SIAM J. OPTIM.
Vol. 35, No. 2, pp. 842–871

BLOCK MAJORIZATION-MINIMIZATION WITH DIMINISHING RADIUS FOR CONSTRAINED NONSMOOTH NONCONVEX OPTIMIZATION*

HANBAEK LYU† AND YUCHEN LI†

Abstract. Block majorization-minimization (BMM) is a simple iterative algorithm for constrained nonconvex optimization that sequentially minimizes majorizing surrogates of the objective function in each block while the others are held fixed. BMM entails a large class of optimization algorithms such as block coordinate descent and its proximal-point variant, expectation-minimization, and block projected gradient descent. We first establish that for general constrained nonsmooth nonconvex optimization, BMM with ρ-strongly convex and L_g-smooth surrogates can produce an ε-approximate first-order optimal point within $\widetilde{O}((1+L_g+\rho^{-1})\varepsilon^{-2})$ iterations and asymptotically converges to the set of first-order optimal points. Next, we show that BMM combined with a trust-region method with diminishing radius has an improved complexity of $\widetilde{O}((1+L_g)\varepsilon^{-2})$, independent of the inverse strong convexity parameter ρ^{-1}, allowing improved theoretical and practical performance with "flat" surrogates. Our results hold robustly even when the convex subproblems are solved inexactly as long as the optimality gaps are summable. Central to our analysis is a novel continuous first-order optimality measure, by which we bound the worst-case suboptimality in each iteration by the first-order improvement the algorithm makes. We apply our general framework to obtain new results on various algorithms such as the celebrated multiplicative update algorithm for nonnegative matrix factorization by Lee and Seung, regularized nonnegative tensor decomposition, and the classical block projected gradient descent algorithm. Lastly, we numerically demonstrate that the additional use of diminishing radius can improve the convergence rate of BMM in many instances.

Key words. block majorization-minimization, block coordinate descent, constrained optimization, nonsmooth nonconvex optimization, trust-region

MSC codes. 90C26, 90C59, 90C60, 90C90

DOI. 10.1137/23M1604515

1. Introduction.

Throughout this paper, we are interested in the minimization of a continuous function $F:\mathcal{E}:=\mathbb{R}^{I_1}\times\cdots\times\mathbb{R}^{I_m}\to[0,\infty)$ on a Cartesian product of closed convex sets $\boldsymbol{\Theta}=\Theta^{(1)}\times\cdots\times\Theta^{(m)}$:

$$\boldsymbol{\theta}^*\in\mathop{\arg\min}_{\boldsymbol{\theta}=[\theta^{(1)},\dots,\theta^{(m)}]\in\boldsymbol{\Theta}}\left(F(\boldsymbol{\theta}):=f(\boldsymbol{\theta})+p(\boldsymbol{\theta})\right). \tag{1.1}$$

The objective function F is the sum of a smooth (possibly nonconvex) part f and a nonsmooth (continuous and convex) part p. Under a minimal set of assumptions, we investigate how to obtain first-order optimal points of (1.1) from an arbitrary initialization.

To obtain first-order optimal solutions to (1.1), we consider algorithms based on *block majorization-minimization* (BMM) [17]. The high-level idea of BMM is that, in order to minimize a multiblock objective, one can minimize a majorizing surrogate of the objective in each block, say, in a cyclic order: For $n\geq 1$ and $i=1,\dots,m$,

*Received by the editors September 24, 2023; accepted for publication (in revised form) January 21, 2025; published electronically April 25, 2025.
https://doi.org/10.1137/23M1604515
Funding: The first author is partially supported by NSF awards DMS-2206296 and DMS-2010035. The research of the second author is supported in part by NSF award DMS-2023239 through the Institute for Foundations of Data Science.
†Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 USA (hlyu@math.wisc.edu, yli966@wisc.edu).

$$
\text{(1.2)} \quad \textbf{BMM} \quad \begin{cases} g_n^{(i)} \leftarrow \begin{bmatrix} \text{Majorizing surrogate of} \\ \theta \mapsto f_n^{(i)}(\theta) := f\left(\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \ldots, \theta_{n-1}^{(m)}\right) \end{bmatrix}, \\ p_n^{(i)}(\theta) := p\left(\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \ldots, \theta_{n-1}^{(m)}\right), \\ \theta_n^{(i)} \in \arg\min_{\theta \in \Theta^{(i)}} \left(G_n^{(i)}(\theta) := g_n^{(i)}(\theta) + p_n^{(i)}(\theta)\right). \end{cases}
$$

While we consider BMM only with cyclic updates, other block selection rules (e.g., random i.i.d. sampling) can be used; see [17] for other update rules.

BMM entails numerous well-known algorithms for constrained nonconvex minimization. First, when the smooth part f of the objective function is convex in each block (i.e., block multiconvex) and the surrogate $g_n^{(i)}$ in (1.2) is identical to $f_n^{(i)}$, then BMM reduces to block coordinate descent (BCD), also known as nonlinear Gauss–Seidel [38], where one sequentially minimizes the objective in each block coordinate while the others are held fixed:

$$
\text{(1.3)} \quad \textbf{BCD} \qquad \theta_n^{(i)} \in \arg\min_{\theta \in \Theta^{(i)}} \left(F_n^{(i)}(\theta) = f_n^{(i)}(\theta) + p_n^{(i)}(\theta)\right),
$$

where $f_n^{(i)}$ and $p_n^{(i)}$ are defined in (1.2). Due to its simplicity, BCD has been widely used in various optimization problems such as nonnegative matrix or tensor factorization [23, 20]. Using proximal surrogates in (1.2), BMM becomes BCD with proximal regularization (BCD-PR):

$$
\text{(1.4)} \quad \textbf{BCD-PR} \qquad \theta_n^{(i)} \in \arg\min_{\theta \in \Theta^{(i)}} \left(G_n^{(i)}(\theta) := f_n^{(i)}(\theta) + \frac{\lambda}{2}\|\theta - \theta_{n-1}^{(i)}\|^2 + p_n^{(i)}(\theta)\right),
$$

where $\lambda \geq 0$ is a fixed constant. If we use prox-linear surrogates in (1.2), then BMM becomes the block prox-linear minimization [39], which is equivalent to the block projected gradient descent (BPGD) [37] when the nonsmooth part p is nonexistent:

$$
\text{(1.5)} \quad \textbf{BPGD} \quad \begin{cases} g_n^{(i)}(\theta) := f_n^{(i)}(\theta_{n-1}^{(i)}) + \langle \nabla f_n^{(i)}(\theta_{n-1}^{(i)}), \theta - \theta_{n-1}^{(i)} \rangle + \frac{\rho}{2}\|\theta - \theta_{n-1}^{(i)}\|^2, \\ \theta_n^{(i)} \leftarrow \arg\min_{\theta \in \Theta^{(i)}} \left(G_n^{(i)}(\theta) = g_n^{(i)}(\theta) + p_n^{(i)}(\theta)\right) \\ = \mathrm{Proj}_{\Theta^{(i)}}\left(\theta_{n-1}^{(i)} - \frac{1}{\rho}\nabla f_n^{(i)}(\theta_{n-1}^{(i)})\right) \ \text{ if } p = 0. \end{cases}
$$

The function $g_n^{(i)}$ in (1.5) is indeed a majorizing surrogate of $f_n^{(i)}$ when the smooth part f of the objective has L-Lipschitz gradient and $\rho \geq L$. BPGD has applications in nonnegative matrix factorization (NMF) [25], nonnegative tensor completion [26], and many other problems where the objective function is generally nonconvex and the constraint set is convex in each block.

A key advantage of BMM over BCD is that one can work with user-constructed majorizing surrogates $g_n^{(i)}$ that are strongly convex, while the smooth part $f_n^{(i)}$ of the marginal block objective may not even be convex. This advantage is implicit in BPGD but becomes apparent if we view it as a BMM with prox-linear surrogates in (1.5), which is λ-strongly convex. For instance, it ensures the uniqueness of their minimizer, which is a key property that warrants asymptotic convergence to stationary points. In addition, strong convexity also plays a key role in iteration complexity analysis [39, 18, 21] since it often implies square-summability of one-step parameter changes.

In this work, we establish that the iteration complexity of BMM with ρ-strongly convex and L_g-smooth surrogates for general nonsmooth nonconvex constrained optimization problem is $\widetilde{O}((1+L_g+\rho^{-1})\varepsilon^{-2})$, where $\widetilde{O}(\cdot)$ is the usual $O(\cdot)$ hiding a polylogarithmic factor in ε. Since $\rho \le L_g$, our result indicates that choosing surrogates that are not too steep (L_g large) or too flat (ρ^{-1} large) will be beneficial. We move one step further beyond what the classical BMM can offer. Can we modify BMM in such a way that the complexity bound does not involve the undesirable factor of ρ^{-1}? The reason why this factor appears in the bound is as follows. When the surrogates are nearly flat ($\rho > 0$ but small), one seeks to make aggressive parameter changes whenever possible, but it becomes ineffective when the objective cannot be greatly improved. Thus, one may try to improve BMM by keeping the nearly flat surrogates if possible but gradually limiting the range of parameter changes as the algorithm proceeds. From this motivation, we propose BMM in conjunction with trust-region techniques. Namely, fix a sequence $(r_n)_{n\ge1}$ of numbers in $(0,\infty]$ (including ∞) that acts as the radii of the trust-region. We then generalize (1.2) as

$$\text{(1.6)}\quad \textbf{BMM-DR}\begin{cases} g_n^{(i)} \leftarrow \text{Majorizing surrogate of } f_n^{(i)} \text{ at } \theta_{n-1}^{(i)} \text{ as in (1.2)}, \\ \theta_n^{(i)} \in \arg\min_{\theta\in\Theta^{(i)},\, \|\theta-\theta_n^{(i)}\|\le r_n} \left(G_n^{(i)}(\theta) = g_n^{(i)}(\theta) + p_n^{(i)}(\theta)\right). \end{cases}$$

The majorizing surrogate $g_n^{(i)}$ in (1.6) is assumed to satisfy the following properties:

(1) (Majorization) $g_n^{(i)}(\theta) - f_n^{(i)}(\theta) \ge 0$ for all $\theta\in\Theta^{(i)}$;

(2) (Sharpness) $g_n^{(i)}(\theta_{n-1}^{(i)}) = f_n^{(i)}(\theta_{n-1}^{(i)})$ and $\nabla g_n^{(i)}(\theta_{n-1}^{(i)}) = \nabla f_n^{(i)}(\theta_{n-1}^{(i)})$;

(3) (Strong convexity) $g_n^{(i)}$ is ρ-strongly convex on $\Theta^{(i)}$ for some $\rho\ge 0$.

Note that (1.6) is identical to BMM (1.2) except that we restrict the range of parameter search within a radius r_n from the previous estimation. When $r_n \equiv \infty$, then this additional radius constraint becomes vacuous and we recover the standard BMM (1.2). The resulting algorithm, which we call BMM with diminishing radius (BMM-DR), is stated in algorithm (1.6).

Our key finding is that the additional trust-region constraint in BMM-DR improves the complexity bound to $\widetilde{O}((1+L_g)\varepsilon^{-2})$, removing the dependence on the inverse strong convexity parameter ρ^{-1}, even allowing convex surrogates ($\rho=0$) (see Theorem 2.1). The improvement is significant for BMM with "nearly flat surrogates" (e.g., linear surrogates for concave objectives [8]). We also note that unlike the classical use of trust-region where the radii are computed adaptively depending on the algorithm's progress [13], we use a simple nonadaptive sequence of radii (e.g., $r_n = O(1/\sqrt{n})$). We also establish asymptotic convergence to first-order optimal points of BMM-DR. Our theoretical establishments are verified by numerical experiments in section 7.

Related work. There are several stylized examples of BMM in a wide range of problems. For the single block case ($m=1$), BMM reduces to the well-known majorization-minimization algorithm [22], which entails the EM algorithm for maximum likelihood estimation, forward-backward splitting [12], iterative reweighted least squares [14], and the classical proximal point algorithm [6, sect. 3.4.3]. With multiple blocks ($m\ge2$), BMM entails Multiplicative update for NMF by Lee and Seung [23], the convex-concave procedure for the difference of convex programs [40], and alternating least squares (ALS) for nonnegative CANDECOMP/PARAFAC (CP) decomposition [10].

