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):

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

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):

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

$\forall \mathbf{x}, \eta, i = 1, ..., n$. 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:

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

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 *nonseparable* 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 D-C 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). ***(v)*** We also provide several important discussions of the proposed method (See Section D in the Appendix).

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. $\mathrm{diag}(\mathbf{c})$ is a diagonal matrix with $\mathbf{c}$ as the main diagonal entries. We define $\|\mathbf{d}\|_{\mathbf{c}}^2 = \sum_i \mathbf{c}_i \mathbf{d}_i^2$. $\mathrm{sign}(\cdot)$ 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: $F'(\mathbf{x}; \mathbf{d}) \triangleq \lim_{t \downarrow 0} \frac{1}{t}(F(\mathbf{x} + t\mathbf{d}) - F(\mathbf{x}))$. $\mathrm{dist}(\Omega, \Omega') \triangleq \inf_{\mathbf{v} \in \Omega, \mathbf{v}' \in \Omega'} \|\mathbf{v} - \mathbf{v}'\|$ 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:

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

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

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

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 Section D.1 in 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):

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

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):

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

$$\min_{\|\mathbf{x}\|_\infty \le 1} \ \tfrac{1}{2}\|\mathbf{G}\mathbf{x}-\mathbf{y}\|_2^2 + \rho(\sqrt{n} - \|\mathbf{x}\|). \tag{7}$$

• **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 Non-Convex 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:

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

► **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):

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

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.

$$\bar{\eta}^t \in \bar{\mathcal{M}}_{i^t}(\mathbf{x}^t) \triangleq \arg\min_{\eta} \ \mathcal{M}_{i^t}(\mathbf{x}^t, \eta) \tag{13}$$
$$\text{with } \mathcal{M}_i(\mathbf{x}, \eta) \triangleq \mathcal{S}_i(\mathbf{x}, \eta) + h_i(\mathbf{x} + \eta e_i) - g(\mathbf{x} + \eta e_i) + \tfrac{\theta}{2}\|(\mathbf{x} + \eta e_i) - \mathbf{x}\|_2^2$$

• Option II: Sequential Convex Approximation (***SCA***) strategy.

$$\bar{\eta}^t \in \bar{\mathcal{P}}_{i^t}(\mathbf{x}^t) \triangleq \arg\min_{\eta} \ \mathcal{P}_{i^t}(\mathbf{x}^t, \eta) \tag{14}$$
$$\mathcal{P}_i(\mathbf{x}, \eta) \triangleq \mathcal{S}_i(\mathbf{x}, \eta) + h_i(\mathbf{x} + \eta e_i) + \mathcal{G}_i(\mathbf{x}, \eta) + \tfrac{\theta}{2}\|(\mathbf{x} + \eta e_i) - \mathbf{x}\|_2^2$$

(**S3**) $\mathbf{x}^{t+1} = \mathbf{x}^t + \bar{\eta}^t \cdot e_{i^t} \ \ (\Leftrightarrow \mathbf{x}_{i^t}^{t+1} = \mathbf{x}_{i^t}^t + \bar{\eta}^t)$
(**S4**) Increment t by 1.
end while

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 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 5.11 and Theorem 5.13, 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 Section A in the appendix.

We introduce the following useful definition.

Definition 5.1. (**Globally or Locally Bounded Nonconvexity**) 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 $\rho < +\infty$. 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: $\xi^{t-1} \triangleq \{i^0, \ i^1, ..., i^{t-1}\}$.

We now develop the following technical lemma that will be used to analyze Algorithm 1 subsequently.

Lemma 5.3. *For any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{d} \in \mathbb{R}^n$, $\bar{\mathbf{c}} \in \mathbb{R}^n$, we define $h'(\mathbf{x}) \triangleq \sum_{i=1}^n h(\mathbf{x} + \mathbf{d}_i e_i)$, $g'(\mathbf{x}) \triangleq \sum_{i=1}^n g(\mathbf{x} + \mathbf{d}_i e_i)$, and $f'(\mathbf{x}) \triangleq \sum_{i=1}^n f(\mathbf{x} + \mathbf{d}_i e_i)$. We have:*

$$\textstyle\sum_{i=1}^n \|\mathbf{x} + \mathbf{d}_i e_i\|_{\mathbf{c}}^2 = \|\mathbf{x} + \mathbf{d}\|_{\mathbf{c}}^2 + (n-1)\|\mathbf{x}\|_{\mathbf{c}}^2 \tag{15}$$
$$h'(\mathbf{x}) = h(\mathbf{x} + \mathbf{d}) + (n-1)h(\mathbf{x}) \tag{16}$$
$$f'(\mathbf{x}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{d}\rangle + \tfrac{1}{2}\|\mathbf{d}\|_{\mathbf{c}}^2 + (n-1)f(\mathbf{x}) \tag{17}$$
$$-g'(\mathbf{x}) \leq -g(\mathbf{x}) - \langle \partial g(\mathbf{x}), \mathbf{d}\rangle - (n-1)g(\mathbf{x}) \tag{18}$$

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 5.4. (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 5.5. (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 5.6. (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 grow condition* for any coordinate-wise stationary point.

Lemma 5.7. *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 $\grave{\mathbf{x}}$ is said to be a local minima if $F(\grave{\mathbf{x}}) \leq F(\grave{\mathbf{x}} + \mathbf{d})$ 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 5.8. ***(Optimality Hierarchy between the Optimality Conditions).*** *Assume that the assumption made in Lemma 5.7 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: $\min_x (x-1)^2 - 4|x|$. This problem contains three critical points $\{-1, 0, 3\}$, two directional points / local minima $\{-1, 3\}$, and a unique coordinate-wise stationary point $\{3\}$. This unique coordinate-wise stationary point can be found using a clever breakpoint searching method (discussed later in Section 6).

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 5.9. (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 5.10. (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 5.6.

Theorem 5.11. *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 $T \leq \lceil \frac{2n(F(\mathbf{x}^0) - F(\bar{\mathbf{x}}))}{\theta\epsilon} \rceil = O(\epsilon^{-1})$. **(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 $T \leq \lceil \frac{2n(F(\mathbf{x}^0) - F(\bar{\mathbf{x}}))}{\beta\epsilon} \rceil = O(\epsilon^{-1})$.*

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 5.8).

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 5.12. (*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 5.13. ***(Convergence Rate for CD-SNCA)***. *Let $\ddot{\mathbf{x}}$ be any coordinate-wise stationary point. We define $\ddot{q}^t \triangleq F(\mathbf{x}^t) - F(\ddot{\mathbf{x}})$, $\ddot{r}^t \triangleq \frac{1}{2}\|\mathbf{x}^t - \ddot{\mathbf{x}}\|^2_{\bar{\mathbf{c}}}$, $\bar{\mathbf{c}} \triangleq \mathbf{c} + \theta$, $\bar{\rho} = \frac{\rho}{\min(\bar{\mathbf{c}})}$, $\gamma \triangleq 1 + \frac{\rho}{\theta}$, 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 5.14. ***(Convergence Rate for CD-SCA)***. *Let $\check{\mathbf{x}}$ be any critical point. We define $\check{q}^t \triangleq F(\mathbf{x}^t) - F(\check{\mathbf{x}})$, $\check{r}^t \triangleq \frac{1}{2}\|\mathbf{x}^t - \check{\mathbf{x}}\|^2_{\bar{\mathbf{c}}}$, $\bar{\mathbf{c}} \triangleq \mathbf{c} + \theta$, 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).

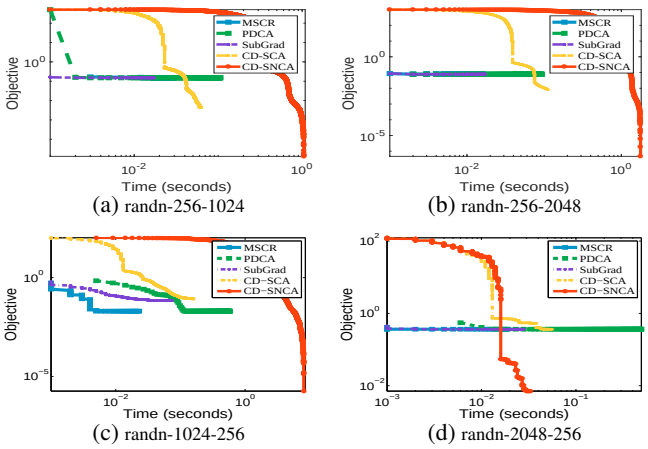
7 Experiments

This section demonstrates the effectiveness and efficiency of Algorithm 1 on two statistical learning tasks, namely the ℓ_p norm generalized eigenvalue problem and the approximate sparse optimization problem. For more experiments, please refer to Section C in the Appendix.

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 supplemental material.



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

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

Figure 1: The convergence curve of the compared methods for solving the ℓ_p norm generalized eigenvalue problem on different data sets.

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 $g(\mathbf{x}) \triangleq \|\mathbf{G}\mathbf{x}\|_1$. 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: $\min_{-\mathbf{1}\le\mathbf{y}\le\mathbf{1}} \min_{\mathbf{x}} f(\mathbf{x}) - \langle \mathbf{G}\mathbf{x}, \mathbf{y} \rangle$. 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: $\min_{-\mathbf{1}\le\mathbf{y}\le\mathbf{1}} -\frac{1}{2\alpha}\mathbf{y}^T\mathbf{G}\mathbf{G}^T\mathbf{y}$. 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).

	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 1^{st}, 2^{nd}, and 3^{rd} best results are colored with **red**, green and blue, respectively.

	MSCR	PDCA	SubGrad	***CD-SCA***	***CD-SNCA***
randn-256-1024	0.090 ± 0.017	0.090 ± 0.016	0.775 ± 0.040	0.092 ± 0.018	**0.034 ± 0.004**
randn-256-2048	0.052 ± 0.009	0.052 ± 0.010	1.485 ± 0.030	0.061 ± 0.012	**0.027 ± 0.002**
randn-1024-256	1.887 ± 0.353	1.884 ± 0.352	2.215 ± 0.379	1.881 ± 0.337	**1.681 ± 0.346**
randn-2048-256	3.795 ± 0.518	3.794 ± 0.518	4.127 ± 0.525	3.772 ± 0.522	**3.578 ± 0.484**
e2006-256-1024	0.217 ± 0.553	0.217 ± 0.553	0.597 ± 0.391	0.218 ± 0.556	**0.087 ± 0.212**
e2006-256-2048	0.050 ± 0.068	0.050 ± 0.068	0.837 ± 0.209	0.050 ± 0.068	**0.025 ± 0.032**
e2006-1024-256	3.078 ± 2.928	3.078 ± 2.928	3.112 ± 2.844	3.097 ± 2.960	**2.697 ± 2.545**
e2006-2048-256	1.799 ± 1.453	1.799 ± 1.453	1.918 ± 1.518	1.805 ± 1.456	**1.688 ± 1.398**
randn-256-1024-C	0.086 ± 0.012	0.087 ± 0.012	0.775 ± 0.038	0.083 ± 0.011	**0.033 ± 0.002**
randn-256-2048-C	0.043 ± 0.006	0.044 ± 0.006	1.472 ± 0.027	0.051 ± 0.009	**0.026 ± 0.001**
randn-1024-256-C	1.997 ± 0.250	1.998 ± 0.250	2.351 ± 0.297	1.979 ± 0.265	**1.781 ± 0.244**
randn-2048-256-C	3.618 ± 0.681	3.617 ± 0.682	3.965 ± 0.717	3.619 ± 0.679	**3.420 ± 0.673**
e2006-256-1024-C	0.031 ± 0.031	0.031 ± 0.031	0.339 ± 0.073	0.030 ± 0.028	**0.015 ± 0.014**
e2006-256-2048-C	0.217 ± 0.575	0.217 ± 0.575	0.596 ± 0.418	0.215 ± 0.568	**0.071 ± 0.176**
e2006-1024-256-C	3.789 ± 4.206	3.798 ± 4.213	3.955 ± 4.363	3.851 ± 4.339	**3.398 ± 3.855**
e2006-2048-256-C	4.480 ± 6.916	4.482 ± 6.918	4.710 ± 7.292	4.461 ± 6.844	**4.200 ± 6.608**

Table 2: Comparisons of objective values of all the methods for solving the approximate sparse optimization problem. The 1^{st}, 2^{nd}, and 3^{rd} best results are colored with **red**, green and blue, respectively.

7.3 Approximate Sparse Optimization

We consider solving Problem (6). To generate the original signal $\ddot{\mathbf{x}}$ of s-sparse structure, we randomly select a support set S with $|S| = 200$ and set $\ddot{\mathbf{x}}_{\{1,\dots,n\}\setminus S} = \mathbf{0}$, $\ddot{\mathbf{x}}_S = \text{randn}(|S|, 1)$. The observation vector is generated via $\mathbf{y} = \mathbf{A}\ddot{\mathbf{x}} + \text{randn}(m, 1) \times 0.1 \times \|\mathbf{A}\ddot{\mathbf{x}}\|$. This problem is consistent with Problem (1) with $f(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{G}\mathbf{x} - \mathbf{y}\|_2^2$, $h(\mathbf{x}) \triangleq \rho\|\mathbf{x}\|_1$, and $g(\mathbf{x}) \triangleq \rho\sum_{i=1}^s |\mathbf{x}^t_{[i]}|$. $\nabla f(\mathbf{x})$ is L-Lipschitz with $L = \|\mathbf{G}\|_2^2$ and coordinate-wise Lipschitz with $\mathbf{c}_i = (\mathbf{G}^T\mathbf{G})_{ii}$, $\forall i$. The subgradient of $g(\mathbf{x})$ at $\mathbf{x}^t$ can be computed as: $\mathbf{g}^t = \rho \cdot \arg\max_{\mathbf{y}} \langle \mathbf{y}, \mathbf{x}^t \rangle, s.t. \|\mathbf{y}\|_\infty \le 1, \|\mathbf{y}\|_1 \le k$. We set $\rho = 1$.

We compare with the following methods. ***(i)*** Multi-Stage Convex Relaxation (MSCR). It generate a sequence $\{\mathbf{x}^t\}$ as: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{G}\mathbf{x} - \mathbf{y}\|_2^2 + \rho\|\mathbf{x}\|_1 - \langle \mathbf{x} - \mathbf{x}^t, \mathbf{g}^t \rangle$. ***(ii)*** Proximal DC algorithm (PDCA). It generates the new iterate using: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \frac{L}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 + \langle \mathbf{x} - \mathbf{x}^t, \nabla f(\mathbf{x}) \rangle + \rho\|\mathbf{x}\|_1 - \langle \mathbf{x} - \mathbf{x}^t, \mathbf{g}^t \rangle$. ***(iii)*** Subgradient method (SubGrad). It uses the following iteration: $\mathbf{x}^{t+1} = \mathbf{x}^t - \frac{0.1}{t} \cdot (\nabla f(\mathbf{x}) + \rho\text{sign}(\mathbf{x}^t) - \mathbf{g}^t)$. ***(iv) CD-SCA*** solves a convex problem: $\bar{\eta}^t = \arg\min_{\eta} 0.5(\mathbf{c}_{i^t} + \theta)\eta^2 + \rho|\mathbf{x}^t_{i^t} + \eta| + [\nabla f(\mathbf{x}^t) - \mathbf{g}^t]_{i^t} \cdot \eta$ 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 the top-s norm function (see Section 6.2) as: $\bar{\eta}^t = \arg\min_{\eta} \frac{\mathbf{c}_i+\theta}{2}\eta^2 + \nabla_{i^t} f(\mathbf{x}^t)\eta + \rho|\mathbf{x}^t_{i^t} + \eta| - \rho\sum_{i=1}^s |(\mathbf{x}^t + \eta e_i)_{[i]}|$ 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 2, ***CD-SNCA*** consistently gives the best performance.

7.4 Computational Efficiency

Figure 1 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.

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.; 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.

Appendix

The appendix is organized as follows.
Section A presents the mathematical proofs for the theoretical analysis.
Section B shows more examples of the breakpoint searching methods for proximal operator computation.
Section C demonstrates some more experiments.
Section D provides some discussions of our methods.

