{\grave{\mathbf{x}}\} \overset{(c)}{\subseteq} \{\check{\mathbf{x}}\}$.*

Remarks. ***(i)*** The coordinate-wise optimality condition is stronger than the critical point condition (Gotoh, Takeda, and Tono 2018; Zhang 2010; Bi, Liu, and Pan 2014) and the directional point condition (Pang, Razaviyayn, and Alvarado 2017) when the function $(-g(\mathbf{x}))$ is locally ρ-bounded nonconvex. ***(ii)*** Our optimality analysis can be also applied to the equivalent dual problem which is also a DC program as in (10). ***(iii)*** We explain the optimality of coordinate-wise stationary point is stronger than that of previous definitions using the following one-dimensional example: $\min_x \frac{1}{2}(x-1)^2 - 3|x|$. This problem contains three critical points $\{-2, 0, 4\}$, two directional points / local minima $\{-2, 4\}$, and a unique coordinate-wise stationary point $\{4\}$. This unique coordinate-wise stationary point can be found using a clever breakpoint searching method (discussed later in Section 6). Figure 1 demonstrates the optimality hierarchy between different optimality conditions and the geometric interpretation for this one-dimensional example.

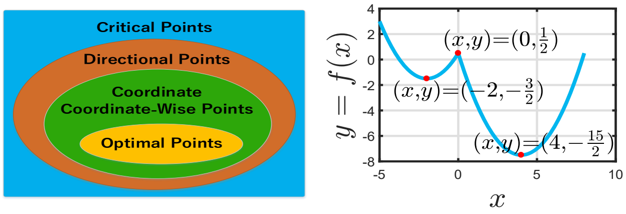


Figure 1: Left: optimality hierarchy. Right: one dimensional example: $\min_x \frac{1}{2}(x-1)^2 - 3|x|$.

5.2 Convergence Analysis

We provide a convergence analysis for ***CD-SNCA*** and ***CD-SCA***. First, we define the approximate critical point and approximate coordinate-wise stationary point as follows.

Definition 5.8. (Approximate Critical Point) Given any constant $\epsilon > 0$, a point $\check{\mathbf{x}}$ is called a ϵ-approximate critical point if: $\text{dist}(\nabla f(\check{\mathbf{x}}), \partial g(\check{\mathbf{x}}) - \partial h(\check{\mathbf{x}}))^2 \leq \epsilon$.

Definition 5.9. (Approximate Coordinate-Wise Stationary Point) Given any constant $\epsilon > 0$, a point $\ddot{\mathbf{x}}$ is called a ϵ-approximate coordinate-wise stationary point if: $\frac{1}{n}\sum_{i=1}^n \text{dist}(0, \arg\min_\eta \mathcal{M}_i(\ddot{\mathbf{x}}, \eta))^2 \leq \epsilon$, where $\mathcal{M}_i(\mathbf{x}, \eta)$ is defined in Definition 5.5.

Theorem 5.10. *We have the following results.* ***(a)*** *For* ***CD-SNCA****, it holds that $F(\mathbf{x}^{t+1}) - F(\mathbf{x}^t) \leq -\frac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2$. Algorithm 1 finds an ϵ-approximate* ***coordinate-wise stationary point*** *of Problem (1) in at most T iterations in the sense of expectation, where $T \leq \lceil \frac{2n(F(\mathbf{x}^0) - F(\bar{\mathbf{x}}))}{\theta\epsilon} \rceil = \mathcal{O}(\epsilon^{-1})$.* ***(b)*** *For* ***CD-SCA****, it holds that $F(\mathbf{x}^{t+1}) - F(\mathbf{x}^t) \leq -\frac{\beta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2$ with $\beta \triangleq \min(\mathbf{c}) + 2\theta$. Algorithm 1 finds an ϵ-approximate* ***critical point*** *of Problem (1) in at most T iterations in the sense of expectation, where $T \leq \lceil \frac{2n(F(\mathbf{x}^0) - F(\bar{\mathbf{x}}))}{\beta\epsilon} \rceil = \mathcal{O}(\epsilon^{-1})$.*

Remarks. While existing methods only find critical points or directional points of Problem (1), ***CD-SNCA*** is guaranteed to find a coordinate-wise stationary point which has stronger optimality guarantees (See Theorem 5.7).

To achieve stronger convergence result for Algorithm 1, we make the following *Luo-Tseng error bound assumption*, which has been extensively used in all aspects of mathematical optimization (cf. (Dong and Tao 2021; Yue, Zhou, and So 2019)).