Asymptotic convergence to stationary points of BCD for nonconvex objectives has been extensively studied in the literature [27]. It is well-known that BCD does not always converge to the stationary points of the nonconvex objective function that is convex in each block [32], but such global convergence is guaranteed under additional assumptions such as two-block ($m = 2$), strict quasiconvexity for $m - 2$ blocks [16], or uniqueness of minimizer per block [7, sect. 2.7]. Due to the additional proximal regularization, BCD-PR is guaranteed to converge to stationary points as long as the proximal surrogates (see (1.4)) are strongly convex [16, 39, 21]. In [33], BMM (1.2) for minimizing smooth objectives is known to converge to the set of stationary points when the surrogates $g_n^{(i)}$ have unique minimizer over the constraint sets $\Theta^{(i)}$. For nonsmooth nonconvex constrained optimization, Xu and Yin [39] showed that BCD, BCD-PR, and BPGD converge asymptotically to the set of Nash equilibria (a weaker notion than stationary points) when the nonsmooth part of the objective (possibly in conjunction with the indicator function of convex constraint set) is not necessarily continuous but block-separable. However, asymptotic convergence to stationary points in the nonconvex nonsmooth constrained setting is still unknown.

For minimizing convex objectives, BMM reduces the gap between the current objective value and the global minimum at rate $O(1/n)$ in n iterations [18], assuming strong convexity of the surrogates. A series of works including [9, 4] proved the complexity of BMM and its variants for convex objectives under different settings. A summary of some techniques used in the proofs can also be found in [3].

Compared to the convex minimization case, the iteration complexity for BMM (1.2) for the constrained nonconvex nonsmooth setting is more limited. Here by "iteration complexity," we mean the worst-case number of iterations until an ε-approximate first-order optimal point is obtained, using a suitable measure of suboptimality. Xu and Yin [39] obtain the local rate of convergence of BCD, BCD-PR, and BPGD under the additional assumption that the objective function satisfies the Kurdyka–Łojasiewicz inequality. Recently, Lyu and Kwon showed that BCD-PR has iteration complexity of $\widetilde{O}((1 + L_g + \rho^{-1})\varepsilon^{-2})$ [21] for both the constrained and unconstrained settings, but their result does not cover nonsmooth objectives and general surrogates. A Riemannian counterpart of BPGD for compact manifolds was recently shown to have iteration complexity of $\widetilde{O}(\varepsilon^{-2})$ [31]. However, this result does not hold for the Euclidean setting with or without constraints, as the underlying manifold should be compact without boundary. Razaviyayn et al. [34] show that BMM with randomized coordinate update has complexity of $O_{\mathbb{E}}((1 + \frac{L_g^2}{\rho} + \rho^{-1})\varepsilon^2)$, where the subscript $\mathbb{E}$ means that the iteration complexity holds after taking the expectation over the randomness of the block selection during the algorithm. However, for the cyclic update rule, the (almost sure) complexity of BMM is not yet known.

We are the first to propose BMM-DR so there are no prior results on its iteration complexity or asymptotic convergence properties. See Table 1 for a concise summary of prior results.

Contribution. In this work, we analyze BMM with the optional trust-region (1.6). Our main results are summarized below:

(1) We obtain a worst-case (anytime) bound of $\widetilde{O}((1 + L_g + \rho^{-1})\varepsilon^{-2})$ on the number of iterations to achieve ε-approximate "stationary-Nash" points (see Definition 2.4) using ρ-strongly convex and L_g-smooth surrogates. (If p is block-separable, stationary-Nash points are simply stationary points.)

(2) Using an optional trust-region with diminishing radius with the same surrogates in (1), we obtain an improved iteration complexity $\widetilde{O}((1+L_g)\varepsilon^{-2})$ that is independent of the strong convexity parameter ρ of the surrogates.

TABLE 1

Survey known results on the iteration complexity of BMM for multiblock minimization problems with ρ-strongly convex and L_g-smooth surrogates. C = convex, NC = nonconvex, S = smooth, and NS = nonsmooth.

Method	Objective	Block update	Complexity	Asymp. conv.	Inexact computation
BPGD [39]	C & NS	cyclic	Depends on KL-ineq.	✓	✗
BPGD [4]	C & S	cyclic	$\widetilde{O}(\varepsilon^{-1})$	✗	✗
BCD-PR [21]	NC & S	cyclic	$\widetilde{O}((1+L_g+\rho^{-1})\varepsilon^{-2})$	✓	✓
BMM [34]	NC & NS	random	$O_{\mathbb{E}}((1+\frac{L_g^2}{\rho}+\rho^{-1})\varepsilon^{-2})$	✓	✗
BMM [34]	NC & NS	cyclic	✗	✓	✗
BMM (Ours)	NC & NS	cyclic	$\widetilde{O}((1+L_g+\rho^{-1})\varepsilon^{-2})$	✓	✓
BMM-DR (Ours)	NC & NS	cyclic	$\widetilde{O}((1+L_g)\varepsilon^{-2})$	✓	✓

(3) We obtain global asymptotic convergence of BMM(-DR) to stationary points from arbitrary initialization assuming block-separability of the nonconvex part p.

(4) All the aforementioned results hold under inexact execution of the algorithm.

To the best of our knowledge, we believe that our work provides the first result on the global rate of convergence and iteration complexity of BMM(-DR) for minimizing nonsmooth nonconvex objectives under convex constraints. Especially, our rate of convergence does not depend on ρ^{-1} if a trust-region with diminishing radius is used. For gradient descent methods with an unconstrained nonconvex objective, it is known that such a rate of convergence cannot be faster than $\widetilde{O}(\varepsilon^{-2})$ [11], so our rate bound matches the optimal result up to a polylogarithmic factor. Furthermore, prior results on asymptotic convergence on BMM for nonconvex nonsmooth optimization show that the limit points are only Nash but not stationary, even with the block-separability assumption on the nonsmooth part [39]. We also show that these results continue to hold if convex subproblems are solved inexactly, allowing for easier practical implementation of the algorithm with the same theoretical guarantees. Such robustness results are not often provided in the literature.

We apply our general framework to various stylized examples such as NMF, nonnegative CANDECOMP/PARAFAC decomposition (NCPD), and BPGD and get the following results:

(5) For BPGD, we obtain iteration complexity $\widetilde{O}((1+\rho+\rho^{-1})\varepsilon^{-2})$.

(6) We obtain a regularized version of the multiplicative update algorithm for nonnegative matrix/tensor factorization with guaranteed asymptotic convergence to stationary points and iteration complexity of $\widetilde{O}(\varepsilon^{-2})$, where the implicit constant depends on the hyperparameters.

We believe that these are the first iteration complexity results for NMF and NCPD as well as BPGD for nonconvex objectives. We experimentally validate our theoretical results with both synthetic and real-world data. We show that using trust-regions improves the performance of BMM on both matrix factorization and tensor decomposition problems. Moreover, we find that our algorithms outperform the existing ones especially when the matrix and tensors to be factorized are sparse.

Central to our analysis is a novel continuous first-order optimality measure (2.4), by which we bound the worst-case suboptimality in each iteration by the first-order improvement the algorithm makes.

1.1. Notation. Denote $a \wedge b = \min(a \wedge b)$ for $a, b \in \mathbb{R}$. For a convex function $p : \boldsymbol{\Theta} \to \mathbb{R}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $\partial p(\theta)$ denote the *subdifferential set*

$$\partial p(\boldsymbol{\theta}) := \{\eta \in \mathbb{R}^p : p(\boldsymbol{\theta}') - p(\boldsymbol{\theta}) \geq \langle \eta, \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle \text{ for all } \theta \in \boldsymbol{\Theta}\}.$$

By a slight abuse of notation, we also use $\partial p(\boldsymbol{\theta})$ to denote a generic subdifferential in $\partial p(\boldsymbol{\theta})$. Throughout this paper, we will denote by $(\boldsymbol{\theta}_n)_{n\geq 1}$ a (possibly inexact) output of algorithm (1.6). Fix $n \geq 1$ and $i = 1, \ldots, m$. Write $\boldsymbol{\theta}_n = [\theta_n^{(1)}, \ldots, \theta_n^{(m)}]$ and for each $\theta \in \mathbb{R}^{I_i}$, define

$$\begin{aligned} \boldsymbol{\theta}_{n;i} &:= \left(\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta_n^{(i)}, \theta_{n-1}^{(i+1)}, \ldots, \theta_{n-1}^{(m)}\right), \\ (\theta, \boldsymbol{\theta}_{n;i}) &:= \left(\theta_n^{(1)}, \ldots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \ldots, \theta_{n-1}^{(m)}\right). \end{aligned}$$

Then we can write the function $f_n^{(i)}$ in (1.2) as $f_n^{(i)}(\theta) := f(\theta, \boldsymbol{\theta}_{n;i})$. Also, we denote

$$\Theta_n^{(i)} := \{\theta \in \Theta^{(i)} \mid \|\theta - \theta_{n-1}^{(i)}\| \leq r_n\}, \tag{1.7}$$

which is the constraint set that appears in algorithm (1.6). Denote $\Lambda := \{\boldsymbol{\theta}_n \mid n \geq 1\} \subseteq \boldsymbol{\Theta}$. Also denote $\Lambda_n^\star$ the set of the exact output of one step of algorithm (1.6):

$$\Lambda_n^\star := \left\{ \boldsymbol{\theta}_n^\star = [\theta_n^{(1\star)}, \ldots, \theta_n^{(m\star)}] : \begin{array}{c} \theta_n^{(i\star)} \text{ is an exact minimizer of } g_n^{(i)} \\ \text{over } \Theta_n^{(i)} \text{ for } i = 1, \ldots, m \end{array} \right\}.$$

We will denote a generic element of $\Lambda_n^\star$ by $\boldsymbol{\theta}_n^\star$.

1.2. Organization. This paper is organized as follows. We state the main results in section 2. Section 3 gives a sketch of the analysis of the complexity results. In section 4, we prove the iteration complexity results stated in Theorem 2.1(i)–(ii). In section 5, we prove the asymptotic stationary result stated in Theorem 2.1(iii). Then we provide some applications of our theory in section 6. Finally, we present the experimental results of the applications in section 7.

2. Statement of main results.

2.1. Measure of stationarity. Recall that a necessary condition for a point $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ to be a local minimizer of the objective $F = f + p$ over $\boldsymbol{\Theta}$ is the following first-order optimality condition:

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1} \langle -\nabla f(\boldsymbol{\theta}^*) - \partial p(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq 0, \tag{2.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product on $\mathbb{R}^{I_1 + \cdots + I_m} \supseteq \boldsymbol{\Theta}$. Points satisfying the above condition are called the *stationary points* of F over $\boldsymbol{\Theta}$. We propose an equivalent condition for first-order stationarity as

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 1} \left[V(\boldsymbol{\theta}^*, \boldsymbol{\theta}) := \langle -\nabla f(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle + p(\boldsymbol{\theta}^*) - p(\boldsymbol{\theta}) \right] \leq 0. \tag{2.2}$$

To see the equivalence, first note that the convexity of p yields

$$p(\boldsymbol{\theta}^*) - p(\boldsymbol{\theta}) \leq \langle -\partial p(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{2.3}$$

Hence (2.1) implies (2.2). Conversely, suppose (2.2) holds. Let $\varphi(\boldsymbol{\theta}) = V(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ denote the concave function in the supremum in (2.2). Since $\varphi(\boldsymbol{\theta}^*) = 0$, (2.2) implies that $\boldsymbol{\theta}^*$ is a local maximizer of φ over $\boldsymbol{\Theta}$. Then writing the first-optimality condition for $\boldsymbol{\theta}^*$

being a local minimizer of $-\varphi$ over $\boldsymbol{\Theta}$ and noting $-\partial\varphi(\boldsymbol{\theta}^*) = \nabla f(\boldsymbol{\theta}^*) + \partial p(\boldsymbol{\theta}^*)$ gives exactly (2.1).

An important advantage in using the equivalent stationary measure in (2.2) is that the bivariate function V is *continuous* whenever the nonsmooth part p is continuous (e.g., ℓ_1-regularization), whereas the corresponding function in (2.1) is not. Note that we can still incorporate hard convex constraints $\boldsymbol{\Theta}$ while keeping p continuous instead of viewing the discontinuous indicator function of $\boldsymbol{\Theta}$ as part of p (see, e.g., [39]).