A Mathematical Proofs

A.1 Proof for Lemma 5.2

Proof. Recall that the function $\tilde{z}(\mathbf{x}) \triangleq \|\mathbf{x}\|_p$ is convex when $p \geq 1$, and its subgradient w.r.t. $\mathbf{x}$ can be computed as $\partial\tilde{z}(\mathbf{x}) = \|\mathbf{x}\|_p^{1-p}\text{sign}(\mathbf{x})\odot|\mathbf{x}|^{p-1}$. Therefore, the function $z(\mathbf{x}) = -\|\mathbf{x}\|_p$ with $p \geq 1$ is concave, and $\partial z(\mathbf{x}) = -\|\mathbf{x}\|_p^{1-p}\text{sign}(\mathbf{x})\odot|\mathbf{x}|^{p-1}$.

As the two reference points are different with $\mathbf{x} \neq \mathbf{y}$, we assume that there exists a constant $\epsilon > 0$ satisfying $\|\mathbf{x}-\mathbf{y}\| \geq \epsilon$. We consider two cases for $p \geq 1$ and derive the following results.

(a) When $p \geq 2$, we have:

$$\begin{aligned}
& z(\mathbf{x}) - z(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \partial z(\mathbf{x})\rangle \\
\overset{(a)}{=} & -\|\mathbf{x}\|_p + \|\mathbf{y}\|_p + \langle \mathbf{x}-\mathbf{y}, \|\mathbf{x}\|_p^{1-p}\text{sign}(\mathbf{x})\odot|\mathbf{x}|^{p-1}\rangle \\
\overset{(b)}{\leq} & -\|\mathbf{x}\|_p + \|\mathbf{y}\|_p + \|\mathbf{x}\|_p^{1-p}\|\mathbf{y}-\mathbf{x}\|\|\text{sign}(\mathbf{x})\odot|\mathbf{x}|^{p-1}\| \\
\overset{(c)}{\leq} & \|\mathbf{y}-\mathbf{x}\|_p + \|\mathbf{y}-\mathbf{x}\|\|\mathbf{x}\|_p^{1-p}\|\mathbf{x}\|_p^{p-1} \\
= & \|\mathbf{x}-\mathbf{y}\|_p + \|\mathbf{x}-\mathbf{y}\|_2 \\
\overset{(d)}{\leq} & 2\|\mathbf{x}-\mathbf{y}\|_2 \\
= & \frac{4}{\epsilon}\cdot\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,
\end{aligned} \tag{19}$$

where step (a) uses $z(\mathbf{x}) = -\|\mathbf{x}\|_p$ and $\partial z(\mathbf{x}) = -\|\mathbf{x}\|_p^{1-p}\text{sign}(\mathbf{x})\odot|\mathbf{x}|^{p-1}$; step (b) uses the Cauchy-Schwarz inequality; step (c) uses triangle inequality and the fact that $\||\mathbf{x}|^{p-1}\|_2 \leq \|\mathbf{x}\|_p^{p-1}$ when $p \geq 2$; step (d) uses $\|\mathbf{x}-\mathbf{y}\|_p \leq \|\mathbf{x}-\mathbf{y}\|$ for all $p \geq 2$.

(b) When $1 \leq p < 2$, we have:

$$\begin{aligned}
& z(\mathbf{x}) - z(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \partial z(\mathbf{x})\rangle \\
= & -\|\mathbf{x}\|_p + \|\mathbf{y}\|_p + \langle \mathbf{x}-\mathbf{y}, \|\mathbf{x}\|_p^{1-p}\text{sign}(\mathbf{x})\odot|\mathbf{x}|^{p-1}\rangle \\
\leq & -\|\mathbf{x}\|_p + \|\mathbf{y}\|_p + \|\mathbf{x}\|_p^{1-p}\|\mathbf{y}-\mathbf{x}\|\|\text{sign}(\mathbf{x})\odot|\mathbf{x}|^{p-1}\| \\
\overset{(a)}{\leq} & \|\mathbf{y}-\mathbf{x}\|_p + \|\mathbf{x}-\mathbf{y}\|\|\mathbf{x}\|_p^{1-p}\|\mathbf{x}\|_p^{p-1}\cdot n^{1/p} \\
\overset{(b)}{\leq} & \|\mathbf{x}-\mathbf{y}\|\cdot n^{1/p} + \|\mathbf{x}-\mathbf{y}\|\cdot n^{1/p} \\
= & 2\|\mathbf{x}-\mathbf{y}\|\cdot n^{1/p} \\
\leq & \frac{4}{\epsilon}\cdot n^{1/p}\cdot\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,
\end{aligned} \tag{20}$$

where step (a) uses $\||\mathbf{x}|^{p-1}\|_2 \leq n^{1/p}\|\mathbf{x}\|_p^{p-1}$ for all $1 \leq p < 2$; step (b) uses $\|\mathbf{y}-\mathbf{x}\|_p \leq \|\mathbf{y}-\mathbf{x}\|\cdot n^{1/p}$.

Combining the two inequalities as in (19) and (20), we conclude that there exists $\rho < +\infty$ such that $z(\mathbf{x}) - z(\mathbf{y}) - \langle \mathbf{x}-\mathbf{y}, \partial z(\mathbf{x})\rangle \leq \frac{\rho}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ with $\rho = \begin{cases} 4/\epsilon, & p \geq 2; \\ n^{1/p}\cdot 4/\epsilon, & 1 \leq p \leq 1. \end{cases}$. In other words, $z(\mathbf{x}) = -\|\mathbf{x}\|_p$ is locally ρ-bounded nonconvex. □

A.2 Proof for Lemma 5.3

Proof. **(a)** For any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{d} \in \mathbb{R}^n$, and $\bar{\mathbf{c}} \in \mathbb{R}^n$, we derive the following equalities:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}+\mathbf{d}_i e_i\|_{\bar{\mathbf{c}}}^2 &= \frac{1}{n}\|\mathbf{d}\|_{\bar{\mathbf{c}}}^2 + \frac{2}{n}\langle \mathbf{x}, \bar{\mathbf{c}}\odot\mathbf{d}\rangle + \|\mathbf{x}\|_{\bar{\mathbf{c}}}^2 \\
&= \frac{1}{n}\|\mathbf{d}\|_{\bar{\mathbf{c}}}^2 + \frac{2}{n}\langle \mathbf{x}, \mathbf{d}\odot\bar{\mathbf{c}}\rangle + (\frac{1}{n}\|\mathbf{x}\|_{\bar{\mathbf{c}}}^2 - \frac{1}{n}\|\mathbf{x}\|_{\bar{\mathbf{c}}}^2) + \|\mathbf{x}\|_{\bar{\mathbf{c}}}^2 \\
&= \frac{1}{n}\|\mathbf{d}+\mathbf{x}\|_{\bar{\mathbf{c}}}^2 + (1-\frac{1}{n})\|\mathbf{x}\|_{\bar{\mathbf{c}}}^2.
\end{aligned}$$

(b) The proof for this equality is almost the same as Lemma 1 in (Lu and Xiao 2015). For completeness, we include the proof here. We have the following results:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}+\mathbf{d}_i e_i) &= \frac{1}{n}\sum_{i=1}^{n}\left(h_i(\mathbf{x}_i+\mathbf{d}_i)+\sum_{j\neq i} h_j(\mathbf{x}_j)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}(h_i(\mathbf{x}_i+\mathbf{d}_i))+\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} h_j(\mathbf{x}_j)\\ &= \frac{1}{n}h(\mathbf{x}+\mathbf{d})+\frac{n-1}{n}h(\mathbf{x}).\end{aligned}$$

(c) We obtain the following results:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}+\mathbf{d}_i e_i)\\ \overset{(a)}{\leq}\ &\frac{1}{n}\left(\sum_{i=1}^{n} f(\mathbf{x})+\langle\nabla f(\mathbf{x}),\mathbf{d}_i e_i\rangle+\frac{1}{2}\|\mathbf{d}\|_{\mathbf{c}}^2\right)\\ \overset{(b)}{=}\ &f(\mathbf{x})+\frac{1}{n}[\langle\nabla f(\mathbf{x}),\mathbf{d}\rangle+\frac{1}{2}\|\mathbf{d}\|_{\mathbf{c}}^2]\\ =\ &(1-\frac{1}{n})f(\mathbf{x})+\frac{1}{n}[f(\mathbf{x})+\langle\nabla f(\mathbf{x}),\mathbf{d}\rangle+\frac{1}{2}\|\mathbf{d}\|_{\mathbf{c}}^2],\end{aligned}$$

where step (a) uses the coordinate-wise Lipschitz continuity of $\nabla f(\mathbf{x})$ as in (2); step (b) uses $\sum_{i=1}^{n}\langle\nabla f(\mathbf{x}^t),\ \mathbf{d}_i e_i\rangle = \langle\nabla f(\mathbf{x}^t),\ \mathbf{d}\rangle$ and $\sum_{i=1}^{n}\mathbf{c}_i\mathbf{d}_i^2 = \|\mathbf{d}\|_{\mathbf{c}}^2$.
(d) We have the following inequalities:

$$\begin{aligned}-\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}+\mathbf{d}_i e_i) &\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\left(-g(\mathbf{x})-\langle\partial g(\mathbf{x}),\ \mathbf{d}_i e_i\rangle\right),\\ &\overset{(b)}{=} -g(\mathbf{x})-\frac{1}{n}\langle\partial g(\mathbf{x}),\ \mathbf{d}\rangle,\end{aligned}\tag{21}$$

where step (a) uses the fact $g(\mathbf{x})$ is convex that $\forall\mathbf{x},\mathbf{y},\ -g(\mathbf{y})\leq -g(\mathbf{x})-\langle\partial g(\mathbf{x}),\ \mathbf{y}-\mathbf{x}\rangle$; step (b) uses $\sum_{i=1}^{n}\langle\mathbf{y},\mathbf{d}_i e_i\rangle = \langle\mathbf{y},\mathbf{d}\rangle$. □

A.3 Proof for Lemma 5.7

First, since $z(\mathbf{x})\triangleq -g(\mathbf{x})$ is locally ρ-bounded nonconvex at the point $\ddot{\mathbf{x}}$, we have:

$$-g(\ddot{\mathbf{x}}) \leq -g(\mathbf{y})-\langle\ddot{\mathbf{x}}-\mathbf{y},\ \partial g(\ddot{\mathbf{x}})\rangle+\frac{\rho}{2}\|\ddot{\mathbf{x}}-\mathbf{y}\|_2^2,\ \forall\mathbf{y}.$$

Applying the inequality above with $\mathbf{y}=\ddot{\mathbf{x}}+\mathbf{d}$ for any $\mathbf{d}\in\mathbb{R}^n$, we obtain:

$$\begin{aligned}\forall\mathbf{d},\ -g(\ddot{\mathbf{x}}) &\leq -g(\ddot{\mathbf{x}}+\mathbf{d})-\langle\ddot{\mathbf{x}}-(\ddot{\mathbf{x}}+\mathbf{d}),\ \partial g(\ddot{\mathbf{x}})\rangle+\frac{\rho}{2}\|\ddot{\mathbf{x}}-(\ddot{\mathbf{x}}+\mathbf{d})\|_2^2\\ &= -g(\ddot{\mathbf{x}}+\mathbf{d})+\langle\mathbf{d},\ \partial g(\ddot{\mathbf{x}})\rangle+\frac{\rho}{2}\|\mathbf{d}\|_2^2\\ &\overset{(a)}{\leq} -g(\ddot{\mathbf{x}}+\mathbf{d})+\sum_{i=1}^{n} g(\ddot{\mathbf{x}}+\mathbf{d}_i e_i)-ng(\ddot{\mathbf{x}})+\frac{\rho}{2}\|\mathbf{d}\|_2^2,\end{aligned}\tag{22}$$

where step (a) uses claim **(d)** in Lemma 5.3 that $-\sum_{i=1}^{n} g(\ddot{\mathbf{x}}+\mathbf{d}_i e_i)\leq -g(\mathbf{x})-\langle\partial g(\mathbf{x}),\mathbf{d}\rangle-(n-1)g(\mathbf{x})$.

Second, by the optimality of $\ddot{\mathbf{x}}$, we obtain:

$$h(\ddot{\mathbf{x}})-g(\ddot{\mathbf{x}})\leq\langle\mathbf{d}_i e_i,\ \nabla f(\ddot{\mathbf{x}})\rangle+\frac{\mathbf{c}_i+\theta}{2}\mathbf{d}_i^2+h(\ddot{\mathbf{x}}+\mathbf{d}_i e_i)-g(\ddot{\mathbf{x}}+\mathbf{d}_i e_i),\ \forall\mathbf{d}_i.$$

Summing the inequality above over $i = 1, ..., n$, we have:

$$\begin{aligned}
\forall \mathbf{d},\ 0 \quad &\leq \quad ng(\ddot{\mathbf{x}}) - nh(\ddot{\mathbf{x}}) + \frac{1}{2}\|\mathbf{d}\|^2_{(\mathbf{c}+\theta)} + \langle \mathbf{d},\ \nabla f(\ddot{\mathbf{x}})\rangle + \sum_{i=1}^{n} h(\ddot{\mathbf{x}} + \mathbf{d}_i e_i) - \sum_{i=1}^{n} g(\ddot{\mathbf{x}} + \mathbf{d}_i e_i) \\
&\overset{(a)}{\leq} \quad \frac{1}{2}\|\mathbf{d}\|^2_{(\mathbf{c}+\theta)} + \langle \mathbf{d},\ \nabla f(\ddot{\mathbf{x}})\rangle - h(\ddot{\mathbf{x}}) + h(\ddot{\mathbf{x}} + \mathbf{d}) + g(\ddot{\mathbf{x}}) - g(\ddot{\mathbf{x}} + \mathbf{d}) + \frac{\rho}{2}\|\mathbf{d}\|_2^2 \\
&\overset{(b)}{\leq} \quad \frac{1}{2}\|\mathbf{d}\|^2_{(\mathbf{c}+\theta)} + f(\ddot{\mathbf{x}} + \mathbf{d}) - f(\ddot{\mathbf{x}}) - h(\ddot{\mathbf{x}}) + h(\ddot{\mathbf{x}} + \mathbf{d}) + g(\ddot{\mathbf{x}}) - g(\ddot{\mathbf{x}} + \mathbf{d}) + \frac{\rho}{2}\|\mathbf{d}\|_2^2 \\
&\overset{(c)}{=} \quad \frac{1}{2}\|\mathbf{d}\|^2_{(\mathbf{c}+\theta)} + F(\ddot{\mathbf{x}} + \mathbf{d}) - F(\ddot{\mathbf{x}}) + \frac{\rho}{2}\|\mathbf{d}\|_2^2,
\end{aligned}$$

where step (a) uses (16) in Lemma 5.3 and (22); step (b) uses the convexity of $f(\cdot)$ that:

$$\forall \mathbf{d},\ \langle \nabla f(\ddot{\mathbf{x}}), (\ddot{\mathbf{x}} + \mathbf{d}) - \ddot{\mathbf{x}}\rangle \leq f(\ddot{\mathbf{x}} + \mathbf{d}) - f(\ddot{\mathbf{x}});$$

step (c) uses the definition of $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x}) - g(\mathbf{x})$. Rearranging terms, we obtain:

$$\forall \mathbf{d},\ F(\ddot{\mathbf{x}}) \leq F(\ddot{\mathbf{x}} + \mathbf{d}) + \frac{1}{2}\|\mathbf{d}\|^2_{(\mathbf{c}+\theta+\rho)}.$$

A.4 Proof for Theorem 5.8

Proof. **(a)** We show that any optimal point $\bar{\mathbf{x}}$ is a coordinate-wise stationary point $\ddot{\mathbf{x}}$, i.e., $\{\bar{\mathbf{x}}\} \subseteq \{\ddot{\mathbf{x}}\}$. By the optimality of $\bar{\mathbf{x}}$, we have:

$$f(\bar{\mathbf{x}}) + h(\bar{\mathbf{x}}) - g(\bar{\mathbf{x}}) \leq f(\mathbf{x}) + h(\mathbf{x}) - g(\mathbf{x}),\ \forall \mathbf{x}.$$

