The Multi-Block DC Function Class: Theory, Algorithms, and Applications

[1,2] **Pouria Fatemi*** [3,4,5] **Hoomaan Maskan*** [3] **Alp Yurtsever** [1,2] **Suvrit Sra**

[1] Technical University of Munich, School of CIT, Dept. of Mathematics
[2] Munich Center for Machine Learning
[3] Umeå University, Dept. of Mathematics and Mathematical Statistics
[4] Uppsala University, Dept. of Information Technology
[5] Uppsala University, Science for Life Laboratory

Abstract

We present the Multi-Block DC (BDC) class, a rich class of structured nonconvex functions that admit a DC ("difference-of-convex") decomposition across parameter blocks. This multi-block class not only subsumes the usual DC programming, but also turns out to be provably more powerful. Specifically, we demonstrate how standard models (e.g., polynomials and tensor factorization) *must have* DC decompositions of exponential size, while their BDC formulation is polynomial. This separation in complexity also underscores another *key aspect*: unlike DC formulations, obtaining BDC formulations for problems is vastly easier and constructive. We illustrate this aspect by presenting explicit BDC formulations for modern tasks such as deep ReLU networks, a result with no known equivalent in the DC class. Moreover, we complement the theory by developing algorithms with nonasymptotic convergence theory, including both batch and stochastic settings, and demonstrate the broad applicability of our method through several applications.

1 Introduction

Difference-of-convex (DC) programming provides a powerful framework for nonconvex optimization, but existing formulations are difficult to adapt to the block-structured nature of modern machine learning. To this end, we introduce a multi-block difference-of-convex (BDC) framework, in which the objective function admits a DC decomposition with respect to each block when the remaining variables are fixed. The shift from single block to multiple blocks unravels a wealth of new considerations, of which key drivers for our work are complexity theory, algorithms, and empirical behavior.

DC formulations have a long history and have recently received renewed attention for structure-aware modeling in modern applications (Khamaru and Wainwright, 2018; Davis et al., 2022; Maskan et al., 2025b). However, identifying practically useful DC decompositions for a given problem is difficult. Although the existence of DC decompositions can be guaranteed under mild smoothness assumptions (Tuy, 2016), such results are non-constructive and provide limited guidance. As shown by Ahmadi and Hall (2018), identifying suitable DC decompositions, even for polynomials of degree four, is strongly NP-hard.

These challenges call for a more flexible perspective. In this work, we depart from seeking monolithic DC decompositions and pursue BDC formulations instead. These formulations strictly generalize classical DC functions, meaning that there exist BDC functions that do not admit any global DC decomposition. Moreover, BDC formulations elegantly capture the structure of a broad range of

*Authors contributed equally.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

applications, including layered machine learning models and factorized representations. We show that BDC decompositions are relatively simpler to construct and thus they embody a versatile modeling class with strong theoretical foundations. Furthermore, even when a function is DC, often its DC representation can be exponentially more complex than its BDC representation.

Main Contributions. We make the following main contributions in this work:

(i) We introduce BDC, the class of multi-block difference-of-convex functions, which provides a principled framework for modeling structured nonconvex optimization problems. We establish fundamental properties of this class, most notably, we prove that for a specific atom dictionary, BDC representations can be ***exponentially smaller*** in complexity than DC (Theorem 3.2). Furthermore, we provide concrete tools for constructing BDC formulations, highlighting BDC's flexibility (Proposition 3.1).

(ii) We derive explicit BDC representations for deep ReLU networks, showing how the network admits a layer-wise DC structure (Theorem 3.6). Using this result, we construct BDC decompositions for standard supervised-learning objectives with mean-squared error and cross-entropy losses using a general composition principle (Proposition 3.4) that may be of independent interest.

(iii) We study multi-block variants of the DC Algorithm tailored to BDC functions, exploiting block structure for efficient updates, and establish new convergence guarantees under generalized smoothness and stochastic gradients assumptions.

2 Problem Setup and Related Work

We consider the general optimization problem

$$\min_{\boldsymbol{\theta}\in\mathcal{X}} \; f(\boldsymbol{\theta}), \tag{2.1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is possibly nonconvex and $\mathcal{X} \subseteq \mathbb{R}^d$ is the non-empty open convex domain with $f^\star := \inf_{\boldsymbol{\theta}\in\mathcal{X}} f(\boldsymbol{\theta}) > -\infty$. We assume that $\mathcal{X}$ admits a Cartesian product decomposition $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, where each $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ and $d = \sum_{i=1}^n d_i$. We also define $\bar{\mathcal{X}}_i = \mathcal{X}_1 \times \cdots \times \{0\}^{d_i} \times \cdots \times \mathcal{X}_n$, the set obtained from $\mathcal{X}$ by forcing the ith block to be zero. Let $D_i \in \mathbb{R}^{d_i \times d}$ be the selection matrix that extracts the ith block of $\boldsymbol{\theta}$. Equivalently, D_i is obtained by taking d_i distinct rows of the identity matrix I_d, so that $\{D_i\}_{i=1}^n$ forms a non-overlapping partition of the coordinates and satisfies $\sum_{i=1}^n D_i^\top D_i = I_d$. For clarity, we use boldface letters (e.g., $\boldsymbol{\theta} \in \mathcal{X}$) to denote full decision variables, and non-boldface symbols (e.g., $\theta_i \in \mathcal{X}_i$) to denote individual blocks. We define, for each $i \in [n]$, $\theta_i := D_i\boldsymbol{\theta}$, so that $\theta_i \in \mathbb{R}^{d_i}$ represents the ith block of $\boldsymbol{\theta}$. We also define the block-extended vector $\boldsymbol{\theta}_i := D_i^\top D_i\boldsymbol{\theta}$, which coincides with $\boldsymbol{\theta}$ on block i and is zero elsewhere. Its complement is $\bar{\boldsymbol{\theta}}_i := (I_d - D_i^\top D_i)\boldsymbol{\theta}$, so that $\boldsymbol{\theta} = \boldsymbol{\theta}_i + \bar{\boldsymbol{\theta}}_i$. We are now ready to state our key structural assumption on f.

Assumption 1. We assume that $f : \mathcal{X} \to \mathbb{R}$ admits a DC decomposition with respect to each block θ_i when all other blocks are fixed. Formally, this means that for each $i \in [n]$, there exist functions $g_i, h_i : \mathcal{X}_i \times \bar{\mathcal{X}}_i \to \mathbb{R}$, such that for every $\boldsymbol{\theta} \in \mathcal{X}$,

$$f(\boldsymbol{\theta}) = g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i(\theta_i; \bar{\boldsymbol{\theta}}_i), \tag{2.2}$$

where $g_i(\cdot\,; \bar{\boldsymbol{\theta}}_i)$ and $h_i(\cdot\,; \bar{\boldsymbol{\theta}}_i)$ are convex in θ_i. We refer to this property as a BDC decomposition.

2.1 BDC versus DC

A reader familiar with DC programming might wonder about the precise relation between BDC functions and classical DC functions. Note that BDC is a strict superset of DC: there exist BDC functions that do not admit a global DC decomposition. For instance, Veselý and Zajíček (2018) construct a function in $\mathbb{R}^2$ that is DC on every convex curve but not globally DC, and hence is BDC.

The more important distinction between the BDC and DC classes lies in their *representation efficiency*. In particular, Theorem 3.2 shows that for a fixed atom dictionary, DC representations require exponentially many components even for simple monomials, whereas BDC representations need only polynomially many. Moreover, BDC decompositions are typically much easier to construct in practice than global DC decompositions as they can often be obtained constructively and therefore translate directly into algorithms. This flexibility is valuable for complex models such as deep ReLU

networks, where global DC decompositions are hard to derive, but BDC decompositions can be built systematically (Theorem 3.6).

Example 2.1. *Let $\mathcal{T} \in \mathbb{R}^{m_1 \times \cdots \times m_n}$ be an nth-order tensor, and let $\theta_i \in \mathbb{R}^{m_i \times r}$ denote the ith factor matrix. The canonical polyadic (CP) decomposition solves*

$$\min_{\theta_1,\ldots,\theta_n} \frac{1}{2}\left\|\mathcal{T} - [\![\theta_1, \ldots, \theta_n]\!]\right\|_F^2,$$

where $[\![\theta_1, \ldots, \theta_n]\!]$ denotes the rank-r CP reconstruction. This problem is nonconvex jointly in all θ_i's, but convex in each θ_i when the others are fixed. This gives a BDC *decomposition with $h_i = 0$. This observation underlies the classical alternating least-squares (ALS) algorithm, which performs exact multi-block minimization steps. Although this is a purely multi-block convex structure ($h_i = 0$), it illustrates how* BDC *decompositions are easier to obtain than* DC *decompositions. In particular, in Theorem 3.2 we show that the DC decomposition requires exponentially many terms for this task.*

2.2 Related Work

DC programming has been employed in a wide range of machine learning applications from kernel selection (Argyriou et al., 2006) to discrepancy estimation for domain adaptation (Awasthi et al., 2024). The classical method for solving DC problems is the DC Algorithm (DCA), introduced by Tao and Souad (1986). The first asymptotic convergence results for DCA were established by Tao (1997), with a simplified analysis under differentiability assumptions later provided by Lanckriet and Sriperumbudur (2009). More recently, non-asymptotic convergence rates of $\mathcal{O}(1/k)$ have been established (Khamaru and Wainwright, 2018; Yurtsever and Sra, 2022; Abbaszadehpeivasti et al., 2023). For a comprehensive survey, we refer to (Le Thi and Pham Dinh, 2018, 2024).

Despite its generality, the BDC function class remains largely unexplored. The only prior study we are aware of is (Pham Dinh et al., 2022) that considers only the two-block case (termed partial DC decomposition) and proposes the alternating DC algorithm along with its convergence guarantee to weak and Fréchet/Clarke critical points under proper assumptions. However, they investigate neither the constructive structure (algebra) of BDC functions nor their broader application potential, topics that we address through a general BDC formulation and algorithms with convergence guarantees. Note that for a fixed block partition, BDC is more general than the class of differences of multiconvex functions (Xu and Yin, 2013). A representation $f = W - Q$, with both W and Q multiconvex, requires a single global pair whose components are convex in every block separately whereas BDC permits a different pair (g_i, h_i) for each block, requiring convexity only in the active block. Consequently, every multiconvex difference is BDC, but the converse fails (see (Pham Dinh et al., 2022, Section 4) for a two-block example). Additionally, the BDC class has a better expressability than the difference of multiconvex class. The detail of this comparison is deferred to Appendix A.1.

Finally, our framework should not be confused with the block-coordinate DCA of Maskan et al. (2025a), which tackles the simpler classical DC problem with a fixed global decomposition and develops a block-coordinate algorithm. In contrast, we introduce and study the BDC problem class which induces a broader and more flexible formulation. Our work, moreover, calls for a conceptual shift: rather than seeking a global DC decomposition, we advocate a multi-block decomposition, as this is vastly easier to construct, more expressive, and often algorithmically advantageous.

3 The BDC Function Class

In this section, we motivate and study the BDC function class. First, we prove that for a fixed atom class, the complexity of a DC decomposition for monomials is exponentially higher than its BDC counterpart. Then, we represent a BDC formulation of MSE and CE loss functions. Finally, we propose an explicit BDC decomposition for deep ReLU networks using a similar approach to (Berrada et al., 2017). Now, we present our closure proposition. The proof is given in Appendix B.1. This result is used in later sections.

Proposition 3.1. *Let f_i be* BDC *functions for $i = 1, \ldots, m$. Then, the following functions are also* BDC*: (i) $\sum_{i=1}^m \alpha_i f_i$, for $\alpha_i \in \mathbb{R}$, (ii) $\max_{i=1,\ldots m} f_i$, (iii) $\min_{i=1,\ldots m} f_i$.*

3.1 DC and BDC Complexity of a Monomial

As discussed in Example 2.1, here we show the exponential separation in monomial representation complexity between BDC and DC classes. Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$ and $f(\boldsymbol{\theta}) = \theta_1^{b_1}\theta_2^{b_2}\cdots\theta_n^{b_n}$ with $s = \sum_{i=1}^n b_i$. We measure DC and BDC complexities by the minimum number of atoms needed to represent a decomposition. Take DC function $f(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) - h(\boldsymbol{\theta})$ with $g(\boldsymbol{\theta}) = \sum_{i=1}^r \alpha_i\, \phi_i(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta}) = \sum_{i=r+1}^{r+q} \alpha_i\, \phi_i(\boldsymbol{\theta})$, where each $\alpha_i > 0$ and ϕ_i is a convex atom: $\phi_i(\boldsymbol{\theta}) = (u_i^\top \boldsymbol{\theta})^s$ if s is even, and $\phi_i(\boldsymbol{\theta}) = (u_i^\top \boldsymbol{\theta} + \kappa_i)^{s+1}$ if s is odd. The choice of the simple power atoms was motivated by their link to the notion of *Waring rank* (Carlini et al., 2012) and the *polarization property* (Drápal and Vojtěchovskỳ, 2009) used in the proof of Theorem 3.2. We denote by N the *minimum atom count*, i.e., the minimum of $r+q$ over all such decompositions. We bound N in the following theorem. The detailed proof and definitions needed for this result are given in Appendix B.2.

Theorem 3.2. *Consider $f(\boldsymbol{\theta}) = \prod_{i=1}^n \theta_i^{b_i}$ with $1 \le b_1 \le \cdots \le b_n$ and $s = \sum_i b_i$. Then the minimum atom count N for* DC *decomposition is either of the following:*

- *If s is even and atoms are of the form $(u^\top \boldsymbol{\theta})^s$, then $\prod_{i=2}^n (b_i+1) \le N \le \left\lfloor \frac{1}{2} \prod_{i=1}^n (b_i+1) \right\rfloor$.*
- *If s is odd and atoms are of the form $(u^\top \boldsymbol{\theta} + \kappa)^{s+1}$, then $N = \prod_{i=1}^n (b_i+1)$.*

As Theorem 3.2 shows, the DC decomposition of a monomial requires a large number of atoms. In contrast, a BDC decomposition can be significantly more compact. In the simplest case, each $\theta_i^{b_i}$ is treated as a standalone block, reducing the atom count *exponentially* compared to the DC decomposition. More generally, one may split the monomial into a few larger blocks, decompose each block, and then multiply the resulting sums, thereby reducing the complexity. For instance, the monomial $\theta_1\theta_2\theta_3^2\theta_4^4$ requires at least 30 atoms in a DC representation, which matches the lower bound of Theorem 3.2. Instead taking the trivial blocks θ_1, θ_2, θ_3^2, and θ_4^4, yields a BDC decomposition with only 4 atoms. Alternatively, splitting into two blocks, $\theta_1\theta_2$ and $\theta_3^2\theta_4^4$, results in $2 + 7 = 9$ atoms in total through Theorem 3.2. An explicit BDC decomposition in this case is

$$\begin{aligned}\theta_1\theta_2\theta_3^2\theta_4^4 = \frac{1}{14400}\left[(\theta_1+\theta_2)^2 - (\theta_1-\theta_2)^2\right] \times \Big[5\left((\theta_3+\theta_4)^6 + (\theta_3-\theta_4)^6\right) \\ + 3\left((\theta_3+3\theta_4)^6 + (\theta_3-3\theta_4)^6\right) - 8\left((\theta_3+2\theta_4)^6 + (\theta_3-2\theta_4)^6 + 420\theta_4^6\right)\Big].\end{aligned}$$

This example illustrates the flexibility of block-dependent representations.

3.2 BDC Formulation of MSE Loss

For a label $y \in \mathbb{R}$ and the BDC function $f(\boldsymbol{\theta}) = g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$, with nonnegative g_i and h_i, the MSE loss is $\mathcal{L}_y^{\mathrm{MSE}}(\boldsymbol{\theta}) := \big(f(\boldsymbol{\theta}) - y\big)^2$. This yields the explicit BDC decomposition

$$\mathcal{L}_y^{\mathrm{MSE}}(\boldsymbol{\theta}) = 2\big(g_i^2(\theta_i; \bar{\boldsymbol{\theta}}_i) + (h_i(\theta_i; \bar{\boldsymbol{\theta}}_i) + y)^2\big) - (g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) + h_i(\theta_i; \bar{\boldsymbol{\theta}}_i) + y)^2, \tag{3.1}$$

a difference of two multi-block convex functions if $y \ge 0$.

Remark 3.3. *If labels y are not guaranteed to be nonnegative, one can shift labels and outputs by a constant $c \ge 0$ so that $y + c \ge 0$. This translation does not affect the* BDC *structure, so the assumption $y \ge 0$ is not restrictive.*

3.3 BDC Formulation of Cross Entropy Loss

Before we can establish a BDC formulation of the CE loss, we need a general result that extends BDC decompositions to more complex structures. Specifically, we develop a composition principle ensuring that if the input admits a BDC decomposition, then the expression obtained through a conjugate function can also be written explicitly in BDC form. The following proposition establishes this principle (see Appendix B.4 for the proof). In contrast to many DC composition rules that only guarantee existence, our result is *constructive*.

Proposition 3.4. *Let $U \subset \mathbb{R}^m$ be non-empty compact and let $f : U \to \mathbb{R}$ be lower semicontinuous real-valued. Define the conjugate function $f^*(t) = \max_{u \in U}\{\langle u, t\rangle - f(u)\}$. Suppose $E(\boldsymbol{\theta}) =$*

$(E_1(\boldsymbol{\theta}), \ldots, E_m(\boldsymbol{\theta}))$, *and that for each component* $j = 1, \ldots, m$ *and each block* $i \in [n]$, *the mapping* $\boldsymbol{\theta} \mapsto E_j(\boldsymbol{\theta})$ *admits a* BDC *decomposition of the form* $E_j(\boldsymbol{\theta}) = a_{ij}(\theta_i; \bar{\boldsymbol{\theta}}_i) - b_{ij}(\theta_i; \bar{\boldsymbol{\theta}}_i)$, *where* $a_{ij}(\cdot; \bar{\boldsymbol{\theta}}_i)$ *and* $b_{ij}(\cdot; \bar{\boldsymbol{\theta}}_i)$ *are convex in* θ_i. *For* $j \in [m]$ *define*

$$\underline{u}_j := \min_{u \in U} u_j, \bar{u}_j := \max_{u \in U} u_j, c_j^+ := \max\{-\underline{u}_j, 0\}, \text{and } d_j^+ := \max\{\bar{u}_j, 0\}.$$

Let $c^+, d^+ \in \mathbb{R}^m$ *denote the vectors with components* c_j^+, d_j^+. *Then* $f^* \circ E$ *is* BDC, *with the explicit decomposition* $f^*(E(\boldsymbol{\theta})) = g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$, *where*

$$h_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := \langle c^+, a_i(\theta_i; \bar{\boldsymbol{\theta}}_i) \rangle + \langle d^+, b_i(\theta_i; \bar{\boldsymbol{\theta}}_i) \rangle \quad \text{and} \quad g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := f^*(E(\boldsymbol{\theta})) + h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$$

with $a_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := (a_{i1}(\theta_i; \bar{\boldsymbol{\theta}}_i), \ldots, a_{im}(\theta_i; \bar{\boldsymbol{\theta}}_i))$ *and* $b_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := (b_{i1}(\theta_i; \bar{\boldsymbol{\theta}}_i), \ldots, b_{im}(\theta_i; \bar{\boldsymbol{\theta}}_i))$.