For block nonconvex nonsmooth optimization problems as we consider in (1.1), BCD-type algorithms may not always converge to the stationary points [2, p. 94], but they converge to *Nash points* that are first-order optimal with respect to perturbing a single block coordinate [39]. Accordingly, we introduce the following notion of *stationary-Nash* points for the points $\boldsymbol{\theta}^*$ satisfying the following condition:

$$\sup_{\substack{\boldsymbol{\theta}\in\boldsymbol{\Theta}\\ \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le 1}} \left[\widetilde{V}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) := \langle -\nabla f(\boldsymbol{\theta}^*), \boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle + \sum_{i=1}^{m} p(\boldsymbol{\theta}^*) - p\big(\boldsymbol{\theta}^* + (\boldsymbol{\theta}-\boldsymbol{\theta}^*)\mathbf{e}^{(i)}\big)\right] \le 0, \tag{2.4}$$

where $\theta^{(i)}$ denotes the ith block coordinate of $\boldsymbol{\theta}$ and $\mathbf{e}^{(i)}$ is the indicator vector for the ith block coordinate. However, it is important to notice that if the nonsmooth part p is block-separable (i.e., $p(\theta^{(1)},\dots,\theta^{(m)}) = \sum_{i=1}^m p_i(\theta^{(i)})$ for p_i convex), then our notion of stationary-Nash points agrees with that of stationary points since (2.2) and (2.4) coincide. While Xu and Yin [39] and Razaviyayn et al. [34] make such block-separability assumption for the nonsmooth part of the objective, we do not make such an assumption unless otherwise mentioned and aim for the most general setting.

For iterative algorithms, a first-order optimality condition may hardly be satisfied exactly in a finite number of iterations, so it is often important to know how many iterations are required until an ε-approximate solution is guaranteed to be obtained. Accordingly, we say $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is an *ε-approximate stationary-Nash point* of F over $\boldsymbol{\Theta}$ if

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\, \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|\le 1} \widetilde{V}(\boldsymbol{\theta}^*,\boldsymbol{\theta}) \le \varepsilon. \tag{2.5}$$

For each $\varepsilon > 0$ we define the *worst-case iteration complexity* N_ε of an algorithm for solving (1.1) to be the worst-case (w.r.t. initialization $\boldsymbol{\theta}_0$) number of iterations to guarantee an ε-approximate stationary-Nash point of F over $\boldsymbol{\Theta}$. Again, recall that $\widetilde{V} = V$ for block-separable p so ε-approximate stationary-Nash points are ε-approximate stationary points in that case.

It is worth expanding on the choice of optimality measures for nonconvex nonsmooth constrained problems since there are multiple ways to choose them. They are typically obtained by relaxing various equivalent conditions for first-order stationarity. For example, the ε-relaxation of the stationarity condition

$$\operatorname{dist}(\mathbf{0}, \partial F(\boldsymbol{\theta}^*) + \mathcal{N}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)) \le 0 \tag{2.6}$$

is standard, where $\mathcal{N}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ denotes the normal cone of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$. However, Davis and Drusvyatskiy, in their celebrated work [15], noted that the resulting stationarity measure is difficult to work with due to its highly discontinuous nature. They proposed to use the norm of the Moreau envelope as an alternative measure, since its being small implies the existence of an approximate stationary point (in the sense of (2.6)) near the estimated parameter.

It appears to us that our suboptimality measure $\widetilde{V}$ in (2.5) strikes a nice balance between generality and manageability with many desirable properties. First, we have already noted that it is continuous, which allows us to push the convergence analysis much further for nonconvex nonsmooth optimization. Second, it is a direct measure of the suboptimality of the parameter being evaluated in contrast to the near-stationary measure of Davis and Drusvyatskiy. Third, for nonconvex smooth optimization ($p = 0$), it agrees with (2.6) (see [1, Prop. B.1]). Fourth, if p is block-separable, then $\widetilde{V}$ coincides with V (2.3) so the stationary-Nash point becomes the usual stationary point. Furthermore, if $f = 0$, then $\widetilde{V}$ reduces to the commonly used function value gap measure in convex optimization problems.

We acknowledge that it is still open to obtain iteration complexity for BMM for general nonconvex nonsmooth optimization problems with the stationary measure from (2.6). The only other related work we are aware of is Razaviyain et al. [34], which uses the ε-relaxation of the gradient mapping being zero (also a near-stationarity measure) with block-separable nonsmooth part.

2.2. Assumptions. Throughout this paper, we assume the following conditions:

A1. The constraint sets $\Theta^{(i)} \subseteq \mathbb{R}^{I_i}$, $i = 1, \ldots, m$, are nonempty, closed, and convex (but not necessarily compact) subsets in $\mathbb{R}^{I_i}$.

A2. The objective function $F = f + p : \mathbf{\Theta} \to \mathbb{R}$, where f is continuously differentiable and p is convex and possibly nonsmooth. For each compact subset $\mathbf{\Theta}_0 \subseteq \mathbf{\Theta}$, there exist a constant $L = L(\mathbf{\Theta}_0)$ such that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ and $|p(x) - p(y)| \leq L\|x - y\|$ for all $x, y \in \mathbf{\Theta}_0$. Also, $F^* := \inf_{\boldsymbol{\theta} \in \mathbf{\Theta}} F(\boldsymbol{\theta}) > -\infty$. Furthermore, the sublevel sets $F^{-1}((-\infty, a]) = \{\boldsymbol{\theta} \in \mathbf{\Theta} : F(\boldsymbol{\theta}) \leq a\}$ for $a \in \mathbb{R}$ are compact.

In A1, we allow the constraint set $\Theta^{(i)}$ to be the whole space $\mathbb{R}^{I_i}$. The C^1-assumption of f in A2 is weaker than the L-smoothness assumption that is standard in the literature of BCD (see, e.g., [39]).

Next, we define the *majorization gap* as the function $h_n^{(i)} := g_n^{(i)} - f_n^{(i)}$ for each $n \geq 0$ and $i = 1, \ldots, m$. Note that $h_n^{(i)} \geq 0$, $h_n^{(i)}(\theta_{n-1}^{(i)}) = 0$. Hence if we assume $h_n^{(i)}$ is differentiable, then necessarily $\nabla h_n^{(i)}(\theta_{n-1}^{(i)}) = 0$.

A3. The surrogates $g_n^{(i)}$ for all n, i have L_g-Lipschitz continuous gradients for some constant $L_g > 0$: For all $\theta, \theta^* \in \Theta^{(i)}$,

$$(2.7) \qquad \text{(Lipschitz gradients)} \qquad \|\nabla g_n^{(i)}(\theta) - \nabla g_n^{(i)}(\theta^*)\| \leq L_g\|\theta - \theta^*\|.$$

Furthermore, the surrogates $g_n^{(i)}$ are ρ-strongly convex for some $\rho \geq 0$ (allowing $\rho = 0$). Also, assume that either of the following holds:

(a) (Trust-region used) $\sum_{n=1}^{\infty} r_n^2 < \infty$ and $r_{n+1}/r_n = O(1)$; or

(b) (Trust-region not used) $r_n \equiv \infty$ and $\rho > 0$.

It is straightforward to extend our analysis to the case where the smoothness parameter L_g in (2.7) depends on the block index i. For simplicity of presentation, we do not pursue this straightforward generalization.

Note that the subproblem of block minimization in algorithm (1.6) amounts to minimizing convex majorizing surrogate $G_n^{(i)}$ over the constraint set $\Theta^{(i)}$ if $r_n = \infty$ or the intersection $\Theta^{(i)} \cap \{\theta : \|\theta - \theta_{n-1}^{(i)}\| \leq r_n\}$ if $r_n < \infty$, which are both convex sets. Hence each iteration of algorithm (1.6) can be readily executed using standard convex optimization procedures (see, e.g., [7]). For instance, each iteration of BPGD (1.5) for smooth objectives can be exactly computed given that projection onto the convex constraint set $\Theta^{(i)}$ has a closed-form expression (e.g., nonnegativity constraints or threshold).

In the literature of BMM, it is often assumed that $\Delta_n \equiv 0$ [33, 34, 39]. However, for many instances of algorithm (1.6), it could be the case that the convex subproblems can only be solved approximately. Fortunately, our analysis of algorithm (1.6) allows *inexact computation* of solutions to the convex subproblems, as long as the "optimality gaps" are summable. To be precise, we define the *optimality gap* at iteration n as

$$\Delta_n = \Delta_n(\boldsymbol{\theta}_0) := \max_{1\le i\le m} \left(G_n^{(i)}(\theta_n^{(i)}) - G_n^{(i)}(\theta_n^{(i\star)}) \right), \quad \text{where } \theta_n^{(i\star)} \in \operatorname*{arg\,min}_{\theta\in\Theta^{(i)},\, \|\theta-\theta_{n-1}^{(i)}\|\le r_n} G_n^{(i)}(\theta). \tag{2.8}$$

Then we require the summability of optimality gaps as in A4.

A4. The optimality gaps are summable: $\sum_{n=1}^{\infty} \Delta_n < \infty$.

We remark that when the surrogates are ρ-strongly convex and L_g-smooth, then one can satisfy A4 by using the well-known complexity result for proximal gradient descent [3, Thm. 10.29]. Namely, compute each $\theta_n^{(i)}$ by running proximal gradient descent for k_n subiterations with stepsize $\tau < 1/L_g$ and initialization $\theta_{n-1}^{(i)}$. Then $\Delta_n \le \frac{L_g}{2}(1-(\rho/L_g))^{k_n} \sum_{i=1}^{m} \|\theta_{n-1}^{(i)} - \theta_n^{(i\star)}\|^2$. In (4.3), we show that the sum of $\sum_{i=1}^{m} \|\theta_{n-1}^{(i)} - \theta_n^{(i\star)}\|^2$ over $n\ge 1$ is finite. Hence A4 is verified if $(1-(\rho/L_g))^{k_n} = O(n^{-2})$. For this, it is enough to have $k_n \approx \log n$. The subiterations of $\log n$ steps for $n=1,\dots,T$ contribute only a $\log T$ factor to the total complexity, which is negligible.

2.3. Statement of main results. Now we state the main result, Theorem 2.1.

THEOREM 2.1. *Assume* A1–A4 *hold. Let* $(\boldsymbol{\theta}_n)_{n\ge 0}$ *be an (possibly inexact) output of algorithm* (1.6)*. Then the following hold:*

(i) *(Rate of convergence) There exist constants* $M, c > 0$ *such that for* $n \ge 1$*,*

$$\min_{1\le k\le n} \left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\, \|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|\le 1} \tilde{V}(\boldsymbol{\theta}_k, \boldsymbol{\theta}) \right] \le c \frac{M + L_g + \left(\rho^{-1} \wedge \sum_{k=1}^{n} r_k^2\right)}{(\sqrt{n}/\log n) \wedge \sum_{k=1}^{n} r_k}.$$

(See (4.8) *and below for explicit expressions for the constants* M, c*.)*

(ii) *(Worst-case iteration complexity) If* $r_n \equiv \infty$*, then the worst-case iteration complexity* N_ε *for algorithm* (1.6) *satisfies* $N_\varepsilon = O\big((1+L_g+\rho^{-1})\varepsilon^{-2}(\log\varepsilon^{-1})^2\big)$*. If* $r_n = 1/(\sqrt{n}\log n)$ *for* $n\ge 1$*, then the iteration complexity improves to* $N_\varepsilon = O((1+L_g)\varepsilon^{-2}(\log\varepsilon^{-1})^2)$*.*

(iii) *(Asymptotic stationarity) Further assume that* $\sum_{n=1}^{\infty} r_n = \infty$*. Then* $(\boldsymbol{\theta}_n)_{n\ge 1}$ *converges to the set of stationary-Nash points of* F *over* $\boldsymbol{\Theta}$*. In particular, if* p *is block-separable, then* $(\boldsymbol{\theta}_n)_{n\ge 1}$ *converges to the set of stationary points of* F *over* $\boldsymbol{\Theta}$*.*

Theorem 2.1(i) provides a bound on the rate of convergence in terms of the stationarity measure introduced in (2.5). The result covers both options when trust-regions of square-summable radii are used or not throughout the iterations. When trust-region is not used, the asymptotic rate is $O(n^{-1/2}\log n)$. However, the constant $1+L_g+\rho^{-1}$ grows unbounded when the strong convexity parameter ρ for the surrogates is vanishingly small. Therefore, using "flat" surrogates may hinder the rate of convergence. One may try to circumvent this issue by using "steep" surrogates (e.g., by adding large proximal terms), but this will be penalized by L_g in the constant of convergence rate since $L_g \ge \rho$. See also Figure 2 for numerical results.