Assumption 5.11. (*Luo-Tseng Error Bound* (Luo and Tseng 1993; Tseng and Yun 2009)) We define a residual function as $\mathcal{R}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n |\text{dist}(0, \bar{\mathcal{M}}_i(\mathbf{x}))|$ or $\mathcal{R}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n |\text{dist}(0, \bar{\mathcal{P}}_i(\mathbf{x}))|$, where $\bar{\mathcal{M}}_i(\mathbf{x})$ and $\bar{\mathcal{P}}_i(\mathbf{x})$ are respectively defined in (13) and (14). For any $\varsigma \geq \min_{\mathbf{x}} F(\mathbf{x})$, there exist scalars $\delta > 0$ and $\varrho > 0$ such that:

$$\forall \mathbf{x}, \ \text{dist}(\mathbf{x}, \mathcal{X}) \leq \delta \mathcal{R}(\mathbf{x}), \ \text{whenever } F(\mathbf{x}) \leq \varsigma, \mathcal{R}(\mathbf{x}) \leq \varrho.$$

Here, $\mathcal{X}$ is the set of stationary points satisfying $\mathcal{R}(\mathbf{x}) = 0$.

We have the following theorems regarding to the convergence rate of ***CD-SNCA*** and ***CD-SCA***.

Theorem 5.12. ***(Convergence Rate for CD-SNCA)****. Let $\ddot{\mathbf{x}}$ be any coordinate-wise stationary point. We define $\ddot{q}^t \triangleq F(\mathbf{x}^t) - F(\ddot{\mathbf{x}})$, $\ddot{r}^t \triangleq \frac{1}{2}\|\mathbf{x}^t - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2$, $\bar{\mathbf{c}} \triangleq \mathbf{c} + \theta$, $\bar{\rho} = \frac{\rho}{\min(\bar{\mathbf{c}})}$, $\gamma \triangleq 1 + \frac{\rho}{\theta}$, and $\varpi \triangleq 1 - \bar{\rho}$. Assume that $z(\mathbf{x}) \triangleq -g(\mathbf{x})$ is globally ρ-bounded non-convex.* ***(a)*** *We have $\varpi\mathbb{E}[\ddot{r}^{t+1}] + \gamma\mathbb{E}[\ddot{q}^{t+1}] \leq (\varpi + \frac{\bar{\rho}}{n})\ddot{r}^t + (\gamma - \frac{1}{n})\ddot{q}^t$.* ***(b)*** *If θ is sufficiently large such that $\varpi \geq 0$, $\mathcal{M}_{i^t}(\mathbf{x}^t, \eta)$ in (13) is convex w.r.t. η for all t, and it holds that: $\mathbb{E}[\ddot{q}^{t+1}] \leq (\frac{\kappa_1 - \frac{1}{n}}{\kappa_1})^{t+1}\ddot{q}^0$, where $\kappa_0 \triangleq \max(\bar{\mathbf{c}})\frac{\delta^2}{\theta}$ and $\kappa_1 \triangleq n\kappa_0(\varpi + \frac{\bar{\rho}}{n}) + \gamma$.*

Theorem 5.13. ***(Convergence Rate for CD-SCA)****. Let $\check{\mathbf{x}}$ be any critical point. We define $\check{q}^t \triangleq F(\mathbf{x}^t) - F(\check{\mathbf{x}})$, $\check{r}^t \triangleq \frac{1}{2}\|\mathbf{x}^t - \check{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2$, $\bar{\mathbf{c}} \triangleq \mathbf{c} + \theta$, and $\bar{\rho} = \frac{\rho}{\min(\bar{\mathbf{c}})}$. Assume that $z(\mathbf{x}) \triangleq -g(\mathbf{x})$ is globally ρ-bounded non-convex.* ***(a)*** *We have $\mathbb{E}[\check{r}^{t+1}] + \mathbb{E}[\check{q}^{t+1}] \leq (1 + \frac{\bar{\rho}}{n})\check{r}^t + (1 - \frac{1}{n})\check{q}^t$.* ***(b)*** *It holds that: $\mathbb{E}[\check{q}^{t+1}] \leq (\frac{\kappa_2 - \frac{1}{n}}{\kappa_2})^{t+1}\check{q}^0$, where $\kappa_0 \triangleq \max(\bar{\mathbf{c}})\frac{\delta^2}{\theta}$ and $\kappa_2 = n\kappa_0(1 + \frac{\bar{\rho}}{n}) + 1$.*