The CE loss for a label $y \in \{1, \ldots, C\}$ is $\mathcal{L}_y^{\mathrm{CE}}(\boldsymbol{\theta}) = \mathrm{LSE}\big(E(\boldsymbol{\theta})\big) - E_y(\boldsymbol{\theta})$, where $\mathrm{LSE}(\cdot)$ is the log-sum-exp function with variational form

$$\mathrm{LSE}(t) = \max_{p \in \Delta_C} \big\{\langle p, t \rangle - \mathrm{Ent}(p)\big\}, \quad \Delta_C := \{p \geq 0,\ 1^\top p = 1\},$$

and $\mathrm{Ent}(p) := \sum_{c=1}^C p_c \log p_c$. Since LSE is the conjugate function of Ent, we utilize Proposition 3.4 to obtain a BDC decomposition for the CE loss.

Corollary 3.5. *Applying Proposition 3.4 with* $U = \Delta_C$ *and* $f = \mathrm{Ent}$ *gives* $\underline{u}_j = 0$ *and* $\bar{u}_j = 1$ *for all* j, *hence* $c^+ = 0$ *and* $d^+ = 1$. *Therefore, for every* y, $\mathcal{L}_y^{\mathrm{CE}}(\boldsymbol{\theta}) = g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$ *where*

$$h_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := a_{iy}(\theta_i; \bar{\boldsymbol{\theta}}_i) + 1^\top b_i(\theta_i; \bar{\boldsymbol{\theta}}_i), \quad g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := \mathrm{LSE}(E(\boldsymbol{\theta})) + 1^\top b_i(\theta_i; \bar{\boldsymbol{\theta}}_i) + b_{iy}(\theta_i; \bar{\boldsymbol{\theta}}_i).$$

3.4 BDC Formulation of a Deep ReLU Network

Consider an L-layer ReLU network parameterized by $\boldsymbol{\theta} = \big(W_1, b_1, \ldots, W_L, b_L\big)$. For input $x \in \mathbb{R}^d$, define $a_0(x) = x$, and for $l = 1, \ldots, L-1$

$$F_x(\boldsymbol{\theta}) = W_L a_{L-1}(x) + b_L, \quad a_l(x) = \sigma\big(W_l\, a_{l-1}(x) + b_l\big),$$

where W_l are weight matrices, b_l are bias vectors, and $\sigma(\cdot)$ denotes the ReLU activation. Here $W_L \in \mathbb{R}^{C \times d_L}$ and $b_L \in \mathbb{R}^C$ represent the weights of the output layer. For regression, we take $C = 1$; for classification, C is the number of classes.

Now, we aim to express the network output as a BDC function in each class. We begin by writing each activation using two nonnegative multi-block component-wise convex functions $a_l = Z_l^+ - Z_l^-$. Using $\sigma(a - b) = \max\{a, b\} - b$, we obtain $Z_{l+1}^+ - Z_{l+1}^- = \sigma(W_{l+1} a_l + b_{l+1}) = a_{l+1}(x)$, which shows that our construction is consistent with $a_l(x)$. Similar decomposition rule was adopted in (Berrada et al., 2017). The network output is expressed as a BDC function through the following initialization and forward recursion:

ReLU BDC Decomposition

Initialization $(l = 1)$**.** $\quad Z_1^+ = \sigma(W_1 x + b_1), \qquad Z_1^- = 0.$

Forward recursion $(l \to l{+}1)$**.** Given (Z_l^+, Z_l^-), define

$$\begin{aligned} p_{l+1} &= \sigma(W_{l+1})\, Z_l^+ + \sigma(-W_{l+1})\, Z_l^- + b_{l+1}, \\ Z_{l+1}^- &= \sigma(W_{l+1})\, Z_l^- + \sigma(-W_{l+1})\, Z_l^+, \\ Z_{l+1}^+ &= \max\big\{p_{l+1},\, Z_{l+1}^-\big\}. \end{aligned}$$

Output layer. Define $F_x(\boldsymbol{\theta}) = A(\boldsymbol{\theta}) - B(\boldsymbol{\theta})$ where

$$\begin{aligned} A(\boldsymbol{\theta}) &= \sigma(W_L) Z_{L-1}^+ + \sigma(-W_L) Z_{L-1}^- + \sigma(b_L), \\ B(\boldsymbol{\theta}) &= \sigma(W_L) Z_{L-1}^- + \sigma(-W_L) Z_{L-1}^+ + \sigma(-b_L). \end{aligned} \tag{3.2}$$

This recursion guarantees $Z_l^\pm \geq 0$ and that each component of $Z_l^\pm$ is convex in the chosen block $\theta_l = (W_l, b_l)$ when all other parameters are fixed; the used operations (nonnegative linear maps

and coordinatewise maxima) preserve convexity and nonnegativity layer by layer. The following Theorem 3.6 proves that each component of $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ in (3.2) is a convex function in every block (See Appendix B.3 for the proof).

Theorem 3.6. *For each block $\theta_l = (W_l, b_l)$, the construction in (3.2) gives functions $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ such that each coordinate of $A(\cdot\,; \bar{\boldsymbol{\theta}}_l)$ and $B(\cdot\,; \bar{\boldsymbol{\theta}}_l)$ is nonnegative and convex in θ_l, and we have $F_x(\boldsymbol{\theta}) = A(\boldsymbol{\theta}) - B(\boldsymbol{\theta})$.*

Our result in Theorem 3.6 provides an explicit BDC formulation for deep ReLU networks. Combining this with the results from Section 3.2 and Section 3.3 provides a framework for training. While it is known (as an existence result) that deep ReLU networks are DC, explicit DC decompositions are currently available mainly for *shallow* networks (Askarizadeh et al., 2024; Tschernutter et al., 2024).

4 BDC Algorithm

In this section, we pair the BDC formulation with suitable algorithmic tools by analyzing two natural variants of the classical DC algorithm and establish convergence guarantees under generalized smoothness conditions and in stochastic settings. Throughout, we assume exact solutions to the BDCA subproblems. All of our guarantees are extendable to approximate solution regimes under common error-summability assumptions (See Xu et al. (2019)).

To define subgradients of $f(\boldsymbol{\theta})$ as in (2.2) for continuously differentiable g_i and locally Lipschitz h_i, we use the Clarke subdifferential, denoted by $\overline{\partial}$. Then, the partial Clarke subdifferential of f satisfies $\overline{\partial}_i f(\boldsymbol{\theta}) = \nabla g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - \partial h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$, for all $i \in \{1, \ldots, n\}$. We measure convergence using a weak Clarke residual

$$\mathcal{G}^2(\boldsymbol{\theta}) := \sum_{i=1}^{n} \operatorname{dist}(0, \overline{\partial}_i f(\boldsymbol{\theta}))^2.$$

In particular, $\mathcal{G}^2(\boldsymbol{\theta}) = 0$ implies $\nabla g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) \in \partial h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$ for all $i = 1, \ldots, n$.

4.1 BDCA under L-smoothness Assumption

Consider BDC problem (2.1). The BDC algorithm at the k^{th} iteration selects a block i_k uniformly at random and updates this block by minimizing a surrogate function as follows:

$$\theta_{i_k}^{k+1} \in \operatorname*{argmin}_{\theta_{i_k} \in \mathcal{X}_{i_k}} g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) - \langle u_{i_k}^k,\, \theta_{i_k} \rangle. \tag{4.1}$$

where $u_{i_k}^k \in \partial_{i_k} h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$ denotes the convex subdifferential of h_{i_k} with respect to the i_k^{th} block. We set $\theta_{i_k}^{k+1}$ as above and keep all other blocks unchanged. The following proposition establishes the convergence rate of the block update (4.1) under regularity assumptions.

Assumption 2. For each $i \in [n]$, the function $g_i(\cdot; \bar{\boldsymbol{\theta}}_i)$ is L_i-smooth on $\mathcal{X}_i$, with $L := \max_{i \in [n]} L_i$.

Assumption 3. For each $i \in [n]$, the function h_i is Lipschitz continuous with constant R on $\mathcal{X}$.

Proposition 4.1. *Under Assumptions 1, 2, and 3, the sequence $\{\boldsymbol{\theta}^k\}_{k \geq 1}$ generated by the update (4.1) with initialization $\boldsymbol{\theta}^0$ for $\mathcal{X}_i = \mathbb{R}^{d_i}$ and any $i \in [n]$ satisfies*

$$\min_{0 \leq k < K} \mathbb{E}\left[\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \leq \frac{2Ln}{K}\left(f(\boldsymbol{\theta}^0) - f^\star\right), \tag{4.2}$$

where the expectation is taken with respect to the random block selections $i_1, \ldots, i_K$.

In Appendix A.2 we present an extension of this result to the case where the problem is constrained to a block-separable, closed and convex set.

4.2 Proximal BDCA under Generalized Smoothness Assumption

Many optimization objectives fail to satisfy a global Lipschitz smoothness assumption. Nevertheless, recent works have identified broader smoothness conditions that hold in important learning settings (Zhang et al., 2019; Crawshaw et al., 2022). These conditions relax classical L-smoothness

by allowing the Hessian norm to grow with the gradient magnitude rather than remain uniformly bounded. Motivated by these developments, and supported by empirical observations that a multi-block version of the generalized smoothness holds in neural network training (see Figure 1), we adopt a generalized smoothness assumption on the components of $g(\boldsymbol{\theta})$. We formalize this observation through the notions of ℓ-smoothness and its first-order counterpart (r, ℓ)-smoothness from (Li et al., 2023).

Definition 1 (ℓ-smoothness)**.** Let $\ell : [0, +\infty) \to (0, +\infty)$ be continuous and non-decreasing. A twice differentiable function $\phi : \mathcal{D} \to \mathbb{R}$ is said to be ℓ-smooth if it satisfies $\|\nabla^2 \phi(\boldsymbol{\theta})\| \leq \ell(\|\nabla \phi(\boldsymbol{\theta})\|)$ at almost every point $\boldsymbol{\theta} \in \mathcal{D}$ with respect to the Lebesgue measure.

Definition 2 ((r, ℓ)-smoothness)**.** Let $r, \ell : [0, +\infty) \to (0, +\infty)$ be continuous functions such that ℓ is non-decreasing and r is non-increasing. A differentiable function $\phi : \mathcal{D} \to \mathbb{R}$ is said to be (r, ℓ)-smooth if for all $\boldsymbol{\theta} \in \mathcal{D}$, we have $\mathcal{B}(\boldsymbol{\theta}, r(\|\nabla \phi(\boldsymbol{\theta})\|)) \subseteq \mathcal{D}$, and that $\|\nabla \phi(\boldsymbol{\theta}^1) - \nabla \phi(\boldsymbol{\theta}^2)\| \leq \ell\left(\|\nabla \phi(\boldsymbol{\theta})\|\right) \|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|$ holds for all $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ in this ball.

It is worth noting that these two notions are related. Specifically, every twice differentiable (r, ℓ)-smooth function is ℓ-smooth. Conversely, an ℓ-smooth function that is closed on its open domain can be shown to satisfy (r, m)-smoothness for appropriately constructed functions r and m (see Appendix A.5).

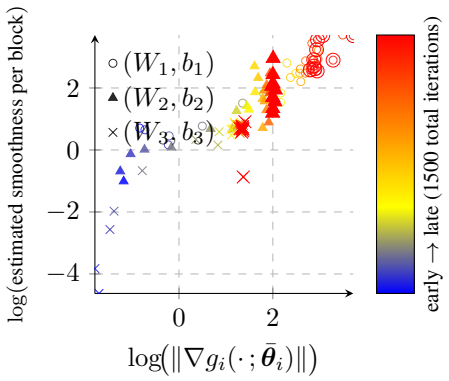


Figure 1: Estimated smoothness constant of $g_i(\cdot\,; \bar{\boldsymbol{\theta}}_i)$ in (3.1) vs its gradient norm. For more details of the experiment see Appendix A.7.1

A typical convergence analysis under (r, ℓ)-smoothness requires control of the update magnitude $\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|$, so that successive iterates remain within the local neighborhood where the smoothness bound is valid. In contrast with algorithms like Gradient Descent (GD), where this follows directly from the explicit step size and bounded gradient norms, such a bound does not arise automatically for BDCA in general. To enforce bounded iterate differences, we add and subtract $\frac{\rho}{2}\|\theta_{i_k}\|^2$ to (2.1) on each block by exploiting the non-uniqueness of the DC decomposition. This leads to a proximal regularization ensuring bounded iterate differences. At k^{th} iteration, the *proximal* BDC *algorithm* selects a block i_k uniformly at random, and updates the selected block by solving the surrogate problem

$$\theta_{i_k}^{k+1} \in \underset{\boldsymbol{\theta}_{i_k} \in \mathcal{X}_{i_k}}{\operatorname{argmin}} g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) - \langle u_{i_k}^k, \theta_{i_k} \rangle + \frac{\rho}{2}\|\theta_{i_k}^k - \theta_{i_k}\|^2, \tag{4.3}$$

where $u_{i_k}^k \in \partial_{i_k} h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$. All other blocks remain unchanged. The next theorem establishes a convergence rate for (4.3).

Assumption 4. For every $i \in [n]$, the function g_i is differentiable and for any fixed $\bar{\boldsymbol{\theta}}_i$, $g_i(\cdot; \bar{\boldsymbol{\theta}}_i)$ is closed within its open domain $\mathcal{X}_i$.

Theorem 4.2. *Suppose Assumptions 1, 3, and 4 hold, and let $\{\boldsymbol{\theta}^k\}$ be the sequence generated by (4.3) from an initial point $\boldsymbol{\theta}^0 \in \mathcal{X}$. Assume that for any fixed $\bar{\boldsymbol{\theta}}_i$, the function $g_i(\cdot; \bar{\boldsymbol{\theta}}_i) : \mathcal{X}_i \to \mathbb{R}$ is ℓ-smooth where ℓ is subquadratic. Further assume that $h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_0}(\theta_{i_0}^0; \bar{\boldsymbol{\theta}}_{i_0}^0) \leq H$ for some constant $H \geq 0$. Define $G := \max_j g_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) - g^* + H$, $E := \sup\{u > 0 : u^2 \leq 2\ell(2u)G\}$, and $L := \ell(2E)$. Choose $\rho \geq L\,\frac{2(E+R)}{E}$. Then the iterates satisfy*

$$\min_{0 \leq k < K} \mathbb{E}\left[\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \leq \frac{2n(L+\rho)}{K}\left(f(\boldsymbol{\theta}^0) - f^\star\right). \tag{4.4}$$

The proof and further discussion are deferred to Appendix A.3. Compared to (Li et al., 2023), the rate is scaled by a factor of n, reflecting the random block selection scheme (4.3).

4.3 Stochastic Proximal BDCA under Generalized Smoothness

In this section, we consider a stochastic variant of (2.1). Assume that, for each block i, the functions g_i and h_i admit stochastic representations of the form $f(\boldsymbol{\theta}) = g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$ with

$$g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := \mathbb{E}_{s\sim\mathbb{P}}\big[g_i(\theta_i; \bar{\boldsymbol{\theta}}_i, s)\big], \quad h_i(\theta_i; \bar{\boldsymbol{\theta}}_i) := \mathbb{E}_{s\sim\mathbb{P}}\big[h_i(\theta_i; \bar{\boldsymbol{\theta}}_i, s)\big], \tag{4.5}$$

where $(\Omega, \Sigma_\Omega, \mathbb{P})$ is a probability space and, for each $s \in \Omega$, the functions $g_i(\cdot;\cdot, s)$ and $h_i(\cdot;\cdot, s)$ are convex functions on $\mathcal{X}_i$ w.r.t θ_i when $\bar{\boldsymbol{\theta}}_i$ is fixed. At iteration k, we draw an i.i.d. sample $s^k \sim \mathbb{P}$ and select a block i_k uniformly at random. We then form an unbiased estimator $\hat{u}^k_{i_k}$ of $u^k_{i_k} \in \partial_{i_k} h_{i_k}(\theta^k_{i_k}; \bar{\boldsymbol{\theta}}^k_{i_k})$ using s^k. Replacing the deterministic update (4.3), the stochastic update is given by

$$\theta^{k+1}_{i_k} \in \operatorname*{argmin}_{\boldsymbol{\theta}_{i_k}\in\mathcal{X}_{i_k}} g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}^k_{i_k}, s^k) - \langle \hat{u}^k_{i_k}, \theta_{i_k}\rangle + \frac{\rho}{2}\|\theta^k_{i_k} - \theta_{i_k}\|^2. \tag{SBDC}$$

SBDC was previously studied for the global DC function class in (Nitanda and Suzuki, 2017; Xu et al., 2019). Here, we focus on the BDC class. In this section, we make the following assumption.

Assumption 5. Let $\{\mathcal{F}_k\}_{k\geq 0}$ be the natural filtration generated by the algorithm, and let $s^k \sim \mathbb{P}$ denote the random sample used at iteration k independent of $\mathcal{F}_k$. For each block $i \in [n]$ and each iteration k, let $u^k_i \in \partial_i h_i(\theta^k_i, \bar{\boldsymbol{\theta}}^k_i)$. Assume that the stochastic estimators $\hat{u}^k_i$ and $\nabla_i \hat{g}_i(\theta^k_i, \bar{\boldsymbol{\theta}}^k_i)$, computed using s^k, satisfy $\mathbb{E}\big[\hat{u}^k_i \mid \mathcal{F}_k\big] = u^k_i$ and $\mathbb{E}\big[\nabla_i \hat{g}_i(\theta^k_i, \bar{\boldsymbol{\theta}}^k_i) \mid \mathcal{F}_k\big] = \nabla_i g_i(\theta^k_i, \bar{\boldsymbol{\theta}}^k_i)$. Furthermore, there exists $\sigma \geq 0$ such that, for all $k \geq 0$ and all $i = 1, \ldots, n$, $\mathbb{E}\Big[\Big\|\big(\nabla_i \hat{g}_i(\theta^k_i, \bar{\boldsymbol{\theta}}^k_i) - \hat{u}^k_i\big) - \big(\nabla_i g_i(\theta^k_i, \bar{\boldsymbol{\theta}}^k_i) - u^k_i\big)\Big\|^2 \,\Big|\, \mathcal{F}_k\Big] \leq \sigma^2$.

The following theorem formulates the convergence of the stochastic BDC (SBDC) algorithm.