The upper bound on the rate of convergence in the case of a trust-region with diminishing radius in Theorem 2.1(i) suggests that we can get rid of the unfavorable dependence on ρ^{-1} *with the same surrogates* and *without affecting the rate of convergence.* Indeed, by choosing radii $r_n = 1/(\sqrt{n}\log n)$ for $n\ge 1$, we get

$1/\sum_{k=1}^{n}\min\{r_k,1\} = O(n^{-1/2}\log n)$ and the numerator in the bound is the same constant $1+L_g$ without the term depending on ρ^{-1}. To our knowledge, the best convergence rate for nonconvex nonsmooth block optimization was $O_{\mathbb{E}}((1+L_g^2\rho^{-1}+\rho^{-1})n^{-1/2})$ for a randomized BCD in [34], and there is no known bound on convergence rate that holds almost surely, especially without the dependence on ρ^{-1}.

Theorem 2.1(ii) gives a worst-case iteration complexity of algorithm (1.6) of producing an ε-stationary point. This can be easily obtained from Theorem 2.1(i) by setting the upper bound to be less than ε.

Last, Theorem 2.1(iii) states that the iterates produced by algorithm (1.6), possibly solving the subproblems inexactly with summable optimality gap, asymptotically converge to the set of stationary-Nash points of the problem (1.1). In particular, when the nonsmooth part p is block-separable, this result shows the first asymptotic stationarity of BMM(-DR) iterates in the literature.

The most technical part of our asymptotic analysis is to handle inexact computation when bounded trust-regions are used. Roughly speaking, for asymptotic analysis with trust-region, we need to show that the additional trust-region constraints "vanish" in the limit in the sense that any convergent subsequence of the iterates cannot touch the trust-region boundaries indefinitely. Allowing inexact computation of the surrogate minimization within the trust-region poses an additional challenge. The analysis is given in section 5.

3. Sketch of analysis. In this section, we give a high-level description of our analysis and discuss the key difficulties and how we will handle them.

We first discuss key challenges in the analysis, especially for BMM with trust-regions with square-summable radii. For most iteration complexity analysis of first-order methods (e.g., projected gradient descent), one uses first-order optimality of each update w.r.t. the corresponding subproblem and relates the optimality measure of the overall objective at each iteration with the amount of parameter change [39]. Using this approach, we can deduce the following bound:

$$(3.1)\qquad \sup_{\substack{\boldsymbol{\theta}\in\boldsymbol{\Theta}\\ \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le(r_n\wedge 1)}} \sum_{i=1}^{m}\left(-\langle\nabla_i f(\boldsymbol{\theta}_n)+\partial p_n^{(i)}(\theta_n^{(i)}),\theta^{(i)}-\theta_n^{(i)}\rangle\right) = O\left(\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n-1}\|\right),$$

where we have assumed an exact subproblem solution (i.e., $\Delta_n\equiv 0$) for simplicity of presentation. When trust-region is not used (i.e., $r_n\equiv\infty$), we can use square-summability of the parameter changes (i.e., the right-hand side (RHS) of (3.1)) to deduce that the minimum of the LHS of (3.1) among the first n iterations decay as $\tilde{O}(n^{-1/2})$.

However, if we do use trust-region (i.e., $r_n<\infty$), then the supremum in the LHS of (3.1) is taken over a vanishingly small ball around $\boldsymbol{\theta}_n$. For instance, for square-summable radii, the small ball on the LHS has a vanishing radius of order $\tilde{O}(n^{-1/2})$, which is the same order of the RHS of (3.1). Hence, one cannot deduce a similar rate of convergence result as in the case when trust-region is not used. A different approach to analysis is needed to establish the rate of convergence of BMM-DR.

In order to circumvent the above issue, we establish the complexity results by first showing the finite first-order variation between consecutive iterates as in Proposition 4.5. Then in the key Lemma 4.6, we connect the optimality measure in (2.5) with the first-order variation between iterates. This gives an improved upper bound of the LHS in (3.1), which decays at the rate of $\tilde{O}(n^{-1})$. Therefore, even when trust-region is used, we are still able to conclude the complexity result as in Theorem 2.1.

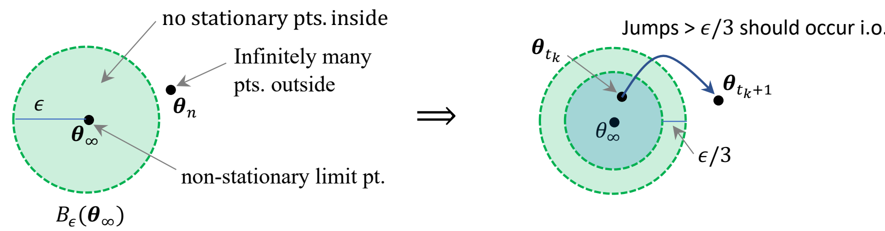

FIG. 1. *Illustration of the proof of Theorem* 2.1(iii) *with diminishing radius.*

For the asymptotic analysis with trust-region, we seek a contradiction after supposing there exists a nonstationary limit point of the set of all estimates $\{\boldsymbol{\theta}_n : n \geq 0\}$. Such a nonstationary limit point should be contained in an open ball that does not contain any other stationary points and there must be infinitely many iterates outside such a ball (Proposition 5.6). This implies that there are infinitely many "crossings" from near the nonstationary limit point to outside of the ball (see Figure 1). By using techniques developed for the complexity analysis, we can deduce from this that there exists a stationary point inside the open ball around the nonstationary point (Proposition 5.5), which is a contradiction.

4. Proof of iteration complexity. We start by recalling a classical lemma on the first-order approximation of functions with Lipschitz gradients.

LEMMA 4.1 (first-order approximation of functions with Lipschitz gradient). *Let $f : \Omega(\subseteq \mathbb{R}^p) \to \mathbb{R}$ be differentiable and ∇f be L-Lipschitz continuous on Ω. Then for each $\theta, \theta' \in \Omega$, $\left|f(\theta') - f(\theta) - \nabla f(\theta)^T(\theta' - \theta)\right| \leq \frac{L}{2}\|\theta - \theta'\|^2$.*

Proof. This is a classical lemma. See [29, Lem. 1.2.3]. ∎

Next, we will show the iterates are asymptotically exact given the surrogates are strongly convex.

PROPOSITION 4.2. *Suppose* A1 *holds and $g_n^{(i)}$ is ρ-strongly convex for some $\rho > 0$. Then $\frac{\rho}{2}\|\theta_n^{(i)} - \theta_n^{(i\star)}\|^2 \leq \Delta_n$.*

Proof. Fix $i \in \{1, \ldots, m\}$. There are two cases to consider. First, when A3(a) holds, the assertion follows from a triangle inequality and that both $\theta_n^{(i)}$ and $\theta_n^{(i\star)}$ are within distance r_n from $\theta_{n-1}^{(i)}$ since $r_n = o(1)$. Second, suppose A3(b) holds. Then by the first-order optimality (recall that $\Theta_n^{(i)}$ in (1.7) is convex)

$$\langle \partial G_n^{(i)}(\theta_n^{(i\star)}), \theta - \theta_n^{(i\star)} \rangle \geq 0 \quad \text{for all } \theta \in \Theta_n^{(i)} \text{ for all } i = 1, \ldots, m.$$

Then by ρ-strong convexity of $g_n^{(i)}$ in A3(b) and convexity of $p_n^{(i)}$,

$$\frac{\rho}{2}\|\theta_n^{(i)} - \theta_n^{(i\star)}\|^2 \leq G_n^{(i)}(\theta_n^{(i)}) - G_n^{(i)}(\theta_n^{(i\star)}) - \langle \partial G_n^{(i)}(\theta_n^{(i\star)}), \theta_n^{(i)} - \theta_n^{(i\star)} \rangle \leq \Delta_n. \tag{4.1}$$ ∎

PROPOSITION 4.3 (monotonicity of objective and stability of iterates). *Suppose* A1, A2, *and* A4 *hold. Then the following hold:*

(i) $F(\boldsymbol{\theta}_{n-1}) - F(\boldsymbol{\theta}_n) \geq -m\Delta_n$.
(ii) $\sum_{n=1}^\infty \sum_{i=1}^m G_n^{(i)}(\theta_{n-1}^{(i)}) - G_n^{(i)}(\theta_n^{(i\star)}) \leq F(\boldsymbol{\theta}_0) - F^* + m\sum_{n=1}^\infty \Delta_n < \infty$.
(iii) *If $\sum_{n=1}^\infty r_n^2 < \infty$, then $\|\theta_n^{(i)} - \theta_{n-1}^{(i)}\| \leq r_n$ for all i, n and $\sum_{n=1}^\infty \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2 \leq \sum_{n=1}^\infty r_n^2 < \infty$. If $g_n^{(i)}$ is ρ-strongly convex for some $\rho > 0$ for all i, n, then*

$$\frac{\rho}{4}\sum_{n=1}^{\infty}\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n-1}\|^2 \le F(\boldsymbol{\theta}_0)-F^*+2m\sum_{n=1}^{\infty}\Delta_n<\infty. \tag{4.2}$$

In particular, in both cases, $\|\theta_{n-1}^{(i)}-\theta_n^{(i)}\|=o(1)$ *for all* $i=1,\dots,m$.

Proof. Fix $i\in\{1,\dots,m\}$. Note that

$$\begin{aligned}F_n^{(i)}(\theta_{n-1}^{(i)})-F_n^{(i)}(\theta_n^{(i)}) &\overset{(a)}{=} G_n^{(i)}(\theta_{n-1}^{(i)})-G_n^{(i)}(\theta_n^{(i)})+G_n^{(i)}(\theta_n^{(i)})-F_n^{(i)}(\theta_n^{(i)})\\ &\overset{(b)}{\ge} G_n^{(i)}(\theta_{n-1}^{(i)})-G_n^{(i)}(\theta_n^{(i\star)})-\Delta_n,\end{aligned}$$

where (a) follows from $G_n^{(i)}(\theta_{n-1}^{(i)})=F_n^{(i)}(\theta_{n-1}^{(i)})$ and that $g_n^{(i)}$ majorizes $f_n^{(i)}$ so that $G_n^{(i)}$ majorizes $F_n^{(i)}$; (b) follows from the definition of Δ_n in (2.8). Summing over all $i=1,\dots,m$, it follows that

$$\begin{aligned}F(\boldsymbol{\theta}_{n-1})-F(\boldsymbol{\theta}_n) &= \sum_{i=1}^{m}F_n^{(i)}(\theta_{n-1}^{(i)})-F_n^{(i)}(\theta_n^{(i)})\\ &\ge -m\Delta_n+\sum_{i=1}^{m}G_n^{(i)}(\theta_{n-1}^{(i)})-G_n^{(i)}(\theta_n^{(i\star)})\ge -m\Delta_n,\end{aligned}$$

where the last inequality uses the definition of $\theta_n^{(i\star)}$. This shows (i). Note that (ii) follows by adding up the first inequality above for $n\ge 1$.

Last, we show (iii). If $\sum_{n=1}^{\infty}r_n^2<\infty$, then the assertion follows immediately. Otherwise, suppose the surrogates $g_n^{(i)}$ are ρ-strongly convex for some $\rho>0$. Then $G_n^{(i)}=g_n^{(i)}+p_n^{(i)}$ is also ρ-strongly convex since $p_n^{(i)}$ is convex, so by the second-order growth property (see (4.1)) and the definition of Δ_n,

$$\frac{\rho}{2}\|\theta_{n-1}^{(i)}-\theta_n^{(i\star)}\|^2\le G_n^{(i)}(\theta_{n-1}^{(i)})-G_n^{(i)}(\theta_n^{(i\star)}).$$

Then by (ii),

$$\sum_{n=1}^{\infty}\frac{\rho}{2}\|\theta_{n-1}^{(i)}-\theta_n^{(i\star)}\|^2\le F(\boldsymbol{\theta}_0)-F^*+m\sum_{n=1}^{\infty}\Delta_n<\infty. \tag{4.3}$$

Then by Young's inequality and Proposition 4.2, we can deduce (4.2). ∎

Proposition 4.4 (boundedness of iterates). *Assume* A1, A2, A4, *and either* A3(a) *or* A3(b) *hold. Then there exist compact and convex subsets* $S^{(i)}\subseteq\Theta^{(i)}$ *for* $i=1,\dots,m$ *such that* $\boldsymbol{\Theta}_0:=S^{(1)}\times\cdots\times S^{(m)}$ *contains* $\bigcup_{n=0}^{\infty}B_{\le 1}(\boldsymbol{\theta}_n)$, *where* $B_{\le 1}(x):=\{y\in\boldsymbol{\Theta}:\|x-y\|\le 1\}$. *Consequently,* ∇f *and* p *are* L*-Lipschitz continuous on* $\boldsymbol{\Theta}_0$ *for some* $L>0$.