Remarks. **(i)** Under the *Luo-Tseng error bound assumption*, ***CD-SNCA*** (or ***CD-SCA***) converges to the coordinate-wise stationary point (or critical point) Q-linearly. **(ii)** Note that the convergence rate κ_1 of ***CD-SNCA*** and κ_2 of ***CD-SCA*** depend on the same coefficients κ_0. When n is large, the terms $n\kappa_0(\varpi + \frac{\bar{\rho}}{n})$ and $n\kappa_0(1 + \frac{\bar{\rho}}{n})$ respectively dominate the value of κ_1 and κ_2. If we choose $0 \leq \varpi < 1$ for ***CD-SNCA***, we have $\kappa_1 \ll \kappa_2$. Thus, the convergence rate of ***CD-SNCA*** could be much faster than that of ***CD-SCA*** for high-dimensional problems.

6 A Breakpoint Searching Method for Proximal Operator Computation

This section presents a new breakpoint searching method to solve Problem (3) exactly and efficiently for different $h(\cdot)$ and $g(\cdot)$. This method first identifies all the possible critical points / breakpoints Θ for $\min_{\eta \in \mathbb{R}} p(\eta)$ as in Problem (3), and then picks the solution that leads to the lowest value as the optimal solution. We denote $\mathbf{A} \in \mathbb{R}^{m \times n}$ be an arbitrary matrix, and define $\mathbf{g} = \mathbf{A}e_i \in \mathbb{R}^m, \mathbf{d} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$.

6.1 When $g(\mathbf{y}) = \|\mathbf{A}\mathbf{y}\|_1$ and $h_i(\cdot) \triangleq 0$

Consider the problem: $\min_\eta \frac{a}{2}\eta^2 + b\eta - \|\mathbf{A}(\mathbf{x} + \eta e_i)\|_1$. It can be rewritten as: $\min_\eta p(\eta) \triangleq \frac{a}{2}\eta^2 + b\eta - \|\mathbf{g}\eta + \mathbf{d}\|_1$. Setting the gradient of $p(\cdot)$ to zero yields: $0 = a\eta + b - \langle \text{sign}(\eta\mathbf{g} + \mathbf{d}), \mathbf{g}\rangle = a\eta + b - \langle \text{sign}(\eta + \mathbf{d} \div |\mathbf{g}|), \mathbf{g}\rangle$, where we use: $\forall \rho > 0, \text{sign}(\mathbf{x}) = \text{sign}(\rho\mathbf{x})$. We assume $\mathbf{g}_i \neq 0$. If this does not hold and there exists $\mathbf{g}_j = 0$ for some j, then $\{\mathbf{g}_j, \mathbf{d}_j\}$ can be removed since it does not affect the minimizer of the problem. We define $\mathbf{z} \triangleq \{+\frac{\mathbf{d}_1}{\mathbf{g}_1}, -\frac{\mathbf{d}_1}{\mathbf{g}_1}, ..., +\frac{\mathbf{d}_m}{\mathbf{g}_m}, -\frac{\mathbf{d}_m}{\mathbf{g}_m}\} \in \mathbb{R}^{2m \times 1}$, and assume $\mathbf{z}$ has been sorted in ascending order. The domain $p(\eta)$ can be divided into $2m + 1$ intervals: $(-\infty, \mathbf{z}_1)$, $(\mathbf{z}_1, \mathbf{z}_2)$,..., and $(\mathbf{z}_{2m}, +\infty)$. There are $2m + 1$ breakpoints $\boldsymbol{\eta} \in \mathbb{R}^{(2m+1) \times 1}$. In each interval, the sign of $(\eta + \mathbf{d} \div |\mathbf{g}|)$ can be determined. Thus, the i-th breakpoints for the i-th interval can be computed as $\boldsymbol{\eta}_i = (\langle \text{sign}(\eta + \mathbf{d} \div |\mathbf{g}|), \mathbf{g}\rangle - b)/a$. Therefore, Problem (3) contains $2m + 1$ breakpoints $\Theta = \{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_{(2m+1)}\}$ for this example.