Theorem 4.3 (informal)**.** *Suppose Assumptions 1, 3, 4, and 5 hold. Further assume that, for every block i and any fixed $\bar{\boldsymbol{\theta}}_i$, the function $g_i(\cdot; \bar{\boldsymbol{\theta}}_i) : \mathcal{X}_i \to \mathbb{R}$ is ℓ-smooth with subquadratic ℓ. Then, for any initialization $\boldsymbol{\theta}^0 \in \mathcal{X}$, target accuracy $\epsilon > 0$, and $0 < \delta < 1$, the stochastic multi-block DC algorithm (SBDC), with appropriately chosen parameters $\rho = \Theta(\sqrt{K})$ and $\sigma^2 = \mathcal{O}(K^{-1/2})$, requires $K = \mathcal{O}(n^2/\epsilon^4)$ iterations to obtain, with probability at least $1-\delta$ over both the samples and the random block choices,*

$$\min_{0\leq k<K} \mathcal{G}^2(\boldsymbol{\theta}^k) \leq \epsilon^2.$$

The formal version of Theorem 4.3 as well as its proof and a detailed discussion are presented in Appendix A.4. The condition $\sigma^2 = \mathcal{O}(1/\sqrt{K})$ can be achieved through a comparable number of samples in the mini-batch or through variance reduction techniques. In particular, by Lemma A.12 (see Appendix A.6) batch size should be $\Theta(n/\epsilon^2)$ and this means a sample complexity $\mathcal{O}(n^3/\epsilon^6)$. A similar assumption has appeared in previous works such as (Nitanda and Suzuki, 2017; Yurtsever et al., 2019).

5 Applications

We explain the versatility of the BDC framework through illustrative applications and experiments. All numerical experiments were conducted on a Mac equipped with an Apple M5 Pro chip with 48 Gb of memory.

Proximal Alternating Linearized Minimization. The class of nonsmooth nonconvex optimization problems of the form

$$\min_{\theta_i\in\mathcal{X}_{d_i}} \sum_{i=1}^{n} f_i(\theta_i) + Q(\boldsymbol{\theta}), \tag{5.1}$$

was studied by Bolte et al. (2014) under the assumption that Q is L-smooth. Under suitable conditions, PALM can be cast in the form of BDCA. Bolte et al. (2014) established a non-asymptotic convergence guarantees for the PALM algorithm assuming the KL property. We obtain convergence guarantees for this problem class within the BDC framework under generalized smoothness assumptions, without requiring the KL property.

Multiplicative Multitask Feature Learning. MMFL aims to train a neural network that learns shared representations across multiple tasks. A shared vector $\boldsymbol{c} \in \mathbb{R}^d$ modulates feature relevance

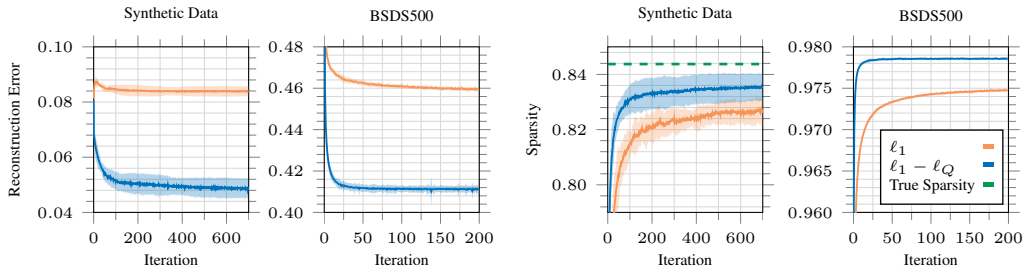


Figure 2: Reconstruction error and sparsity of codes for the ℓ_1 regularizer (orange) and the nonconvex $\ell_1 - \ell_Q$ regularizer (blue) on synthetic data and BSDS500 patches. Solid curves denote the mean over 10 runs, with shaded bands showing 95% confidence interval. The dashed green line indicates the true sparsity level in the synthetic data. The nonconvex $\ell_1 - \ell_Q$ formulation yields both lower reconstruction error and sparser codes.

across T tasks, and is multiplied by the weight vector $\beta_t \in \mathbb{R}^d$, where d is the number of features. Sparse regularization excludes redundant features. For a detailed discussion, see (Wang et al., 2016). The MMFL problem with sparsity imposed on $\boldsymbol{c}$ can be formulated as

$$\min_{\boldsymbol{c}\geq 0,\, \beta_t} \sum_{t=1}^{T} \text{loss}(\text{diag}(\boldsymbol{c})\beta_t, X_t, y_t) + \lambda_1 \sum_{t=1}^{T} \|\beta_t\|_p^p + \lambda_2 \|\boldsymbol{c}\|_0, \tag{5.2}$$

where $X_t \in \mathbb{R}^{n_t \times d}$, $\lambda_1, \lambda_2 \geq 0$ y_t denote the data and labels for task t, and $p \geq 1$. To better approximate hard feature selection, nonconvex surrogates such as $\|\boldsymbol{c}\|_1 - \|\boldsymbol{c}\|_Q$ (largest-Q norm) or the capped ℓ_1 penalty ($\sum_{j=1}^{d} \min\{|\boldsymbol{c}_j|, \gamma\} = \|\boldsymbol{c}\|_1 - \sum_{j=1}^{d} \max\{|c_j| - \gamma, 0\}$ for $\gamma > 0$) are preferred (Gong et al., 2012). Replacing these penalties in (5.2) results in a BDC optimization task.

Rank Regularization. Consider an optimization problem of the following form:

$$\min_{X,Y} \; f(X, Y) + \lambda \, \text{rank}(X) \tag{5.3}$$

where X and Y are two matrices in $\mathbb{R}^{n\times m}$ and the function $f(\cdot)$ is BDC. This type of problem has several applications, such as matrix completion (Hazan et al., 2023). Due to the rank term, (5.3) is NP-hard and a convex surrogate known as the nuclear norm $\|X\|_* = \sum_{i=1}^{\min\{n,m\}} \sigma_i$ is often utilized, where σ_i represents the i-th largest singular value. A tighter non-convex approximation of the rank regularizer is the truncated nuclear norm (TNN), defined as $\sum_{i=r+1}^{\min\{n,m\}} \sigma_i$. TNN can be rewritten as $\|X\|_* - \sum_{i=1}^{r} \sigma_i$, which is a DC function. Thus, replacing it in (5.3) gives a BDC due to the DC regularizer. Note that when $r = 1$, the regularizer is equivalent to $\|X\|_* - \|X\|_2$, which is a special case commonly used as a non-convex regularizer for the rank term (Jiang et al., 2021).

Sparse Dictionary Learning. We illustrate the applicability of our theoretical framework on SDL problem. Given a data matrix $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{m\times n}$, SDL seeks a dictionary $\mathbf{D} = [d_1, \ldots, d_l] \in \mathbb{R}^{m\times l}$ and sparse codes $X = [x_1, \ldots, x_n] \in \mathbb{R}^{l\times n}$ by solving

$$\min_{\mathbf{D}\in\mathcal{C},\, X\in\mathbb{R}^{l\times n}} \sum_{i=1}^{n} \frac{1}{2} \left\| y_i - \mathbf{D}\, x_i \right\|_2^2 + \alpha \sum_{i=1}^{n} \|x_i\|_0, \tag{5.4}$$

where $\mathcal{C} = \left\{ \mathbf{D} \in \mathbb{R}^{m\times l} \mid \|d_j\|_2 \leq 1 \; \forall j \right\}$. Since the ℓ_0 penalty is NP-hard to optimize, it is often replaced with ℓ_1-norm. More recently, nonconvex regularizers have been used to yield a tighter approximation to sparsity. Following (Deng and Lan, 2020; Maskan et al., 2025a), we consider

$$\min_{\mathbf{D}\in\mathcal{C},\, X} \sum_{i=1}^{n} \frac{1}{2} \left\| y_i - \mathbf{D}\, x_i \right\|_2^2 + \alpha \sum_{i=1}^{n} \left(\|x_i\|_1 - \|x_i\|_Q \right). \tag{5.5}$$

Problem (5.5) is BDC. Indeed, fixing either $\mathbf{D}$ or X yields a DC problem. We conducted numerical simulations to solve the SDL problem with ℓ_1 and nonconvex $\ell_1 - \ell_Q$ regularizers (Eq. 5.5) via BDCA (4.1). Performance is measured by reconstruction error $\|Y - \mathbf{D}X\|_F^2$ and the proportion of zeros in X. We compare using synthetic data and Berkeley segmentation dataset (Martin et al., 2001).

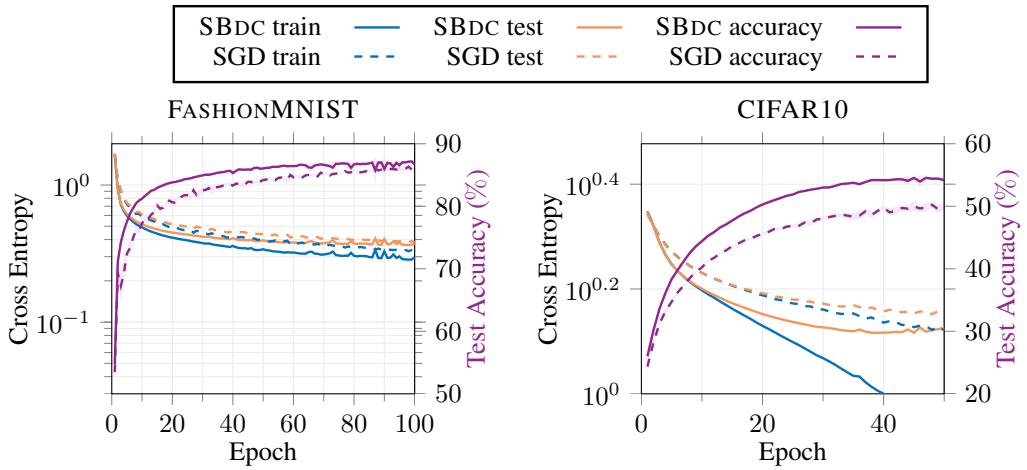


Figure 3: Comparison of SBDC with SGD on FashionMNIST and CIFAR10 using batch size 4096 over 10 Monte-Carlo runs. Solid and dashed curves correspond to SBDC and SGD, respectively. The shaded bands show two-sided 90% confidence intervals for test accuracy.

The results are shown in Figure 2. For more detail and a comparison between GD and the BDCA (4.1), see Appendix A.7.2.

Application to Neural Networks. In Section 3.4, we obtained explicit BDC formulations of neural-network training objectives with MSE and cross-entropy losses. Using these formulations and (SBDC), we train ReLU networks on CIFAR10 (Krizhevsky, 2009) and FASHIONMNIST (Xiao et al., 2017), and compare with SGD. We use a batch size of 4096 and report the mean over 10 Monte-Carlo runs. The motivation for this high batch size regime is due to the noise variance requirement $\sigma^2 = \mathcal{O}(K^{-1/2})$ in Theorem 4.3. One way to ensure this condition is through higher batch sizes. Figure 3 shows the training and test cross-entropy together with test accuracy. On FASHIONMNIST, (SBDC) and SGD attain comparable test performance. On CIFAR10, (SBDC) attains lower test cross-entropy and higher test accuracy under the reported settings. Further implementation details are provided in Appendix A.7.3. Our experiments use a simple ReLU network and are meant as a proof of concepts. Nevertheless, these BDC-aware updates appear to induce distinct optimization dynamics, motivating further investigation beyond the scope of this work.

6 Conclusion and Limitations

We introduced and motivated the *multi-block* DC class—strictly richer than classical DC—and demonstrated its practicality from two angles: (i) compared to DC decompositions, BDC formulations are far cheaper to construct (e.g., exponentially cheaper for monomials), and (ii) obtaining BDC decompositions for modern problems (e.g., training deep ReLU networks) is vastly easier and constructive. Subsequently, after developing foundational properties of the BDC class, we leveraged multi-block convexity to propose a Gauss–Seidel–type BDC algorithm with non-asymptotic guarantees under L-smoothness, generalized smoothness, and stochasticity. Applications to MMFL, rank regularization, sparse dictionary learning, and neural network training illustrate the framework's practicality and breadth.

We conclude by outlining future directions and current limitations. On the theoretical side, an important direction is to further investigate the representation-complexity gap between the BDC and DC classes (e.g., for ReLU networks). On the algorithmic side, although (4.3) guarantees monotone descent, our convergence analysis assumes boundedness of $g_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$ along the trajectory, which we ensure by bounding $h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$ at the update points; removing this assumption would strengthen the theoretical guarantees. In addition, our analysis under the generalized-smoothness assumption applies only to unconstrained BDC problems; extending it to constrained settings remains an open challenge. Finally, in the context of neural networks, we only demonstrated the modeling applicability of BDC functions to ReLU architectures within a simple algorithmic framework. A more comprehensive empirical evaluation and comparison with state-of-the-art training methods are left for future work.

Acknowledgments

Pouria Fatemi and Suvrit Sra acknowledge generous support from the Alexander von Humboldt Foundation. Hoomaan Maskan acknowledges support from the Science for Life Laboratory (SciLifeLab). Hoomaan Maskan and Alp Yurtsever acknowledge support from the Wallenberg AI, Autonomous Systems, and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. The authors would like to thank Manish Krishan Lal for thoughtful discussions. The authors declare no competing interests.

References

Abbaszadehpeivasti, H., de Klerk, E., and Zamani, M. (2023). On the rate of convergence of the difference-of-convex algorithm (DCA). *Journal of Optimization Theory and Applications*, pages 1–22.

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

Argyriou, A., Hauser, R., Micchelli, C. A., and Pontil, M. (2006). A DC-programming algorithm for kernel selection. In *Proceedings of the 23rd international conference on Machine learning*, pages 41–48.

Askarizadeh, M., Morsali, A., Tofigh, S., and Nguyen, K. K. (2024). Convex-concave programming: An effective alternative for optimizing shallow neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–14.

Awasthi, P., Cortes, C., and Mohri, M. (2024). Best-effort adaptation. *Annals of Mathematics and Artificial Intelligence*, 92(2):393–438.

Berrada, L., Zisserman, A., and Kumar, M. P. (2017). Trusting SVM for piecewise linear CNNs. In *International Conference on Learning Representations*.

Bolte, J., Sabach, S., and Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494.

Carlini, E., Catalisano, M. V., and Geramita, A. V. (2012). The solution to the waring problem for monomials and the sum of coprime monomials. *Journal of algebra*, 370:5–14.

Carlini, E., Kummer, M., Oneto, A., and Ventura, E. (2017). On the real rank of monomials. *Mathematische Zeitschrift*, 286(1):571–577.

Crawshaw, M., Liu, M., Orabona, F., Zhang, W., and Zhuang, Z. (2022). Robustness to unbounded smoothness of generalized SignSGD. *Advances in Neural Information Processing Systems*, 35:9955–9968.

Davis, D., Drusvyatskiy, D., Lee, Y. T., Padmanabhan, S., and Ye, G. (2022). A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. *Advances in Neural Information Processing Systems*, 35:6692–6703.

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*, pages 74–89. Springer.

Drápal, A. and Vojtěchovskỳ, P. (2009). Symmetric multilinear forms and polarization of polynomials. *Linear algebra and its applications*, 431(5-7):998–1012.

Gong, P., Ye, J., and Zhang, C.-s. (2012). Multi-stage multi-task feature learning. *Advances in Neural Information Processing Systems*, 25.

Hazan, E., Kalai, A. T., Kanade, V., Mohri, C., and Sun, Y. J. (2023). Partial matrix completion. *Advances in Neural Information Processing Systems*, 36:30134–30145.

Jiang, Z., Zhao, X., and Ding, C. (2021). A proximal DC approach for quadratic assignment problem. *Computational Optimization and Applications*, 78:825–851.

Kan, R. (2008). From moments of sum to moments of product. *Journal of Multivariate Analysis*, 99(3):542–554.

Khamaru, K. and Wainwright, M. (2018). Convergence guarantees for a class of non-convex and non-smooth optimization problems. In *International Conference on Machine Learning*, pages

2601–2610. PMLR.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report.

Lanckriet, G. and Sriperumbudur, B. K. (2009). On the convergence of the concave-convex procedure. *Advances in Neural Information Processing Systems*, 22.

Le Thi, H. A. and Pham Dinh, T. (2018). DC programming and DCA: Thirty years of developments. *Mathematical Programming*, 169(1):5–68.

Le Thi, H. A. and Pham Dinh, T. (2024). Open issues and recent advances in DC programming and DCA. *Journal of Global Optimization*, 88(3):533–590.

Li, H., Qian, J., Tian, Y., Rakhlin, A., and Jadbabaie, A. (2023). Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems*, volume 36, pages 40238–40271.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423.

Maskan, H., Halvachi, P., Sra, S., and Yurtsever, A. (2025a). Randomized block coordinate DC algorithm. *EURO Journal on Computational Optimization*, 13.

Maskan, H., Hou, Y., Sra, S., and Yurtsever, A. (2025b). Revisiting Frank-Wolfe for structured nonconvex optimization. *Advances in Neural Information Processing Systems*, 38.

Nitanda, A. and Suzuki, T. (2017). Stochastic difference of convex algorithm and its application to training deep Boltzmann machines. In *Artificial intelligence and statistics*, pages 470–478. PMLR.

Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.

Pham Dinh, T., Huynh, V. N., Le Thi, H. A., and Ho, V. T. (2022). Alternating DC algorithm for partial DC programming problems. *Journal of Global Optimization*, 82(4):897–928.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. (2016). Stochastic Frank-Wolfe methods for nonconvex optimization. In *2016 54th annual Allerton conference on communication, control, and computing (Allerton)*, pages 1244–1251. IEEE.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. (2018). How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, volume 31.

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

Tao, P. D. and Souad, E. B. (1986). Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients. In *North-Holland Mathematics Studies*, volume 129, pages 249–271. Elsevier.

Tschernutter, D., Kraus, M., and Feuerriegel, S. (2024). A globally convergent algorithm for neural network parameter optimization based on difference-of-convex functions. *Trans. Mach. Learn. Res.*, 2024.

Tuy, H. (2016). *Convex analysis and global optimization*. Springer.

Veselý, L. and Zajíček, L. (2018). A non-DC function which is DC along all convex curves. *Journal of Mathematical Analysis and Applications*, 463(1):167–175.

Wang, X., Bi, J., Yu, S., Sun, J., and Song, M. (2016). Multiplicative multitask feature learning. *Journal of Machine Learning Research*, 17(80):1–33.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

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., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6942–6951. PMLR.

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.

Yurtsever, A. and Sra, S. (2022). CCCP is Frank-Wolfe in disguise. *Advances in Neural Information Processing Systems*, 35:35352–35364.

Yurtsever, A., Sra, S., and Cevher, V. (2019). Conditional gradient methods via stochastic path-integrated differential estimator. In *International Conference on Machine Learning*, pages 7282–7291. PMLR.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*.

A Discussions

In this section, we provide an example showing the expressability of BDC against difference of multiconvex functions, more general results under smoothness assumption in Appendix A.2, background on generalized smoothness in Appendix A.5, useful lemmas in stochastic gradient estimator's variance in Appendix A.6, and more detail on numerical results in Appendix A.7. All the proofs are given in Appendix B.