Letting $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{d}_i e_i$, we have:

$$\begin{aligned}
f(\bar{\mathbf{x}}) + h(\bar{\mathbf{x}}) - g(\bar{\mathbf{x}}) \quad &\leq \quad f(\bar{\mathbf{x}} + \mathbf{d}_i e_i) + h(\bar{\mathbf{x}} + \mathbf{d}_i e_i) - g(\bar{\mathbf{x}} + \mathbf{d}_i e_i),\ \forall \mathbf{d}_i, \forall i \\
&\overset{(a)}{\leq} \quad f(\bar{\mathbf{x}}) + \langle \nabla_i f(\bar{\mathbf{x}}),\ \mathbf{d}_i e_i\rangle + \frac{\mathbf{c}_i}{2}\mathbf{d}_i^2 + h(\bar{\mathbf{x}} + \mathbf{d}_i e_i) - g(\bar{\mathbf{x}} + \mathbf{d}_i e_i),\ \forall \mathbf{d}_i, \forall i, \qquad (23)
\end{aligned}$$

where step (a) uses the coordinate-wise Lipschitz continuity of $\nabla f(\cdot)$ that:

$$f(\bar{\mathbf{x}} + \mathbf{d}_i e_i) \leq f(\bar{\mathbf{x}}) + \langle \nabla_i f(\bar{\mathbf{x}}),\ \mathbf{d}_i e_i\rangle + \frac{\mathbf{c}_i}{2}\mathbf{d}_i^2,\ \forall \mathbf{d}_i.$$

We denote $\bar{\eta}_i$ as the minimizer of the following problem:

$$\forall i,\ \bar{\eta}_i \in \arg\min_{\eta}\ \mathcal{M}_i(\bar{\mathbf{x}}, \eta).$$

Rearranging terms for (23) and using the fact that $\theta \geq 0$, we have:

$$\begin{aligned}
& h(\bar{\mathbf{x}}) - g(\bar{\mathbf{x}}) \leq h(\bar{\mathbf{x}} + \mathbf{d}_i e_i) - g(\bar{\mathbf{x}} + \mathbf{d}_i e_i) + \langle \nabla_i f(\bar{\mathbf{x}}),\ \mathbf{d}_i e_i\rangle + \frac{\mathbf{c}_i + \theta}{2}\mathbf{d}_i^2,\ \forall \mathbf{d}_i, \forall i \\
\overset{(a)}{\Rightarrow} \quad & h(\bar{\mathbf{x}}) - g(\bar{\mathbf{x}}) \leq h(\bar{\mathbf{x}} + \bar{\eta}_i e_i) - g(\bar{\mathbf{x}} + \mathbf{d}_i e_i) + \langle \nabla_i f(\bar{\mathbf{x}}),\ \bar{\eta}_i e_i\rangle + \frac{\mathbf{c}_i + \theta}{2}(\bar{\eta}_i)^2, \forall i \\
\overset{(b)}{\Rightarrow} \quad & \mathcal{M}_i(\bar{\mathbf{x}}, 0) \leq \min_{\eta}\ \mathcal{M}_i(\bar{\mathbf{x}}, \eta), \forall i, \\
\Rightarrow \quad & 0 \in \bar{\mathcal{M}}_i(\bar{\mathbf{x}}), \forall i.
\end{aligned}$$

where step (a) uses the choice $\mathbf{d}_i = \bar{\eta}_i$ for all i; step (b) uses the fact that $\forall i,\ \mathcal{M}_i(\bar{\mathbf{x}}, 0) = h(\bar{\mathbf{x}}) - g(\bar{\mathbf{x}})$ and the definition of $\min_\eta\ \mathcal{M}_i(\bar{\mathbf{x}}, \eta)$. Therefore, any optimal point $\bar{\mathbf{x}}$ is also a coordinate-wise stationary point $\ddot{\mathbf{x}}$.

(b) We show that any coordinate-wise stationary point $\ddot{\mathbf{x}}$ is a directional point $\dot{\mathbf{x}}$, i.e., $\{\ddot{\mathbf{x}}\} \subseteq \{\dot{\mathbf{x}}\}$. Applying the inequality in Lemma 5.7 with $\mathbf{d} = t(\mathbf{y} - \ddot{\mathbf{x}})$, we directly obtain the following results:

$$\begin{aligned}
\lim_{t \downarrow 0} \frac{F(\ddot{\mathbf{x}} + t(\mathbf{y} - \ddot{\mathbf{x}})) - F(\ddot{\mathbf{x}})}{t} \quad &\geq \quad \lim_{t \downarrow 0} \frac{-\frac{1}{2}\|t(\mathbf{y} - \ddot{\mathbf{x}})\|^2_{(\mathbf{c}+\theta+\rho)}}{t} \\
&\overset{(a)}{=} \quad \lim_{t \downarrow 0} \frac{-t^2\frac{1}{2}\|\mathbf{y} - \ddot{\mathbf{x}}\|^2_{(\mathbf{c}+\theta+\rho)}}{t} = 0,
\end{aligned}$$

where step (a) uses the boundedness of ρ that $\rho < +\infty$. Therefore, any coordinate-wise stationary point $\ddot{\mathbf{x}}$ is also a directional point $\dot{\mathbf{x}}$.

(c) We show that any directional point $\grave{\mathbf{x}}$ is a critical point $\breve{\mathbf{x}}$, i.e., $\{\grave{\mathbf{x}}\} \subseteq \{\breve{\mathbf{x}}\}$. Noticing $f(\mathbf{x})$, $h(\mathbf{x})$, and $g(\mathbf{x})$ are convex, we have:

$$\begin{aligned} f(\mathbf{z}) &\leq f(\grave{\mathbf{x}}) - \langle \grave{\mathbf{x}} - \mathbf{z},\ \nabla f(\mathbf{z})\rangle,\ \forall \mathbf{z}, \\ h(\mathbf{z}) &\leq h(\grave{\mathbf{x}}) - \langle \grave{\mathbf{x}} - \mathbf{z},\ \partial h(\mathbf{z})\rangle,\ \forall \mathbf{z}, \\ -g(\mathbf{z}) &\leq -g(\grave{\mathbf{x}}) - \langle \mathbf{z} - \grave{\mathbf{x}},\ \partial g(\grave{\mathbf{x}})\rangle,\ \forall \mathbf{z}. \end{aligned}$$

Adding these three inequalities together, we obtain:

$$F(\mathbf{z}) - F(\grave{\mathbf{x}}) \leq \langle \mathbf{z} - \grave{\mathbf{x}}, -\partial g(\grave{\mathbf{x}}) + \nabla f(\mathbf{z}) + \partial h(\mathbf{z})\rangle,\ \forall \mathbf{z}. \tag{24}$$

We derive the following inequalities:

$$\begin{aligned} \forall \mathbf{y} \in \mathrm{dom}(F),\ 0 \quad &\leq \quad \lim_{t\downarrow 0} \frac{F(\grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}})) - F(\grave{\mathbf{x}})}{t} \\ &\overset{(a)}{\leq} \quad \lim_{t\downarrow 0} \frac{\langle (\grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}})) - \grave{\mathbf{x}}, -\partial g(\grave{\mathbf{x}}) + \nabla f(\grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}})) + \partial h(\grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}}))\rangle}{t}, \\ &= \quad \lim_{t\downarrow 0} \langle \mathbf{y} - \grave{\mathbf{x}}, -\partial g(\grave{\mathbf{x}}) + \nabla f(\grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}})) + \partial h(\grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}}))\rangle, \\ &\overset{(b)}{=} \quad \lim_{t\downarrow 0} \langle \mathbf{y} - \grave{\mathbf{x}}, \partial F(\grave{\mathbf{x}})\rangle, \end{aligned}$$

where step (a) uses (24) with $\mathbf{z} = \grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}})$; step (b) uses $\grave{\mathbf{x}} + t(\mathbf{y} - \grave{\mathbf{x}}) = \grave{\mathbf{x}}$ as $t \downarrow 0$. Noticing the inequality above holds for all $\mathbf{y}$ only when $0 \in \partial F(\grave{\mathbf{x}})$, we conclude that any directional point $\grave{\mathbf{x}}$ is also a critical point $\breve{\mathbf{x}}$.

□

A.5 Proof for Theorem 5.11

Proof. **(a)** We now focus on ***CD-SNCA***. Since $\bar{\eta}^t$ is the global optimal solution to Problem (13), we have:

$$\begin{aligned} & f(\mathbf{x}^t) + \langle \bar{\eta}^t e_{i^t},\ \nabla f(\mathbf{x}^t)\rangle + \frac{\mathbf{c}_{i^t} + \theta}{2}(\bar{\eta}^t)^2 + h(\mathbf{x}^t + \bar{\eta}^t e_{i^t}) - g(\mathbf{x}^t + \bar{\eta}^t e_{i^t}) \\ \leq\ & f(\mathbf{x}^t) + \langle \eta e_{i^t},\ \nabla f(\mathbf{x}^t)\rangle + \frac{\mathbf{c}_{i^t} + \theta}{2}\eta^2 + h(\mathbf{x}^t + \eta e_{i^t}) - g(\mathbf{x}^t + \eta e_{i^t}), \forall \eta. \end{aligned}$$

Letting $\eta = 0$ and using the fact that $\mathbf{x}^{t+1} = \mathbf{x}^t + \bar{\eta}^t \cdot e_{i^t}$, we obtain:

$$f(\mathbf{x}^t) + \langle \mathbf{x}^{t+1} - \mathbf{x}^t,\ \nabla f(\mathbf{x}^t)\rangle + \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2_{\mathbf{c}+\theta} + h(\mathbf{x}^{t+1}) - g(\mathbf{x}^{t+1}) \leq F(\mathbf{x}^t) \tag{25}$$

We derive the following results:

$$\begin{aligned} & F(\mathbf{x}^{t+1}) - F(\mathbf{x}^t) \\ \overset{(a)}{\leq}\ & F(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) - \langle \mathbf{x}^{t+1} - \mathbf{x}^t, \nabla f(\mathbf{x}^t)\rangle - \tfrac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2_{\mathbf{c}+\theta} - h(\mathbf{x}^{t+1}) + g(\mathbf{x}^{t+1}), \\ \overset{(b)}{=}\ & f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) - \langle \mathbf{x}^{t+1} - \mathbf{x}^t,\ \nabla f(\mathbf{x}^t)\rangle - \tfrac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2_{\mathbf{c}+\theta}, \\ \overset{(c)}{\leq}\ & -\frac{\theta}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|_2^2, \end{aligned}$$

where step (a) uses (25); step (b) uses the definition $F(\mathbf{x}^{t+1}) = f(\mathbf{x}^{t+1}) + h(\mathbf{x}^{t+1}) - g(\mathbf{x}^{t+1})$; step (c) uses the coordinate-wise Lipschitz continuity of $\nabla f(\cdot)$.

Taking the expectation for the inequality above, we obtain a lower bound on the expected progress made by each iteration for ***CD-SNCA***:

$$\mathbb{E}[F(\mathbf{x}^{t+1})] - F(\mathbf{x}^t) \leq -\mathbb{E}[\tfrac{\theta}{2n}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2]$$

Summing up the inequality above over $t = 0, 1, ..., T-1$, we have:

$$\mathbb{E}[\frac{\theta}{2}\sum_{t=0}^{T-1}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2] \leq n\mathbb{E}[F(\mathbf{x}^0) - F(\mathbf{x}^{T+1})] \leq n\mathbb{E}[F(\mathbf{x}^0) - F(\bar{\mathbf{x}})]$$

As a result, there exists an index $\bar{t}$ with $0 \leq \bar{t} \leq T-1$ such that:

$$\mathbb{E}[\|\mathbf{x}^{\bar{t}+1} - \mathbf{x}^{\bar{t}}\|_2^2] \leq \frac{2n(F(\mathbf{x}^0) - F(\bar{\mathbf{x}}))}{\theta T}. \tag{26}$$

Furthermore, for any t, we have:

$$\mathbb{E}[\|\mathbf{x}^{\bar{t}+1} - \mathbf{x}^{\bar{t}}\|_2^2] = \frac{1}{n}\sum_{i=1}^{n}(\bar{\mathcal{M}}_i(\mathbf{x}^{\bar{t}}))^2. \tag{27}$$

Combining the two inequalities above, we have the following result:

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{\mathcal{M}}_i(\mathbf{x}^{\bar{t}}))^2 \leq \frac{2n(F(\mathbf{x}^0) - F(\bar{\mathbf{x}}))}{\theta T}. \tag{28}$$

Therefore, we conclude that ***CD-SNCA*** finds an ϵ-approximate coordinate-wise stationary point 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) We now focus on ***CD-SCA***. Since $\bar{\eta}^t$ is the global optimal solution to Problem (14), we have:

$$0 \in [\nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1}) - \partial g(\mathbf{x}^{t+1})]_{i^t} + (\mathbf{c}_{i^t} + \theta)\bar{\eta}^t. \tag{29}$$

Using the coordinate-wise Lipschitz continuity of $\nabla f(\cdot)$, we obtain:

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) + \langle \mathbf{x}^{t+1} - \mathbf{x}^t,\ \nabla f(\mathbf{x}^t)\rangle + \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\mathbf{c}}^2 \tag{30}$$

Since both $h(\cdot)$ and $g(\cdot)$ are convex, we have:

$$h(\mathbf{x}^{t+1}) \leq h(\mathbf{x}^t) - \langle \mathbf{x}^t - \mathbf{x}^{t+1},\ \nabla h(\mathbf{x}^{t+1})\rangle, \tag{31}$$

$$-g(\mathbf{x}^{t+1}) \leq -g(\mathbf{x}^t) - \langle \partial g(\mathbf{x}^t), \mathbf{x}^{t+1} - \mathbf{x}^t\rangle. \tag{32}$$

Adding these three inequalities in (30), (31), and (32) together, we have:

$$\begin{aligned}
& F(\mathbf{x}^{t+1}) - F(\mathbf{x}^t) \\
\leq\ & \langle \mathbf{x}^{t+1} - \mathbf{x}^t,\ \nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1}) - \partial g(\mathbf{x}^t)\rangle + \tfrac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\mathbf{c}}^2 \\
\overset{(a)}{=}\ & \langle \bar{\eta}^t e_{i^t},\ \nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1}) - \partial g(\mathbf{x}^t)\rangle + \frac{\mathbf{c}_{i^t}}{2}\|\bar{\eta}^t e_{i^t}\|_2^2 \\
=\ & \bar{\eta}^t(\nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1}) - \partial g(\mathbf{x}^t))_{i^t} + \frac{\mathbf{c}_i}{2}(\bar{\eta}^t)^2 \\
\overset{(b)}{=}\ & -\frac{\mathbf{c} + 2\theta}{2}(\bar{\eta}^t)^2 \\
\overset{(c)}{\leq}\ & -\frac{\min(\mathbf{c}) + 2\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2,
\end{aligned}$$

where step (a) uses the fact that $\mathbf{x}^{t+1} - \mathbf{x}^t = \bar{\eta}^t e_{i^t}$; step (b) uses (29); step (c) uses $(\bar{\eta}^t)^2 = \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2$.

Using similar strategies as in deriving the results for ***CD-SNCA***, we conclude that Algorithm 1 finds an ϵ-approximate critical point of Problem (1) in at most T iterations in the sense of expectation, where $T \leq \lceil \frac{2n(F(\mathbf{x}^0) - F(\bar{\mathbf{x}}))}{\beta\epsilon} \rceil = O(\epsilon^{-1})$.