Proof. Let $T:=m\sum_{k=1}^{\infty}\Delta_k$, which is finite by A4. Recall that by Proposition 4.3, we have $\sup_{n\ge 0}F(\boldsymbol{\theta}_n)\le F(\boldsymbol{\theta}_0)+T<\infty$. It follows that $\{\boldsymbol{\theta}_n : n\ge 0\}$ is a subset of the sublevel set $A_0:=F^{-1}((-\infty,F(\boldsymbol{\theta}_0)+T])$, which is compact by A2. Let $\Pi^{(i)}$ denote the projection from $\boldsymbol{\Theta}$ to its ith block component $\Theta^{(i)}$. Then $\Pi^{(i)}(A_0)$ is a compact subset of $\Theta^{(i)}$. Take $R^{(i)}$ to be the "unit fattening" of this compact subset:

$$R^{(i)}:=\left\{\theta\in\Theta^{(i)}:\|\theta-\theta'\|\le 1 \text{ for some } \theta'\in\Pi^{(i)}(A_0)\right\}.$$

Now let $S^{(i)}$ be the convex hull of $R^{(i)}$ for $i = 1, \ldots, m$. Then $S^{(i)}$ is closed and bounded, so is also a compact subset of $\Theta^{(i)}$. The claimed containment follows from the construction. The second part of the assertion follows from the first part along with A1. ∎

The following proposition shows the summability of $-V(\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}_n)$.

Proposition 4.5 (finite first variation I). *Assume* A1–A4 *hold and let* $L > 0$ *be as in Proposition* 4.4. *Then*

$$\sum_{n=0}^{\infty} -V(\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}_n) \le F(\boldsymbol{\theta}_0) - F^* + \frac{L}{2} \sum_{n=0}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2.$$

Proof. By A2 and Proposition 4.4, ∇f is L-Lipschitz continuous on the compact subset $\boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}$ (in Proposition 4.4) that contains all iterates $\boldsymbol{\theta}_n$, $n \ge 0$. Hence by Lemma 4.1, for all $n \ge 0$,

$$\begin{aligned}\langle \nabla f(\boldsymbol{\theta}_{n+1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1} \rangle &\le f(\boldsymbol{\theta}_n) - f(\boldsymbol{\theta}_{n+1}) + \frac{L}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2 \\ &= F(\boldsymbol{\theta}_n) - F(\boldsymbol{\theta}_{n+1}) - p(\boldsymbol{\theta}_n) + p(\boldsymbol{\theta}_{n+1}) + \frac{L}{2} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2.\end{aligned}$$

Adding up the above inequality for all $n \ge 0$ shows the assertion. ∎

Next, we show a key lemma for establishing the iteration complexity of our algorithm. This lemma gives an upper bound of $\widetilde{V}(\boldsymbol{\theta}_n, \boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le 1$.

Lemma 4.6 (key lemma for iteration complexity). *Assume* A1–A4 *hold. Let* $b_n \in [0, \min\{r_n, 1\}]$ *for all* $n \ge 1$. *Then*

(4.4)

$$\begin{aligned} b_{n+1} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le 1} \widetilde{V}(\boldsymbol{\theta}_n, \boldsymbol{\theta}) &\le -V(\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}_n) + 3mLb_{n+1} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\| \\ &\quad + \left(\frac{L}{2} + mL\right) \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 + \frac{L_g b_{n+1}^2}{2} + m\Delta_{n+1} \end{aligned}$$

for all $n \ge 1$ *for some constant* $L > 0$.

Proof. Fix $n \ge 0$ and let $\boldsymbol{\theta} = [\theta^{(1)}, \ldots, \theta^{(m)}] \in \boldsymbol{\Theta}$ be such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le \min\{b_{n+1}, 1\}$. Then we have

$$G_{n+1}^{(i)}(\theta_{n+1}^{(i)}) - \Delta_{n+1} \le G_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) \le G_{n+1}^{(i)}(\theta^{(i)}).$$

Let $\boldsymbol{\Theta}_0 = S^{(1)} \times \cdots \times S^{(m)}$ and $L > 0$ be as in Proposition 4.4. Recall that $\nabla g_{n+1}^{(i)}(\theta_n^{(i)}) = \nabla f_{n+1}^{(i)}(\theta_n^{(i)})$ by definition of the majorizing surrogates. Hence by subtracting $f_{n+1}^{(i)}(\theta_n^{(i)})$ from both sides and applying the L_g-smoothness of $g_{n+1}^{(i)}$ on $S^{(i)}$ and Lemma 4.1, we get

$$\begin{aligned} \left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}), \, \theta_{n+1}^{(i)} - \theta_n^{(i)} \right\rangle &\le \left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}), \, \theta^{(i)} - \theta_n^{(i)} \right\rangle + p(\theta; \boldsymbol{\theta}_{n+1;i}) - p(\boldsymbol{\theta}_{n+1;i-1}) \\ &\quad + p(\boldsymbol{\theta}_{n+1;i-1}) - p(\boldsymbol{\theta}_{n+1;i}) + \frac{L}{2} \|\theta_{n+1}^{(i)} - \theta_n^{(i)}\|^2 \\ &\quad + \frac{L_g}{2} \|\theta^{(i)} - \theta_n^{(i)}\|^2 + \Delta_{n+1}. \end{aligned}$$

Denote $\nabla f_{n+1}(\boldsymbol{\theta}_n) := [\nabla f_{n+1}^{(1)}(\theta_n^{(1)}), \dots, \nabla f_{n+1}^{(m)}(\theta_n^{(m)})]$. Adding up these inequalities for $i = 1, \dots, m$,

$$\begin{aligned}\langle \nabla f_{n+1}(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\rangle &\le \langle \nabla f_{n+1}(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n\rangle + \sum_{i=1}^m p(\theta; \boldsymbol{\theta}_{n+1;i}) - p(\boldsymbol{\theta}_{n+1;i-1}) \\ &\quad + p(\boldsymbol{\theta}_n) - p(\boldsymbol{\theta}_{n+1}) + \frac{L}{2}\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \\ &\quad + \frac{L_g}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|^2 + m\Delta_{n+1}.\end{aligned}$$

By Lipschitz continuity of ∇f on $\boldsymbol{\Theta}_0$,

$$\|\nabla_i f(\boldsymbol{\theta}_n) - \nabla f_{n+1}^{(i)}(\theta_n^{(i)})\|, \|\nabla_i f(\boldsymbol{\theta}_{n+1}) - \nabla f_{n+1}^{(i)}(\theta_n^{(i)})\| \le L\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|.$$

Noting that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le b_{n+1}$, we can deduce

$$\begin{aligned}V(\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}_n) &\le \langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n\rangle + \sum_{i=1}^m p(\theta; \boldsymbol{\theta}_{n+1;i}) - p(\boldsymbol{\theta}_{n+1;i-1}) \\ &\quad + \left(\frac{L}{2} + mL\right)\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 + \frac{L_g b_{n+1}^2}{2} + m\Delta_{n+1} \\ &\quad + mLb_{n+1}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|.\end{aligned} \tag{4.5}$$

Now fix arbitrary $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ such that $\|\boldsymbol{\theta}' - \boldsymbol{\theta}_n\| \le 1$. By convexity of $\boldsymbol{\Theta}$, $\boldsymbol{\theta} := b_{n+1}\boldsymbol{\theta}' + (1 - b_{n+1})\boldsymbol{\theta}_n \in \boldsymbol{\Theta}$ and $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le b_{n+1}$. Then by convexity of $p_{n+1}^{(i)}$, we have $p_{n+1}^{(i)}(\theta^{(i)}) \le b_{n+1}p(\theta'; \boldsymbol{\theta}_{n+1;i}) + (1 - b_{n+1})p(\boldsymbol{\theta}_{n+1;i-1})$. Therefore

$$\sum_{i=1}^m (p(\boldsymbol{\theta}_{n+1;i-1}) - p(\theta; \boldsymbol{\theta}_{n+1;i}) \ge b_{n+1}\sum_{i=1}^m (p(\boldsymbol{\theta}_{n+1;i-1}) - p(\theta'; \boldsymbol{\theta}_{n+1;i})).$$

Hence by noting $\boldsymbol{\theta} - \boldsymbol{\theta}_n = b_{n+1}(\boldsymbol{\theta}' - \boldsymbol{\theta}_n)$, (4.5) yields for all $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\theta}' - \boldsymbol{\theta}_n\| \le 1$,

$$\begin{aligned}&b_{n+1}\left(\left\langle -\nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta}' - \boldsymbol{\theta}_n\right\rangle + \sum_{i=1}^m (p(\boldsymbol{\theta}_{n+1;i-1}) - p(\theta'; \boldsymbol{\theta}_{n+1;i}))\right) \\ &\quad \le -V(\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}_n) + mLb_{n+1}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\| \\ &\qquad + \left(\frac{L}{2} + mL\right)\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 + \frac{L_g b_{n+1}^2}{2} + m\Delta_{n+1}.\end{aligned}$$

Last, by using L-Lipschitz continuity of p on $\boldsymbol{\Theta}_0$, we can replace the left-hand side above by $b_{n+1}\widetilde{V}(\boldsymbol{\theta}_n, \boldsymbol{\theta}')$ with an additive error of $2b_{n+1}L\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|$. This gives the assertion. ∎

LEMMA 4.7. *Let $(a_n)_{n\ge 0}$ and $(b_n)_{n\ge 0}$ be sequences of nonnegative real numbers such that $\sum_{n=0}^\infty a_n b_n < \infty$. Then $\min_{1\le k\le n} b_k \le \frac{\sum_{k=0}^\infty a_k b_k}{\sum_{k=1}^n a_k} = O((\sum_{k=1}^n a_k)^{-1})$.*

Proof. The assertion follows from noting that

$$\left(\sum_{k=1}^n a_k\right)\min_{1\le k\le n} b_k \le \sum_{k=1}^n a_k b_k \le \sum_{k=1}^\infty a_k b_k < \infty.$$ ∎

Now we are ready to derive the iteration complexity in Theorem 2.1(i)–(ii).

Proof of Theorem 2.1(i)–(ii). Suppose $b_n \in [0, \min\{r_n, 1\}]$ is such that $\sum_{n=1}^{\infty} b_n^2 < \infty$. Introduce the following notations: $A := F(\boldsymbol{\theta}_0) - F^*$ and

$$B := \sum_{n=1}^{N} \Delta_n,\ C := \sum_{n=1}^{N} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\|^2,\ D := \sum_{n=1}^{N} b_n^2,\ E := \sum_{n=1}^{N} r_n^2.$$

Then $A \in [0, \infty)$ by A2, $B \in [0, \infty)$ by A4, and $D \in [0, \infty)$ by the hypothesis. Furthermore, by the Cauchy–Schwarz inequality,

$$\sum_{n=0}^{N} b_{n+1} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\| \le \sqrt{CD}.$$

Now summing the inequalities in (4.4) in Lemma 4.6 for all $n \ge 0$ and using Proposition 4.5, we get

$$\sum_{n=0}^{N} b_{n+1} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta},\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \le 1} \widetilde{V}(\boldsymbol{\theta}_n, \boldsymbol{\theta}) \le M_0, \tag{4.6}$$

where the constant M_0 is defined as

$$M_0 := A + mB + (m+1)LC + \frac{L_g}{2} D + 3mL\sqrt{CD}. \tag{4.7}$$

Next, by Proposition 4.3 and A3, we have

$$C \le \begin{cases} E \wedge \dfrac{4}{\rho}(A + 2mB) & \text{if A3(a) holds,} \\ \dfrac{4}{\rho}(A + 2mB) & \text{if A3(b) holds.} \end{cases} \tag{4.8}$$

Therefore, for either case we have $C \in [0, \infty)$ and hence $M_0 < \infty$. Now Theorem 2.1(i) is a direct consequence of (4.6) and Lemma 4.7. Note that when A3(a) holds we have $D \le E$. Combining (4.7) and (4.8) gives

$$M_0 \le \begin{cases} M_1 + \frac{L(m+1)(4A+8mB)}{\rho} + \frac{6mL\sqrt{(A+2mB)D}}{\rho^{1/2}} + \frac{L_g}{2} D & \text{if A3(b) holds,} \\ M_1 + \min\left\{(4m+1)LE, H(\rho)\right\} + \frac{L_g}{2} E & \text{if A3(a) holds,} \end{cases} \tag{4.9}$$

where $H(\rho)$ denotes the sum of the second and third terms in the first case, and $M_1 = A + mB$ is a constant independent of L_g and ρ.