A.1 BDC is More Flexible than Difference of Multiconvex

Consider

$$f(\theta) = \theta_1\theta_2\theta_3^2\theta_4^4, \qquad \theta \in \mathbb{R}^4,$$

with blocks $u = (\theta_1, \theta_2)$ and $v = (\theta_3, \theta_4)$. This example illustrates how BDC permits different decompositions for different blocks, even when a common multiconvex pair exists.

Define

$$A_\pm(u) = \frac{1}{4}(\theta_1 \pm \theta_2)^2, \qquad P_t(v) = (\theta_3 + t\theta_4)^6 + (\theta_3 - t\theta_4)^6,$$

and

$$B_+(v) = \frac{5P_1(v) + 3P_3(v)}{3600}, \qquad B_-(v) = \frac{8\big(P_2(v) + 420\theta_4^6\big)}{3600}.$$

Each component is nonnegative and convex in its respective block. Moreover,

$$A_+(u) - A_-(u) = \theta_1\theta_2, \qquad B_+(v) - B_-(v) = \theta_3^2\theta_4^4.$$

A common multiconvex decomposition is therefore

$$\begin{aligned} W(u,v) &= A_+(u)B_+(v) + A_-(u)B_-(v), \\ Q(u,v) &= A_+(u)B_-(v) + A_-(u)B_+(v). \end{aligned}$$

Indeed, $f = W - Q$, and fixing either block makes each expression a nonnegative linear combination of convex functions of the other block. Thus, both W and Q are multiconvex.

BDC permits alternative decompositions tailored to the active block. For the u-block, write $b(v) = \theta_3^2\theta_4^4 \geq 0$ and take

$$g_1(u;v) = b(v)A_+(u), \qquad h_1(u;v) = b(v)A_-(u).$$

These components are convex in u for every fixed v, and $f = g_1 - h_1$.

For the v-block, let $a(u) = \theta_1\theta_2$ and $a_\pm(u) = \max\{\pm a(u), 0\}$. Then

$$\begin{aligned} g_2(v;u) &= a_+(u)B_+(v) + a_-(u)B_-(v), \\ h_2(v;u) &= a_+(u)B_-(v) + a_-(u)B_+(v). \end{aligned}$$

For each fixed u, the coefficients $a_\pm(u)$ are nonnegative constants. Hence both components are convex in v, and $f = g_2 - h_2$. Their dependence on the frozen block need not be convex.

The distinction can be seen directly in the first block. The common multiconvex pair satisfies

$$W = g_1 + R, \qquad Q = h_1 + R, \qquad R(u,v) = \big(A_+(u) + A_-(u)\big)B_-(v).$$

The block-specific decomposition removes this shared term while preserving convexity in u. It does not preserve convexity in v. Indeed, at $u = (1,0)$, both g_1 and h_1 equal $b(v)/4$, and b is nonconvex since

$$b(1,0) = b(0,1) = 0, \qquad b\big(\tfrac{1}{2}, \tfrac{1}{2}\big) = \tfrac{1}{64} > 0.$$

Thus, these components are admissible for the u-block under the BDC definition but cannot serve as a common multiconvex pair. This illustrates the freedom to simplify each decomposition using the frozen variables and the requirements of the active block. The example concerns flexibility of representation, otherwise the monomial itself belongs to both function classes.

Algorithm 1 BDC Algorithm (L-smooth)

Input: set $k = 0$, initialization $\boldsymbol{\theta}^0$ and number of blocks n
REPEAT:
Randomly choose i_k in $[1, ..., n]$ with uniform distribution
Evaluate $u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$,
Solve

$$\theta_{i_k}^{k+1} \in \underset{\theta_{i_k} \in \mathcal{M}^{i_k}}{\operatorname{argmin}} \; g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) - \langle u_{i_k}^k, \, \theta_{i_k} \rangle \tag{A.4}$$

Update $\boldsymbol{\theta}^{k+1} = \bar{\boldsymbol{\theta}}_{i_k}^k + \boldsymbol{\theta}_{i_k}^{k+1}$,
Set $k = k + 1$,
UNTIL Stopping criterion.

A.2 Multi-Block DCA Under Smoothness Assumption

Here, we focus on a more general problem of the form:

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \; f(\boldsymbol{\theta}), \tag{A.1}$$

where for each block $\boldsymbol{\theta}_i$

$$f(\boldsymbol{\theta}) := g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i(\theta_i; \bar{\boldsymbol{\theta}}_i), \tag{A.2}$$

when $g_i(\cdot\,; \bar{\boldsymbol{\theta}}_i)$ is an L-smooth function and we have constraint set $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_n$ and each $\mathcal{M}_i \subseteq \mathbb{R}^{d_i}$ is a closed convex set and $d = \sum_{i=1}^n d_i$. The rest of the setup is similar to the unconstrained setting in Section 2. Problem (A.1) was addressed for DC objective function $f(\boldsymbol{\theta})$ in Maskan et al. (2025a). Here, we show that our formulation is capable of solving such problem formulation under a multi-block DC assumption on the objective $f(\boldsymbol{\theta})$. Under this assumption, we propose a multi-block DCA (BDCA) algorithm, shown in Algorithm 1. The following Theorem shows the convergence of this method. The proof of this theorem is given in Appendix B.11.

Theorem A.1. *Suppose assumptions 1, 2, and 3. Then, the sequence generated by the update* (A.4) *for solving problem* (A.1)*, satisfies*

$$\min_{0 \le k < K} \mathbb{E}\left[\sum_{i=1}^n \operatorname{gap}_{\mathcal{M}_i}^L(\boldsymbol{\theta}^k)\right] \le \frac{n}{K}\left(f(\boldsymbol{\theta}^0) - f^\star\right) \;, \tag{A.3}$$

where the expectation is with respect to the random choice of blocks and

$$\operatorname{gap}_{\mathcal{M}_i}^L(\boldsymbol{\theta}) := \max_{\boldsymbol{x} \in \mathcal{M}_i} \min_{\boldsymbol{z} \in \bar{\partial}_i f(\boldsymbol{\theta})} \left\{ \langle \boldsymbol{z}, \, \theta_i - \boldsymbol{x} \rangle - \frac{L}{2}\|\boldsymbol{x} - \theta_i\|^2 \right\},$$

is a gap measure ensuring convergence to blockwise critical points.

This result, is more general than the one presented in Proposition 4.1. Specifically, when $\mathcal{M}$ becomes $\mathbb{R}^d$ (no constraint), the sum of the gap measure over all blocks becomes $\frac{1}{2L}\mathcal{G}^2(\boldsymbol{y})$ which results in Proposition 4.1.

A.3 Analysis of Multi-Block Proximal DCA under Generalized Smoothness Assumption

In this section, we provide a more detailed discussion and prove the results in Section 4.2. Recall that we assumed a more relaxed assumption on the component $g_i(\cdot\,; \bar{\boldsymbol{\theta}}_i)$, known as ℓ-smoothness. A first order reminiscent of the ℓ-smoothness is the (r, ℓ)-smoothness. To simplify presentation, here we use ∇g_i and ∂h_i instead of $\nabla_i g_i$ and $\partial_i h_i$ which are the gradient and the subdifferential set of the i^{th} decomposition with respect to the i^{th} block. We try to avoid this replacement whenever it raises confusion.

Definition 3 ((r, ℓ)-smoothness,(Li et al., 2023))**.** A real-valued differentiable function $g_i : \mathcal{X}_i \times \bar{\mathcal{X}}_i \to \mathbb{R}$ is (r, ℓ)-smooth for continuous functions $r, \ell : [0, +\infty) \to (0, +\infty)$ where ℓ is non-decreasing and r is non-increasing, if we have

$$\mathcal{B}(\theta_i, r(\|\nabla g_i(\theta_i; \bar{\boldsymbol{\theta}}_i)\|)) \subseteq \mathcal{X}_i, \quad \forall \theta_i \in \mathcal{X}_i,$$

and, for all $\theta_i^1, \theta_i^2 \in \mathcal{B}(\theta_i, r(\|\nabla g_i(\theta_i; \bar{\boldsymbol{\theta}}_i)\|))$ it holds that

$$\|\nabla g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i) - \nabla g_i(\theta_i^2; \bar{\boldsymbol{\theta}}_i)\| \leq \ell(\|\nabla g_i(\theta_i; \bar{\boldsymbol{\theta}}_i)\|)\|\theta_i^1 - \theta_i^2\|.$$

Due to $\|\nabla g_i(\theta_i; \bar{\boldsymbol{\theta}}_i)\| \leq \|\nabla g_j(\theta_j, \bar{\boldsymbol{\theta}}_j)\|$ for $j := \arg\max_k \|\nabla g_k(\theta_k, \bar{\boldsymbol{\theta}}_k)\|$ and the fact that r is a non-increasing function, we get $\mathcal{B}(\boldsymbol{\theta}_i, r(\|\nabla g_j(\theta_j, \bar{\boldsymbol{\theta}}_j)\|)) \subseteq \mathcal{B}(\boldsymbol{\theta}_i, r(\|\nabla g_i(\theta_i; \bar{\boldsymbol{\theta}}_i)\|))$. Therefore, for any $\theta_i^1, \theta_i^2 \in \mathcal{B}(\theta_i, r(\|\nabla g_j(\theta_j, \bar{\boldsymbol{\theta}}_j)\|))$ that satisfy (r, ℓ)-smoothness, we have:

$$\|\nabla g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i) - \nabla g_i(\theta_i^2; \bar{\boldsymbol{\theta}}_i)\| \leq \ell(\|\nabla g_j(\theta_j, \bar{\boldsymbol{\theta}}_j)\|)\|\theta_i^1 - \theta_i^2\|.$$

It is possible to relate these two definitions, i.e., we can show that an ℓ-smooth function is (r, ℓ)-smooth and vice-versa under specific choices for r and ℓ. This connection, investigated by Li et al. (2023), with more discussion and related results are given in Appendix A.5.

A necessary condition for (r, ℓ)-smoothness is that the iterates of our sequential algorithm have a bounded distance $\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|$ so that they stay in a local ball. Usually, this is satisfied through bounded gradient norm condition and the sequential form of the algorithm. For example, in GD we have $\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| = \|\eta \nabla f(\boldsymbol{\theta}^k)\|$. In DCA, such a connection does not have trivial validity. Using the non-uniqueness of the DC decomposition, we add and subtract $\frac{\rho}{2}\|\theta_{i_k}\|^2$ to (2.1) on each block. This gives the subproblems (4.3) after applying DCA, which are proximal-type updates. The expected convergence rate of (4.3) is finalized in the following proposition:

Proposition A.2. *Consider Assumptions 3 and 4 when $\boldsymbol{\theta}^k$ is the output of (4.3) for any initialization $\boldsymbol{\theta}^0 \in \mathcal{X}$. Then, for any ℓ-smooth g_i with subquadratic ℓ, if*

$$\begin{aligned} &h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_0}(\theta_{i_0}^0; \bar{\boldsymbol{\theta}}_{i_0}^0) \leq H \\ &E := \sup\{u > 0 : u^2 \leq 2\ell(2u).G\} < \infty \\ &G := \max_j g_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) - g^* + H \end{aligned}$$

for a constant $H \geq 0$, and $L := \ell(2E)$, then the sequence $\boldsymbol{\theta}^k$ generated by (4.3) with $\rho \geq L\frac{2(E+R)}{E}$ will satisfy

$$\min_{0 \leq k < K} \mathbb{E}\left[\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \leq \frac{2n(L+\rho)}{K}\left(f(\boldsymbol{\theta}^0) - f^\star\right), \tag{A.5}$$

where the expectation is with respect to the random choice of blocks.

Proof. We begin by bounding the updates through the following lemma.

Lemma A.3. *For any starting point $\boldsymbol{\theta}^k$ the update generated by (4.3) is in $\mathcal{B}\left(\boldsymbol{\theta}^k, \frac{2}{\rho}\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\|\right)$.*

See Appendix B.5 for the proof. This result guarantees $\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \leq \frac{2}{\rho}\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\|$. Due to $\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\| \leq \|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)\| + \|u_{i_k}^k\|$ for any $u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$ and R-Lipschitz h_{i_k} (see Assumption 3), we need to bound $\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)\|$ in order to have a bounded $\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\|$. When g_{i_k} is ℓ-smooth with bounded $g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) - g^*$ for some $\theta_{i_k} \in \mathcal{X}_{i_k}$, we get $\|\nabla g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k)\| \leq E$ for $E > 0$ (see Corollary A.11 in Appendix A.5). The following Lemma bounds $g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g^*$ and proposes a choice for ρ such that we have local bound on the gradients.

Lemma A.4. *Consider Assumptions 4 and 3 when $\boldsymbol{\theta}^{k+1}$ is the output of (4.3) for any initialization $\boldsymbol{\theta}^0 \in \mathcal{X}$. Then, if for any randomly selected i_0, we have $h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_0}(\theta_{i_0}^0; \bar{\boldsymbol{\theta}}_{i_0}^0) \leq H$ for a constant $H \geq 0$, then*

$$g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g^* \leq g_{i_0}(\theta_{i_0}^0; \bar{\boldsymbol{\theta}}_{i_0}^0) - g^* + H.$$

Additionally, for any ℓ-smooth g_i with subquadratic ℓ, if $\rho \geq \ell(2E)\frac{2(E+R)}{E}$, then for any $i \in [n]$ and $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \in \mathcal{B}(\boldsymbol{\theta}^k, {}^{2(E+R)}/_{\rho})$ we have:

$$\begin{aligned} &\|\nabla g_i(\theta_i^2; \bar{\boldsymbol{\theta}}_i^2) - \nabla g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i^1)\| \leq L\|\theta_i^1 - \theta_i^2\|, \\ &g_i(\theta_i^2; \bar{\boldsymbol{\theta}}_i^2) \leq g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i^1) + \langle \nabla g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i^1), \theta_i^2 - \theta_i^1 \rangle + \frac{L}{2}\|\theta_i^1 - \theta_i^2\|^2, \end{aligned} \tag{A.6}$$

where $L = \ell(2E)$ is the effective smoothness for some $E > 0$.

See Appendix B.6 for the proof. Note that $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2$ in Lemma A.4 differ only in their i_k^{th} block selected on iteration k of (4.3). Now building on the previous lemmas, we propose our main convergence result for (4.3). The proof of this result is given in Appendix B.7

Proposition A.5. *Assume the conditions in Lemma A.4 and take*

$$E := \sup\{u > 0 : u^2 \leq 2\ell(2u).G\} < \infty$$
$$G := \max_i g_i(\theta_i^0; \bar{\boldsymbol{\theta}}_i^0) - g^* + H,$$

and $L := \ell(2E)$. Then, the sequence $\boldsymbol{\theta}^k$ generated by (4.3) with $\rho \geq L\frac{2(E+R)}{E}$ will satisfy

$$\min_{k\in\{1,\ldots,K\}} \mathbb{E}_i\left[\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \leq \frac{2n(L+\rho)}{K}\left(f(\boldsymbol{\theta}^1) - f^\star\right), \tag{A.7}$$

where the expectation is with respect to the random choice of blocks.

□

A.4 Detailed Analysis of Stochastic Multi-Block Proximal DCA under Generalized Smoothness Assumption

In order to show the convergence of (SBDC), we start by ensuring the boundedness of the updates as in the following lemma. See Appendix B.9 for the proof. To simplify presentation, here we use ∇g_i and ∂h_i instead of $\nabla_i g_i$ and $\partial_i h_i$ which are the gradient and the subdifferential set of the i^{th} decomposition with respect to the i^{th} block.

Lemma A.6. *Denote the sequence generated by (SBDC) as $\boldsymbol{\theta}^k$. Then, for any $u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$, if $\nabla \hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k), \hat{u}_{i_k}^k$ are the respective stochastic approximations of $\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$ $u_{i_k}^k$ and , we have:*

$$\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \leq \frac{2}{\rho}\left(\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\| + \|\nabla \hat{g}_{i_k}(\boldsymbol{\theta}^k) - \hat{u}_{i_k}^k - (\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k)\|\right).$$

Note that the bound in Lemma A.6 does not immediately imply that the solutions to the subproblems (SBDC) will fall inside a ball. For this, we take $\|\nabla \hat{g}_{i_k}(\boldsymbol{\theta}^k) - \hat{u}_{i_k}^k - (\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k)\| \leq F'$ for some $F' > 0$. Later, we find the value of F' such that the bound $\|\nabla \hat{g}_{i_k}(\boldsymbol{\theta}^k) - \hat{u}_{i_k}^k - (\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k)\| \leq F'$ holds with high probability. Then, a similar result to Lemma A.4 holds in the stochastic setting.

Lemma A.7. *Consider Assumptions 4 and 3 when $\boldsymbol{\theta}^k$ is the output of (SBDC) for any initialization $\boldsymbol{\theta}^0 \in \mathcal{X}$. Then, for any ℓ-smooth g_i with subquadratic ℓ if $g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g^* \leq G$ and $\|\nabla \hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \hat{u}_{i_k}^k - (\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k)\| \leq F'$ for $G, F' > 0$ and $\rho \geq L\frac{2(E+R+F')}{E}$ for $L := \ell(2E)$, we have:*

$$\begin{aligned}
&\|\nabla g_i(\theta_i^2; \bar{\boldsymbol{\theta}}_i^2) - \nabla g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i^1)\| \leq L\|\theta_i^1 - \theta_i^2\|, \\
&g_i(\theta_i^2; \bar{\boldsymbol{\theta}}_i^2) \leq g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i^1) + \langle \nabla g_i(\theta_i^1; \bar{\boldsymbol{\theta}}_i^1), \theta_i^2 - \theta_i^1\rangle + \frac{L}{2}\|\theta_i^1 - \theta_i^2\|^2,
\end{aligned} \tag{A.8}$$

for any $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \in \mathcal{B}(\boldsymbol{\theta}^k, {}^{2(E+R+F')}/_{\rho})$.

See Appendix B.10 for the proof. Note that if $F' = 0$, we get $\rho \geq {}^{2L(E+R)}/_{E}$ which was in Lemma A.4. In order to use Lemma A.7, we need to show $g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g^* \leq G$ and $\|\nabla \hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \hat{u}_{i_k}^k - (\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k)\| \leq F'$. Due to stochasticity, it is not possible to directly bound these values for all the iterations. Instead, we will show that the probabilities of the following events are low up to time K:

$$\begin{aligned}
t_1 &:= \min\left\{\min\{k | \max_{i\in[n]} g_i(\theta_i^{k+1}; \bar{\boldsymbol{\theta}}_i^{k+1}) - g^* > G\}, K\right\}, \\
t_2 &:= \min\left\{\min\{k | \|\nabla \hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \hat{u}_{i_k}^k - (\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k)\| > F'\}, K\right\}, \\
t &:= \min\{t_1, t_2\},
\end{aligned} \tag{A.9}$$

In (A.9), the event $t_1 = K$ will ensure $g_{i_k}(\theta^k_{i_k}; \bar{\boldsymbol{\theta}}^k_{i_k}) - g^* \leq G$ before time $k < K$ and the event $t_2 = K$ will ensure $\|\nabla \hat{g}_{i_k}(\theta^k_{i_k}; \bar{\boldsymbol{\theta}}^k_{i_k}) - \hat{u}^k_{i_k} - (\nabla g_{i_k}(\theta^k_{i_k}; \bar{\boldsymbol{\theta}}^k_{i_k}) - u^k_{i_k})\| \leq F'$ before time $k < K$. Next, we should show that the probability of the event $\{t < K\}$ is low. Alternatively, we can show a low probability for the event $\{t = t_2 < K\} \cup \{t = t_1 < K, t_2 = K\}$. This is a similar technique to (Li et al., 2023) in order to show convergence in the stochastic setting. Compared to their work, our proposed method in (SBDC) targets a more general class of functions (BDC). Although the generality of our function class, our guarantee in Theorem 4.3 requires only the first components of our BDC structure to be ℓ-smoothness. In this sense, our work generalizes the prior result by Li et al. (2023). The main convergence result is given in the following theorem (see Appendix B.8 for the proof). This result is the formal statement of Theorem 4.3.