6.2 When $g(\mathbf{y}) = \sum_{i=1}^s |\mathbf{y}_{[i]}|$ and $h_i(\mathbf{y}) \triangleq |\mathbf{y}_i|$

Consider the problem: $\min_\eta \frac{a}{2}\eta^2 + b\eta + |\mathbf{x}_i + \eta| - \sum_{i=1}^s |(\mathbf{x} + \eta e_i)_{[i]}|$. Since the variable η only affects the value of $\mathbf{x}_i$, we consider two cases for $\mathbf{x}_i + \eta$. ***(i)*** $\mathbf{x}_i + \eta$ belongs to the top-s subset. This problem reduces to $\min_\eta \frac{a}{2}\eta^2 + b\eta$, which contains one unique breakpoint: $\{-b/a\}$. ***(ii)*** $\mathbf{x}_i + \eta$ does not belong to the top-s subset. This problem reduces to $\min_\eta \frac{a}{2}\eta^2 + bt + |\mathbf{x}_i + \eta|$, which contains three breakpoints $\{-\mathbf{x}_i, (-1 - b)/a, (1 - b)/a\}$. Therefore, Problem (3) contains 4 breakpoints $\Theta = \{-b/a, -\mathbf{x}_i, (-1 - b)/a, (1 - b)/a\}$ for this example.

When we have found the breakpoint set Θ, we pick the solution that results in the lowest value as the global optimal solution $\bar{\eta}$, i.e., $\bar{\eta} = \arg\min_\eta p(\eta),\ s.t.\ \eta \in \Theta$. Note that the coordinate-wise separable function $h_i(\cdot)$ does not bring much difficulty for solving Problem (3).

7 Experiments

This section demonstrates the effectiveness and efficiency of Algorithm 1 on the ℓ_p norm generalized eigenvalue problem. For more experiment results, please refer to the full version of this paper (Yuan 2023).

7.1 Experimental Settings

We consider the following four types of data sets for the sensing/channel matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$. ***(i)*** 'randn-m-n': $\mathbf{G} = \text{randn}(m, n)$. ***(ii)*** 'e2006-m-n': $\mathbf{G} = \mathbf{X}$. ***(iii)*** 'randn-m-n-C': $\mathbf{G} = \mathcal{N}(\text{randn}(m, n))$. ***(iv)*** 'e2006-m-n-C': $\mathbf{G} = \mathcal{N}(\mathbf{X})$. Here, $\text{randn}(m, n)$ is a function that returns a standard Gaussian random matrix of size $m \times n$. $\mathbf{X}$ is generated by sampling from the original real-world data set 'e2006-tfidf'. $\mathcal{N}(\mathbf{G})$ is defined as: $[\mathcal{N}(\mathbf{G})]_I = 100 \cdot \mathbf{G}_I, [\mathcal{N}(\mathbf{G})]_{\bar{I}} = \mathbf{G}_{\bar{I}}$, where I is a random subset of $\{1, ..., mn\}$, $\bar{I} = \{1, ..., mn\} \setminus I$, and $|I| = 0.1 \cdot mn$. The last two types of data sets are designed to verify the robustness of the algorithms.

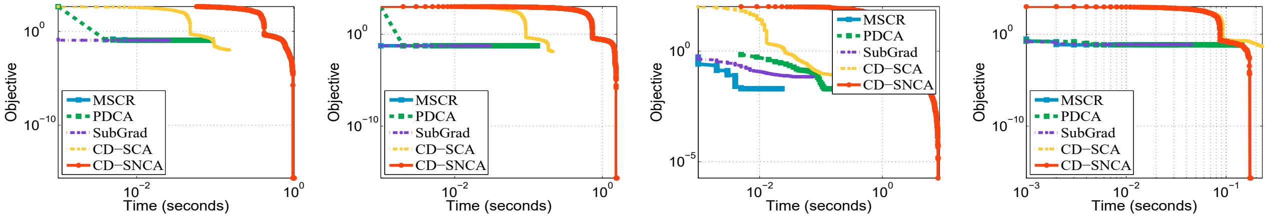
All methods are implemented in MATLAB on an Intel 2.6 GHz CPU with 32 GB RAM. Only our breakpoint searching procedure is developed in C and wrapped into the MATLAB code, since it requires elementwise loops that are less efficient in native MATLAB. We keep a record of the relative changes of the objective by $\mathbf{z}_t = [F(\mathbf{x}^t) - F(\mathbf{x}^{t+1})]/F(\mathbf{x}^t)$, and let all algorithms run up to T seconds and stop them at iteration t if $\text{mean}([\mathbf{z}_{t-\min(t,\upsilon)+1}, \mathbf{z}_{t-min(t,\upsilon)+2}, ..., \mathbf{z}_t]) \leq \epsilon$. The default value $(\theta, \epsilon, \upsilon, T) = (10^{-6}, 10^{-10}, 500, 60)$ is used. All methods are executed 10 times and the average performance is reported. Some Matlab code can be found in the authors' research webpages.