Next, (ii) is a direct consequence of (i). Indeed, if $r_n = 1/(\sqrt{n} \log n)$, then the upper bound on the rate of convergence in (i) is of order $O(1/\sum_{k=1}^{n} k^{-1/2}/(\log k)) = O(\log n / 2 \int_1^n x^{-1/2}\, dx) = O(\frac{\log n}{2\sqrt{n}})$. Similarly, if $r_n \equiv \infty$, then we can choose $b_n = n^{-1/2}/(\log n)$ for $n \ge 1$. Then we have the same rate of convergence in (i). Then one can conclude by using the fact that $n \ge 2\varepsilon^{-1}(\log \varepsilon^{-1})^2$ implies $(\log n)^2/n \le \varepsilon$ for all sufficiently small $\varepsilon > 0$. ∎

Remark 4.8 (remark on diminishing radius). A direct comparison of the constants in (4.9) shows the advantage of diminishing radius for improving the complexity bound by eliminating terms with ρ^{-1}, which is significant when ρ is small. When ρ is large, the algorithm with diminishing radius performs at least as well as the one without diminishing radius. Note that a large ρ will be penalized by L_g since $L_g \ge \rho$. Moreover, when implementing diminishing radius, one cannot take r_n arbitrarily small since $\sum_{n=1}^{\infty} b_n$ appears in the denominator of the complexity. Furthermore, the choice of diminishing radius r_n (and the corresponding b_n) in the proof of Theorem 2.1 is not limited to $r_n = 1/(\sqrt{n} \log n)$.

5. Proof of asymptotic stationarity. Recall that after the update $\boldsymbol{\theta}_{n-1} \mapsto \boldsymbol{\theta}_n$, each block coordinate of $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_n^\star$ are within distance r_n from the corresponding block coordinate of $\boldsymbol{\theta}_{n-1}$. For each $n \geq 1$, we say $\boldsymbol{\theta}_n^\star$ is a *short point* if all of its block coordinates are strictly within r_n from the corresponding block coordinate of $\boldsymbol{\theta}_{n-1}$, and $\boldsymbol{\theta}_n^\star$ is said to be a *long point* otherwise. Observe that if $\boldsymbol{\theta}_n^\star$ is a short point, then imposing the search radius restriction in algorithm (1.6) has no effect and $\boldsymbol{\theta}_n^\star$ is obtained from $\boldsymbol{\theta}_{n-1}$ by a single cycle of exact BMM on the constraint set $\boldsymbol{\Theta}$. In particular, this holds if $r_n = \infty$ since then every $\boldsymbol{\theta}_n^\star \in \Lambda^\star$ must be a short point.

For simplicity of notation, let $h_n^{(i)} = g_n^{(i)} - f_n^{(i)}$ denote the majorization gap. In the following proposition, we show the majorization gap is vanishing.

Proposition 5.1 (vanishing gradients of the majorization gap). *Suppose* A1–A3 *hold. Then there exists a constant $L_h > 0$ such that for all $n \geq 1$ and $i = 1, \ldots, m$,*

$$\|\nabla h_n^{(i)}(\theta_n^{(i)})\| \leq L_h \|\theta_n^{(i)} - \theta_{n-1}^{(i)}\|.$$

Proof. Write $h = h_n^{(i)}$. Let $\boldsymbol{\Theta}_0 = S^{(1)} \times \cdots \times S^{(m)}$ be as in Proposition 4.4. Then $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}_0$ for all $n \geq 1$. Since $\theta_n^{(i)}, \theta_{n-1}^{(i)} \in S^{(i)}$ and $\nabla h(\theta_{n-1}^{(i)}) = 0$, we get

$$\begin{aligned}\|\nabla h(\theta_n^{(i)})\| &= \|\nabla h(\theta_n^{(i)}) - \nabla h(\theta_{n-1}^{(i)})\| \\ &= \|\nabla g_n^{(i)}(\theta_n^{(i)}) - \nabla f_n^{(i)}(\theta_n^{(i)}) - \nabla g_n^{(i)}(\theta_{n-1}^{(i)}) + \nabla f_n^{(i)}(\theta_{n-1}^{(i)})\| \\ &\leq (L_g + L)\|\theta_n^{(i)} - \theta_{n-1}^{(i)}\|. \end{aligned}$$ ◻

Proposition 5.2 (sufficient condition for stationarity I). *Assume* A1–A4 *hold. If $(\boldsymbol{\theta}_{n_k}^\star)_{k\geq 1}$ is a convergent subsequence of $(\boldsymbol{\theta}_n^\star)_{n\geq 1}$ consisting of short points, then* $\lim_{k\to\infty} \boldsymbol{\theta}_{n_k} = \lim_{k\to\infty} \boldsymbol{\theta}_{n_k}^\star =: \boldsymbol{\theta}_\infty$ *and $\boldsymbol{\theta}_\infty$ is a stationary-Nash point of F over $\boldsymbol{\Theta}$.*

Proof. Note that if $\boldsymbol{\theta}_n^\star$ is a short point, then each block coordinate $\theta_n^{(i\star)}$ lies in the interior of $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}^{(i)}\| < r_n$, the trust-region of radius r_n. Hence by the first-order optimality condition for $\theta_n^{(i\star)}$, for all $\theta \in \Theta^{(i)}$ for all $i = 1, \ldots, m$,

$$\langle \nabla g_n^{(i)}(\theta_n^{(i\star)}), \theta - \theta_n^{(i\star)} \rangle + p_n^{(i)}(\theta) - p_n^{(i)}(\theta_n^{(i\star)}) \geq \langle \partial G_n^{(i)}(\theta_n^{(i\star)}), \theta - \theta_n^{(i\star)} \rangle \geq 0. \tag{5.1}$$

We wish to show that $\boldsymbol{\theta}_\infty$ is a stationary-Nash point of F over $\boldsymbol{\Theta}$. Denote $\boldsymbol{\theta}_\infty = [\theta_\infty^{(1)}, \ldots, \theta_\infty^{(m)}]$. Fix $i \in \{1, \ldots, m\}$ and $\theta \in \Theta^{(i)}$ with $\|\theta - \theta_\infty^{(i)}\| \leq 1$. By Proposition 4.4 and continuous differentiability of f in A2, there exists a constant $C > 0$ such that $\|\nabla f_n^{(i)}(\theta_n^{(i)})\| \leq C$ for all n, i. Then by using the Cauchy–Schwarz inequality and Proposition 5.1,

$$\begin{aligned}&\langle \nabla f_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i)} \rangle \\ &\quad \geq \langle \nabla f_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i\star)} \rangle - \|\nabla f_n^{(i)}(\theta_n^{(i)})\| \, \|\theta_n^{(i)} - \theta_n^{(i\star)}\| \\ &\quad \geq \langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i\star)} \rangle - \|\nabla h_n^{(i)}(\theta_n^{(i)})\| \, \|\theta - \theta_n^{(i)}\| - C \, \|\theta_n^{(i)} - \theta_n^{(i\star)}\| \\ &\quad \geq \langle \nabla g_n^{(i)}(\theta_n^{(i)}), \theta - \theta_n^{(i\star)} \rangle - L_h \|\theta_n^{(i)} - \theta_{n-1}^{(i)}\| \, \|\theta - \theta_n^{(i)}\| - C \, \|\theta_n^{(i)} - \theta_n^{(i\star)}\|. \end{aligned}$$

We apply the last inequality for n replaced with n_k. By Proposition 4.2, we have $\boldsymbol{\theta}_{n_k} \to \boldsymbol{\theta}_\infty$ as $k \to \infty$. Also by Proposition 4.3, we have $\|\boldsymbol{\theta}_{n-1} - \boldsymbol{\theta}_n\| = o(1)$. So $\boldsymbol{\theta}_{n_k - 1} \to \boldsymbol{\theta}_\infty$ as $k \to \infty$. Since $\|\theta - \theta_\infty^{(i)}\| \leq 1$, it follows that $\|\theta - \theta_{n_k}^{(i)}\| \leq 2$ for all sufficiently large k. Then by using continuity of ∇f, $\nabla g_n^{(i)}$, and $p_n^{(i)}$ as well as (5.1),

$$\begin{aligned}&\langle \nabla_i f(\boldsymbol{\theta}_\infty),\, \theta-\theta_\infty^{(i)}\rangle + p_\infty^{(i)}(\theta) - p_\infty^{(i)}(\theta_\infty^{(i)}) \\ &\quad = \liminf_{k\to\infty} \left(\langle \nabla f_{n_k}^{(i)}(\theta_{n_k}^{(i)}),\, \theta-\theta_{n_k}^{(i)}\rangle + p_{n_k}^{(i)}(\theta) - p_{n_k}^{(i)}(\theta_{n_k}^{(i)}) \right) \\ &\quad \ge \liminf_{k\to\infty} \left(\langle \nabla g_{n_k}^{(i)}(\theta_{n_k}^{(i\star)}),\, \theta-\theta_{n_k}^{(i\star)}\rangle + p_{n_k}^{(i)}(\theta) - p_{n_k}^{(i)}(\theta_{n_k}^{(i)}) \right) \ge 0,\end{aligned}$$

where $p_\infty^{(i)}(\theta) := p(\theta_\infty^{(1)},\dots,\theta_\infty^{(i-1)},\theta,\theta_\infty^{(i+1)},\dots,\theta_\infty^{(m)})$. This holds for all i so we can conclude. ∎

Now we can deduce Theorem 2.1(iii) for the case of BMM with strongly convex surrogates without diminishing radius.

Proof of Theorem 2.1(iii) *under* A3(b). By Proposition 4.4, the sequence of iterates $(\boldsymbol{\theta}_n)_{n\ge1}$ is bounded. Hence it suffices to show that every limit point of this sequence is a stationary-Nash point of F over $\boldsymbol{\Theta}$. To this effect, let $(\boldsymbol{\theta}_{n_k})_{k\ge1}$ denote an arbitrary convergent subsequence of $(\boldsymbol{\theta}_n)_{n\ge1}$. Denote $\boldsymbol{\theta}_\infty := \lim_{k\to\infty}\boldsymbol{\theta}_{n_k}$. By Proposition 4.2, we have $\boldsymbol{\theta}_{n_k}^\star \to \boldsymbol{\theta}_\infty$ as $k\to\infty$. Since $r_n\equiv\infty$ under A3(b), each $\boldsymbol{\theta}_{n_k}^\star$ is a short point. Thus by Proposition 5.2, $\boldsymbol{\theta}_\infty$ is a stationary-Nash point of F over $\boldsymbol{\Theta}$, as desired. ∎

In the remainder of this section, we prove Theorem 2.1(iii) under A3(a). The key technical difficulty is to handle a sequence of inexact iterates $(\boldsymbol{\theta}_n)_{n\ge1}$ where the exact iterates $\boldsymbol{\theta}_n^\star$ can be either short or long points due to the nondegenerate radius $r_n<\infty$ of the trust-region.

PROPOSITION 5.3 (finite first-order variation II). *Assume* A1, A2, A3(b), *and* A4 *hold. Then* $\sum_{n=0}^\infty \widetilde{V}(\boldsymbol{\theta}_n,\boldsymbol{\theta}_{n+1}^\star)<\infty$.

Proof. Let $L_g := L + L_h > 0$ as in Proposition 4.4. We first note that since $\nabla f_{n+1}^{(i)}(\theta_n^{(i)}) = \nabla g_{n+1}^{(i)}(\theta_n^{(i)})$,

$$\begin{aligned}&\left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta_n^{(i)}-\theta_{n+1}^{(i\star)} \right\rangle \\ &\quad = \left\langle \nabla g_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta_n^{(i)}-\theta_{n+1}^{(i\star)} \right\rangle \\ &\quad \overset{(a)}{\le} \frac{L_g}{2}\|\theta_n^{(i)}-\theta_{n+1}^{(i\star)}\|^2 + G_{n+1}^{(i)}(\theta_n^{(i)}) - G_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) - p_{n+1}^{(i)}(\theta_n^{(i)}) + p_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) \\ &\quad \overset{(b)}{\le} \frac{L_g r_{n+1}^2}{2} + G_{n+1}^{(i)}(\theta_n^{(i)}) - G_{n+1}^{(i)}(\theta_{n+1}^{(i)}) + \Delta_{n+1} - p_{n+1}^{(i)}(\theta_n^{(i)}) + p_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}),\end{aligned}$$

where (a) follows from Lemma 4.1 and A3 and the definition of $G_{n+1}^{(i)}$ and (b) follows from the trust-region constraint and the definition of Δ_{n+1} in (2.8).