Proposition A.8. *Suppose Assumptions 3, 4, and 5 hold, and let $\{\boldsymbol{\theta}^k\}_{k=0}^K$ be generated by (SBDC) from $\boldsymbol{\theta}^0 \in \mathcal{X}$. Assume that, for every block i and any fixed $\bar{\boldsymbol{\theta}}_i$, $g_i(\cdot; \bar{\boldsymbol{\theta}}_i)$ is ℓ-smooth with subquadratic ℓ. Let $g^* > -\infty$ be a common lower bound for all g_i, and assume that*

$$h_{i_k}(\theta^k_{i_k}; \bar{\boldsymbol{\theta}}^k_{i_k}) - h_{i_0}(\theta^0_{i_0}; \bar{\boldsymbol{\theta}}^0_{i_0}) \leq H$$

for a constant $H \geq 0$, for any chosen block i_k, i_0 and every iterate $0 \leq k \leq K$. For $\epsilon > 0$, $0 < \delta < 1$, and $K \geq 1$, choose positive G, E, L, ρ, F' satisfying

$$C' := \frac{K\sigma^2}{\rho} + H, \qquad G := \frac{8\left(\max_j g_j(\theta^0_j; \bar{\boldsymbol{\theta}}^0_j) - g^* + C'\right)}{\delta},$$
$$E := \sup\{u > 0 : u^2 \leq 2\ell(2u)G\} < \infty, \qquad L := \ell(2E),$$
$$F' := \frac{E\rho}{9L} - (E + R), \qquad \sigma^2 = \mathcal{O}(K^{-1/2}),$$
$$\rho := \left(18L + \frac{9ER}{G} + \frac{81L}{4}\frac{C' - H}{C'}\right)\sqrt{K}.$$

When $C' = 0$, the last term in parentheses is taken to be zero. Suppose, in addition, that

$$K \geq \frac{4n(L + \rho)}{\delta\epsilon^2}\left[\frac{G\delta}{8} + \frac{3(E + R)^2 + (2K + 1)\sigma^2}{2\rho}\right].$$

Then, with probability at least $1 - \delta$ over both the samples and the random block choices,

$$\min_{0 \leq k < K} \mathcal{G}^2(\boldsymbol{\theta}^k) \leq \epsilon^2.$$

Now, we state the formal version of Theorem 4.3.

A.5 Background on (r, ℓ)-Smoothness and ℓ-Smoothness

Here, we discuss the required background and results on ℓ-smoothness. We mainly represent the results from (Li et al., 2023) and briefly explain the results and connections with this work.

We start with the following lemma characterizing a local descent condition for any $x \in \mathcal{X}$ when g is (r, ℓ)-smooth:

Lemma A.9 (Li et al. (2023))**.** *If g is (r, ℓ)-smooth, for any $x \in \mathcal{X}$ satisfying $\|\nabla g(x)\| \leq E$ we have $\mathcal{B}(x, r(E)) \subset \mathcal{X}$, and for any $x_1, x_2 \in \mathcal{B}(x, r(E))$,*

$$\|\nabla g(x_2) - \nabla g(x_1)\| \leq L\|x_1 - x_2\| \quad g(x_2) \leq g(x_1) + \langle \nabla g(x_1), x_2 - x_1\rangle + \frac{L}{2}\|x_1 - x_2\|^2$$

where $L = \ell(E)$ is the effective smoothness.

The following proposition, bridges ℓ-smoothness and (r, ℓ)-smoothness. The importance of this result is due to the fact that it shows applicability of the descent Lemma A.9 on ℓ-smooth functions.

Proposition A.10 (Li et al. (2023))**.** *An (r, l)-smooth function is l-smooth; and an l-smooth function is (r, m)-smooth with $m(u) := l(u + a)$ and $r(u) := a/m(u)$ for any $a > 0$ if f is a closed function within its open domain $\mathcal{X}$.*

With this result, one can use Lemma A.9 on an ℓ-smooth function which satisfies the conditions in Lemma A.9: bounded gradients and (r, ℓ)-smoothness. Also, we need to ensure that the updates remain inside a ball. Despite the convexity of the function g in our problem setup, DCA updates do not guarantee the boundedness of its gradients. Therefore, we use the following corollary which provides such bound when the function ℓ is sub-quadratic in the sense that $\lim_{t\to\infty} {}^{\ell(t)}/_{t^2} = 0$.

Corollary A.11 (Li et al. (2023))**.** *Suppose Assumption 4 and take g as ℓ-smooth with sub-quadratic ℓ. If $g(x) - inf_{y\in\mathcal{X}} g(y) \leq G$ for some $x \in \mathcal{X}$ and $G \geq 0$, then $E^2 = 2\ell(2E)G$ and $\|\nabla g(x)\| \leq E < \infty$ for $E := \sup\{u \geq 0 | u^2 \leq 2\ell(2u)G\}$*

With Corollary A.11, if we can show that the updates remain inside a ball, then the descent condition in Lemma A.9 holds.

A.6 Useful Lemma on Gradient Estimation Variance

The following lemma is a classical result on the variance in terms of the mini-batch size. We used this Lemma for the discussions on our reduced variance assumption in Theorem 4.3.

Lemma A.12 (Lemma 2 from (Reddi et al., 2016))**.** *Suppose that $\mathcal{S}^k$ is a subset that samples S^k i.i.d realizations from the distribution $\mathcal{P}$. Let the stochastic estimator $\nabla f(\boldsymbol{\theta}^k, s^k)$ satisfy the bounded variance condition Assumption 5. Then, the following bound holds:*

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}^k, s^k) - \nabla f(\boldsymbol{\theta}^k)\|^2\right] \leq \frac{\sigma^2}{S^k}, \quad \forall \boldsymbol{\theta} \in \mathcal{X}. \tag{A.10}$$

A.7 Additional Details on Numerical Examples

In this section, we provide the reader with more detail on our implementation settings and parameter choices.

A.7.1 Generalized Smoothness on Deep Networks.

The relationship between the Hessian of the objective function in training language models and the norm of its gradient was already observed in (Zhang et al., 2019). This relationship was later extended to more general cases by Li et al. (2023). The previous analyses heavily relied on the trajectory of the optimization guided by GD updates. Here, we want to show that a similar relationship exists between the estimated smoothness of the first BDC component and its gradient norm when the updates are done by the BDCA. In order to do this, we use the same smoothness estimator as in (Santurkar et al., 2018) defined below:

$$\hat{L}_{g_i}(\boldsymbol{\theta}^k) = \max_{\gamma\in\{\delta, 2\delta, \ldots, 1\}} \frac{\|\nabla g_i(\theta_i^k + \gamma d) - \nabla g_i(\theta_i^k)\|}{\|\gamma d\|}, \tag{A.11}$$

for a small value δ and $d = \theta_i^{k+1} - \theta_i^k$. This value determines the variations along d on the block i. Note that unlike previous results, in BDCA we do not necessarily decrease the value of the first component along the update trajectory. To show this, we conducted numerical simulations on a regression task using a three-layer ReLU network of size $(8 \times 64 \times 32 \times 1)$ on the California Housing dataset (Pace and Barry, 1997). We considered training for 30 epochs, a learning rate of 0.5×10^{-3} with 10 oracle calls to the BDCA sub-problem solver. We set $\delta = 0.25$. For each active block, we evaluate the first convex component on the current mini-batch, holding the complementary parameters fixed. We used the block subgradient returned by automatic differentiation in (A.11) to estimate $\hat{L}_{g_i}(\boldsymbol{\theta}^k)$. Numerical diagnostics indicate that g_i was differentiable at all evaluated points. The logarithm of the estimated smoothness constant of the first BDC component in (3.1) was depicted against its gradient norm for each block in Figure 1. This figure motivates that a sub-quadratic relationship between the layer-wise smoothness constant and their gradient norms exists. Such condition is required for our convergence result in Theorem 4.2 and Theorem 4.3.

A.7.2 Sparse Dictionary Learning

Here, we explain the implementation structures of the sparse dictionary learning problem with more detail. Note that the structure of SDL problem fits with the more general analysis provided in Appendix A.2.

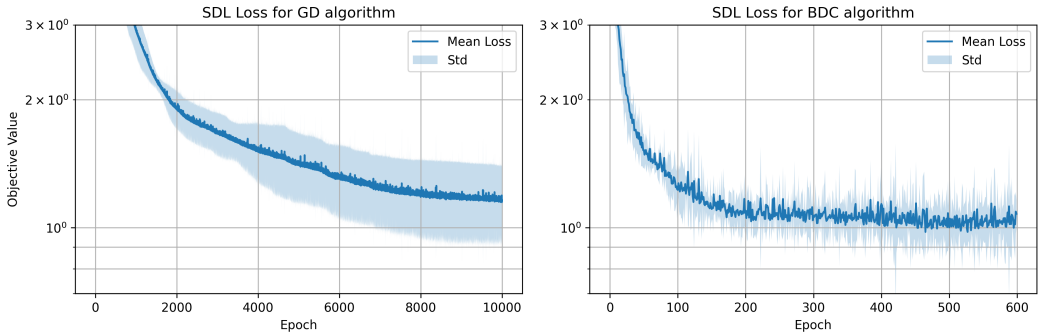


Figure 4: SDL loss evolution with the nonconvex $\ell_1 - \ell_Q$ regularizer (Eq. 5.5) on synthetic data. The left panel shows the loss trajectory of the joint full-batch gradient descent method with an adaptive step size, while the right panel shows the loss evolution of the BDC algorithm. Each curve reports the mean and ± 2 standard deviations over 3 independent runs.

Implementation Details. Both formulations (with ℓ_1 norm and (5.5)) are solved via alternating minimization. In each iteration, we first update X: for the ℓ_1 model, we use GD; for the nonconvex model (5.5), we employ the DC algorithm by linearizing the $\|\cdot\|_Q$ term and then applying GD to the resulting convex surrogate. Next, we update $\mathbf{D}$ using a Frank–Wolfe procedure, projecting onto $\mathcal{C}$ to enforce the unit-ℓ_2 constraints. A line search determines the optimal step size in each Frank–Wolfe update. We evaluate performance by the reconstruction error $\|Y - \mathbf{D}X\|_F^2$ and the proportion of zeros in X. Each experiment is repeated 10 times, and we report a 95% probabilistic bound in our plots. We compare the formulations on synthetic data and Berkeley segmentation dataset Martin et al. (2001).

Synthetic Data. We set $m = 10$, $l = 32$, and $n = 100$. A ground-truth dictionary $\mathbf{D}^* \in \mathbb{R}^{m \times l}$ is generated by sampling each entry i.i.d. from $\mathcal{N}(0, 1)$ and normalizing each column to unit ℓ_2-norm. The true sparse code matrix $X^* \in \mathbb{R}^{l \times n}$ has exactly five nonzero entries per column, drawn i.i.d. from $\mathcal{N}(0, 1)$. We synthesize the data as $Y = \mathbf{D}^* X^*$, using $\alpha = 0.1$ and $Q = 5$. Results are shown in Figure 2.

In addition to the previous experiment, we also compare the loss behavior of the SDL problem under the GD and BDC algorithms for the nonconvex $\ell_1 - \ell_Q$ regularizer (Eq. 5.5). At each iteration, the gradient descent algorithm performs a single joint full-batch update of both the dictionary D and the code matrix X, using the adaptive step size

$$\eta = \frac{1}{\|D\|_2^2 + \|X\|_2^2}$$

which is motivated by the block Lipschitz constants of the smooth reconstruction term. Dictionary feasibility is maintained by projecting each column of D onto the unit ℓ_2-ball after every update. The resulting loss curves for both GD and BDC are shown in Figure 4. As shown in Figure 4, BDCA converges faster and attains a lower objective value compared to GD.

Berkeley Segmentation Dataset. Martin et al. (2001) From the BSDS500 training set (200 images), we randomly extract 50 grayscale patches of size 8×8 from each image. Any patch that is identically zero is discarded; the remaining patches are demeaned and normalized to unit ℓ_2-norm, then assembled as columns of Y. For this experiment we use $\alpha = 0.2$, $Q = 5$, and $l = 256$. Results are shown in Figure 2.

A.7.3 Training Neural Networks

Here, we explain the implementation structures of the training problem with more detail.

Implementation details (Classification Task). We evaluate CIFAR10 and FASHIONMNIST using a batch size of 4096 and 10 random seeds. For FASHIONMNIST, we use a three-layer ReLU network

with sequential input-output dimensions $(28 \times 28, 512, 64, 10)$ and train for 100 epochs. For SGD, the step size is 8×10^{-2}. For (SBDC), each selected-block subproblem is solved by 50 inner gradient steps with step size 10^{-2}, and we set $\rho = 10^{-3}$. For CIFAR10, we use a four-layer ReLU network with sequential input-output dimensions $(3 \times 32 \times 32, 256, 128, 64, 10)$ and train for 50 epochs. For SGD, the step size is 10^{-1}. For (SBDC), each selected-block subproblem is solved by 150 inner gradient steps with step size 5×10^{-3}, and we set $\rho = 10^{-1}$. Figure 3 reports the mean curves and two-sided 90% Student-t confidence intervals across the 10 runs.

B Proofs

B.1 Proof of Proposition 3.1

Fix any block $i \in [n]$ and fix an arbitrary complement $\bar{\boldsymbol{\theta}}_i \in \bar{\mathcal{X}}_i$. We work with the i-th block (with $\bar{\boldsymbol{\theta}}_i$ fixed and θ_i free). By BDC assumption, for each $r \in \{1, \dots, m\}$ there exist functions

$$g_i^{(r)}(\cdot\,; \bar{\boldsymbol{\theta}}_i),\ h_i^{(r)}(\cdot\,; \bar{\boldsymbol{\theta}}_i) :\ \mathcal{X}_i \to \mathbb{R}$$

that are convex in θ_i such that

$$f_r(\boldsymbol{\theta}) \ = \ g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) \ - \ h_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i), \qquad \boldsymbol{\theta}_i \in \mathcal{X}_i.$$

We show that each operation preserves this BDC form.

Linear combinations. Let $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ and write $\alpha_r = \alpha_r^+ - \alpha_r^-$ with $\alpha_r^\pm \geq 0$. Then, for every $\boldsymbol{\theta}_i \in \mathcal{X}_i$,

$$\begin{aligned}
\sum_{r=1}^m \alpha_r\, f_r(\boldsymbol{\theta}) &= \sum_{r=1}^m \alpha_r^+ \big(g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i)\big) \\
&\quad - \sum_{r=1}^m \alpha_r^- \big(g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i)\big) \\
&= \underbrace{\Big(\sum_{r=1}^m \alpha_r^+\, g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) + \sum_{r=1}^m \alpha_r^-\, h_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i)\Big)}_{\text{convex in } \boldsymbol{\theta}_i} \\
&\quad - \underbrace{\Big(\sum_{r=1}^m \alpha_r^+\, h_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) + \sum_{r=1}^m \alpha_r^-\, g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i)\Big)}_{\text{convex in } \boldsymbol{\theta}_i}.
\end{aligned}$$

Each bracket is a nonnegative sum of convex functions of θ_i, hence convex. Therefore $\sum_{r=1}^m \alpha_r f_r$ is BDC.

Maximum. Using the BDC decompositions of all f_r, for every $\theta_i \in \mathcal{X}_i$,

$$\begin{aligned}
\max_{1 \leq r \leq m} f_r(\boldsymbol{\theta}) &= \max_{1 \leq r \leq m} \Big\{ g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) \Big\} \\
&= \max_{1 \leq r \leq m} \Big\{ g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) + \sum_{\substack{s=1 \\ s \neq r}}^m h_i^{(s)}(\theta_i; \bar{\boldsymbol{\theta}}_i) \Big\} \ - \ \sum_{k=1}^m h_i^{(k)}(\theta_i; \bar{\boldsymbol{\theta}}_i). \qquad \text{(B.1)}
\end{aligned}$$

For the fixed $\bar{\boldsymbol{\theta}}_i$, each inner map

$$\theta_i \ \mapsto \ g_i^{(r)}(\theta_i; \bar{\boldsymbol{\theta}}_i) + \sum_{\substack{s=1 \\ s \neq r}}^m h_i^{(s)}(\theta_i; \bar{\boldsymbol{\theta}}_i)$$

is convex in θ_i (sum of convex functions); the pointwise maximum over finitely many convex functions is convex in θ_i; and the final sum $\sum_{k=1}^m h_i^{(k)}(\theta_i; \bar{\boldsymbol{\theta}}_i)$ is convex in θ_i. Hence the right-hand side of (B.1) is a difference of two convex functions of θ_i, proving that $\max_r f_r$ is BDC.

Minimum. By the linear combination with $\alpha_r = -1$, the function $-f_r$ is BDC for each r. Applying the Maximum property to $\{-f_r\}_{r=1}^m$ and using

$$\min_{1 \leq r \leq m} f_r(\boldsymbol{\theta}) \ = \ - \max_{1 \leq r \leq m} \big(-f_r(\boldsymbol{\theta})\big),$$

we conclude that $\min_r f_r$ is BDC.

Since the block i was arbitrary, all three operations preserve the BDC property.

B.2 Proof of Theorem 3.2

We prove the theorem by treating the *even-degree* and *odd-degree* cases separately. In each case we first derive an *upper bound* via the polarization identity of monomials together with a precise pairing argument that halves the raw atom count, and then obtain a *lower bound* by relating any DC decomposition to a (real) Waring decomposition and invoking known rank formulas for monomials.

Preliminaries

Polarization identity of monomials. We use the following polynomial identity.