□

A.6 Proof for Theorem 5.13

Proof. Let $\ddot{\mathbf{x}}$ be any coordinate-wise stationary point. First, the optimality condition for the nonconvex subproblem as in (13) can be written as:

$$0 \in \nabla_{i^t} f(\mathbf{x}^t) + \bar{\mathbf{c}}_{i^t}\bar{\eta}^t + \partial_{i^t} h(\mathbf{x}^{t+1}) - \partial_{i^t} g(\mathbf{x}^{t+1}). \tag{33}$$

Second, for any $\mathbf{x}^t$, $\mathbf{x}^{t+1}$, and $\ddot{\mathbf{x}}$, since i^t is chosen uniformly and randomly, we have:

$$\mathbb{E}[\|\mathbf{x}^{t+1} - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2] = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}^t + (\mathbf{x}^{t+1} - \mathbf{x}^t)_i e_i - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2. \tag{34}$$

Applying the inequality in (15) with $\mathbf{x} = \mathbf{x}^t - \ddot{\mathbf{x}}$ and $\mathbf{d} = \mathbf{x}^{t+1} - \mathbf{x}^t$, we have:

$$\frac{1}{n}\|\mathbf{x}^{t+1} - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2 + \frac{1}{n}(n-1)\|\mathbf{x}^t - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2 = \frac{1}{n}\sum_{i=1}^{n}\|(\mathbf{x}^t - \ddot{\mathbf{x}}) + (\mathbf{x}^{t+1} - \mathbf{x}^t)_i e_i\|_{\bar{\mathbf{c}}}^2. \tag{35}$$

Combining the two inequalities in (34) and (35), we have:

$$\mathbb{E}[\|\mathbf{x}^{t+1}-\ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2]=\frac{1}{n}\|\mathbf{x}^{t+1}-\ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2+\frac{1}{n}(n-1)\|\mathbf{x}^{t}-\ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2. \tag{36}$$

Third, since $z(\mathbf{x})\triangleq -g(\mathbf{x})$ is globally ρ-bounded nonconvex, we have

$$\forall \mathbf{x},\mathbf{y},\ -g(\mathbf{x})\le -g(\mathbf{y})-\langle \mathbf{x}-\mathbf{y},\ \partial g(\mathbf{x})\rangle+\frac{\rho}{2}\|\mathbf{x}-\mathbf{y}\|_2^2.$$

Applying this inequality with $\mathbf{y}=\mathbf{x}+\mathbf{d}$, we have:

$$\begin{aligned}\forall \mathbf{x},\forall \mathbf{d}, -g(\mathbf{x})+g(\mathbf{x}+\mathbf{d})-\frac{\rho}{2}\|\mathbf{d}\|_2^2 &\le \langle \mathbf{d},\ \partial g(\mathbf{x})\rangle\\ &\overset{(a)}{\le} \sum_{i=1}^n g(\mathbf{x}+\mathbf{d}_i e_i)-g(\mathbf{x})-(n-1)g(\mathbf{x})\\ &= \sum_{i=1}^n g(\mathbf{x}+\mathbf{d}_i e_i)-ng(\mathbf{x}),\end{aligned} \tag{37}$$

where step (a) uses (18) in Lemma 5.3.
We apply (16), (17), and (37) with $\mathbf{x}=\mathbf{x}^t$ and $\mathbf{d}=\mathbf{x}^{t+1}-\mathbf{x}^t$, and obtain the following inequalities:

$$\mathbb{E}[h(\mathbf{x}^{t+1})] = \frac{1}{n}h(\mathbf{x}^{t+1})+(1-\frac{1}{n})h(\mathbf{x}^t) \tag{38}$$

$$\mathbb{E}[f(\mathbf{x}^{t+1})] \le \frac{1}{n}f(\mathbf{x}^t)+\frac{1}{n}\langle\nabla f(\mathbf{x}^t),\mathbf{x}^{t+1}-\mathbf{x}^t\rangle+\frac{1}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2+(1-\frac{1}{n})f(\mathbf{x}^t) \tag{39}$$

$$-\mathbb{E}[g(\mathbf{x}^{t+1})] \le -\frac{1}{n}g(\mathbf{x}^{t+1})+\tfrac{\rho}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2-(1-\frac{1}{n})g(\mathbf{x}^t). \tag{40}$$

(a) We derive the following results:

$$\begin{aligned}&\mathbb{E}[\frac{1}{2}\|\mathbf{x}^{t+1}-\ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2]-\frac{1}{2}\|\mathbf{x}^t-\ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2\\ \overset{(a)}{=}\ &\mathbb{E}[\langle \mathbf{x}^{t+1}-\ddot{\mathbf{x}},\bar{\mathbf{c}}\odot(\mathbf{x}^{t+1}-\mathbf{x}^t)\rangle]-\mathbb{E}[\frac{1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2]\\ \overset{(b)}{=}\ &\mathbb{E}[\langle \mathbf{x}^{t+1}-\ddot{\mathbf{x}},-(\nabla_{i^t}f(\mathbf{x}^t)+\partial_{i^t}h(\mathbf{x}^{t+1})-\partial_{i^t}g(\mathbf{x}^{t+1}))\cdot e_{i^t}\rangle]-\mathbb{E}[\frac{1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2]\\ \overset{(c)}{=}\ &\frac{1}{n}\langle \ddot{\mathbf{x}}-\mathbf{x}^{t+1},\nabla f(\mathbf{x}^t)+\partial h(\mathbf{x}^{t+1})-\partial g(\mathbf{x}^{t+1})\rangle-\frac{1}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2,\end{aligned} \tag{41}$$

where step (a) uses uses the Pythagoras relation that: $\forall \bar{\mathbf{c}},\mathbf{x},\mathbf{y},\mathbf{z}$, $\frac{1}{2}\|\mathbf{y}-\mathbf{z}\|_{\bar{\mathbf{c}}}^2-\frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_{\bar{\mathbf{c}}}^2=\langle \mathbf{y}-\mathbf{z},\bar{\mathbf{c}}\odot(\mathbf{y}-\mathbf{x})\rangle-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_{\bar{\mathbf{c}}}^2$; step (b) uses the optimality condition in (33); step (c) uses the fact that $\mathbb{E}[\langle \mathbf{x}_{i^t}e_{i^t},\mathbf{y}\rangle]=\frac{1}{n}\sum_{j=1}^n \mathbf{x}_j\mathbf{y}_j=\frac{1}{n}\langle \mathbf{x},\mathbf{y}\rangle$.
We now bound the term $\frac{1}{n}\langle \ddot{\mathbf{x}}-\mathbf{x}^{t+1},-\partial g(\mathbf{x}^{t+1})\rangle$ in (41) by the following inequalities:

$$\begin{aligned}&\frac{1}{n}\langle \ddot{\mathbf{x}}-\mathbf{x}^{t+1},-\partial g(\mathbf{x}^{t+1})\rangle\\ \overset{(a)}{\le}\ &-\frac{1}{n}g(\ddot{\mathbf{x}})+\frac{1}{n}g(\mathbf{x}^{t+1})+\frac{\rho}{2n}\|\ddot{\mathbf{x}}-\mathbf{x}^{t+1}\|_2^2\\ \overset{(b)}{\le}\ &-\frac{1}{n}g(\ddot{\mathbf{x}})+\frac{1}{n}g(\mathbf{x}^{t+1})+\frac{\bar{\rho}}{2n}\|\ddot{\mathbf{x}}-\mathbf{x}^{t+1}\|_{\bar{\mathbf{c}}}^2\\ \overset{(c)}{=}\ &-\frac{1}{n}g(\ddot{\mathbf{x}})+\frac{1}{n}g(\mathbf{x}^{t+1})+\frac{\bar{\rho}}{2}\left(\mathbb{E}[\|\ddot{\mathbf{x}}-\mathbf{x}^{t+1}\|_{\bar{\mathbf{c}}}^2]-(1-\frac{1}{n})\|\ddot{\mathbf{x}}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2\right)\\ \overset{(d)}{\le}\ &-\frac{1}{n}g(\ddot{\mathbf{x}})+\mathbb{E}[g(\mathbf{x}^{t+1})]+\frac{\rho}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2-(1-\frac{1}{n})g(\mathbf{x}^t)+\frac{\bar{\rho}}{2}\left(\mathbb{E}[\|\ddot{\mathbf{x}}-\mathbf{x}^{t+1}\|_{\bar{\mathbf{c}}}^2]-(1-\frac{1}{n})\|\ddot{\mathbf{x}}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2\right),\end{aligned} \tag{42}$$

where step (a) uses the globally ρ-bounded nonconvexity of $-g(\cdot)$; step (b) uses the fact that $\rho\|\mathbf{v}\|_2^2\le\rho\cdot\frac{1}{\min(\bar{\mathbf{c}})}\|\mathbf{v}\|_{\bar{\mathbf{c}}}^2=\bar{\rho}\|\mathbf{v}\|_{\bar{\mathbf{c}}}^2$ for all $\mathbf{v}$; step (c) uses (36); step (d) uses (40).

We now bound the term $\frac{1}{n}\langle \ddot{\mathbf{x}} - \mathbf{x}^{t+1}, \nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1})\rangle$ in (41) by the following inequalities:

$$
\begin{aligned}
& \frac{1}{n}\langle \ddot{\mathbf{x}} - \mathbf{x}^{t+1}, \nabla f(\mathbf{x}^t) + \partial h(\mathbf{x}^{t+1})\rangle \\
= \quad & \left(\frac{1}{n}\langle \ddot{\mathbf{x}} - \mathbf{x}^{t}, \nabla f(\mathbf{x}^t)\rangle + \frac{1}{n}\langle \mathbf{x}^t - \mathbf{x}^{t+1}, \nabla f(\mathbf{x}^t)\rangle\right) + \frac{1}{n}\langle \ddot{\mathbf{x}} - \mathbf{x}^{t+1}, \partial h(\mathbf{x}^{t+1})\rangle \\
\overset{(a)}{\leq} \quad & \frac{1}{n}\left(f(\ddot{\mathbf{x}}) - f(\mathbf{x}^t)\right) - \frac{1}{n}\langle \mathbf{x}^{t+1} - \mathbf{x}^t, \nabla f(\mathbf{x}^t)\rangle + \frac{1}{n}\left(h(\ddot{\mathbf{x}}) - h(\mathbf{x}^{t+1})\right) \\
\overset{(b)}{\leq} \quad & \frac{1}{n}f(\ddot{\mathbf{x}}) - \mathbb{E}[f(\mathbf{x}^{t+1})] + \frac{1}{2n}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\mathbf{c}}^2 + (1 - \frac{1}{n})f(\mathbf{x}^t) + \frac{1}{n}\left(h(\ddot{\mathbf{x}}) - h(\mathbf{x}^{t+1})\right) \\
\overset{(c)}{\leq} \quad & \frac{1}{n}f(\ddot{\mathbf{x}}) - \mathbb{E}[f(\mathbf{x}^{t+1})] + \tfrac{1}{2n}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\mathbf{c}}^2 + (1 - \frac{1}{n})f(\mathbf{x}^t) + \frac{1}{n}h(\ddot{\mathbf{x}}) - \mathbb{E}[h(\mathbf{x}^{t+1})] + (1 - \frac{1}{n})h(\mathbf{x}^t), \qquad (43)
\end{aligned}
$$

where step (a) uses the convexity of $f(\cdot)$ and $h(\cdot)$ that:

$$
\begin{aligned}
\langle \ddot{\mathbf{x}} - \mathbf{x}^t, \nabla f(\mathbf{x}^t)\rangle &\leq f(\ddot{\mathbf{x}}) - f(\mathbf{x}^t), \\
\langle \ddot{\mathbf{x}} - \mathbf{x}^{t+1}, \partial h(\mathbf{x}^{t+1})\rangle &\leq h(\ddot{\mathbf{x}}) - h(\mathbf{x}^{t+1});
\end{aligned}
$$

step (b) uses (39); step (c) uses (38).

Combining (41), (42), and (43) together, and using the fact that $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x}) - g(\mathbf{x})$, we obtain:

$$
\begin{aligned}
& \mathbb{E}[\frac{1-\bar{\rho}}{2}\|\mathbf{x}^{t+1} - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2] - \frac{1 - \bar{\rho} + \frac{\bar{\rho}}{n}}{2}\|\mathbf{x}^t - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2 \\
\leq \quad & \frac{1}{n}(F(\ddot{\mathbf{x}}) - F(\mathbf{x}^t)) - \mathbb{E}[F(\mathbf{x}^{t+1})] + F(\mathbf{x}^t) + \mathbb{E}[\frac{\rho}{2n}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2] \\
\overset{(a)}{\leq} \quad & \frac{1}{n}(F(\ddot{\mathbf{x}}) - F(\mathbf{x}^t)) - \mathbb{E}[F(\mathbf{x}^{t+1})] + F(\mathbf{x}^t) + \frac{\rho}{\theta}\mathbb{E}[F(\mathbf{x}^t) - F(\mathbf{x}^{t+1})] \\
\overset{(b)}{=} \quad & -\frac{1}{n}\ddot{q}^t + (1 + \frac{\rho}{\theta})(\ddot{q}^t - \mathbb{E}[\ddot{q}^{t+1}]), \qquad (44)
\end{aligned}
$$

where step (a) uses the sufficient decrease condition that $\mathbb{E}[\frac{1}{2n}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2] \leq \frac{1}{\theta}\mathbb{E}[F(\mathbf{x}^t) - F(\mathbf{x}^{t+1})]$; step (b) uses the definition of $\ddot{q}^t \triangleq F(\mathbf{x}^t) - F(\ddot{\mathbf{x}})$ and the fact that $F(\mathbf{x}^t) - F(\mathbf{x}^{t+1}) = \ddot{q}^t - \ddot{q}^{t+1}$. Using the definitions that $\ddot{r}^{t+1} \triangleq \frac{1}{2}\|\mathbf{x}^t - \ddot{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2$, $\varpi \triangleq 1 - \bar{\rho}$, and $\gamma \triangleq (1 + \frac{\rho}{\theta})$, we rewrite (44) as:

$$
\begin{aligned}
& \mathbb{E}[(1-\bar{\rho})\ddot{r}^{t+1}] - (1 - \bar{\rho} + \tfrac{\bar{\rho}}{n})\ddot{r}^t \leq -\frac{1}{n}\ddot{q}^t + (1 + \frac{\rho}{\theta})(\ddot{q}^t - \mathbb{E}[\ddot{q}^{t+1}]) \\
\Leftrightarrow \quad & \varpi\mathbb{E}[\ddot{r}^{t+1}] + \gamma\mathbb{E}[\ddot{q}^{t+1}] \leq (\varpi + \frac{\bar{\rho}}{n})\ddot{r}^t + (\gamma - \frac{1}{n})\ddot{q}^t. \qquad (45)
\end{aligned}
$$

(b) We now discuss the situation when $\varpi \geq 0$. We notice that the function $\mathcal{S}_{i^t}(\mathbf{x}, \eta) + h_{i^t}(\mathbf{x} + \eta e_{i^t}) + \frac{\theta}{2}\|(\mathbf{x} + \eta e_{i^t}) - \mathbf{x}\|_2^2$ is $(\min(\mathbf{c}) + \theta)$-strongly convex w.r.t. η and the term $-g(\mathbf{x} + \eta e_{i^t})$ is globally ρ-bounded nonconvex w.r.t. η for all t. Therefore, $\mathcal{M}_{i^t}(\mathbf{x}^t, \eta)$ in (13) is convex if:

$$
\min(\mathbf{c}) + \theta - \rho \geq 0 \quad \Leftrightarrow \quad \varpi \geq 0.
$$