Next, recalling $G_{n+1}^{(i)}(\theta_n^{(i)}) = F_{n+1}^{(i)}(\theta_n^{(i)})$ and $G_n^{(i)} \ge F_n^{(i)}$, we get

$$\begin{aligned}\sum_{i=1}^m G_{n+1}^{(i)}(\theta_n^{(i)}) - G_{n+1}^{(i)}(\theta_{n+1}^{(i)}) &= \sum_{i=1}^m F_{n+1}^{(i)}(\theta_n^{(i)}) - G_{n+1}^{(i)}(\theta_{n+1}^{(i)}) \\ &\le \sum_{i=1}^m F_{n+1}^{(i)}(\theta_n^{(i)}) - F_{n+1}^{(i)}(\theta_{n+1}^{(i)}) = F(\boldsymbol{\theta}_n) - F(\boldsymbol{\theta}_{n+1}).\end{aligned}$$

Then summing the above inequality over $n\ge0$ is at most $F(\boldsymbol{\theta}_0)-F^*$, so this shows

$$\sum_{n=0}^\infty\sum_{i=1}^m G_{n+1}^{(i)}(\theta_n^{(i)}) - G_{n+1}^{(i)}(\theta_{n+1}^{(i)}) \le F(\boldsymbol{\theta}_0) - F^*.$$

Last, using L-Lipschitz continuity of ∇f on $\boldsymbol{\Theta}_0$ (see Proposition 4.4 and A2),

$$\begin{aligned}\sum_{i=1}^m &\left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta_n^{(i)} - \theta_{n+1}^{(i\star)} \right\rangle \\ &= \left\langle \left[\nabla f_{n+1}^{(1)}(\theta_n^{(1)}), \dots, \nabla f_{n+1}^{(m)}(\theta_n^{(m)})\right], \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^\star \right\rangle \\ &\geq \left\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^\star \right\rangle - mL\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|\,\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^\star\| \\ &\geq \left\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^\star \right\rangle - mLr_{n+1}^2.\end{aligned}$$

Therefore, combining the above inequalities, we conclude

$$\begin{aligned}\sum_{n=0}^\infty &\left(\widetilde{V}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}^\star) = \left\langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}^\star \right\rangle + \sum_{i=1}^m \left(p_{n+1}^{(i)}(\theta_n^{(i)}) - p_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) \right) \right) \\ &\leq \left(\frac{mL_g}{2} + mL \right) \sum_{n=1}^\infty r_n^2 + (F(\boldsymbol{\theta}_0) - F^*) + m\sum_{n=1}^\infty \Delta_n < \infty.\end{aligned}$$ ∎

Lemma 5.4 (key lemma for asymptotic stationarity for inexact trust-region). *Assume* A1, A2, A3(b), *and* A4. *Then there exists $c > 0$ such that for all $n \geq 0$,*

$$\min\{r_n, 1\} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\, \|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\leq 1} \widetilde{V}(\boldsymbol{\theta}_n, \boldsymbol{\theta}) \leq \widetilde{V}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1}^\star) + cr_{n+1}^2.$$

Proof. Let $b_n := \min\{r_n, 1\}$ for all $n \geq 1$. Fix $n \geq 0$ and let $\boldsymbol{\theta} = [\theta^{(1)}, \dots, \theta^{(m)}] \in \boldsymbol{\Theta}$ be such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq b_{n+1} \leq r_{n+1}$. Then we have

$$G_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) \leq G_{n+1}^{(i)}(\theta^{(i)}).$$

Let $\boldsymbol{\Theta}_0 = S^{(1)} \times \cdots \times S^{(m)}$ be as in Proposition 4.4. From A3, each surrogate $g_n^{(i)}$ has L_g-Lipschitz continuous gradient on $S^{(i)}$. Also, $\boldsymbol{\theta}, \boldsymbol{\theta}_{n+1}^\star, \boldsymbol{\theta}_n \in \boldsymbol{\Theta}_0$ for all $n \geq 1$. Recall that $\nabla g_{n+1}^{(i)}(\theta_n^{(i)}) = \nabla f_{n+1}^{(i)}(\theta_n^{(i)})$ since $g_n^{(i)} \geq f_n^{(i)}$ and $g_{n+1}^{(i)}(\theta_n^{(i)}) = f_{n+1}^{(i)}(\theta_n^{(i)})$. Hence by subtracting $g_{n+1}^{(i)}(\theta_n^{(i)})$ from both sides and applying the L_g-smoothness of $g_n^{(i)}$ on $S^{(i)}$ (see Lemma 4.1),

$$\begin{aligned}&\left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta_{n+1}^{(i\star)} - \theta_n^{(i)} \right\rangle + p_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) \\ &\quad\leq \left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta^{(i)} - \theta_n^{(i)} \right\rangle + \frac{L_g}{2}\|\theta_{n+1}^{(i\star)} - \theta_n^{(i)}\|^2 + \frac{L_g}{2}\|\theta^{(i)} - \theta_n^{(i)}\|^2 + p_{n+1}^{(i)}(\theta^{(i)}) \\ &\quad\leq \left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta^{(i)} - \theta_n^{(i)} \right\rangle + L_g r_{n+1}^2 + p_{n+1}^{(i)}(\theta^{(i)}).\end{aligned}$$

Adding up these inequalities for $i = 1, \dots, m$,

$$\begin{aligned}\sum_{i=1}^m &\left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta_{n+1}^{(i\star)} - \theta_n^{(i)} \right\rangle + p_{n+1}^{(i)}(\theta_{n+1}^{(i\star)}) - p_{n+1}^{(i)}(\theta^{(i)}) \\ &\leq \left\langle \left[\nabla f_{n+1}^{(1)}(\theta_n^{(1)}), \dots, \nabla f_{n+1}^{(m)}(\theta_n^{(m)})\right], \boldsymbol{\theta} - \boldsymbol{\theta}_n \right\rangle + mL_g r_{n+1}^2 \\ &\overset{(a)}{\leq} \langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle + mL\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| + mL_g r_{n+1}^2 \\ &\overset{(b)}{\leq} \langle \nabla f(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle + cr_{n+1}^2\end{aligned}$$

for $c := m^2L + mL_g$, where (a) uses that $\nabla_i f$ is L-Lipschitz in the ith block coordinate (Proposition 4.4) for each $i = 1, \dots, m$ and (b) follows since $\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| \leq mr_{n+1}$

and $\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le r_{n+1}$. The above inequality holds for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ with $\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le b_{n+1}$. Furthermore, note that

$$\begin{aligned}\sum_{i=1}^{m}&\left\langle \nabla f_{n+1}^{(i)}(\theta_n^{(i)}),\, \theta_{n+1}^{(i\star)}-\theta_n^{(i)}\right\rangle\\ &=\left\langle \nabla f_{n+1}^{(1)}(\theta_n^{(1)}),\dots,\nabla f_{n+1}^{(m)}(\theta_n^{(1)}),\, \boldsymbol{\theta}_{n+1}^{\star}-\boldsymbol{\theta}_n\right\rangle\\ &\ge \left\langle \nabla f(\boldsymbol{\theta}_n),\, \boldsymbol{\theta}_{n+1}^{\star}-\boldsymbol{\theta}_n\right\rangle - mL\|\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n+1}\|\,\|\boldsymbol{\theta}_{n+1}^{\star}-\boldsymbol{\theta}_n\|\\ &\ge \left\langle \nabla f(\boldsymbol{\theta}_n),\, \boldsymbol{\theta}_{n+1}^{\star}-\boldsymbol{\theta}_n\right\rangle - mLr_{n+1}^2.\end{aligned}$$

Thus it follows that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le b_{n+1}} \widetilde{V}(\boldsymbol{\theta}_n,\boldsymbol{\theta})\le \widetilde{V}(\boldsymbol{\theta}_n,\boldsymbol{\theta}_{n+1}^{\star})+(c+mL)r_{n+1}^2.$$

Then using the same argument as in Lemma 4.6, we can conclude the assertion. ∎

Proposition 5.5 (sufficient condition for stationarity II). *Assume* A1*,* A2*,* A3(b)*, and* A4*. If there exists a subsequence $(\boldsymbol{\theta}_{n_k})_{k\ge1}$ such that $\sum_{k=1}^{\infty}\|\boldsymbol{\theta}_{n_k}-\boldsymbol{\theta}_{n_k+1}^{\star}\|=\infty$, then there is a further subsequence $(s_k)_{k\ge1}$ of $(n_k)_{k\ge1}$ so that $\lim_{k\to\infty}\boldsymbol{\theta}_{s_k}$ exists and is a stationary-Nash point of F over $\boldsymbol{\Theta}$.*

Proof. Denote $\eta_n:=\boldsymbol{\theta}_n-\boldsymbol{\theta}_{n+1}^{\star}$. Then under the hypothesis, by Proposition 5.3

$$\liminf_{k\to\infty}\sum_{k=1}^{\infty}\frac{1}{\|\eta_{n_k}\|}\widetilde{V}(\boldsymbol{\theta}_{n_k},\boldsymbol{\theta}_{n_k+1}^{\star})=0. \tag{5.2}$$

Let $(t_k)_{k\ge1}$ be a subsequence of $(n_k)_{k\ge1}$ for which the liminf in (5.2) is achieved. According to Proposition 4.4, by taking a further subsequence, we may assume that $\boldsymbol{\theta}_\infty=\lim_{k\to\infty}\boldsymbol{\theta}_{t_k}$ exists. Then $\boldsymbol{\theta}_\infty\in\boldsymbol{\Theta}$ since $\boldsymbol{\Theta}$ is closed by A1.

Now suppose for contradiction that $\boldsymbol{\theta}_\infty$ is not a stationary-Nash point of F over $\boldsymbol{\Theta}$. Then there exists $\boldsymbol{\theta}'\in\boldsymbol{\Theta}$ and $\delta>0$ such that $\widetilde{V}(\boldsymbol{\theta}_\infty,\boldsymbol{\theta}')>\delta>0$. Denote $\boldsymbol{\theta}^t:=t\boldsymbol{\theta}'+(1-t)\boldsymbol{\theta}_\infty$ for $t\in[0,1]$. Then by convexity of p,

$$\widetilde{V}(\boldsymbol{\theta}_\infty,\boldsymbol{\theta}^t)\ge t\widetilde{V}(\boldsymbol{\theta}_\infty,\boldsymbol{\theta}')>t\,\delta>0 \quad \text{for all } t\in[0,1].$$

Choose t sufficiently small so that $\|\boldsymbol{\theta}^t-\boldsymbol{\theta}_\infty\|<1/2$ and denote $\boldsymbol{\theta}^*$ for such $\boldsymbol{\theta}^t$. Note that $\boldsymbol{\theta}^*\in\boldsymbol{\Theta}$ by convexity of $\boldsymbol{\Theta}$.

By A1, the function $\widetilde{V}(\cdot,\cdot)$ in (2.4) is continuous. Hence we have $\widetilde{V}(\boldsymbol{\theta}_{t_k},\boldsymbol{\theta}^*)\to\widetilde{V}(\boldsymbol{\theta}_\infty,\boldsymbol{\theta}^*)>t\delta>0$ as $k\to\infty$.

Hence by Lemma 5.4, there is a constant $c>0$ such that for all $n\ge1$ and k sufficiently large,

$$\frac{t\delta}{2}\le\widetilde{V}(\boldsymbol{\theta}_{t_k},\boldsymbol{\theta}^*)\le\sup_{\substack{\boldsymbol{\theta}\in\boldsymbol{\Theta}\\ \|\boldsymbol{\theta}-\boldsymbol{\theta}_{t_k}\|\le1}}\widetilde{V}(\boldsymbol{\theta}_{t_k},\boldsymbol{\theta})\le\frac{\|\eta_{t_k}\|}{\min\{r_{t_k},1\}}\frac{\widetilde{V}(\boldsymbol{\theta}_{t_k},\boldsymbol{\theta}_{t_k+1}^{\star})}{\|\eta_{t_k}\|}+\frac{cr_{t_k+1}^2}{\min\{r_{t_k},1\}}.$$

Since $\|\boldsymbol{\theta}_{n+1}^{\star}-\boldsymbol{\theta}_n\|\le mr_{n+1}$, by using the hypothesis, the right-hand side above converges to zero as $k\to\infty$, which is a contradiction. ∎

The following result is crucial in establishing global convergence to stationary-Nash points (Theorem 2.1(i)) and it is the only place where we use nonsummability of the radii r_n's in the proof of Theorem 2.1.