Lemma B.1 (Polarization identity of monomials Kan (2008))**.** *Let $b_1, \dots, b_M \in \mathbb{Z}_{\geq 0}$ with $S = \sum_{i=1}^{M} b_i$ and variables $\theta_1, \dots, \theta_M$. Then*

$$\prod_{i=1}^{M} \theta_i^{b_i} = \frac{1}{S!} \sum_{v_1=0}^{b_1} \cdots \sum_{v_M=0}^{b_M} (-1)^{\sum_{i=1}^{M} v_i} \prod_{i=1}^{M} \binom{b_i}{v_i} \Big(\sum_{i=1}^{M} (\tfrac{b_i}{2} - v_i)\theta_i \Big)^S.$$

Waring decompositions and ranks. A *Waring decomposition* of a degree-S homogeneous polynomial (form) F is an identity

$$F(\boldsymbol{\theta}) = \sum_{j=1}^{r} c_j \, \ell_j(\boldsymbol{\theta})^S, \qquad \ell_j(\boldsymbol{\theta}) = a_{j1}\theta_1 + \cdots + a_{jn}\theta_n.$$

We distinguish two notions:

- **Complex Waring rank** $\mathrm{rk}_{\mathbb{C}}(F)$: the minimal r for which there exist real scalars c_j and linear forms ℓ_j with complex coefficients such that $F = \sum_{j=1}^{r} c_j \, \ell_j^S$.
- **Real Waring rank** $\mathrm{rk}_{\mathbb{R}}(F)$: the minimal r for which there exist *real* c_j and *real-coefficient* ℓ_j such that $F = \sum_{j=1}^{r} c_j \, \ell_j^S$.

Allowing complex coefficients cannot increase the minimum, hence

$$\mathrm{rk}_{\mathbb{C}}(F) \;\leq\; \mathrm{rk}_{\mathbb{R}}(F).$$

Case 1: s even (linear-even atoms $(u^\top \boldsymbol{\theta})^s$).

First, we upper bound N via the polarization identity. Apply Lemma B.1 with $M = n$, $S = s$. This expresses f as a linear combination of $\prod_{i=1}^{n}(b_i + 1)$ degree-s powers of linear forms:

$$f(\boldsymbol{\theta}) = \frac{1}{s!} \sum_{v_1=0}^{b_1} \cdots \sum_{v_n=0}^{b_n} (-1)^{\sum_{i=1}^{n} v_i} \Big(\textstyle\prod_{i=1}^{n} \binom{b_i}{v_i} \Big) \Big(\sum_{i=1}^{n} (\tfrac{b_i}{2} - v_i)\theta_i \Big)^s.$$

Write $v = (v_1, \dots, v_n)$ and let the *complement* be $b - v = (b_1 - v_1, \dots, b_n - v_n)$. Since s is even, we have

$$\Big(\sum_{i=1}^{n} (\tfrac{b_i}{2} - (b_i - v_i))\theta_i \Big)^s = \Big(-\sum_{i=1}^{n} (\tfrac{b_i}{2} - v_i)\theta_i \Big)^s = \Big(\sum_{i=1}^{n} (\tfrac{b_i}{2} - v_i)\theta_i \Big)^s,$$

so the two atoms coincide. Thus each complementary pair $\{v, b - v\}$ contributes *twice* the same atom, and pairing halves the count to $\frac{1}{2}\prod_{i=1}^{n}(b_i + 1)$. If all b_i are even, when $v_i = b_i/2$, we have $\frac{b_i}{2} - v_i = 0$, so the exact number of nonzero atoms equals $\big(\prod_{i=1}^{n}(b_i+1) - 1\big)/2 = \big\lfloor \frac{1}{2}\prod_{i=1}^{n}(b_i+1) \big\rfloor$. Therefore,

$$N \;\leq\; \Big\lfloor \tfrac{1}{2} \prod_{i=1}^{n} (b_i + 1) \Big\rfloor.$$

Second, we find the lower bound N via Waring rank. Any DC split in this model can be written as

$$f(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) - h(\boldsymbol{\theta}) = \sum_{i=1}^{r} \alpha_i \, (u_i^\top \boldsymbol{\theta})^s \;-\; \sum_{i=r+1}^{r+q} \alpha_i \, (u_i^\top \boldsymbol{\theta})^s = \sum_{i=1}^{N} c_i \, (u_i^\top \boldsymbol{\theta})^s,$$

with $\alpha_i > 0$, $c_i = \pm\alpha_i$, and $N = r + q$. This is a Waring decomposition with real coefficients and real linear forms, hence

$$N \;\geq\; \mathrm{rk}_{\mathbb{R}}(f) \;\geq\; \mathrm{rk}_{\mathbb{C}}(f).$$

For monomials the complex rank is known exactly:

Lemma B.2 (Complex Waring rank of a monomial Carlini et al. (2012))**.** *Let $b_1, \ldots, b_n \in \mathbb{Z}_{\geq 0}$ with $s = \sum_{i=1}^n b_i$, and $1 \leq b_1 \leq \cdots \leq b_n$. For the monomial $f(\boldsymbol{\theta}) = \theta_1^{b_1} \cdots \theta_n^{b_n}$,*

$$\mathrm{rk}_{\mathbb{C}}(f) = \prod_{i=2}^{n} (b_i + 1).$$

Combining $N \geq \mathrm{rk}_{\mathbb{C}}(f)$ with Lemma B.2 yields

$$N \;\geq\; \prod_{i=2}^{n} (b_i + 1).$$

Finally, the relationship between real and complex ranks clarifies tightness:

Theorem B.3 (Carlini et al. (2017))**.** *Let $f(\boldsymbol{\theta}) = \theta_1^{b_1} \cdots \theta_n^{b_n}$ with $1 \leq b_1 \leq \cdots \leq b_n$. Then $\mathrm{rk}_{\mathbb{R}}(f) = \mathrm{rk}_{\mathbb{C}}(f)$ if and only if $b_1 = 1$.*

Hence, when $b_1 = 1$ the lower bound $\prod_{i=2}^{n}(b_i + 1)$ equals the complex (and real) rank, and together with the polarization upper bound we obtain matching bounds; for $b_1 > 1$ a strict gap can remain.

Case 2: s odd (affine-even atoms $(u^\top \boldsymbol{\theta} + \kappa)^{s+1}$). We now handle s odd, where $(u^\top \boldsymbol{\theta})^s$ is not convex. To remain in a convex-atom setting we use *even-degree affine atoms*, obtained via degree $(s + 1)$ homogenization.

Define

$$F(\boldsymbol{\theta}, t) \;=\; t\, f(\boldsymbol{\theta}) \;=\; t\, \theta_1^{b_1} \cdots \theta_n^{b_n},$$

which is homogeneous of even degree $S = s + 1$ in the variables $(\boldsymbol{\theta}, t) \in \mathbb{R}^n \times \mathbb{R}$. Any atom of the form $(u^\top \boldsymbol{\theta} + \kappa\, t)^S$ is convex (even power of an affine form). Evaluating any homogeneous decomposition of F at $t = 1$ yields atoms $(u^\top \boldsymbol{\theta} + \kappa)^{s+1}$, which remain convex in $\boldsymbol{\theta}$.

Now, we upper bound N via the polarization identity and pairing. Apply Lemma B.1 to the $(n+1)$-variate monomial $t\, \theta_1^{b_1} \cdots \theta_n^{b_n}$ with $(z_0, \ldots, z_n) = (t, \theta_1, \ldots, \theta_n)$, $(b_0, \ldots, b_n) = (1, b_1, \ldots, b_n)$, and $S = s + 1$. We obtain

$$F(\boldsymbol{\theta}, t) = \frac{1}{S!} \sum_{v_0=0}^{1} \sum_{v_1=0}^{b_1} \cdots \sum_{v_n=0}^{b_n} (-1)^{\sum_{i=0}^n v_i} \left(\prod_{i=0}^n \binom{b_i}{v_i} \right) \left(\sum_{i=0}^n \left(\tfrac{b_i}{2} - v_i \right) z_i \right)^S,$$

a signed sum of even powers of affine forms $(\kappa\, t + u^\top \boldsymbol{\theta})^S$. Pair each index $v = (v_0, \ldots, v_n)$ with its complement $b - v$. Hence the raw count $(b_0 + 1) \prod_{i=1}^n (b_i + 1) = 2 \prod_{i=1}^n (b_i + 1)$ collapses *exactly* by a factor 2, yielding

$$\#\text{atoms in } F \;=\; \prod_{i=1}^{n} (b_i + 1).$$

Setting $t = 1$ gives a DC decomposition of f with convex affine-power atoms $(u^\top \boldsymbol{\theta} + \kappa)^{s+1}$ and

$$N \;\leq\; \prod_{i=1}^{n} (b_i + 1).$$

Affine vs. homogeneous decompositions. In the odd-degree case we use degree-$d = s + 1$ powers of *affine* forms $(\ell(\boldsymbol{\theta}) + \beta)^d$. It is crucial that sums of such affine powers correspond *exactly* to homogeneous sums of degree-d powers of linear forms in one extra variable, with a *term-by-term* correspondence that preserves the number of terms.

Lemma B.4 (Affine–homogeneous correspondence)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be of degree d, and let its degree-d homogenization be $F(X_0, X) = X_0^d\, f(X/X_0)$, so that F is homogeneous of degree d and $F(1, \boldsymbol{\theta}) = f(\boldsymbol{\theta})$. Then the following are equivalent:*

1. $f(\boldsymbol{\theta}) = \sum_{j=1}^{r} c_j \, (\ell_j(\boldsymbol{\theta}) + \beta_j)^d$ *(affine sum of degree-d powers).*

2. $F(X_0, X) = \sum_{j=1}^{r} c_j \, (\beta_j X_0 + \ell_j(X))^d$ *(homogeneous sum of degree-d powers).*

Moreover, the number of terms r is preserved in both directions.

Proof. (1) $\Rightarrow$ (2): Substitute $\boldsymbol{\theta} = X/X_0$ and multiply by X_0^d, then expand: $F(X_0, X) = X_0^d f(X/X_0) = \sum_j c_j \, (\beta_j X_0 + \ell_j(X))^d$. (2) $\Rightarrow$ (1): Evaluate at $X_0 = 1$ to get $f(\boldsymbol{\theta}) = F(1, \boldsymbol{\theta}) = \sum_j c_j \, (\ell_j(\boldsymbol{\theta}) + \beta_j)^d$. Thus the atoms correspond bijectively and the count r is unchanged. □

By Lemma B.4 with $d = S = s + 1$, every DC decomposition of f into affine atoms $(\ell(\boldsymbol{\theta}) + \beta)^{s+1}$ induces a homogeneous Waring decomposition of F into atoms $(\beta t + \ell(\boldsymbol{\theta}))^{s+1}$ with the *same* number of terms, and conversely any homogeneous decomposition of F restricts at $t = 1$ to a decomposition of f with the *same* number of terms. Thus the minimal atom count in our odd-s DC model equals the affine Waring rank of f at degree $s + 1$, which by Lemma B.4 equals the Waring rank of F.

We lower bound N via Waring rank of the lifted monomial. Any such DC decomposition of f induces a decomposition of F of the form

$$F(\boldsymbol{\theta}, t) = \sum_{j=1}^{N} c_j \left(u_j^\top \boldsymbol{\theta} + \kappa_j t \right)^{s+1},$$

which is a real Waring decomposition of the $(n+1)$-variate monomial $t^1 \theta_1^{b_1} \cdots \theta_n^{b_n}$ of degree $S = s + 1$. The complex Waring rank of this monomial equals

$$\mathrm{rk}_{\mathbb{C}}(t^1 \theta_1^{b_1} \cdots \theta_n^{b_n}) = \prod_{i=1}^{n} (b_i + 1),$$

and, since the smallest exponent is 1, real and complex ranks coincide (see Theorem B.3). Therefore

$$N \;\geq\; \mathrm{rk}_{\mathbb{R}}(F) \;=\; \mathrm{rk}_{\mathbb{C}}(F) \;=\; \prod_{i=1}^{n} (b_i + 1).$$

Together with the upper bound we conclude

$$N \;=\; \prod_{i=1}^{n} (b_i + 1).$$

Conclusion. For even s, the polarization identity and complementary-index pairing yield $N \leq \left\lfloor \frac{1}{2} \prod_{i=1}^{n} (b_i + 1) \right\rfloor$, while the Waring-rank argument gives $N \geq \prod_{i=2}^{n} (b_i + 1)$; when $b_1 = 1$ these bounds are tight. For odd s, the degree-$(s + 1)$ homogenization $F(\boldsymbol{\theta}, t) = t f(\boldsymbol{\theta})$, the polarization identity in $n+1$ variables, and the corresponding Waring-rank lower bound match exactly, giving $N = \prod_{i=1}^{n} (b_i + 1)$.

B.3 Proof of Theorem 3.6

Fix an arbitrary block θ_l and hold all other blocks fixed. When we refer to 'convex' for a vector-valued function in the proof, we mean it in the componentwise sense. We consider two cases.

Case 1: Hidden block $\theta_l = (W_l, b_l)$.

First, We prove by induction on k that $Z_k^{\pm}$ are componentwise convex in (W_l, b_l) and satisfy $Z_k^{\pm} \geq 0$.

Base ($k < l$): For $k < l$, the quantities $Z_k^{\pm}$ do not depend on (W_l, b_l) and are thus constant (hence convex) w.r.t. (W_l, b_l). It remains to justify nonnegativity for these layers. at $k = 1$,

$$Z_1^+ = \sigma(W_1 x + b_1) \geq 0, \qquad Z_1^- = 0.$$

Assume $Z_s^{\pm} \geq 0$ for some $s < l - 1$. Then

$$Z_{s+1}^- = \sigma(W_{s+1}) Z_s^- + \sigma(-W_{s+1}) Z_s^+ \;\geq\; 0,$$

because $\sigma(W_{s+1})$ and $\sigma(-W_{s+1})$ are entrywise nonnegative. Moreover,

$$Z_{s+1}^+ = \max\Big\{ \underbrace{\sigma(W_{s+1})Z_s^+ + \sigma(-W_{s+1})Z_s^- + b_{s+1}}_{p_{s+1}},\ Z_{s+1}^- \Big\} \ \geq\ Z_{s+1}^- \ \geq\ 0.$$

By induction, $Z_k^\pm \geq 0$ for all $k < l$.

Layer $k = l$: With $Z_{l-1}^\pm \geq 0$ fixed,

$$p_l = \sigma(W_l)Z_{l-1}^+ + \sigma(-W_l)Z_{l-1}^- + b_l, \quad Z_l^- = \sigma(W_l)Z_{l-1}^- + \sigma(-W_l)Z_{l-1}^+.$$

Entrywise $w \mapsto \sigma(\pm w)$ are convex and nonnegative; multiplying by fixed nonnegative vectors $Z_{l-1}^\pm$ and adding the affine term b_l preserve convexity. Hence p_l and Z_l^- are convex, with $Z_l^- \geq 0$. Set

$$Z_l^+ \ =\ \max\{p_l, Z_l^-\},$$

which is convex (pointwise max preserves convexity) and satisfies $Z_l^+ \geq Z_l^- \geq 0$.

Induction ($k \to k{+}1$ for $k \geq l$): Assume $Z_k^\pm$ are convex in (W_l, b_l) and $Z_k^\pm \geq 0$. For fixed (W_{k+1}, b_{k+1}), the matrices $\sigma(W_{k+1})$ and $\sigma(-W_{k+1})$ are entrywise nonnegative constants. Thus

$$p_{k+1} = \sigma(W_{k+1})Z_k^+ + \sigma(-W_{k+1})Z_k^- + b_{k+1}, \quad Z_{k+1}^- = \sigma(W_{k+1})Z_k^- + \sigma(-W_{k+1})Z_k^+$$

are nonnegative linear images of (Z_k^+, Z_k^-) plus a constant; hence $Z_{k+1}^- \geq 0$ and both p_{k+1}, Z_{k+1}^- are convex. Finally,

$$Z_{k+1}^+ = \max\{p_{k+1}, Z_{k+1}^-\}$$

is convex and satisfies $Z_{k+1}^+ \geq Z_{k+1}^- \geq 0$. By induction, this holds for all $k \geq l$, in particular for $k = L - 1$. Using $\sigma(a - b) = \max\{a, b\} - b$ coordinatewise,

$$Z_{k+1}^+ - Z_{k+1}^- \ =\ \sigma\big(W_{k+1}(Z_k^+ - Z_k^-) + b_{k+1}\big),$$

so $a_{k+1} = Z_{k+1}^+ - Z_{k+1}^-$ and in particular $a_{L-1} = Z_{L-1}^+ - Z_{L-1}^-$.

At the end, keep (W_L, b_L) fixed. For each class c,

$$A_c(\boldsymbol{\theta}) = \sigma(W_{L,c})Z_{L-1}^+ + \sigma(-W_{L,c})Z_{L-1}^- + \sigma(b_{L,c}),$$
$$B_c(\boldsymbol{\theta}) = \sigma(W_{L,c})Z_{L-1}^- + \sigma(-W_{L,c})Z_{L-1}^+ + \sigma(-b_{L,c}).$$

Here $\sigma(\pm W_{L,c}) \geq 0$ and $\sigma(\pm b_{L,c}) \geq 0$ are constants; therefore $A_c(\cdot\,; \bar{\boldsymbol{\theta}}_l), B_c(\cdot\,; \bar{\boldsymbol{\theta}}_l)$ are nonnegative linear combinations of the convex functions $Z_{L-1}^\pm$ plus constants, hence are convex and nonnegative in $\theta_l = (W_l, b_l)$. Using $\sigma(t) - \sigma(-t) = t$ entrywise and $a_{L-1} = Z_{L-1}^+ - Z_{L-1}^-$, we obtain

$$A(\boldsymbol{\theta}) - B(\boldsymbol{\theta}) = \big(\sigma(W_L) - \sigma(-W_L)\big)(Z_{L-1}^+ - Z_{L-1}^-) + \big(\sigma(b_L) - \sigma(-b_L)\big) = W_L a_{L-1} + b_L = F_x(\boldsymbol{\theta}).$$

Case 2: Output block $\theta_L = (W_L, b_L)$.

Here $Z_{L-1}^\pm$ are fixed and nonnegative. The entrywise maps $W_L \mapsto \sigma(\pm W_L)$ and $b_L \mapsto \sigma(\pm b_L)$ are convex and nonnegative. Hence each component of $A(\cdot\,; \bar{\boldsymbol{\theta}}_L), B(\cdot\,; \bar{\boldsymbol{\theta}}_L)$ in (3.2) is a nonnegative linear combination of nonnegative convex functions, and is therefore convex and nonnegative in (W_L, b_L).

B.4 Proof of Proposition 3.4

Fix an arbitrary block $i \in [n]$ and fix $\bar{\boldsymbol{\theta}}_i$. By the componentwise BDC assumption, for each coordinate $j = 1, \ldots, m$ there exist convex functions $a_{ij}(\cdot\,; \bar{\boldsymbol{\theta}}_i)$ and $b_{ij}(\cdot\,; \bar{\boldsymbol{\theta}}_i)$ in θ_i such that $E_j(\boldsymbol{\theta}) = a_{ij}(\theta_i; \bar{\boldsymbol{\theta}}_i) - b_{ij}(\theta_i; \bar{\boldsymbol{\theta}}_i)$.