7.2 ℓ_p Norm Generalized Eigenvalue Problem

We consider Problem (4) with $p = 1$ and $\mathbf{Q} = \mathbf{I}$. We have the following problem: $\min_\mathbf{x} \frac{\alpha}{2}\|\mathbf{x}\|_2^2 - \|\mathbf{G}\mathbf{x}\|_1$. It is consistent with Problem (1) with $f(\mathbf{x}) \triangleq \frac{\alpha}{2}\|\mathbf{x}\|_2^2$, $h(\mathbf{x}) \triangleq 0$, and

	MSCR	PDCA	T-DUAL	***CD-SCA***	***CD-SNCA***
randn-256-1024	-1.329 ± 0.038	-1.329 ± 0.038	-1.329 ± 0.038	-1.426 ± 0.056	**-1.447 ± 0.053**
randn-256-2048	-1.132 ± 0.021	-1.132 ± 0.021	-1.132 ± 0.021	-1.192 ± 0.019	**-1.202 ± 0.016**
randn-1024-256	-5.751 ± 0.163	-5.751 ± 0.163	-5.664 ± 0.173	-5.755 ± 0.108	**-5.817 ± 0.129**
randn-2048-256	-9.364 ± 0.183	-9.364 ± 0.183	-9.161 ± 0.101	-9.405 ± 0.182	**-9.408 ± 0.164**
e2006-256-1024	-28.031 ± 37.894	-28.031 ± 37.894	-27.996 ± 37.912	-27.880 ± 37.980	**-28.167 ± 37.826**
e2006-256-2048	-22.282 ± 24.007	-22.282 ± 24.007	-22.282 ± 24.007	-22.113 ± 23.941	**-22.448 ± 23.908**
e2006-1024-256	-43.516 ± 77.232	-43.516 ± 77.232	-43.364 ± 77.265	-43.283 ± 77.297	**-44.269 ± 76.977**
e2006-2048-256	-44.705 ± 47.806	-44.705 ± 47.806	-44.705 ± 47.806	-44.633 ± 47.789	**-45.176 ± 47.493**
randn-256-1024-C	-1.332 ± 0.019	-1.332 ± 0.019	-1.332 ± 0.019	-1.417 ± 0.027	**-1.444 ± 0.029**
randn-256-2048-C	-1.161 ± 0.024	-1.161 ± 0.024	-1.161 ± 0.024	-1.212 ± 0.022	**-1.219 ± 0.023**
randn-1024-256-C	-5.650 ± 0.141	-5.650 ± 0.141	-5.591 ± 0.145	-5.716 ± 0.159	**-5.808 ± 0.134**
randn-2048-256-C	-9.236 ± 0.125	-9.236 ± 0.125	-9.067 ± 0.137	-9.243 ± 0.145	**-9.377 ± 0.233**
e2006-256-1024-C	-4.841 ± 6.410	-4.841 ± 6.410	-4.840 ± 6.410	-4.837 ± 6.411	**-5.027 ± 6.363**
e2006-256-2048-C	-4.297 ± 2.825	-4.297 ± 2.825	-4.297 ± 2.823	-4.259 ± 2.827	**-4.394 ± 2.814**
e2006-1024-256-C	-6.469 ± 3.663	-6.469 ± 3.663	-6.469 ± 3.663	-6.470 ± 3.663	**-6.881 ± 3.987**
e2006-2048-256-C	-31.291 ± 60.597	-31.291 ± 60.597	-31.291 ± 60.597	-31.284 ± 60.599	**-32.026 ± 60.393**

Table 1: Comparisons of objective values of all the methods for solving the ℓ_1 norm PCA problem. The best results are bolded.



(a) randn-256-1024 (b) randn-256-2048 (c) randn-1024-256 (d) randn-2048-256

Figure 2: The convergence curve of the compared methods for solving the ℓ_p norm generalized eigenvalue problem on different data sets.

$g(\mathbf{x}) \triangleq \|\mathbf{G}\mathbf{x}\|_1$. The subgradient of $g(\mathbf{x})$ at $\mathbf{x}^t$ can be computed as $\mathbf{g}^t \triangleq \mathbf{G}^T\text{sign}(\mathbf{G}\mathbf{x}^t)$. $\nabla f(\mathbf{x})$ is L-Lipschitz with $L = 1$ and coordinate-wise Lipschitz with $\mathbf{c} = \mathbf{1}$. We set $\alpha = 1$.