We now discuss the case when $F(\cdot)$ satisfies the Luo-Tseng error bound assumption. We bound the term $\ddot{r}^t$ using the following

inequalities:

$$\begin{aligned}
\ddot{r}^t &\triangleq \max(\bar{\mathbf{c}})\frac{1}{2}\|\mathbf{x}^t-\ddot{\mathbf{x}}\|_2^2 \\
&\overset{(a)}{\leq} \max(\bar{\mathbf{c}})\frac{1}{2}\frac{\delta^2}{n^2}(\sum_{i=1}^n|\bar{\mathcal{M}}_i(\mathbf{x}^t)|)^2 \\
&\overset{(b)}{\leq} \max(\bar{\mathbf{c}})\frac{1}{2}\frac{\delta^2}{n^2}n\cdot(\sum_{i=1}^n|\bar{\mathcal{M}}_i(\mathbf{x}^t)|^2) \\
&\overset{(c)}{\leq} \max(\bar{\mathbf{c}})\frac{1}{2}\frac{\delta^2}{n^2}n\cdot\left(n\mathbb{E}[\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2]\right) \\
&= \max(\bar{\mathbf{c}})\frac{\delta^2}{\theta}\cdot\frac{\theta}{2}\mathbb{E}[\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2] \\
&\overset{(d)}{\leq} \max(\bar{\mathbf{c}})\delta^2\frac{n}{\theta}(F(\mathbf{x}^t)-\mathbb{E}[F(\mathbf{x}^{t+1})]) \\
&= \max(\bar{\mathbf{c}})\delta^2\frac{n}{\theta}(\ddot{q}^t-\mathbb{E}[\ddot{q}^{t+1}]), \\
&\overset{(e)}{=} \kappa_0 n(\ddot{q}^t-\mathbb{E}[\ddot{q}^{t+1}]),
\end{aligned} \tag{46}$$

where step (a) uses Assumption 5.12 that $\|\mathbf{x}^t-\ddot{\mathbf{x}}\|_2^2\leq\delta^2(\frac{1}{n}\sum_{i=1}^n|\text{dist}(0,\bar{\mathcal{M}}_i(\mathbf{x}))|)^2$ for any coordinate-wise stationary point $\ddot{\mathbf{x}}$; step (b) uses the fact that $\forall\mathbf{x},\ \|\mathbf{x}\|_1^2\leq n\|\mathbf{x}\|_2^2$; step (c) uses the fact that $\mathbb{E}[\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2]=\mathbb{E}[\|(\mathbf{x}^t+\bar{\mathcal{M}}_{i^t}(\mathbf{x}^t)e_{i^t})-\mathbf{x}^t\|_2^2]=\mathbb{E}[|\bar{\mathcal{M}}_{i^t}(\mathbf{x}^t)|^2]=\frac{1}{n}\sum_{i=1}^n|\bar{\mathcal{M}}_i(\mathbf{x}^t)|^2$; step (d) uses the sufficient decrease condition that $\mathbb{E}[\frac{1}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2]\leq\frac{1}{\theta}\mathbb{E}[F(\mathbf{x}^t)-F(\mathbf{x}^{t+1})]$; step (e) uses the definition of $\kappa_0\triangleq\max(\bar{\mathbf{c}})\frac{\delta^2}{\theta}$. Since $\varpi\geq0$, we have form (45):

$$\begin{aligned}
&\gamma\mathbb{E}[\ddot{q}^{t+1}]\leq\left(\varpi+\frac{\bar{\rho}}{n}\right)\kappa_0 n(\ddot{q}^t-\mathbb{E}[\ddot{q}^{t+1}])+(\gamma-\frac{1}{n})\ddot{q}^t \\
\Rightarrow\ &\underbrace{(\gamma+\kappa_0 n(\varpi+\frac{\bar{\rho}}{n}))}_{\triangleq\kappa_1}\mathbb{E}[\ddot{q}^{t+1}]\leq\underbrace{(\gamma+\kappa_0 n(\varpi+\frac{\bar{\rho}}{n})}_{=\kappa_1}-\frac{1}{n})\ddot{q}^t \\
\Rightarrow\ &\mathbb{E}[\ddot{q}^{t+1}]\leq\frac{\kappa_1-\frac{1}{n}}{\kappa_1}\ddot{q}^t \\
\Rightarrow\ &\mathbb{E}[\ddot{q}^{t+1}]\leq(\frac{\kappa_1-\frac{1}{n}}{\kappa_1})^{t+1}\ddot{q}^0.
\end{aligned}$$

Thus, we finish the proof of this theorem.

□

A.7 Proof for Theorem 5.14

Proof. Let $\check{\mathbf{x}}$ be any coordinate-wise stationary point. First, the optimality condition for the convex subproblem as in (14) can be written as:

$$0\in\nabla_{i^t}f(\mathbf{x}^t)+\bar{\mathbf{c}}_{i^t}\bar{\eta}^t+\partial_{i^t}h(\mathbf{x}^{t+1})-\partial_{i^t}g(\mathbf{x}^t). \tag{47}$$

Second, we apply (16), (17), and (18) in Lemma 5.3 with $\mathbf{x}=\mathbf{x}^t$ and $\mathbf{d}=\mathbf{x}^{t+1}-\mathbf{x}^t$, and obtain the following inequalities:

$$\mathbb{E}[h(\mathbf{x}^{t+1})] = \frac{1}{n}h(\mathbf{x}^{t+1})+(1-\frac{1}{n})h(\mathbf{x}^t) \tag{48}$$

$$\mathbb{E}[f(\mathbf{x}^{t+1})] \leq \frac{1}{n}f(\mathbf{x}^t)+\frac{1}{n}\langle\nabla f(\mathbf{x}^t),\mathbf{x}^{t+1}-\mathbf{x}^t\rangle+\tfrac{1}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\mathbf{c}}^2+(1-\frac{1}{n})f(\mathbf{x}^t) \tag{49}$$

$$-\mathbb{E}[g(\mathbf{x}^{t+1})] \leq -\frac{1}{n}\langle\partial g(\mathbf{x}^t),\mathbf{x}^{t+1}-\mathbf{x}^t\rangle-g(\mathbf{x}^t). \tag{50}$$

(a) We derive the following results:

$$\begin{aligned}
& \mathbb{E}[\frac{1}{2}\|\mathbf{x}^{t+1}-\check{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2]-\frac{1}{2}\|\mathbf{x}^t-\check{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2 \\
\overset{(a)}{=} & \mathbb{E}[\langle \mathbf{x}^{t+1}-\check{\mathbf{x}}, \bar{\mathbf{c}}\odot(\mathbf{x}^{t+1}-\mathbf{x}^t)\rangle]-\mathbb{E}[\frac{1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2] \\
\overset{(b)}{=} & \mathbb{E}[\langle \mathbf{x}^{t+1}-\check{\mathbf{x}}, -(\nabla_{i^t} f(\mathbf{x}^t)+\partial_{i^t} h(\mathbf{x}^{t+1})-\partial_{i^t} g(\mathbf{x}^t))\cdot e_{i^t}\rangle]-\mathbb{E}[\frac{1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2] \\
\overset{(c)}{=} & \frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^{t+1}, \nabla f(\mathbf{x}^t)+\partial h(\mathbf{x}^{t+1})-\partial g(\mathbf{x}^t)\rangle-\frac{1}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\bar{\mathbf{c}}}^2, \qquad (51)
\end{aligned}$$

where step (a) uses uses the Pythagoras relation that: $\forall \bar{\mathbf{c}}, \mathbf{x}, \mathbf{y}, \mathbf{z}$, $\frac{1}{2}\|\mathbf{y}-\mathbf{z}\|_{\bar{\mathbf{c}}}^2-\frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_{\bar{\mathbf{c}}}^2=\langle \mathbf{y}-\mathbf{z}, \bar{\mathbf{c}}\odot(\mathbf{y}-\mathbf{x})\rangle-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_{\bar{\mathbf{c}}}^2$; step (b) uses the optimality condition in (47); step (c) uses the fact that $\mathbb{E}[\langle \mathbf{x}_{i^t} e_{i^t}, \mathbf{y}\rangle]=\frac{1}{n}\sum_{j=1}^n \mathbf{x}_j\mathbf{y}_j=\frac{1}{n}\langle \mathbf{x}, \mathbf{y}\rangle$.
We now bound the term $\frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^{t+1}, -\partial g(\mathbf{x}^t)\rangle$ in (51) by the following inequalities:

$$\begin{aligned}
& \frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^{t+1}, -\partial g(\mathbf{x}^t)\rangle \\
= & \frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^t, -\partial g(\mathbf{x}^t)\rangle+\frac{1}{n}\langle \mathbf{x}^t-\mathbf{x}^{t+1}, -\partial g(\mathbf{x}^t)\rangle \\
\overset{(a)}{\leq} & -\frac{1}{n}g(\check{\mathbf{x}})+\frac{1}{n}g(\mathbf{x}^t)+\frac{\rho}{2n}\|\check{\mathbf{x}}-\mathbf{x}^t\|_2^2+\frac{1}{n}\langle \mathbf{x}^{t+1}-\mathbf{x}^t, \partial g(\mathbf{x}^t) \\
\overset{(b)}{\leq} & -\frac{1}{n}g(\check{\mathbf{x}})+\frac{1}{n}g(\mathbf{x}^t)+\frac{\rho}{2n}\|\check{\mathbf{x}}-\mathbf{x}^t\|_2^2-g(\mathbf{x}^t)+\mathbb{E}[g(\mathbf{x}^{t+1})], \qquad (52)
\end{aligned}$$

where step (a) uses the globally ρ-bounded nonconvexity of $-g(\cdot)$; step (b) uses (50).
We now bound the term $\frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^{t+1}, \nabla f(\mathbf{x}^t)+\partial h(\mathbf{x}^{t+1})\rangle$ in (51) by the following inequalities:

$$\begin{aligned}
& \frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^{t+1}, \nabla f(\mathbf{x}^t)+\partial h(\mathbf{x}^{t+1})\rangle \\
= & \frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^t, \nabla f(\mathbf{x}^t)\rangle+\frac{1}{n}\langle \check{\mathbf{x}}-\mathbf{x}^{t+1}, \partial h(\mathbf{x}^{t+1})\rangle+\frac{1}{n}\langle \mathbf{x}^t-\mathbf{x}^{t+1}, \nabla f(\mathbf{x}^t)\rangle \\
\overset{(a)}{\leq} & \frac{1}{n}[f(\check{\mathbf{x}})-f(\mathbf{x}^t)+h(\check{\mathbf{x}})-h(\mathbf{x}^{t+1})]-\frac{1}{n}\langle \mathbf{x}^{t+1}-\mathbf{x}^t, \nabla f(\mathbf{x}^t)\rangle \\
\overset{(b)}{\leq} & \frac{1}{n}[f(\check{\mathbf{x}})+h(\check{\mathbf{x}})-h(\mathbf{x}^{t+1})]-\mathbb{E}[f(\mathbf{x}^{t+1})]+\tfrac{1}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\mathbf{c}}^2+(1-\frac{1}{n})f(\mathbf{x}^t) \\
\overset{(c)}{\leq} & \frac{1}{n}[f(\check{\mathbf{x}})+h(\check{\mathbf{x}})]-\mathbb{E}[h(\mathbf{x}^{t+1})]+(1-\frac{1}{n})h(\mathbf{x}^t)-\mathbb{E}[f(\mathbf{x}^{t+1})]+\tfrac{1}{2n}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\mathbf{c}}^2+(1-\frac{1}{n})f(\mathbf{x}^t), \qquad (53)
\end{aligned}$$

where step (a) uses the convexity of $f(\cdot)$ and $h(\cdot)$ that:

$$\begin{aligned}
\langle \check{\mathbf{x}}-\mathbf{x}^t, \nabla f(\mathbf{x}^t)\rangle &\leq f(\check{\mathbf{x}})-f(\mathbf{x}^t), \\
\langle \check{\mathbf{x}}-\mathbf{x}^{t+1}, \partial h(\mathbf{x}^{t+1})\rangle &\leq h(\check{\mathbf{x}})-h(\mathbf{x}^{t+1});
\end{aligned}$$

step (b) uses (49); step (c) uses (48).
Combining (51), (52), (53), and using the fact that $F(\mathbf{x})=f(\mathbf{x})+h(\mathbf{x})-g(\mathbf{x})$, we obtain:

$$\begin{aligned}
& \mathbb{E}[\frac{1}{2}\|\mathbf{x}^{t+1}-\check{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2]-\frac{1}{2}\|\mathbf{x}^t-\check{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2 \\
\leq & \frac{\rho}{2n}\|\mathbf{x}^t-\check{\mathbf{x}}\|_2^2+\frac{1}{n}F(\check{\mathbf{x}})-\mathbb{E}[F(\mathbf{x}^{t+1})]+(1-\frac{1}{n})F(\mathbf{x}^t) \\
\overset{(a)}{\leq} & \frac{\bar{\rho}}{2n}\|\mathbf{x}^t-\check{\mathbf{x}}\|_{\bar{\mathbf{c}}}^2+\frac{1}{n}F(\check{\mathbf{x}})-\mathbb{E}[F(\mathbf{x}^{t+1})]+(1-\frac{1}{n})F(\mathbf{x}^t), \qquad (54)
\end{aligned}$$

where step (a) uses $\|\mathbf{x}\|_2^2\leq\frac{1}{\min(\bar{\mathbf{c}})}\|\mathbf{x}\|_{\bar{\mathbf{c}}}^2, \forall \mathbf{x}$. The inequality in (54) can be rewritten as:

$$\mathbb{E}[\check{r}^{t+1}]+\mathbb{E}[\check{q}^{t+1}]\leq\check{r}^t+\frac{\bar{\rho}}{n}\check{r}^t-\frac{1}{n}\check{q}^t+\check{q}^t. \qquad (55)$$

(b) We now discuss the case when $F(\cdot)$ satisfies the Luo-Tseng error bound assumption. We first bound the term $\check{r}^t$:

$$\begin{aligned}\check{r}^t &\triangleq \max(\bar{\mathbf{c}})\frac{1}{2}\|\mathbf{x}^t-\check{\mathbf{x}}\|_2^2\\ &\overset{(a)}{\leq} \max(\bar{\mathbf{c}})\frac{1}{2}\frac{\delta^2}{n^2}(\sum_{i=1}^n|\bar{\mathcal{P}}_i(\mathbf{x}^t)|)^2\\ &\overset{(b)}{\leq} \kappa_0 n(\check{q}^t-\mathbb{E}[\check{q}^{t+1}]),\end{aligned}$$

where step (a) uses Assumption 5.12 with the residual function defining as $\mathcal{R}(\mathbf{x})=\frac{1}{n}\sum_{i=1}^n|\bar{\mathcal{P}}_i(\mathbf{x})|$; step (b) uses the same strategy as in deriving the results in (46). Finally, we have form (55):

$$\begin{aligned}&\mathbb{E}[\check{q}^{t+1}]\leq\left(1+\frac{\bar{\rho}}{n}\right)\kappa_0 n(\check{q}^t-\mathbb{E}[\check{q}^{t+1}])+(1-\frac{1}{n})\check{q}^t\\ \Rightarrow\quad &\underbrace{(1+(1+\frac{\bar{\rho}}{n})\kappa_0 n)}_{\triangleq\kappa_2}\mathbb{E}[\check{q}^{t+1}]\leq\underbrace{((1+\frac{\bar{\rho}}{n})\kappa_0 n+1}_{=\kappa_2}-\frac{1}{n})\check{q}^t\\ \Rightarrow\quad &\mathbb{E}[\check{q}^{t+1}]\leq\frac{\kappa_2-\frac{1}{n}}{\kappa_2}\check{q}^t\\ \Rightarrow\quad &\mathbb{E}[\check{q}^{t+1}]\leq(\frac{\kappa_2-\frac{1}{n}}{\kappa_2})^{t+1}\check{q}^0.\end{aligned}$$

Thus, we finish the proof of this theorem.

□

B More Examples of the Breakpoint Searching Method for Proximal Operator Computation

B.1 When $g(\mathbf{y})=\|\mathbf{A}\mathbf{y}\|_\infty$ and $h_i(\cdot)\triangleq 0$

Consider the problem: $\min_\eta\frac{a}{2}\eta^2+b\eta-\|\mathbf{A}(\mathbf{x}+\eta e_i)\|_\infty$. It can be rewritten as: $\min_\eta\frac{a}{2}\eta^2+b\eta-\|\mathbf{g}\eta+\mathbf{d}\|_\infty$. It is equivalent to $\min_\eta p(\eta)\triangleq\frac{a}{2}\eta^2+b\eta-\max_{i=1}^{2m}(\bar{\mathbf{g}}_i\eta+\bar{\mathbf{d}}_i)$ with $\bar{\mathbf{g}}=[\mathbf{g}_1,\mathbf{g}_2,...,\mathbf{g}_m,-\mathbf{g}_1,-\mathbf{g}_2,...,-\mathbf{g}_m]$ and $\bar{\mathbf{d}}=[\mathbf{d}_1,\mathbf{d}_2,...,\mathbf{d}_m,-\mathbf{d}_1,-\mathbf{d}_2,...,-\mathbf{d}_m]$. Setting the gradient of $p(\cdot)$ to zero yields: $a\eta+b+\bar{\mathbf{g}}_i=0$ with $i=1,2,...,(2m)$. We have $\boldsymbol{\eta}=(-b-\bar{\mathbf{g}})/a$. Therefore, Problem (3) contains $2m$ breakpoints $\Theta=\{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,...,\boldsymbol{\eta}_{2m}\}$ for this example.