Multi-Block convexity of g. From the conjugate definition and $E_j = a_{ij} - b_{ij}$,

$$f^*(E(\boldsymbol{\theta})) = \max_{u \in U} \Big\{ \langle u, a_i(\theta_i; \bar{\boldsymbol{\theta}}_i)\rangle - \langle u, b_i(\theta_i; \bar{\boldsymbol{\theta}}_i)\rangle - f(u) \Big\}.$$

Adding h_i yields the variational form

$$g_i(\theta_i;\bar{\boldsymbol{\theta}}_i) = \max_{u\in U}\Big\{\langle u + c^+,\, a_i(\theta_i;\bar{\boldsymbol{\theta}}_i)\rangle + \langle -u + d^+,\, b_i(\theta_i;\bar{\boldsymbol{\theta}}_i)\rangle - f(u)\Big\}.$$

For any fixed $u \in U$, the map

$$\theta_i \;\mapsto\; \langle u + c^+,\, a_i(\theta_i;\bar{\boldsymbol{\theta}}_i)\rangle + \langle -u + d^+,\, b_i(\theta_i;\bar{\boldsymbol{\theta}}_i)\rangle - f(u)$$

is convex in θ_i since $u_j + c_j^+ \geq 0$ and $-u_j + d_j^+ \geq 0$ for all j, making it a nonnegative linear combination of convex functions. Taking the pointwise maximum over $u \in U$ preserves convexity, so $g_i(\cdot;\bar{\boldsymbol{\theta}}_i)$ is convex.

Multi-Block convexity of h. By definition,

$$h_i(\theta_i;\bar{\boldsymbol{\theta}}_i) = \langle c^+,\, a_i(\theta_i;\bar{\boldsymbol{\theta}}_i)\rangle + \langle d^+,\, b_i(\theta_i;\bar{\boldsymbol{\theta}}_i)\rangle,$$

which is a nonnegative linear combination of convex functions of θ_i, hence convex.

Finally, by construction,

$$f^*(E(\boldsymbol{\theta})) = g_i(\theta_i;\bar{\boldsymbol{\theta}}_i) - h_i(\theta_i;\bar{\boldsymbol{\theta}}_i),$$

so $f^* \circ E$ admits a multi-block DC decomposition. Since block i was arbitrary, $f^* \circ E$ is BDC.

B.5 Proof of Lemma A.3

Due to (4.3), we have

$$u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) \quad \text{and} \quad \langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \leq g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2. \tag{B.2}$$

Using convexity of $g_{i_k}(\cdot\,,\bar{\boldsymbol{\theta}}_{i_k}^k)$ in (B.2), we have

$$\begin{aligned}
&\langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \leq g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2,\\
\implies &\langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \leq -\langle \nabla g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k),\, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2,\\
\implies &\frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \leq \langle \nabla g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k,\, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle,\\
\implies &\frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \leq \|\nabla g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\|\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|,\\
\implies &\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \leq \frac{2}{\rho}\|\nabla g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\|.
\end{aligned}$$

We get that the update $\boldsymbol{\theta}^{k+1}$ is in $\mathcal{B}\left(\boldsymbol{\theta}^k, \frac{2}{\rho}\|\nabla g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\|\right)$.

B.6 Proof of Lemma A.4

From (4.3) we know:

$$u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) \quad \text{and} \quad \langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \leq g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2.$$

Now, using convexity of h_{i_k} we get:

$$\begin{aligned}
h_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) &\leq g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2,\\
g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) &\leq g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k) + h_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k),\\
f(\boldsymbol{\theta}^{k+1}) &\leq f(\boldsymbol{\theta}^k)
\end{aligned}$$

Unrolling this inequality to the initialization gives

$$\begin{aligned}
f(\boldsymbol{\theta}^{k+1}) &\leq f(\boldsymbol{\theta}^0)\\
g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) &\leq g_{i_0}(\theta_{i_0}^0;\bar{\boldsymbol{\theta}}_{i_0}^0) + h_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_0}(\theta_{i_0}^0;\bar{\boldsymbol{\theta}}_{i_0}^0),\\
&\leq g_{i_0}(\theta_{i_0}^0;\bar{\boldsymbol{\theta}}_{i_0}^0) + H,
\end{aligned}$$

where we have used $h_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - h_{i_0}(\theta_{i_0}^{0};\bar{\boldsymbol{\theta}}_{i_0}^{0}) \leq H$ and the fact that $\bar{\boldsymbol{\theta}}_{i_k}^{k+1} = \bar{\boldsymbol{\theta}}_{i_k}^{k}$. Since this result holds for any k, through Corollary A.11 we have $\|\nabla g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k})\| \leq E$. Recall that g_{i_k} is ℓ-smooth with a subquadratic ℓ. Using Proposition A.10, g_{i_k} is also (r,m)-smooth with $r(u) = \frac{a}{m(u)}$ and $m(u) := \ell(u+a)$ for some $a > 0$. Therefore, we can use Lemma A.9 if we ensure the updates are inside $\mathcal{B}(\boldsymbol{\theta}^k, r(E))$. Similar to the proof of Lemma A.3 (Appendix B.5), we know

$$\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \leq \sup_{u \in \partial h_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k})} \frac{2(\|\nabla g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k})\| + \|u\|)}{\rho} \leq \frac{2(E+R)}{\rho}.$$

As a result taking $\rho \geq \frac{2(E+R)}{r(E)} = \ell(2E)\frac{2(E+R)}{E}$ will satisfy the conditions in Lemma A.9. This implies the desired result.

B.7 Proof of Proposition A.5

Using the assumptions in the theorem statement, we know that for any $\boldsymbol{\theta}^k \in \mathcal{X}$, the update $\boldsymbol{\theta}^{k+1} \in \mathcal{B}(\boldsymbol{\theta}^k, r(E))$. For any $\theta_{i_k} \in \mathcal{B}(\theta_{i_k}^k, r(E))$, consider the surrogate function

$$\hat{f}(\theta_{i_k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) := g_{i_k}(\theta_{i_k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - h_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - \langle u_{i_k}^k, \theta_{i_k} - \theta_{i_k}^k\rangle + \frac{\rho}{2}\|\theta_{i_k}^k - \theta_{i_k}\|^2.$$

where $u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k})$ is the subgradient used at iteration k of the algorithm. From (4.3) we know that

$$\hat{f}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) \leq \hat{f}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}),$$

and further considering the descent Lemma A.9 for g_{i_k} with $L = \ell(2E)$, we get

$$\begin{aligned}\hat{f}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) \leq\ & g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) + \langle \nabla g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}), \theta_{i_k} - \theta_{i_k}^k\rangle \\ &+ \frac{L+\rho}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2 - h_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - \langle u_{i_k}^k, \theta_{i_k} - \theta_{i_k}^k\rangle.\end{aligned}$$

By Assumption 1, we get

$$\begin{aligned}&\hat{f}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) \leq f(\boldsymbol{\theta}^k) + \langle \nabla g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - u_{i_k}^k, \theta_{i_k} - \theta_{i_k}^k\rangle + \frac{L+\rho}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2,\\
&\langle \nabla g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}\rangle \leq f(\boldsymbol{\theta}^k) + \frac{L+\rho}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2 - \hat{f}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}),\\
&\langle \nabla g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}\rangle \leq g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) + \frac{L+\rho}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2\\
&- \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 + \langle u_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle,\end{aligned} \tag{B.3}$$

Using the non-positivity of $-\frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \leq 0$, and recalling the fact that here we used ∇g_i and ∂h_i instead of $\nabla_i g_i$ and $\partial_i h_i$, we have

$$\begin{aligned}\langle \nabla_{i_k} g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}\rangle - \frac{L+\rho}{2}&\|\theta_{i_k} - \theta_{i_k}^k\|^2\\
&\leq g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) + \langle u_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle,\\
&\leq g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) + h_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - h_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}),\\
&\leq f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^{k+1}).\end{aligned} \tag{B.4}$$

Define $z_{i_k}^k := \nabla_{i_k} g_{i_k}(\theta_{i_k}^{k};\bar{\boldsymbol{\theta}}_{i_k}^{k}) - u_{i_k}^k$. Using the infimum over $\overline{\partial}_{i_k} f(\boldsymbol{\theta}^k)$ instead of $z_{i_k}^k$ and maximizing the left hand side with respect to θ_{i_k} gives

$$\inf_{z_{i_k}^k \in \overline{\partial}_{i_k} f(\boldsymbol{\theta}^k)} \frac{1}{2(L+\rho)}\|z_{i_k}^k\|^2 \leq f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^{k+1}). \tag{B.5}$$

Let us denote by $\mathbb{E}_{|k}$ the conditional expectation with respect to the random selection of i_k, given all the random choices in the previous iterations. Then, we have

$$\begin{aligned}\mathbb{E}_{|k}\left[\inf_{z_{i_k}^k \in \overline{\partial}_{i_k} f(\boldsymbol{\theta}^k)} \frac{1}{2(L+\rho)}\|z_{i_k}^k\|^2\right] &= \frac{1}{n}\left(\sum_{i=1}^n \inf_{z_i^k \in \overline{\partial}_i f(\boldsymbol{\theta}^k)} \frac{1}{2(L+\rho)}\|z_i^k\|^2\right), \\ &= \frac{1}{n}\left(\frac{1}{2(L+\rho)}\sum_{i=1}^n \operatorname{dist}(0, \overline{\partial}_i f(\boldsymbol{\theta}^k))^2\right),\end{aligned} \tag{B.6}$$

Using (B.5) and (B.6) we have:

$$\frac{1}{2(L+\rho)}\mathcal{G}^2(\boldsymbol{\theta}^k) \leq nf(\boldsymbol{\theta}^k) - n\mathbb{E}_{|k}\left[f(\boldsymbol{\theta}^{k+1})\right]. \tag{B.7}$$

Now, taking expectation w.r.t. all the iterations we have

$$\mathbb{E}\left[\frac{1}{2(L+\rho)}\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \leq n\mathbb{E}\left[f(\boldsymbol{\theta}^k)\right] - n\mathbb{E}\left[f(\boldsymbol{\theta}^{k+1})\right]. \tag{B.8}$$

Finally, we take the average of this inequality over $k = 1, \ldots, K$:

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\frac{1}{2(L+\rho)}\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \leq \frac{n}{K}\left(f(\boldsymbol{\theta}^1) - \mathbb{E}\left[f(\boldsymbol{\theta}^{K+1})\right]\right) \leq \frac{n}{K}\left(f(\boldsymbol{\theta}^1) - f^\star\right). \tag{B.9}$$

which concludes the proof.

B.8 Proof of Proposition A.8

Denote

$$\boldsymbol{\epsilon}_k := \nabla \hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \hat{u}_{i_k}^k - \left(\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\right).$$

Let $\mathcal{F}_k$ contain all information before drawing i_k and s^k. The block i_k is uniform on $[n]$, independently of $\mathcal{F}_k$, and s^k is independent of $(\mathcal{F}_k, i_k)$. The subgradients u_i^k are $\mathcal{F}_k$-measurable, and $\nabla \hat{g}_{i_k}$ is the gradient of the sampled function used in (SBDC). Assumption 5 therefore gives

$$\mathbb{E}[\boldsymbol{\epsilon}_k \mid \mathcal{F}_k, i_k] = 0, \qquad \mathbb{E}[\|\boldsymbol{\epsilon}_k\|^2 \mid \mathcal{F}_k, i_k] \leq \sigma^2.$$

All probabilities and unconditional expectations below are over both samples and block choices.

We retain the stopping rules in (A.9), with t_1 testing $\max_{i\in[n]} g_i(\theta_i^{k+1}; \bar{\boldsymbol{\theta}}_i^{k+1}) - g^* > G$ and t_2 testing $\|\boldsymbol{\epsilon}_k\| > F'$, each capped at K, and $t = \min\{t_1, t_2\}$. Here, every component in the maximum is evaluated at the full updated vector $\boldsymbol{\theta}^{k+1}$. The definition of G ensures $\max_i g_i(\theta_i^0; \bar{\boldsymbol{\theta}}_i^0) - g^* \leq G$. Consequently,

$$\max_{i\in[n]} g_i(\theta_i^k; \bar{\boldsymbol{\theta}}_i^k) - g^* \leq G \qquad \text{for } 0 \leq k \leq t_1.$$

We want to bound the probability of

$$\{t < K\} = \{t_2 < K\} \cup \left(\{t_1 < K\} \cap \{t_2 = K\}\right).$$

For the first event, Chebyshev's inequality and the union bound give

$$\mathbb{P}(t_2 < K) = \mathbb{P}\left(\bigcup_{k<K}\{\|\boldsymbol{\epsilon}_k\| > F'\}\right) \leq \sum_{k<K}\mathbb{P}(\|\boldsymbol{\epsilon}_k\| > F') \leq \frac{K\sigma^2}{F'^2}. \tag{B.10}$$

We choose parameters so that $K\sigma^2/F'^2 \leq \delta/4$.

For $k \leq t_1$ and $k < t_2$, we have $g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g^* \leq G$ and $\|\boldsymbol{\epsilon}_k\| \leq F'$. Thus $\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)\| \leq E$ by Corollary A.11, and the update is bounded by Lemma A.6. Using Lemma A.7, we get

$$\begin{aligned} g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) &\leq \langle \nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k),\, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle + \frac{L}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \\ &\leq \|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)\|\|\theta_{i_k}^{k+1} - \boldsymbol{\theta}_{i_k}^k\| + \frac{L}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \\ &\leq E\frac{2}{\rho}(E + R + F') + \frac{L}{2}\left[\frac{2}{\rho}(E + R + F')\right]^2. \end{aligned} \tag{B.11}$$

Take $F' = {}^{E\rho}/_{9L} - (E+R)$ and note that $E^2 = 2LG$. F' is positive for $\rho > {}^{9L(E+R)}/_{E}$. This is a valid choice of ρ since it satisfies (see Lemma A.7):

$$\rho \geq \frac{2(E+R+F')}{r(E)}.$$

Now, replacing in (B.11) gives:

$$g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^{k}) - g_{i_k}(\theta_{i_k}^{k}; \bar{\boldsymbol{\theta}}_{i_k}^{k}) \leq \frac{G}{2}. \tag{B.12}$$

For the exit probability, we use the population objective so that the argument applies to whichever component attains the maximum. Due to (SBDC), we have $u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$ and

$$\langle \hat{u}_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1} \rangle \leq g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2. \tag{B.13}$$

By adding and subtracting $\langle \hat{u}_{i_k}^k - u_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k \rangle$ and using convexity of the sampled function $g_{i_k}(\cdot; \bar{\boldsymbol{\theta}}_{i_k}^k, s^k)$, we have

$$\begin{aligned}
&\langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1} \rangle \leq \langle \hat{u}_{i_k}^k - u_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k \rangle - \langle \nabla \hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k),\, \theta_{i_k}^{k+1} - \theta_{i_k}^k \rangle - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2, \\
&\implies \langle \nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k), \theta_{i_k}^{k+1} - \theta_{i_k}^k \rangle \leq \langle \boldsymbol{\epsilon}_k, \theta_{i_k}^k - \theta_{i_k}^{k+1} \rangle - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 - \langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1} \rangle,
\end{aligned}$$

For $k \leq t_1$ and $k < t_2$, the conditions of Lemma A.7 are satisfied. Therefore, local smoothness of g_{i_k} together with Young's inequality imply:

$$\begin{aligned}
g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \frac{L}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 & \\
\leq \frac{\rho}{4}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 + \frac{1}{\rho}\|\boldsymbol{\epsilon}_k\|^2 - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 - \langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1} \rangle & \\
= \frac{1}{\rho}\|\boldsymbol{\epsilon}_k\|^2 - \langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1} \rangle - \frac{\rho}{4}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2. &
\end{aligned}$$

Now, using $\rho \geq {}^{9L(E+R)}/_{E}$, we know that $\frac{L}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \leq \frac{\rho}{4}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2$. Therefore, we have:

$$g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) \leq \frac{1}{\rho}\|\boldsymbol{\epsilon}_k\|^2 + h_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k).$$

The last inequality follows from convexity of h_{i_k}. Since $f(\boldsymbol{\theta}) = g_i(\theta_i; \bar{\boldsymbol{\theta}}_i) - h_i(\theta_i; \bar{\boldsymbol{\theta}}_i)$ for every block i, it follows that

$$f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^k) \leq \frac{\|\boldsymbol{\epsilon}_k\|^2}{\rho} \qquad (k \leq t_1,\ k < t_2).$$

The assumed bound on h holds for any chosen block. In particular, $h_i(\theta_i^m; \bar{\boldsymbol{\theta}}_i^m) \leq \max_j h_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) + H$ for every i and every iterate $0 \leq m \leq K$ used below. Therefore,

$$0 \leq \max_i g_i(\theta_i^m; \bar{\boldsymbol{\theta}}_i^m) - g^* \leq f(\boldsymbol{\theta}^m) - f(\boldsymbol{\theta}^0) + \max_j g_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) - g^* + H. \tag{B.14}$$

Summing the population descent inequality along each trajectory, before taking expectations, gives, for $0 \leq m \leq \min\{t_1 + 1, t_2, K\}$,

$$\max_i g_i(\theta_i^m; \bar{\boldsymbol{\theta}}_i^m) - g^* \leq \max_j g_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) - g^* + H + \frac{1}{\rho}\sum_{k<m}\|\boldsymbol{\epsilon}_k\|^2.$$

On $\{t_1 < K, t_2 = K\}$, take $m = t_1 + 1$. The left-hand side exceeds G, and the noise sum is bounded by the sum over all $k < K$. Hence

$$\mathbb{P}(t_1 < K, t_2 = K) \leq \mathbb{P}\left(\max_j g_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) - g^* + H + \frac{1}{\rho}\sum_{k<K}\|\boldsymbol{\epsilon}_k\|^2 > G\right). \tag{B.15}$$

Assumption 5 gives

$$\mathbb{E}_{s,i}\left[\max_j g_j(\theta_j^0;\bar{\boldsymbol{\theta}}_j^0) - g^* + H + \frac{1}{\rho}\sum_{k<K}\|\boldsymbol{\epsilon}_k\|^2\right] \le \max_j g_j(\theta_j^0;\bar{\boldsymbol{\theta}}_j^0) - g^* + C', \tag{B.16}$$

where $C' := K\sigma^2/\rho + H$. The quantity inside the expectation is nonnegative. Markov's inequality and $G = 8(\max_j g_j(\theta_j^0;\bar{\boldsymbol{\theta}}_j^0) - g^* + C')/\delta$ therefore imply

$$\mathbb{P}(t_1 < K, t_2 = K) \le \frac{\max_j g_j(\theta_j^0;\bar{\boldsymbol{\theta}}_j^0) - g^* + C'}{G} = \frac{\delta}{8} \le \frac{\delta}{4}. \tag{B.17}$$

Using $F' = E\rho/(9L) - (E+R)$, the remaining parameter condition is

$$\frac{K\sigma^2}{(\frac{E\rho}{9L} - (E+R))^2} \le \frac{\delta}{4}. \tag{B.18}$$

Since $E^2 = 2LG$, we have

$$\begin{aligned} F'^2 &= \frac{2G\rho^2}{81L} + (E+R)^2 - \frac{2\rho}{9L}(2LG + ER) \\ &\ge \frac{2G\rho^2}{81L} - \frac{2\rho}{9L}(2LG + ER). \end{aligned} \tag{B.19}$$