We compare with the following methods. ***(i)*** Multi-Stage Convex Relaxation (MSCR). It generates the new iterate using: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} f(\mathbf{x}) - \langle \mathbf{x} - \mathbf{x}^t, \mathbf{g}^t\rangle$. ***(ii)*** Toland's dual method (T-DUAL). It rewrite the problem as: $\min_{-\mathbf{1}\leq\mathbf{y}\leq\mathbf{1}} \min_{\mathbf{x}} \; f(\mathbf{x}) - \langle \mathbf{G}\mathbf{x}, \mathbf{y}\rangle$. Setting the gradient of $\mathbf{x}$ to zero, we have: $\alpha\mathbf{x} - \mathbf{G}^T\mathbf{y} = \mathbf{0}$, leading to the following dual problem: $\min_{-\mathbf{1}\leq\mathbf{y}\leq\mathbf{1}} -\frac{1}{2\alpha}\mathbf{y}^T\mathbf{G}\mathbf{G}^T\mathbf{y}$. Toland's dual method uses the iteration: $\mathbf{y}^{t+1} = \text{sign}(\mathbf{G}\mathbf{G}^T\mathbf{y}^t)$, and recovers the primal solution via $\mathbf{x} = \frac{1}{\alpha}\mathbf{G}^T\mathbf{y}$. Note that the method in (Kim and Klabjan 2019) is essentially the Toland's duality method and they consider a constrained problem: $\min_{\|\mathbf{x}\|=1} -\|\mathbf{G}\mathbf{x}\|_1$. ***(iii)*** Subgradient method (SubGrad). It generates the new iterate via: $\mathbf{x}^{t+1} = \mathbf{x}^t - \frac{0.1}{t} \cdot (\nabla f(\mathbf{x}^t) - \mathbf{g}^t)$. ***(iv) CD-SCA*** solves a convex problem: $\bar{\eta}^t = \arg\min_{\eta} \frac{\mathbf{c}_i+\theta}{2}\eta^2 + (\nabla_{i^t} f(\mathbf{x}^t) - \mathbf{g}^t_{i^t})\eta$ and update $\mathbf{x}^t$ via $\mathbf{x}^{t+1}_{i^t} = \mathbf{x}^t_{i^t} + \bar{\eta}^t$. ***(v) CD-SNCA*** computes the nonconvex proximal operator of ℓ_1 norm (see Section 6.1) as: $\bar{\eta}^t = \arg\min_{\eta} \frac{\mathbf{c}_i+\theta}{2}\eta^2 + \nabla_{i^t} f(\mathbf{x}^t)\eta - \|\mathbf{G}(\mathbf{x} + \eta e_i)\|_1$ and update $\mathbf{x}^t$ via $\mathbf{x}^{t+1}_{i^t} = \mathbf{x}^t_{i^t} + \bar{\eta}^t$.

As can be seen from Table 1, the proposed method ***CD-SNCA*** consistently gives the best performance. Such results are not surprising since ***CD-SNCA*** is guaranteed to find stronger stationary points than the other methods (while ***CD-SNCA*** finds a coordinate-wise stationary point, all the other methods only find critical points).

7.3 Computational Efficiency

Figure 2 shows the convergence curve for solving the ℓ_p norm generalized eigenvalue problem. All methods take about 30 seconds to converge. ***CD-SNCA*** generally takes a little more time to converge than the other methods. However, we argue that the computational time is acceptable and pays off as ***CD-SNCA*** generally achieves higher accuracy.

8 Conclusions

We present CD methods for solving DC functions using sequential nonconvex approximation and sequential convex approximation. A novel optimality analysis and a novel convergence analysis for the CD methods are provided. The proposed ***CD-SNCA*** exploits specific structures of the DC function to escape bad local minima and finds stronger stationary points. It has shown superior performance than other existing methods both theoretically and experimentally.

Acknowledgments

This work was supported by NSFC (12271278, 61772570) and Guangdong Natural Science Funds for Distinguished Young Scholar (2018B030306025).

References

Ahmadi, A. A.; and Hall, G. 2018. DC decomposition of nonconvex polynomials with algebraic techniques. *Mathematical Programming*, 169(1): 69–94.

Attouch, H.; Bolte, J.; Redont, P.; and Soubeyran, A. 2010. Proximal Alternating Minimization and Projection Methods for Nonconvex Problems: An Approach Based on the Kurdyka-Lojasiewicz Inequality. *Mathematics of Operations Research*, 35(2): 438–457.