B.2 When $g(\mathbf{y})=\|\max(0,\mathbf{A}\mathbf{y})\|_1$ and $h_i(\cdot)\triangleq 0$

Consider the problem: $\min_\eta\frac{a}{2}\eta^2+b\eta-\|\max(0,\mathbf{A}(\mathbf{x}+\eta e_i))\|_1$. Using the fact that $\max(0,a)=\frac{1}{2}(a+|a|)$, we have the following equivalent problem: $\min_\eta\frac{a}{2}\eta^2+b\eta-\frac{1}{2}\langle\mathbf{1},\mathbf{A}e_i\rangle\eta-\frac{1}{2}\|\mathbf{A}(\mathbf{x}+\eta e_i)\|_1$. Therefore, the proximal operator of $g(\mathbf{x})=\|\max(0,\mathbf{A}\mathbf{x})\|_1$ can be transformed to the proximal operator of $g(\mathbf{x})=\|\mathbf{A}\mathbf{x}\|_1$ with suitable parameters.

B.3 When $g(\mathbf{y})=\|\mathbf{A}\mathbf{y}\|_2$ and $h_i(\cdot)\triangleq 0$

Consider the problem: $\min_\eta\frac{a}{2}\eta^2+b\eta-\|\mathbf{A}(\mathbf{x}+\eta e_i)\|_p$. It can be rewritten as: $\min_\eta p(\eta)\triangleq\frac{a}{2}\eta^2+b\eta-\|\mathbf{g}\eta+\mathbf{d}\|_p$. Setting the gradient of $p(\cdot)$ to zero yields: $0=a\eta+b-\|\mathbf{g}\eta-\mathbf{d}\|_p^{1-p}\langle\mathbf{g},\text{sign}(\mathbf{g}\eta+\mathbf{d})\odot|\mathbf{g}\eta+\mathbf{d}|^{p-1}\rangle$. We only focus on $p=2$. We obtain: $a\eta+b=\frac{\langle\mathbf{g},\mathbf{g}\eta+\mathbf{d}\rangle}{\|\mathbf{g}\eta-\mathbf{d}\|}\Leftrightarrow\|\mathbf{g}\eta-\mathbf{d}\|(a\eta+b)=\langle\mathbf{g},\mathbf{g}\eta+\mathbf{d}\rangle\Leftrightarrow\|\mathbf{g}\eta-\mathbf{d}\|_2^2(a\eta+b)^2=(\langle\mathbf{g},\mathbf{g}\eta+\mathbf{d}\rangle)^2$. Solving this quartic equation we obtain all of its real roots $\{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,...,\boldsymbol{\eta}_c\}$ with $1\leq c\leq 4$. Therefore, Problem (3) at most contains 4 breakpoints $\Theta=\{\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,...,\boldsymbol{\eta}_c\}$ for this example.

C More Experiments

In this section, we present the experiment results of the approximate binary optimization problem and the generalized linear regression problem.

C.1 Approximate Binary Optimization

We consider Problem (7). We generate the observation vector via $\mathbf{y}=\max(0,\mathbf{A}\ddot{\mathbf{x}}+\text{randn}(m,1)\times 0.1\times\|\mathbf{A}\ddot{\mathbf{x}}\|)$ with $\ddot{\mathbf{x}}=\text{randn}(d,1)$. This problem is consistent with $f(\mathbf{x})\triangleq\frac{1}{2}\|\mathbf{G}\mathbf{x}-\mathbf{y}\|_2^2$, $g(\mathbf{x})\triangleq-\rho\|\mathbf{x}\|$, and $h(\mathbf{x})=\sum_i^n h_i(\mathbf{x}_i)$ with $h_i(z)\triangleq I_{[-1,1]}(z)$ where $I_{[-1,1]}(z)$ denotes an indicator function on the box constraint ($h_i(z)=0$ if $-1\leq z\leq 1$, $+\infty$ otherwise). We notice that $\nabla f(\cdot)$ is L-Lipschitz with constant $L=\|\mathbf{G}\|_2^2$ and coordinate-wise Lipschitz with constant $\mathbf{c}=\text{diag}(\mathbf{G}^T\mathbf{G})$. The subgradient of $g(\mathbf{x})$ at $\mathbf{x}^t$ can be computed as: $\partial g(\mathbf{x}^t)=-\frac{\rho\mathbf{x}^t}{\|\mathbf{x}^t\|}\triangleq\mathbf{g}^t$. We set $\rho=5$.

We compare with the following methods. ***(i)*** Multi-Stage Convex Relaxation (MSCR). It solves the following problem: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{G}\mathbf{x} - \mathbf{y}\|_2^2 - \langle \mathbf{x} - \mathbf{x}^t,\ \mathbf{g}^t\rangle, s.t. \|\mathbf{x}\|_\infty \leq 1$. This is essentially equivalent to the alternating minimization method in (Yuan and Ghanem 2017). ***(ii)*** Proximal DC algorithm (PDCA). It considers the following iteration: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \frac{L}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 + \langle \mathbf{x} - \mathbf{x}^t,\ \nabla f(\mathbf{x}^t)\rangle - \langle \mathbf{x} - \mathbf{x}^t,\ \mathbf{g}^t\rangle$. ***(iii)*** Subgradient method (SubGrad). It uses the following iteration: $\mathbf{x}^{t+1} = \mathcal{P}_\Omega(\mathbf{x}^t - \frac{0.1}{t} \cdot (\nabla f(\mathbf{x}) - \mathbf{g}^t))$, where $\mathcal{P}_\Omega(\mathbf{x}) \triangleq \max(-1, \min(\mathbf{x}, 1))$ is the projection operation on the convex set $\Omega \triangleq \{\mathbf{x} | \|\mathbf{x}\|_\infty \leq 1\}$. ***(iv) CD-SCA*** solves a convex problem $\bar{\eta}^t = \arg\min_\eta 0.5(\mathbf{c}_{i^t} + \theta)\eta^2 + [\nabla f(\mathbf{x}^t) - \mathbf{g}^t]_{i^t}\eta,\ s.t.\ -1 \leq \mathbf{x}_{i^t}^t + \eta \leq 1$ and update $\mathbf{x}^t$ via $\mathbf{x}_{i^t}^{t+1} = \mathbf{x}_{i^t}^t + \bar{\eta}^t$. ***(v) CD-SNCA*** computes the nonconvex proximal operator of ℓ_2 norm (see Section B.3) as $\bar{\eta}^t = \arg\min_\eta \frac{\mathbf{c}_i+\theta}{2}\eta^2 + \nabla_{i^t} f(\mathbf{x}^t)\eta - \rho\|\mathbf{x}^t + \eta e_i\|, s.t. -1 \leq \mathbf{x}_{i^t}^t + \eta \leq 1$ and update $\mathbf{x}^t$ via $\mathbf{x}_{i^t}^{t+1} = \mathbf{x}_{i^t}^t + \bar{\eta}^t$.

As can be seen from Table 3, the proposed method ***CD-SNCA*** consistently gives the best performance. This is due to the fact that ***CD-SNCA*** finds stronger stationary points than the other methods. Such results consolidate our previous conclusions.

	MSCR	PDCA	SubGrad	CD-SCA	CD-SNCA
randn-256-1024	1.336 ± 0.108	1.336 ± 0.108	1.280 ± 0.098	1.540 ± 0.236	**0.046 ± 0.010**
randn-256-2048	1.359 ± 0.207	1.359 ± 0.207	1.305 ± 0.199	1.503 ± 0.242	**0.021 ± 0.004**
randn-1024-256	2.275 ± 0.096	2.275 ± 0.096	2.268 ± 0.092	2.380 ± 0.180	**1.203 ± 0.043**
randn-2048-256	3.569 ± 0.144	3.569 ± 0.144	3.561 ± 0.143	3.614 ± 0.162	**2.492 ± 0.084**
e2006-256-1024	1.069 ± 0.313	1.069 ± 0.313	0.605 ± 0.167	0.809 ± 0.222	**0.291 ± 0.025**
e2006-256-2048	0.936 ± 0.265	0.936 ± 0.265	0.640 ± 0.187	0.798 ± 0.255	**0.263 ± 0.028**
e2006-1024-256	2.245 ± 0.534	2.245 ± 0.534	1.670 ± 0.198	1.780 ± 0.238	**1.266 ± 0.057**
e2006-2048-256	3.507 ± 0.529	3.507 ± 0.529	3.053 ± 0.250	3.307 ± 0.396	**2.532 ± 0.191**
randn-256-1024-C	1.357 ± 0.130	1.357 ± 0.130	1.302 ± 0.134	1.586 ± 0.192	**0.051 ± 0.012**
randn-256-2048-C	1.260 ± 0.126	1.261 ± 0.126	1.202 ± 0.122	1.444 ± 0.099	**0.019 ± 0.003**
randn-1024-256-C	2.254 ± 0.097	2.254 ± 0.097	2.226 ± 0.084	2.315 ± 0.154	**1.175 ± 0.045**
randn-2048-256-C	3.531 ± 0.159	3.531 ± 0.159	3.520 ± 0.150	3.544 ± 0.184	**2.445 ± 0.082**
e2006-256-1024-C	1.281 ± 0.628	1.323 ± 0.684	0.473 ± 0.128	0.671 ± 0.257	**0.302 ± 0.043**
e2006-256-2048-C	1.254 ± 0.535	1.254 ± 0.535	0.577 ± 0.144	0.717 ± 0.218	**0.287 ± 0.029**
e2006-1024-256-C	2.308 ± 0.640	2.308 ± 0.640	1.570 ± 0.237	1.837 ± 0.322	**1.303 ± 0.060**
e2006-2048-256-C	3.429 ± 0.687	3.429 ± 0.687	2.693 ± 0.335	2.790 ± 0.287	**2.431 ± 0.150**

Table 3: Comparisons of objective values of all the methods for solving the approximate binary optimization problem. The 1^{st}, 2^{nd}, and 3^{rd} best results are colored with **red**, green and blue, respectively.

	MSCR	PDCA	SubGrad	CD-SCA	CD-SNCA
randn-256-1024	0.046 ± 0.019	0.046 ± 0.019	0.077 ± 0.017	**0.039 ± 0.018**	**0.039 ± 0.019**
randn-256-2048	0.023 ± 0.008	0.022 ± 0.007	0.060 ± 0.006	**0.021 ± 0.007**	**0.021 ± 0.007**
randn-1024-256	0.480 ± 0.063	0.473 ± 0.057	0.771 ± 0.089	0.464 ± 0.059	**0.461 ± 0.060**
randn-2048-256	1.335 ± 0.205	1.330 ± 0.205	1.810 ± 0.262	1.329 ± 0.197	**1.325 ± 0.197**
e2006-256-1024	0.046 ± 0.093	0.047 ± 0.105	0.050 ± 0.088	0.047 ± 0.100	**0.045 ± 0.097**
e2006-256-2048	0.022 ± 0.009	0.025 ± 0.012	0.036 ± 0.040	0.029 ± 0.039	**0.020 ± 0.020**
e2006-1024-256	0.922 ± 0.754	0.925 ± 0.758	0.941 ± 0.792	0.925 ± 0.757	**0.858 ± 0.717**
e2006-2048-256	1.031 ± 0.835	1.035 ± 0.838	1.075 ± 0.867	1.024 ± 0.827	**1.010 ± 0.817**
randn-256-1024-C	0.036 ± 0.012	0.036 ± 0.012	0.069 ± 0.014	0.031 ± 0.012	**0.030 ± 0.010**
randn-256-2048-C	0.019 ± 0.003	0.018 ± 0.003	0.058 ± 0.004	**0.016 ± 0.003**	**0.016 ± 0.003**
randn-1024-256-C	0.462 ± 0.089	0.465 ± 0.092	0.768 ± 0.127	0.463 ± 0.088	**0.457 ± 0.092**
randn-2048-256-C	1.155 ± 0.159	1.157 ± 0.165	1.570 ± 0.238	1.161 ± 0.168	**1.147 ± 0.160**
e2006-256-1024-C	0.023 ± 0.020	0.025 ± 0.023	0.032 ± 0.026	0.031 ± 0.038	**0.019 ± 0.018**
e2006-256-2048-C	0.034 ± 0.029	0.037 ± 0.034	0.036 ± 0.066	0.034 ± 0.052	**0.025 ± 0.043**
e2006-1024-256-C	1.772 ± 2.200	1.788 ± 2.200	1.797 ± 2.294	1.768 ± 2.195	**1.702 ± 2.162**
e2006-2048-256-C	1.474 ± 1.247	1.486 ± 1.249	1.520 ± 1.278	1.446 ± 1.233	**1.431 ± 1.224**

Table 4: Comparisons of objective values of all the methods for solving the generalized linear regression problem. The 1^{st}, 2^{nd}, and 3^{rd} best results are colored with **red**, green and blue, respectively.

C.2 Generalized Linear Regression

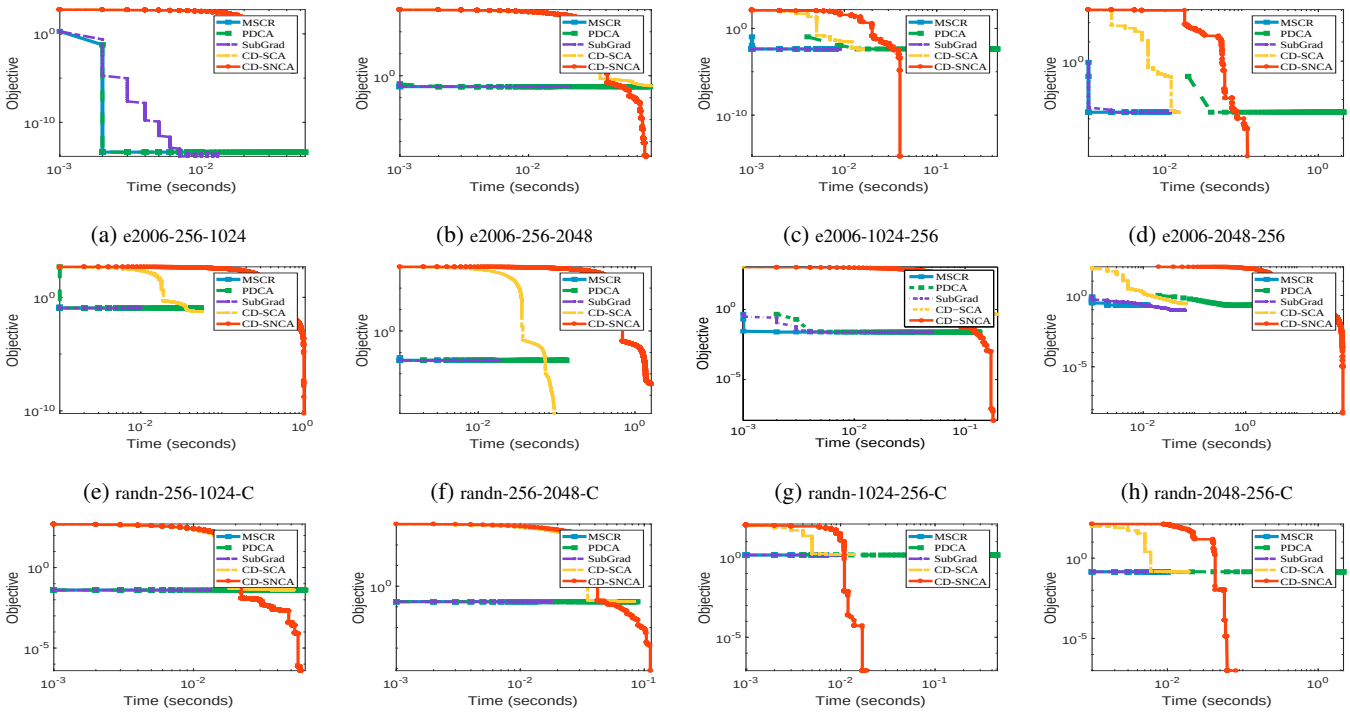
We consider Problem (8). We have the following optimization problem: $\min_{\mathbf{x}} \frac{1}{2}\|\max(0, \mathbf{G}\mathbf{x}) - \mathbf{y}\|_2^2$. We generate the observation vector via $\mathbf{y} = \max(0, \mathbf{A}\ddot{\mathbf{x}} + \text{randn}(m, 1) \times 0.1 \times \|\mathbf{A}\ddot{\mathbf{x}}\|)$ with $\ddot{\mathbf{x}} = \text{randn}(d, 1)$. This problem is consistent with Problem (1) with $f(\mathbf{x}) \triangleq \frac{1}{2}\|\max(0, \mathbf{G}\mathbf{x})\|_2^2$ and $g(\mathbf{x}) \triangleq \|\max(0, \mathbf{A}\mathbf{x})\|_1$ with $\mathbf{A} = \text{diag}(\mathbf{y})\mathbf{G}$. We notice that $\nabla f(\cdot)$ is L-Lipschitz with $L = \|\mathbf{G}\|_2^2$ and coordinate-wise Lipschitz with $\mathbf{c} = \text{diag}(\mathbf{G}^T\mathbf{G})$. The subgradient of $g(\mathbf{x})$ at $\mathbf{x}^t$ can be computed as: $\partial g(\mathbf{x}^t) = \mathbf{A}^T \max(0, \mathbf{A}\mathbf{x}^t) \triangleq \mathbf{g}^t$.