For $C' > 0$, using $C' - H = K\sigma^2/\rho$ and $G\delta \ge 8C'$, a sufficient condition for the last expression to be at least $4K\sigma^2/\delta$ is

$$\begin{aligned} &\rho^2 - \frac{9\rho}{G}(2LG + ER) \ge \frac{81L\rho(C'-H)}{4C'}, \\ &\text{equivalently,} \qquad \rho \ge 18L + \frac{9ER}{G} + \frac{81L}{4}\frac{C'-H}{C'}. \end{aligned} \tag{B.20}$$

This also implies $\rho > 9L(E+R)/E$ and $\rho \ge 2L$. Thus the stated choice $\rho = (18L + 9ER/G + (81L/4)(C'-H)/C')\sqrt{K}$ is sufficient for $K \ge 1$. If $C' = 0$, then $H = \sigma^2 = 0$; the last term in this choice can be omitted. Combining the two exit estimates, we obtain

$$\mathbb{P}(t < K) \le \mathbb{P}(t_2 < K) + \mathbb{P}(t_1 < K, t_2 = K) \le \frac{\delta}{2}.$$

We next prove the stationarity bound using a population comparator. For every block, denote

$$z_i^k := \nabla g_i(\theta_i^k;\bar{\boldsymbol{\theta}}_i^k) - u_i^k \in \overline{\partial}_i f(\boldsymbol{\theta}^k),$$

and define

$$\Psi_k := \langle z_{i_k}^k + \boldsymbol{\epsilon}_k,\, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle + \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2.$$

Convexity of the sampled function and optimality of the update imply, for any feasible comparison point θ_{i_k},

$$\begin{aligned} \Psi_k &\le g_{i_k}(\theta_{i_k}^{k+1};\bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - \langle \hat{u}_{i_k}^k,\, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle + \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \\ &\le g_{i_k}(\theta_{i_k};\bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - g_{i_k}(\theta_{i_k}^k;\bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - \langle \hat{u}_{i_k}^k,\, \theta_{i_k} - \theta_{i_k}^k\rangle + \frac{\rho}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2. \end{aligned} \tag{B.21}$$

In particular, comparison with $\theta_{i_k}^k$ and completion of the square give the bounds

$$-\frac{\|z_{i_k}^k + \boldsymbol{\epsilon}_k\|^2}{2\rho} \le \Psi_k \le 0. \tag{B.22}$$

At a point satisfying $\max_i g_i(\theta_i^k;\bar{\boldsymbol{\theta}}_i^k) - g^* \le G$, we have $\|z_i^k\| \le E + R$ for every block. Condition on $\mathcal{F}_k$ and i_k, and choose the comparison point

$$\theta_{i_k} = \theta_{i_k}^k - \frac{z_{i_k}^k}{L+\rho}.$$

This point is fixed before drawing s^k, and its displacement satisfies

$$\frac{\|z_{i_k}^k\|}{L+\rho} \le \frac{E+R}{L+\rho} < \frac{E}{9L} < r(E) = \frac{E}{L}.$$

Thus it lies in the population smoothness ball. Taking expectation in (B.21), using the population representation of g_i, unbiasedness of $\hat{u}_{i_k}^k$, and population smoothness, gives

$$\begin{aligned}\mathbb{E}[\Psi_k \mid \mathcal{F}_k, i_k] &\le g_{i_k}\left(\theta_{i_k}^k - \frac{z_{i_k}^k}{L+\rho}; \bar{\boldsymbol{\theta}}_{i_k}^k\right) - g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) + \frac{\langle u_{i_k}^k, z_{i_k}^k\rangle}{L+\rho} + \frac{\rho\|z_{i_k}^k\|^2}{2(L+\rho)^2} \\ &\le -\frac{\|z_{i_k}^k\|^2}{L+\rho} + \frac{\|z_{i_k}^k\|^2}{2(L+\rho)} = -\frac{\|z_{i_k}^k\|^2}{2(L+\rho)}.\end{aligned} \tag{B.23}$$

Note that the expectation here uses the full current sample distribution and is not conditioned on $k < t$. The event $\{k \le t\}$ is $\mathcal{F}_k$-measurable, since it only requires that no previous iteration has triggered a stop. On this event, the starting point satisfies the gradient bound for every block. Averaging (B.23) over the current uniform block gives

$$\mathbb{E}\left[\mathbf{1}_{\{k\le t\}}\Psi_k \mid \mathcal{F}_k\right] \le -\frac{\mathbf{1}_{\{k\le t\}}}{2n(L+\rho)}\sum_{i=1}^n \|z_i^k\|^2 \le -\frac{\mathbf{1}_{\{k\le t\}}}{2n(L+\rho)}\mathcal{G}^2(\boldsymbol{\theta}^k). \tag{B.24}$$

Define, for $0 \le k < K$,

$$\xi_k := \mathbf{1}_{\{k\le t\}}\Psi_k - \mathbb{E}\left[\mathbf{1}_{\{k\le t\}}\Psi_k \mid \mathcal{F}_k\right], \qquad \xi_K := 0.$$

By (B.22) and Assumption 5 we know that $\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0$, and $\xi_k = 0$ for $k > t$ which makes these variables integrable. For $k < t$, smoothness and convexity of h_{i_k} give

$$\begin{aligned}f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^k) &\le \langle z_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle + \frac{L}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \\ &= \Psi_k - \langle \boldsymbol{\epsilon}_k, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle - \frac{\rho - L}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \\ &\le \Psi_k + \frac{\|\boldsymbol{\epsilon}_k\|^2}{2(\rho - L)} \le \Psi_k + \frac{\|\boldsymbol{\epsilon}_k\|^2}{\rho},\end{aligned}$$

where $\rho \ge 2L$. Consequently, (B.24) yields

$$f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^{k+1}) \ge \frac{\mathcal{G}^2(\boldsymbol{\theta}^k)}{2n(L+\rho)} - \xi_k - \frac{\|\boldsymbol{\epsilon}_k\|^2}{\rho}, \qquad k < t. \tag{B.25}$$

Since ξ_k is centered and vanishes after t,

$$\mathbb{E}_{s,i}\left[\sum_{k<t}\xi_k\right] = -\mathbb{E}_{s,i}\left[\mathbf{1}_{\{t<K\}}\xi_t\right].$$

The terminal term must be retained because $\{k < t\}$ depends on the current draws. For $k < K$, on $\{k \le t\}$, (B.22), conditional unbiasedness, and $\|z_i^k\| \le E + R$ imply

$$-\frac{(E+R)^2 + \sigma^2}{2\rho} \le \mathbb{E}[\Psi_k \mid \mathcal{F}_k] \le 0, \qquad |\xi_k| \le \frac{\|z_{i_k}^k + \boldsymbol{\epsilon}_k\|^2 + (E+R)^2 + \sigma^2}{2\rho}.$$

Moreover,

$$\mathbb{E}_{s,i}\left[\sum_{k<K}\mathbf{1}_{\{t=k\}}\|\boldsymbol{\epsilon}_k\|^2\right] \le \sum_{k<K}\mathbb{E}_{s,i}[\|\boldsymbol{\epsilon}_k\|^2] \le K\sigma^2.$$

It follows that

$$\mathbb{E}_{s,i}\left[\mathbf{1}_{\{t<K\}}|\xi_t|\right] \le \frac{3(E+R)^2 + (2K+1)\sigma^2}{2\rho}. \tag{B.26}$$

From (B.14), $f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^t) \le \max_j g_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) - g^* + H$. Summing (B.25) and using (B.26), we obtain

$$\begin{aligned}\frac{1}{2n(L+\rho)}\mathbb{E}_{s,i}\left[\sum_{k<t}\mathcal{G}^2(\boldsymbol{\theta}^k)\right] &\le \mathbb{E}_{s,i}\left[f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^t)\right] + \frac{K\sigma^2}{\rho} + \mathbb{E}_{s,i}\left[\mathbf{1}_{\{t<K\}}|\xi_t|\right] \\ &\le \max_j g_j(\theta_j^0; \bar{\boldsymbol{\theta}}_j^0) - g^* + C' + \frac{3(E+R)^2 + (2K+1)\sigma^2}{2\rho} \quad \text{(B.27)} \\ &= \frac{G\delta}{8} + \frac{3(E+R)^2 + (2K+1)\sigma^2}{2\rho}.\end{aligned}$$

Since the squared residuals are nonnegative, this implies

$$\mathbb{E}_{s,i}\left[\mathbf{1}_{\{t=K\}}\frac{1}{K}\sum_{k<K}\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \le \frac{2n(L+\rho)}{K}\left[\frac{G\delta}{8} + \frac{3(E+R)^2 + (2K+1)\sigma^2}{2\rho}\right].$$

Choose K so that

$$K \ge \frac{4n(L+\rho)}{\delta\epsilon^2}\left[\frac{G\delta}{8} + \frac{3(E+R)^2 + (2K+1)\sigma^2}{2\rho}\right]. \quad \text{(B.28)}$$

The preceding expectation is then at most $\delta\epsilon^2/2$. Define the event

$$\varrho := \left\{\frac{1}{K}\sum_{k<K}\mathcal{G}^2(\boldsymbol{\theta}^k) > \epsilon^2\right\}.$$

Markov's inequality gives

$$\mathbb{P}(\{t=K\}\cap\varrho) \le \frac{1}{\epsilon^2}\mathbb{E}_{s,i}\left[\mathbf{1}_{\{t=K\}}\frac{1}{K}\sum_{k<K}\mathcal{G}^2(\boldsymbol{\theta}^k)\right] \le \frac{\delta}{2}.$$

Therefore,

$$\mathbb{P}(\varrho) \le \mathbb{P}(t<K) + \mathbb{P}(\{t=K\}\cap\varrho) \le \delta.$$

In particular, with probability at least $1-\delta$ over samples and block choices,

$$\min_{0\le k<K}\mathcal{G}^2(\boldsymbol{\theta}^k) \le \frac{1}{K}\sum_{k<K}\mathcal{G}^2(\boldsymbol{\theta}^k) \le \epsilon^2.$$

With $\rho = \Theta(\sqrt{K})$ and $\sigma^2 = \mathcal{O}(K^{-1/2})$, both C' and the terminal correction are bounded independently of K. For fixed confidence and problem constants, the sufficient horizon (B.28) therefore implies the iteration order $\mathcal{O}(n^2/\epsilon^4)$.

B.9 Proof of Lemma A.6

Due to (SBDC), we have $u_{i_k}^k \in \partial h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$ and

$$\langle \hat{u}_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \le g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k, s^k) - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2. \quad \text{(B.29)}$$

By adding and subtracting $\langle \hat{u}_{i_k}^k - u_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle$ and using convexity of g_{i_k}, we have

$$\begin{aligned}&\langle u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \le \langle \hat{u}_{i_k}^k - u_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle - \langle \nabla\hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k),\, \theta_{i_k}^{k+1} - \theta_{i_k}^k\rangle - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2, \\ \implies &\langle u_{i_k}^k - \nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k), \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \le \langle \nabla\hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \hat{u}_{i_k}^k - \left(\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\right), \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \\ &\qquad - \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2,\end{aligned}$$

Now, through Cauchy–Schwarz inequality we get

$$\begin{aligned}\frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \le &\langle \nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k,\, \theta_{i_k}^k - \theta_{i_k}^{k+1}\rangle \\ &+ \|\nabla\hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \hat{u}_{i_k}^k - \left(\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\right)\|\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|, \\ \implies \frac{\rho}{2}\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|^2 \le &\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k\|\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\| + \|\hat{z}_{i_k}^k - z_{i_k}^k\|\|\theta_{i_k}^{k+1} - \theta_{i_k}^k\|, \\ \implies \|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \le &\frac{2}{\rho}\left(\|z_{i_k}^k\| + \|\hat{z}_{i_k}^k - z_{i_k}^k\|\right),\end{aligned}$$

where $z_{i_k}^k \in \overline{\partial_{i_k}} f(\boldsymbol{\theta}^k)$, $\hat{z}_{i_k}^k \in \overline{\partial_{i_k}} \hat{f}(\boldsymbol{\theta}^k)$, and the last line holds due to the fact that (SBDC) updates only the i_k^{th} block at each iteration. This implies the desired result.

B.10 Proof of Lemma A.7

By assumption, we know that $g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g^* \leq G$ and $\|\nabla \hat{g}_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \hat{u}_{i_k}^k - (\nabla_{i_k} g(\boldsymbol{\theta}^k) - u_{i_k}^k)\| \leq F'$ for $G, F' > 0$. Since this result holds for any k, through Corollary A.11 we have $\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)\| \leq E$. Recall that g_{i_k} is ℓ-smooth with a subquadratic ℓ. Using Proposition A.10, g is also (r, m)-smooth with $r(u) = \frac{a}{m(u)}$ and $m(u) := \ell(u + a)$ for some $a > 0$. Therefore, we can use Lemma A.9 if we ensure the updates are inside $\mathcal{B}(\boldsymbol{\theta}^k, r(E))$. From the proof of Lemma A.6 (Appendix B.9), we know

$$\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \leq \sup_{u \in \partial h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)} \frac{2(\|\nabla g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)\| + \|u\| + F')}{\rho} \leq \frac{2(E + R + F')}{\rho}.$$

As a result taking $\rho \geq \frac{2(E+R+F')}{r(E)} = \ell(2E)\frac{2(E+R+F')}{E}$ will satisfy the conditions in Lemma A.9. This implies the desired result.

B.11 Proof of Theorem A.1

For any $\boldsymbol{\theta}_{i_k} \in \mathcal{M}^{i_k}$, Consider the surrogate function

$$\hat{f}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) := g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \langle u_{i_k}^k, \theta_{i_k} - \theta_{i_k}^k \rangle.$$

where $u_{i_k}^k \in \partial_{i_k} h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k)$. Also, it is important to mention that $f(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) = f(\boldsymbol{\theta}^k)$ and $\hat{f}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) = \hat{f}(\boldsymbol{\theta}^k)$. From (A.4) we know that

$$\hat{f}(\boldsymbol{\theta}^{k+1}) \leq \hat{f}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) = g_{i_k}(\theta_{i_k}; \bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \langle u_{i_k}^k, \theta_{i_k} - \theta_{i_k}^k \rangle,$$

and further considering the smoothness of g, we get

$$\hat{f}(\boldsymbol{\theta}^{k+1}) \leq g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) + \langle \nabla_{i_k} g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k), \theta_{i_k} - \theta_{i_k}^k \rangle + \frac{L}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2 - h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - \langle u_{i_k}^k, \theta_{i_k} - \theta_{i_k}^k \rangle.$$

By (A.2), we get

$$\hat{f}(\boldsymbol{\theta}^{k+1}) \leq f(\boldsymbol{\theta}^k) + \langle \nabla_{i_k} g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k, \theta_{i_k} - \theta_{i_k}^k \rangle + \frac{L}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2$$

Therefore,

$$\begin{aligned}
\langle \nabla_{i_k} g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k} \rangle \leq & f(\boldsymbol{\theta}^k) - \hat{f}(\boldsymbol{\theta}^{k+1}) + \frac{L}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2 \\
\leq & g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k) + \frac{L}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2 \\
& + \langle u_{i_k}^k, \theta_{i_k}^{k+1} - \theta_{i_k}^k \rangle.
\end{aligned}$$

Therefore, by convexity of $h_{i_k}(\cdot\,; \bar{\boldsymbol{\theta}}_{i_k}^k)$ we have:

$$\begin{aligned}
\langle \nabla_{i_k} g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k} \rangle - \frac{L}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2 \leq & g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - g_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k) \\
& + h_{i_k}(\theta_{i_k}^{k+1}; \bar{\boldsymbol{\theta}}_{i_k}^k) - h_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k), \\
\langle \nabla_{i_k} g_{i_k}(\theta_{i_k}^k; \bar{\boldsymbol{\theta}}_{i_k}^k) - u_{i_k}^k, \theta_{i_k}^k - \theta_{i_k} \rangle - \frac{L}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2 \leq & f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^{k+1}).
\end{aligned}$$

Let us denote by $\mathbb{E}_{|k}$ the conditional expectation with respect to the random selection of i_k, given all the random choices in the previous iterations. Then, maximizing over θ_{i_k}, we have

$$\mathbb{E}_{|k}\left[\max_{\theta_{i_k} \in \mathcal{M}^{i_k}} \inf_{z_{i_k}^k \in \overline{\partial}_{i_k} f(\boldsymbol{\theta}^k)} \langle z_{i_k}^k, \theta_{i_k}^k - \theta_{i_k} \rangle - \frac{L}{2}\|\theta_{i_k} - \theta_{i_k}^k\|^2\right] = \frac{1}{n}\sum_{i=1}^n \text{gap}_{\mathcal{M}^{d_i}}^L(\boldsymbol{\theta}^k) \tag{B.30}$$

for $z_i^k := \nabla_i g_i(\theta_i^k; \bar{\boldsymbol{\theta}}_i^k) - u_i^k \in \overline{\partial}_{i_k} f(\boldsymbol{\theta}^k)$. Using (B.30) we have

$$\sum_{i=1}^{n} \operatorname{gap}_{\mathcal{M}^i}^{L}(\boldsymbol{\theta}^k) \leq n f(\boldsymbol{\theta}^k) - n\mathbb{E}_{|k}\left[f(\boldsymbol{\theta}^{k+1})\right].$$

Now, taking expectation w.r.t. all the iterations we have

$$\mathbb{E}\left[\sum_{i=1}^{n} \operatorname{gap}_{\mathcal{M}^i}^{L}(\boldsymbol{\theta}^k)\right] \leq n\mathbb{E}\left[f(\boldsymbol{\theta}^k)\right] - n\mathbb{E}\left[f(\boldsymbol{\theta}^{k+1})\right].$$

Finally, we take the average of this inequality over $k = 1, \ldots, K$:

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\sum_{i=1}^{n} \operatorname{gap}_{\mathcal{M}^i}^{L}(\boldsymbol{\theta}^k)\right] \leq \frac{n}{K}\left(f(\boldsymbol{\theta}^1) - \mathbb{E}\left[f(\boldsymbol{\theta}^{K+1})\right]\right) \leq \frac{n}{K}\left(f(\boldsymbol{\theta}^1) - f^\star\right).$$

We complete the proof by noting that the minimum of $\sum_{i=1}^{n} \operatorname{gap}_{\mathcal{M}^i}^{L}(\boldsymbol{\theta}^k)$ over $k = 1, \ldots, K$ is smaller than or equal to the average gap summation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contribution is the study of a new class of functions and its applications, which is reflected in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the final section of the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have provided detailed explanations of the proofs such that the reader can follow them step by step. The entire set of assumptions and proofs was checked to make sure they are correct.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, detailed explanation of all the experiments are given in Appendix A.7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We believe in open science and transparency. Therefore we attached the code of the implementation along with detailed explanations on the simulations in the appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The code and its explanation was attached/given in the paper Appendix A.7.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Figures 2 and 3 contain statistical significance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All the explanations were given in the numerics. Additional explanations were provided in the Appendix A.7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the authors made sure that the research done here conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification:[N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The instruction on how to use the code is provided.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.