Beck, A.; and Eldar, Y. C. 2013. Sparsity Constrained Nonlinear Optimization: Optimality Conditions and Algorithms. *SIAM Journal on Optimization*, 23(3): 1480–1509.

Beck, A.; and Hallak, N. 2020. On the Convergence to Stationary Points of Deterministic and Randomized Feasible Descent Directions Methods. *SIAM Journal on Optimization*, 30(1): 56–79.

Beck, A.; and Teboulle, M. 2021. Dual Randomized Coordinate Descent Method for Solving a Class of Nonconvex Problems. *SIAM Journal on Optimization*, 31(3): 1877–1896.

Beck, A.; and Tetruashvili, L. 2013. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4): 2037–2060.

Bi, S.; Liu, X.; and Pan, S. 2014. Exact penalty decomposition method for zero-norm minimization based on MPEC formulation. *SIAM Journal on Scientific Computing*, 36(4): A1451–A1477.

Böhm, A.; and Wright, S. J. 2021. Variable Smoothing for Weakly Convex Composite Functions. *Journal of Optimization Theory and Applications*, 188(3): 628–649.

Breheny, P.; and Huang, J. 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1): 232.

Candès, E. J.; Li, X.; and Soltanolkotabi, M. 2015. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *IEEE Transactions on Information Theory*, 61(4): 1985–2007.

Davis, D.; Drusvyatskiy, D.; MacPhee, K. J.; and Paquette, C. 2018. Subgradient Methods for Sharp Weakly Convex Functions. *Journal of Optimization Theory and Applications*, 179(3): 962–982.

Davis, D.; and Grimmer, B. 2019. Proximally Guided Stochastic Subgradient Method for Nonsmooth, Nonconvex Problems. *SIAM Journal on Optimization*, 29(3): 1908–1930.

Deng, Q.; and Lan, C. 2020. Efficiency of coordinate descent methods for structured nonconvex optimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 74–89. Springer.

Dong, H.; and Tao, M. 2021. On the Linear Convergence to Weak/Standard d-Stationary Points of DCA-Based Algorithms for Structured Nonsmooth DC Programming. *J. Optim. Theory Appl.*, 189(1): 190–220.

Gong, P.; Zhang, C.; Lu, Z.; Huang, J.; and Ye, J. 2013. A General Iterative Shrinkage and Thresholding Algorithm for Non-convex Regularized Optimization Problems. In *International Conference on Machine Learning (ICML)*, volume 28, 37–45.

Gotoh, J.; Takeda, A.; and Tono, K. 2018. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169(1): 141–176.

Horst, R.; and Thoai, N. V. 1999. DC programming: overview. *Journal of Optimization Theory and Applications*, 103(1): 1–43.

Horst, R.; and Tuy, H. 2013. *Global optimization: Deterministic approaches*. Springer Science & Business Media.

Hsieh, C.-J.; Chang, K.-W.; Lin, C.-J.; Keerthi, S. S.; and Sundararajan, S. 2008. A dual coordinate descent method for large-scale linear SVM. In *International Conference on Machine Learning (ICML)*, 408–415.

Hsieh, C.-J.; and Dhillon, I. S. 2011. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 1064–1072.

Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S. M.; and Jordan, M. I. 2017. How to Escape Saddle Points Efficiently. In *International Conference on Machine Learning (ICML)*, volume 70, 1724–1732.

Joki, K.; Bagirov, A. M.; Karmitsa, N.; and Mäkelä, M. M. 2017. A proximal bundle method for nonsmooth DC optimization utilizing nonconvex cutting planes. *Journal of Global Optimization*, 68(3): 501–535.

Joki, K.; Bagirov, A. M.; Karmitsa, N.; Makela, M. M.; and Taheri, S. 2018. Double bundle method for finding Clarke stationary points in nonsmooth DC programming. *SIAM Journal on Optimization*, 28(2): 1892–1919.

Jr., G. D. F. 1972. Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. *IEEE Transactions on Information Theory*, 18(3): 363–378.

Kim, C.; and Klabjan, D. 2019. A simple and fast algorithm for L1-norm kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8): 1842–1855.

Kiryo, R.; Niu, G.; Du Plessis, M. C.; and Sugiyama, M. 2017. Positive-unlabeled learning with non-negative risk estimator. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Li, X.; Chen, S.; Deng, Z.; Qu, Q.; Zhu, Z.; and Man-Cho So, A. 2021. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3): 1605–1634.

Li, Y.; Lu, J.; and Wang, Z. 2019. Coordinatewise descent methods for leading eigenvalue problem. *SIAM Journal on Scientific Computing*, 41(4): A2681–A2716.