We compare with the following methods. ***(i)*** Multi-Stage Convex Relaxation (MSCR). It solves the following problem: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} f(\mathbf{x}) - \langle \mathbf{x} - \mathbf{x}^t,\ \mathbf{g}^t\rangle$. ***(ii)*** Proximal DC algorithm (PDCA). It considers the following iteration: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \langle \mathbf{x} - \mathbf{x}^t,\ \nabla f(\mathbf{x})\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 - \langle \mathbf{x} - \mathbf{x}^t,\ \mathbf{g}^t\rangle$. ***(iii)*** Subgradient method (SubGrad). It uses the following iteration: $\mathbf{x}^{t+1} = \mathbf{x}^t - \frac{0.1}{t} \cdot (\nabla f(\mathbf{x}) - \mathbf{g}^t)$. ***(iv) CD-SCA*** solves a convex problem $\bar{\eta}^t = \arg\min_\eta 0.5(\mathbf{c}_{i^t} + \theta)\eta^2 + [\nabla f(\mathbf{x}^t) - \rho\mathbf{g}^t]_{i^t} \cdot \eta$ with and update $\mathbf{x}^t$ via $\mathbf{x}_{i^t}^{t+1} = \mathbf{x}_{i^t}^t + \bar{\eta}^t$. ***(v) CD-SNCA*** computes the nonconvex proximal operator (see Section B.2) as $\bar{\eta}^t = \arg\min_\eta \frac{\mathbf{c}_i+\theta}{2}\eta^2 + \nabla_{i^t} f(\mathbf{x}^t)\eta - \|\max(0, \mathbf{A}(\mathbf{x}^t + \eta e_i))\|_1$ and update $\mathbf{x}^t$ via $\mathbf{x}_{i^t}^{t+1} = \mathbf{x}_{i^t}^t + \bar{\eta}^t$.

As can be seen from Table 4, the proposed method ***CD-SNCA*** consistently outperforms the other methods.

C.3 More Experiments on Computational Efficiency

Figure 2, Figure 3, Figure 4, and Figure 5 show the convergence curve of the compared methods for solving the ℓ_p norm generalized eigenvalue problem, the approximate sparse optimization problem, the approximate binary optimization problem, and the generalized linear regression problem, respectively. We conclude that ***CD-SNCA*** at least achieves comparable efficiency, if it is not faster than the compared methods. However, it generally achieves lower objective values than the other methods.



(a) e2006-256-1024 (b) e2006-256-2048 (c) e2006-1024-256 (d) e2006-2048-256

(e) randn-256-1024-C (f) randn-256-2048-C (g) randn-1024-256-C (h) randn-2048-256-C

(i) e2006-256-1024-C (j) e2006-256-2048-C (k) e2006-1024-256-C (l) e2006-2048-256-C

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

D Discussions

This section presents some discussions of our method. First, we discuss the equivalent reformulations for the ℓ_p norm generalized eigenvalue problem (see Section D.1). Second, we use several examples to explain the optimality hierarchy between the optimality conditions (see Section D.2).

D.1 Equivalent Reformulations for the ℓ_p Norm Generalized Eigenvalue Problem

First of all, using classical Lagigian dual theory, Problem (1) is equvilent to the following optimization problem.

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) + h(\mathbf{x}), \ s.t. \ g(\mathbf{x}) \geq \lambda,$$
$$\min_{\mathbf{x}\in\mathbb{R}^n} \ h(\mathbf{x}) - g(\mathbf{x}), \ s.t. \ f(\mathbf{x}) \leq \delta,$$

for some suitable λ and δ. For some special problems where the parameters λ and δ that are not important, the two constrained problems above can be solved by finding the solution to Problem (1).

We pay special attention to the following problems with $\mathbf{Q} \succ \mathbf{0}$:

$$\min_{\mathbf{x}} \ F_1(\mathbf{x}) \triangleq \tfrac{\alpha}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \|\mathbf{A}\mathbf{x}\|_p \tag{56}$$
$$\min_{\mathbf{x}} \ F_2(\mathbf{x}) \triangleq -\|\mathbf{A}\mathbf{x}\|_p, \ s.t. \ \mathbf{x}^T\mathbf{Q}\mathbf{x} \leq 1 \tag{57}$$
$$\min_{\mathbf{x}} \ F_3(\mathbf{x}) \triangleq \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}, \ s.t. \ \|\mathbf{A}\mathbf{x}\|_p \geq 1. \tag{58}$$

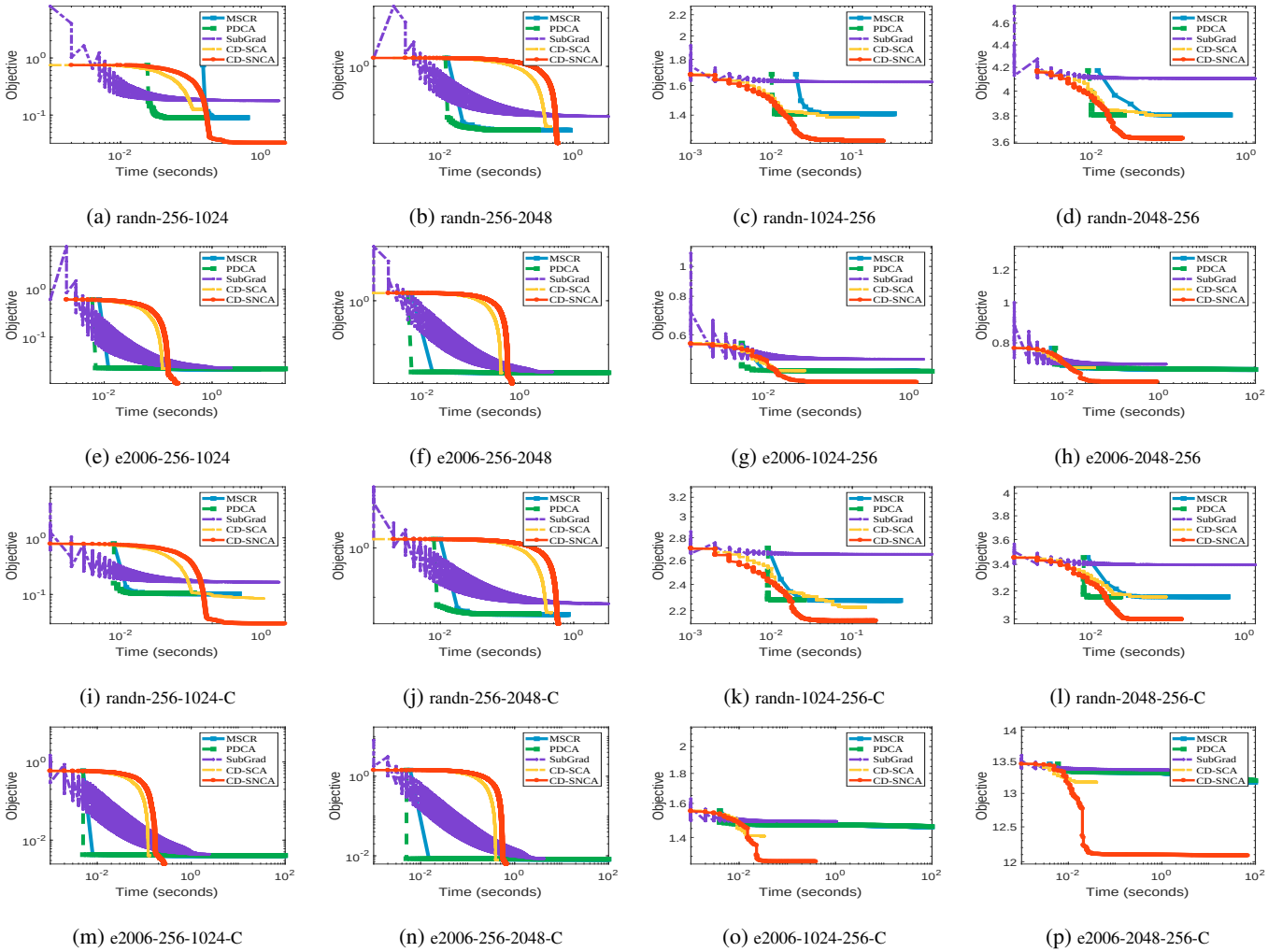
The following proposition establish the relations between Problem (56), Problem (57), and Problem (58).

Proposition D.1. *We have the following results.*

(a) *If $\bar{\mathbf{x}}$ is an optimal solution to (56), then $\pm\bar{\mathbf{x}}(\bar{\mathbf{x}}^T\mathbf{Q}\bar{\mathbf{x}})^{-\frac{1}{2}}$ and $\frac{\pm\bar{\mathbf{x}}}{\|\mathbf{A}\bar{\mathbf{x}}\|_p}$ are respectively optimal solutions to (57) and (58).*

(b) *If $\bar{\mathbf{y}}$ is an optimal solution to (57), then $\frac{\pm\|\mathbf{A}\bar{\mathbf{y}}\|_p\cdot\bar{\mathbf{y}}}{\alpha\bar{\mathbf{y}}^T\mathbf{Q}\bar{\mathbf{y}}}$ and $\frac{\pm\bar{\mathbf{y}}}{\|\mathbf{A}\bar{\mathbf{y}}\|_p}$ are respectively optimal solutions to (56) and (58).*

(c) *If $\bar{\mathbf{z}}$ is an optimal solution to (58), then $\frac{\pm\bar{\mathbf{z}}\|\mathbf{A}\bar{\mathbf{z}}\|_p}{\alpha\bar{\mathbf{z}}^T\mathbf{Q}\bar{\mathbf{z}}}$ and $\pm\bar{\mathbf{z}}(\bar{\mathbf{z}}^T\mathbf{Q}\bar{\mathbf{z}})^{-\frac{1}{2}}$ are respectively optimal solutions to (56) and (57).*



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

(e) e2006-256-1024 (f) e2006-256-2048 (g) e2006-1024-256 (h) e2006-2048-256

(i) randn-256-1024-C (j) randn-256-2048-C (k) randn-1024-256-C (l) randn-2048-256-C

(m) e2006-256-1024-C (n) e2006-256-2048-C (o) e2006-1024-256-C (p) e2006-2048-256-C

Figure 3: The convergence curve of the compared methods for solving the approximate sparse optimization problem on different data sets.

Proof. Using the Lagrangian dual, we introduce a multiplier $\theta_1 > 0$ for the constraint $\mathbf{x}^T\mathbf{Q}\mathbf{x} \leq 1$ as in Problem (57) and a multiplier $\theta_2 > 0$ for the constraint $-\|\mathbf{A}\mathbf{x}\|_p \leq -1$ as in Problem (58), leading to the following optimization problems:

$$\begin{aligned} &\min_{\mathbf{x}} \ \tfrac{\theta_1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \|\mathbf{A}\mathbf{x}\|_p \quad , \\ \min_{\mathbf{x}} \ \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \theta_2\|\mathbf{A}\mathbf{x}\|_p \Leftrightarrow &\min_{\mathbf{x}} \ \tfrac{1}{2\theta_2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \|\mathbf{A}\mathbf{x}\|_p \quad . \end{aligned}$$

These two problems have the same form as Problem (56). It is not hard to notice that the gradient of $F_1(\cdot)$ can be computed as:

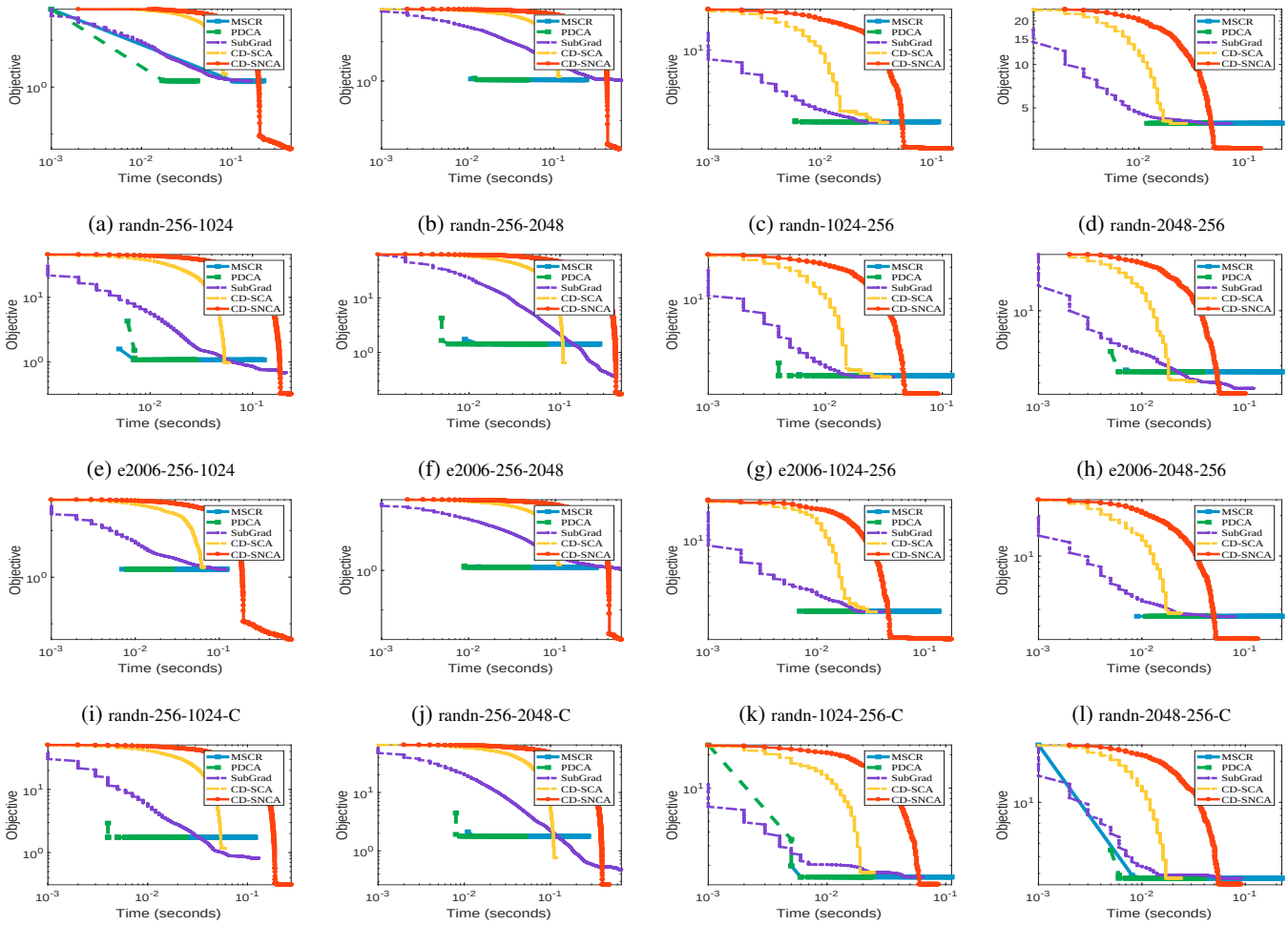
$$\nabla F_1(\mathbf{x}) = \alpha\mathbf{Q}\mathbf{x} - \|\mathbf{A}\mathbf{x}\|_p^{1-p}\mathbf{A}^T(\text{sign}(\mathbf{A}\mathbf{x}) \odot |\mathbf{A}\mathbf{x}|^{p-1}).$$

By the first-order optimality condition, we have:

$$\mathbf{x} = \frac{1}{\alpha}\mathbf{Q}^{-1}\left(\|\mathbf{A}\mathbf{x}\|_p^{1-p}\mathbf{A}^T(\text{sign}(\mathbf{A}\mathbf{x}) \odot |\mathbf{A}\mathbf{x}|^{p-1})\right).$$

Therefore, the optimal solution for Problem (56), Problem (57), and Problem (58) only differ by a scale factor.
(a) Since $\bar{\mathbf{x}}$ is the optimal solution to (56), the optimal solution to Problem (57) and (58) can be respectively computed as $\alpha_2\bar{\mathbf{x}}$ and $\alpha_3\bar{\mathbf{x}}$ with

$$\begin{aligned} \alpha_2 &= \arg\min_{\alpha} \ F_2(\alpha\bar{\mathbf{x}}), \ s.t. \ (\alpha\bar{\mathbf{x}})^T\mathbf{Q}(\alpha\bar{\mathbf{x}}) \leq 1 \\ \alpha_3 &= \arg\min_{\alpha} \ F_3(\alpha\bar{\mathbf{x}}), \ s.t. \ \|\alpha\bar{\mathbf{x}}\|_p \geq 1. \end{aligned}$$



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

(e) e2006-256-1024 (f) e2006-256-2048 (g) e2006-1024-256 (h) e2006-2048-256

(i) randn-256-1024-C (j) randn-256-2048-C (k) randn-1024-256-C (l) randn-2048-256-C

(m) e2006-256-1024-C (n) e2006-256-2048-C (o) e2006-1024-256-C (p) e2006-2048-256-C

Figure 4: The convergence curve of the compared methods for solving the approximate binary optimization problem on different data sets.