Liu, J.; Wright, S. J.; Ré, C.; Bittorf, V.; and Sridhar, S. 2015. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, 16(285-322): 1–5.

Lu, Z.; and Xiao, L. 2015. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2): 615–642.

Lu, Z.; and Zhou, Z. 2019. Nonmonotone Enhanced Proximal DC Algorithms for a Class of Structured Nonsmooth DC Programming. *SIAM Journal on Optimization*, 29(4): 2725–2752.

Luo, Z.-Q.; and Tseng, P. 1993. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1): 157–178.

Maingé, P.-E.; and Moudafi, A. 2008. Convergence of new inertial proximal methods for DC programming. *SIAM Journal on Optimization*, 19(1): 397–413.

Mairal, J. 2013. Optimization with First-Order Surrogate Functions. In *International Conference on Machine Learning (ICML)*, volume 28, 783–791.

Necoara, I. 2013. Random coordinate descent algorithms for multi-agent convex optimization over networks. *IEEE Transactions on Automatic Control*, 58(8): 2001–2012.

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

Nitanda, A.; and Suzuki, T. 2017. Stochastic Difference of Convex Algorithm and its Application to Training Deep Boltzmann Machines. In Singh, A.; and Zhu, X. J., eds., *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, 470–478. PMLR.

Pang, J.; Razaviyayn, M.; and Alvarado, A. 2017. Computing B-Stationary Points of Nonsmooth DC Programs. *Mathematics of Operations Research*, 42(1): 95–118.

Patrascu, A.; and Necoara, I. 2015. Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization. *Journal of Global Optimization*, 61(1): 19–46.

Razaviyayn, M.; Hong, M.; and Luo, Z. 2013. A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization. *SIAM Journal on Optimization*, 23(2): 1126–1153.

Richtárik, P.; and Takávc, M. 2014. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2): 1–38.

Shechtman, Y.; Beck, A.; and Eldar, Y. C. 2014. GESPAR: Efficient phase retrieval of sparse signals. *IEEE Transactions on Signal Processing*, 62(4): 928–938.

Sriperumbudur, B. K.; Torres, D. A.; and Lanckriet, G. R. G. 2007. Sparse eigen methods by D.C. programming. In *International Conference on Machine Learning (ICML)*, volume 227, 831–838.

Tao, P. D.; and An, L. T. H. 1997. Convex analysis approach to DC programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1): 289–355.

Thi, H. A. L.; and Dinh, T. P. 2018. DC programming and DCA: thirty years of developments. *Math. Program.*, 169(1): 5–68.

Toland, J. F. 1979. A duality principle for non-convex optimisation and the calculus of variations. *Archive for Rational Mechanics and Analysis*, 71(1): 41–61.

Tseng, P.; and Yun, S. 2009. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1): 387–423.

Xu, Y.; Qi, Q.; Lin, Q.; Jin, R.; and Yang, T. 2019. Stochastic Optimization for DC Functions and Non-smooth Non-convex Regularizers with Non-asymptotic Convergence. In Chaudhuri, K.; and Salakhutdinov, R., eds., *International Conference on Machine Learning (ICML)*, volume 97, 6942–6951.

Xu, Y.; and Yin, W. 2013. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3): 1758–1789.

Yuan, G. 2023. Coordinate Descent Methods for DC Minimization: Optimality Conditions and Global Convergence. *arXiv,* https://arxiv.org/abs/2109.04228.

Yuan, G.; and Ghanem, B. 2016. Sparsity Constrained Minimization via Mathematical Programming with Equilibrium Constraints. In *arXiv:1608.04430.*

Yuan, G.; and Ghanem, B. 2017. An Exact Penalty Method for Binary Optimization Based on MPEC Formulation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2867–2875.

Yuan, G.; Shen, L.; and Zheng, W.-S. 2020. A block decomposition algorithm for sparse optimization. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 275–285.

Yuan, X.; Li, P.; and Zhang, T. 2017. Gradient Hard Thresholding Pursuit. *Journal of Machine Learning Research*, 18: 166:1–166:43.

Yue, M.; Zhou, Z.; and So, A. M. 2019. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo-Tseng error bound property. *Math. Program.*, 174(1-2): 327–358.

Zhang, T. 2010. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11: 1081–1107.

Zhang, X.; Yu, Y.; Wang, L.; and Gu, Q. 2019. Learning one-hidden-layer relu networks via gradient descent. In *International Conference on Artificial Intelligence and Statistics*, 1524–1534.