After some preliminary calculations, we have: $\alpha_2 = \pm 1/\sqrt{\bar{\mathbf{x}}^T \mathbf{Q} \bar{\mathbf{x}}}$ and $\alpha_3 = \pm 1/\|\mathbf{A}\bar{\mathbf{x}}\|_p$.

(b) Since $\bar{\mathbf{y}}$ is an optimal solution to (57), the optimal solution to Problem (56) and Problem (58) can be respectively computed as $\alpha_1 \bar{\mathbf{y}}$ and $\alpha_3 \bar{\mathbf{y}}$ with

$$\alpha_1 = \arg\min_{\alpha} \ F_1(\alpha \bar{\mathbf{y}})$$
$$\alpha_3 = \arg\min_{\alpha} \ F_3(\alpha \bar{\mathbf{y}}), \ s.t. \ \|\alpha \bar{\mathbf{y}}\|_p \geq 1.$$

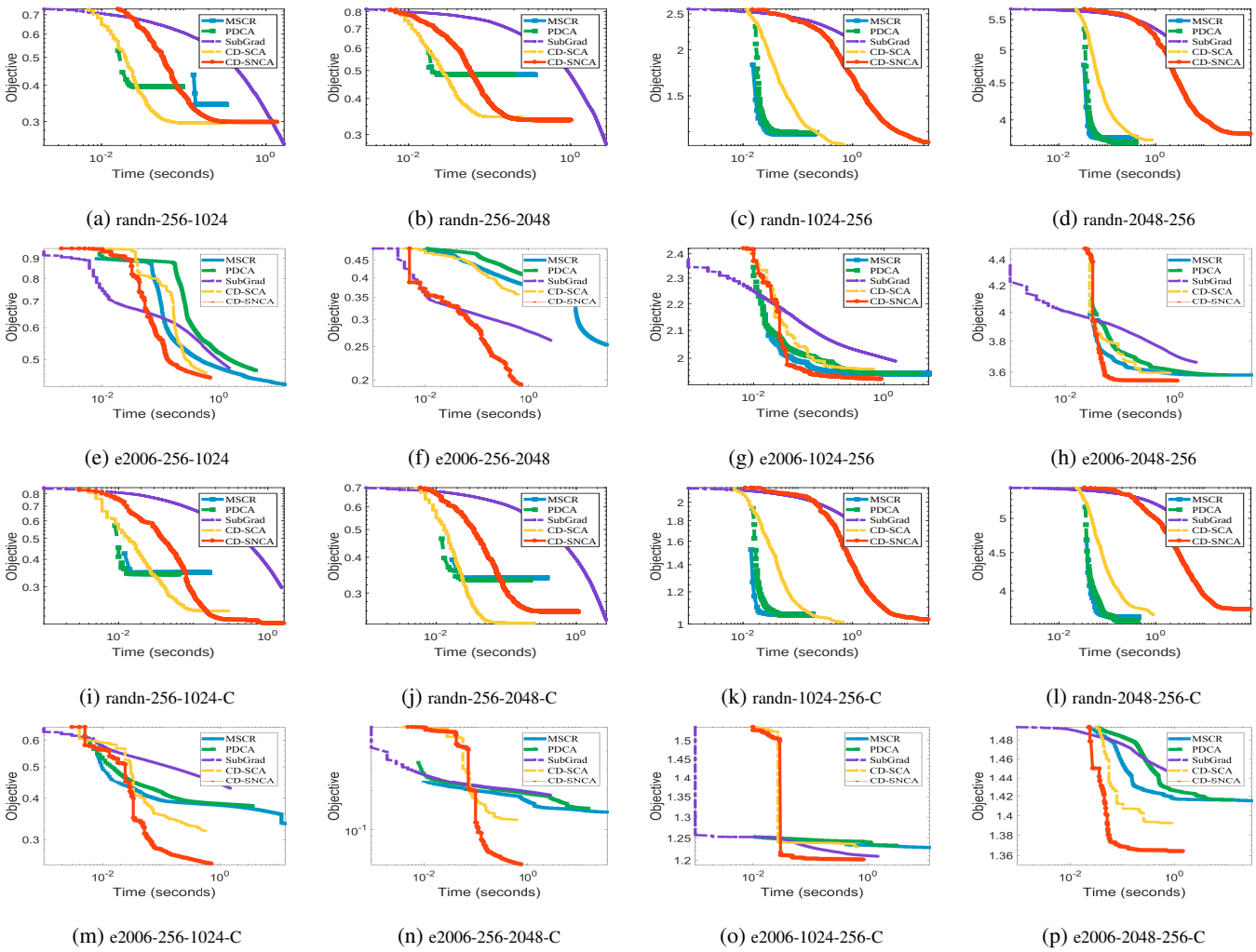
Therefore, $\alpha_1 = \pm \frac{\|\mathbf{A}\bar{\mathbf{y}}\|_p}{\alpha \bar{\mathbf{y}}^T \mathbf{Q} \bar{\mathbf{y}}}$ and $\alpha_3 = \pm 1/\|\mathbf{A}\bar{\mathbf{y}}\|_p$.

(c) Since $\bar{\mathbf{z}}$ is an optimal solution to (58), the optimal solution to Problem (56) and Problem (57) can be respectively computed as $\alpha_1 \bar{\mathbf{z}}$ and $\alpha_2 \bar{\mathbf{z}}$ with

$$\alpha_1 = \arg\min_{\alpha} \ F_1(\alpha \bar{\mathbf{z}})$$
$$\alpha_2 = \arg\min_{\alpha} \ F_2(\alpha \bar{\mathbf{z}}), \ s.t. \ (\alpha \bar{\mathbf{z}})^T \mathbf{Q} (\alpha \bar{\mathbf{z}}) \leq 1.$$

Therefore, $\alpha_1 = \pm \frac{\|\mathbf{A}\bar{\mathbf{z}}\|_p}{\alpha \bar{\mathbf{z}}^T \mathbf{Q} \bar{\mathbf{z}}}$ and $\alpha_2 = \pm 1/\sqrt{\bar{\mathbf{z}}^T \mathbf{Q} \bar{\mathbf{z}}}$.

□



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

(e) e2006-256-1024 (f) e2006-256-2048 (g) e2006-1024-256 (h) e2006-2048-256

(i) randn-256-1024-C (j) randn-256-2048-C (k) randn-1024-256-C (l) randn-2048-256-C

(m) e2006-256-1024-C (n) e2006-256-2048-C (o) e2006-1024-256-C (p) e2006-2048-256-C

Figure 5: The convergence curve of the compared methods for solving the generalized linear regression problem on different data sets.

D.2 Examples for Optimality Hierarchy between the Optimality Conditions

We show some examples to explain the optimality hierarchy between the optimality conditions. Since the condition of directional point is difficult to verify, we only focus on the condition of critical point and coordinate-wise stationary point in the sequel.

• **The First Running Example**. We consider the following problem:

$$\min_{\mathbf{x}} \ \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \langle \mathbf{x}, \mathbf{p}\rangle - \|\mathbf{A}\mathbf{x}\|_1 \tag{59}$$

with using the following parameters:

$$\mathbf{Q} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \ \mathbf{p} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} 1 & -1 & 1 \\ 3 & 1 & 0 \\ 4 & 2 & -1 \end{pmatrix}.$$

First, using the Legendre-Fenchel transform, we can rewrite Problem (59) as the following optimization probelm:

$$\min_{\mathbf{x},\ \mathbf{y}} \ \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \langle \mathbf{x}, \mathbf{p}\rangle - \langle \mathbf{A}\mathbf{x},\ \mathbf{y}\rangle, \ \ -1 \leq \mathbf{y} \leq \mathbf{1}.$$

Second, we have the following first-order optimality condition for Problem (59):

$$\begin{aligned} (\mathbf{Q}\mathbf{x} + \mathbf{p} - \mathrm{sign}(\mathbf{A}\mathbf{x}))_J = \mathbf{0}, \ J \triangleq \{j \mid (\mathbf{A}\mathbf{x})_j \neq 0\} \\ -\mathbf{1} \leq (\mathbf{Q}\mathbf{x} + \mathbf{p})_I \leq \mathbf{1}, \ I \triangleq \{i \mid (\mathbf{A}\mathbf{x})_i = 0\}. \end{aligned} \tag{60}$$

$\mathbf{y}$	$\mathbf{x}$	Function Value	Critical Point	CWS Point
$[1;1;1]$	$[1.75;0;-1]$	-6.625	**Yes**	No
$[1;1;[-1,1]]$	NA	NA	No	No
$[1;1;-1]$	$[-0.25;-2;-1]$	-8.125	No	No
$[1;[-1,1];1]$	NA	NA	No	No
$[1;[-1,1];[-1,1]]$	NA	NA	No	No
$[1;[-1,1];-1]$	NA	NA	No	No
$[1;-1;1]$	$[0.25;-2;-3]$	-4.1250	No	No
$[1;-1;[-1,1]]$	$[-0.3333;0.2667;-0.1333]$	-1.9956	No	No
$[1;-1;-1]$	$[-1.75;-4;-3]$	-16.1250	No	No
$[[-1,1];1;1]$	NA	NA	No	No
$[[-1,1];1;[-1,1]]$	NA	NA	No	No
$[[-1,1]1;-1]$	$[0;-2;-2]$	-6.0000	No	No
$[[-1,1];[-1,1];1]$	NA	NA	No	No
$[[-1,1];[-1,1];[-1,1]]$	$[0;0;0]$	0	**Yes**	No
$[[-1,1];[-1,1];-1]$	$[0;0;0]$	0	**Yes**	No
$[[-1,1];-1;1]$	NA	NA	No	No
$[[-1,1];-1;[-1,1]]$	$[0;0;0]$	0	**Yes**	No
$[[-1,1];-1;-1]$	$[0;0;0]$	0	**Yes**	No
$[-1;1;1]$	$[1.25;0;-3]$	-7.6250	**Yes**	No
$[-1;1;[-1,1]]$	NA	NA	No	No
$[-1;1;-1]$	$[-0.75;-2;-3]$	-12.1250	No	No
$[-1;[-1,1];1]$	NA	NA	No	No
$[-1;[-1,1];[-1,1]]$	$[0;0;0]$	0	**Yes**	No
$[-1;[-1,1];-1]$	$[0;0;0]$	0	**Yes**	No
$[-1;-1;1]$	$[-0.25;-2;-5]$	-6.6250	No	No
$[-1;-1;[-1,1]]$	$[0;0;0]$	0	**Yes**	No
$[-1;-1;-1]$	$[-2.25;-4;-5]$	-18.625	**Yes**	**Yes**

Table 5: Solutions satisfying optimality conditions for Problem (59).

Third, we notice the following relations between $\mathbf{Ax}$ and $\mathbf{y}$:

$$\begin{aligned}(\mathbf{Ax})_i > 0 \quad &\Rightarrow \quad \mathbf{y}_i = 1\\ (\mathbf{Ax})_i < 0 \quad &\Rightarrow \quad \mathbf{y}_i = -1\\ (\mathbf{Ax})_i = 0 \quad &\Rightarrow \quad \mathbf{y}_i \in [-1,1].\end{aligned}$$

We enumerate all possible solutions for $\mathbf{y}$ (as shown in the first column of Table 5), and then compute the solution satisfying the first-order optimality condition using (60). Table 5 shows the solutions satisfying optimality conditions for Problem (59). The condition of the Coordinate-wise Stationary (CWS) point might be a much stronger condition than the condition of critical point.

$(\lambda_i,\ \mathbf{u}_i)$	$\mathbf{x}$	Function Value	Critical Point	CWS Point
$(0.5468,\ [-0.2934, 0.8139, 0.5015])$	$\pm[-0.2169, 0.6019, 0.3709]$	-5.7418	**Yes**	No
$(7.8324,\ [0.1733, -0.4707, 0.8651])$	$\pm[0.4850, -1.3172, 2.4212]$	-82.2404	**Yes**	No
$(33.6207,\ [-0.9402, -0.3407, 0.0030])$	$\pm[-5.4514, -1.9755, 0.0172]$	-353.0178	**Yes**	**Yes**
	$[0,0,0]$	0	**Yes**	No

Table 6: Solutions satisfying optimality conditions for Problem (61).

• **The Second Running Example**. We consider the following example:

$$\min_{\mathbf{x}} \ \frac{1}{2}\mathbf{x}^T\mathbf{x} - \|\mathbf{Ax}\|_2 \tag{61}$$

with using the following parameter:

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & 2 \\ 3 & 1 & 0 \\ 4 & 2 & -1 \end{pmatrix}.$$

Using the first-order optimality condition, one can show that the basic stationary points are $\{\mathbf{0}\} \cup \{\pm\sqrt{\lambda_i}\mathbf{u}_i\}$, where $(\lambda_i, \mathbf{u}_i)$ are the eigenvalue pairs of the matrix $\mathbf{A}^T\mathbf{A}$. Table 6 shows the solutions satisfying optimality conditions for Problem (61). There exists two coordinate-wise stationary points. Therefore, coordinate-wise-stationary might be a much stronger condition than criticality.

$\mathbf{y}$	$\mathbf{x}$	Function Value	Critical Point	CWS Point
$[1;0;0;0]$	$[1;-1;1]$	-2.5000	**Yes**	No
$[0;1;0;0]$	$[2;0;2]$	-4.0000	**Yes**	No
$[0;0;1;0]$	$[3;1;0]$	-9.0000	**Yes**	No
$[0;0;0;1]$	$[4;2;-1]$	-10.5000	**Yes**	**Yes**
$[-1;0;0;0]$	$[-1;1;-1]$	-2.5000	**Yes**	No
$[0;-1;0;0]$	$[-2;0;-2]$	-4.0000	**Yes**	No
$[0;0;-1;0]$	$[-3;-1;0]$	-9.0000	**Yes**	No
$[0;0;0;-1]$	$[-4;-2;1]$	-10.5000	**Yes**	**Yes**

Table 7: Solutions satisfying optimality conditions for Problem (62).

• **The Third Running Example**. We consider the following example:

$$\min_{\mathbf{x}} \ \frac{1}{2}\mathbf{x}^T\mathbf{x} - \|\mathbf{A}\mathbf{x}\|_\infty \tag{62}$$

with using the same value for $\mathbf{A}$ as in Problem (61). It is equivalent to the following problem:

$$\min_{\mathbf{x},\mathbf{y}} \ \frac{1}{2}\mathbf{x}^T\mathbf{x} - \langle \mathbf{A}\mathbf{x},\ \mathbf{y}\rangle,\ \|\mathbf{y}\|_1 \leq 1.$$

We enumerate some possible solutions for $\mathbf{y}$, and then compute the solution satisfying the first-order optimality condition via the optimality of $\mathbf{x}$ that: $\mathbf{x} = \mathbf{A}^T\mathbf{y}$. Table 7 shows the solutions satisfying optimality conditions for Problem (62). These results consolidate our previous conclusions.