PROPOSITION 5.6 (local structure of a nonstationary-Nash limit point). *Assume* A1, A2, A3(b), *and* A4. *Suppose there exists a nonstationary-Nash limit point* $\boldsymbol{\theta}_\infty$ *of* Λ. *Then there exists* $\varepsilon > 0$ *such that the* ε*-neighborhood* $B_\varepsilon(\boldsymbol{\theta}_\infty) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| < \varepsilon\}$ *with the following properties:*

(a) $B_\varepsilon(\boldsymbol{\theta}_\infty)$ *does not contain any stationary-Nash points of* Λ.

(b) *There exists infinitely many* $\boldsymbol{\theta}_n^\star$*'s outside of* $B_\varepsilon(\boldsymbol{\theta}_\infty)$.

Proof. We first show that there exists $\varepsilon > 0$ such that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ satisfies (a). Suppose for contradiction that there exists no such $\varepsilon > 0$. Then we have a sequence $(\boldsymbol{\theta}_{\infty;k})_{k\ge 1}$ of stationary-Nash points of Λ that converges to $\boldsymbol{\theta}_\infty$. Fix $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Note that by continuity of $\widetilde{V}(\cdot,\cdot)$ and $\widetilde{V}(\boldsymbol{\theta}_{\infty;k}, \boldsymbol{\theta}) \le 0$ since $\boldsymbol{\theta}_{\infty;k}$ is a stationary-Nash point of $F_{\infty;k}$ over $\boldsymbol{\Theta}$, we have $\widetilde{V}(\boldsymbol{\theta}_\infty, \boldsymbol{\theta}) \le 0$. Since $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ was arbitrary, this implies that $\boldsymbol{\theta}_\infty$ is a stationary-Nash point of F over $\boldsymbol{\Theta}$, a contradiction.

Next, we show that we can assume by choosing $\varepsilon > 0$ smaller, if necessary, that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ that does not contain any short points of Λ. Suppose not. Then for all $\varepsilon > 0$ sufficiently small, $B_\varepsilon(\boldsymbol{\theta}_\infty)$ contains some short point of Λ. Then there exists a subsequence of short points of Λ converging to $\boldsymbol{\theta}_\infty$, but then $\boldsymbol{\theta}_\infty$ has to be stationary-Nash by Proposition 5.2, which contradicts the hypothesis.

Now we assume that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ has no short point of Λ and also satisfies (a). We will show that $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ satisfies (b). Then $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ satisfies (a)–(b), as desired. Suppose for contradiction there are only finitely many $\boldsymbol{\theta}_n^\star$'s outside of $B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$. Then there exists an integer $M \ge 1$ such that $\boldsymbol{\theta}_n^\star \in B_{\varepsilon/2}(\boldsymbol{\theta}_\infty)$ for all $n \ge M$. Then each $\boldsymbol{\theta}_n^\star$ for $n \ge M$ is a long point of Λ. By definition, it follows that $\|\boldsymbol{\theta}_n^\star - \boldsymbol{\theta}_{n-1}\| \ge r_n$ for all $n \ge M$. Then since $\sum_{n=1}^\infty r_n = \infty$, by Proposition 5.5 there exists a subsequence $(n_k)_{k\ge 1}$ such that $\boldsymbol{\theta}'_\infty := \lim_{k\to\infty} \boldsymbol{\theta}_{n_k}$ exists and is stationary-Nash. But since $\boldsymbol{\theta}'_\infty \in B_\varepsilon(\boldsymbol{\theta})$, this contradicts (a). This shows the assertion. ∎

We are now ready to give proof of the asymptotic stationarity result stated in Theorem 2.1(iii).

Proof of Theorem 2.1(iii) *under* A3(a). Suppose for contradiction that there exists a non-stationary-Nash limit point $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$ of Λ. We may choose $\varepsilon > 0$ such that $B_\varepsilon(\boldsymbol{\theta}_\infty)$ satisfies the conditions (a)–(b) of Proposition 5.6. Choose $M \ge 1$ large enough so that $mr_n < \varepsilon/3$ whenever $n \ge M$. We call an integer interval $I := [\ell, \ell']$ a *crossing* if $\boldsymbol{\theta}_\ell \in B_{\varepsilon/3}(\boldsymbol{\theta}_\infty)$, $\boldsymbol{\theta}_{\ell'} \in B_{2\varepsilon/3}(\boldsymbol{\theta}_\infty)$, and no proper subset of I satisfies both of these conditions. By definition, two distinct crossings have empty intersections. Fixing a crossing $I = [\ell, \ell']$, it follows that by the triangle inequality,

$$\sum_{n=\ell}^{\ell'-1} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\| \ge \|\boldsymbol{\theta}_{\ell'} - \boldsymbol{\theta}_\ell\| \ge \varepsilon/3. \tag{5.3}$$

Note that since $\boldsymbol{\theta}_\infty$ is a limit point of Λ, $\boldsymbol{\theta}_n$ visits $B_{\varepsilon/3}(\boldsymbol{\theta}_\infty)$ infinitely often. Moreover, by condition (b) of Proposition 5.6, $\boldsymbol{\theta}_n^\star$ also exits $B_\varepsilon(\boldsymbol{\theta}_\infty)$ infinitely often. But since $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_n^\star\| \le mr_n < \varepsilon/3$, it follows that $\boldsymbol{\theta}_n$ exits $B_{2\varepsilon/3}(\boldsymbol{\theta}_\infty)$ infinitely often. It follows that there are infinitely many crossings. Let t_k denote the kth smallest index that appears in some crossing. Then $t_k \to \infty$ as $k \to \infty$, and by (5.3),

$$\sum_{k=1}^\infty \|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\| \ge (\#\text{ of crossings})\,\frac{\varepsilon}{3} = \infty.$$

By the triangle inequality,

$$\|\boldsymbol{\theta}^\star_{t_k+1} - \boldsymbol{\theta}_{t_k}\| \ge \|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\| - \|\boldsymbol{\theta}^\star_{t_k+1} - \boldsymbol{\theta}_{t_k+1}\| \ge \|\boldsymbol{\theta}_{t_k+1} - \boldsymbol{\theta}_{t_k}\| - \Delta_{t_k}.$$

Since Δ_n's are summable by Assumption A4, it follows that $\sum_{k=1}^\infty \|\boldsymbol{\theta}^\star_{t_k+1} - \boldsymbol{\theta}_{t_k}\| = \infty$.

Then by Proposition 5.5, there exists a subsequence $(s_k)_{k\geq 1}$ of $(t_k)_{k\geq 1}$ such that $\boldsymbol{\theta}'_\infty := \lim_{k\to\infty} \boldsymbol{\theta}_{s_k}$ exists and is a stationary-Nash point of F over $\boldsymbol{\Theta}$. However, since $\boldsymbol{\theta}_{t_k} \in B_{2\varepsilon/3}(\boldsymbol{\theta}_\infty)$ for all $k \geq 1$ by the choice of t_k, we have $\boldsymbol{\theta}'_\infty \in B_\varepsilon(\boldsymbol{\theta}_\infty)$. This contradicts condition (a) of Proposition 5.6 since $B_\varepsilon(\boldsymbol{\theta}_\infty)$ cannot contain any stationary-Nash point in Λ. ∎

6. Applications of the main result.

6.1. Nonnegative matrix factorization. Given a $p \times N$ data matrix $X \in \mathbb{R}^{p\times N}$ and an integer parameter $r \geq 1$, consider the following *constrained matrix factorization* problem:

$$\underset{W\in\Theta^{(1)}\subseteq\mathbb{R}^{p\times r},\, H\in\Theta^{(2)}\subseteq\mathbb{R}^{r\times N}}{\arg\min} \|X - WH\|_F^2 + \lambda_1 \|H\|_1, \tag{6.1}$$

where the two factors W and H are called *dictionary* and *code* matrices of X, respectively, and $\lambda_1 \geq 0$ is an ℓ_1-regularizer for the code matrix H. Depending on the application contexts, we may impose some *constraints* $\Theta^{(1)}$ and $\Theta^{(2)}$ on the dictionary and code matrices, respectively, such as nonnegativity or some other convex constraints. An interpretation of this approximate factorization is that the r columns of W give an approximate basis for spanning the N columns of X, where the columns of H give suitable linear coefficients for each approximation [30].

The NMF [23] is a special instance of the constrained matrix factorization problem above where $\Theta^{(1)} = \mathbb{R}^{p\times r}_{\geq 0}$ and $\Theta^{(2)} = \mathbb{R}^{r\times N}_{\geq 0}$ and $\lambda_1 = 0$ and $X \in \mathbb{R}^{p\times N}_{\geq 0}$ is given. NMF has numerous applications in text analysis, image reconstruction, medical imaging, bioinformatics, and many other scientific fields [5, 36]. The use of nonnegativity constraint in NMF is crucial in obtaining a "parts-based" representation of the input signal [23].

In [23], the following multiplicative update is studied for NMF:

$$\textbf{MU} \qquad \begin{cases} H_{n+1} \leftarrow H_n \odot (W_n^T X) \oslash (W_n^T W_n H_n), \\ W_{n+1}^T \leftarrow W_n^T \odot (H_n X^T) \oslash (H_n H_n^T W_n^T), \end{cases} \tag{6.2}$$

where $\odot$ and $\oslash$ denote entrywise multiplication and division. Given a nonnegative initialization (W_0, H_0), the iterate (6.2) generates a sequence of nonnegative factor matrices $(W_n, H_n)_{\geq 0}$. In [23], it was shown that the objective value of the NMF monotonically decreases under the iterate (6.2), but it has not been proven that the convergence is toward the set of stationary points of the NMF problem. There are some works that propose modified versions of (6.2) and show asymptotic convergence to stationary points (e.g., [24]). Furthermore, to the best of our knowledge, there has not been any result on the rate of convergence of any variants of MU (6.2).

Here we propose *MU with regularization* (MUR), which falls under our BMM (algorithm (1.6)) and satisfies the hypothesis of our main result, Theorem 2.1. Fix regularization parameters $\delta \geq 0$ (thresholding parameter) and $\rho \geq 0$ (proximal regularization parameter). Consider the following variant of MU:

$$\textbf{MUR} \begin{cases} \tilde{H}_n \leftarrow H_n \vee \delta, \\ H_{n+1} \leftarrow (\tilde{H}_n \odot (W_n^T X + \rho \tilde{H}_n) \oslash ((W_n^T W_n + \rho \mathbf{I})\tilde{H}_n), \\ \tilde{W}_n \leftarrow \tilde{W}_n \vee \delta, \\ W_{n+1}^T \leftarrow W_n^T \odot (H_n X^T + \rho \tilde{W}_n^T) \oslash ((H_n H_n^T + \rho \mathbf{I})\tilde{W}_n^T), \end{cases} \tag{6.3}$$

where $(\cdot \vee \delta)$ is the operation of taking maximum with δ entrywise and $\mathbf{I}$ denotes the $r \times r$ identity matrix. Note that by setting $\delta = \rho = 0$, (6.3) reduces to the standard MU in (6.2). The following corollary shows that the MUR (6.3) algorithm for NMF, as long as $\rho, \delta > 0$, retains the asymptotic convergence and the rate of convergence stated in Theorem 2.1.

COROLLARY 6.1 (convergence of MUR for NMF). *Fix a matrix* $X \in \mathbb{R}^{p\times N}_{\geq 0}$. *Let* $(\boldsymbol{\theta}_n)_{n\geq 0}$, $\boldsymbol{\theta}_n := (W_n, H_n)$ *be generated by* (6.3) *with arbitrary initialization* $\boldsymbol{\theta}_0$ *in a compact set* $\boldsymbol{\Theta} \subseteq \mathbb{R}^{p\times r}_{\geq 0} \times \mathbb{R}^{r\times N}_{\geq 0}$. *Denote* $f(W,H) := \frac{1}{2}\|X - WH\|_F^2$. *Suppose the thresholding and proximal regularization parameters* δ, ρ *are strictly positive. Then Theorem* 2.1(i)–(iii) *holds for* $(\boldsymbol{\theta}_n)_{n\geq 0}$.

While we restrict the factors (W,H) for NMF to live inside a compact subset $\boldsymbol{\Theta}$ (Corollary 6.1) for a technical reason, this does not lose any generality in terms of the objective values if $\boldsymbol{\Theta}$ consists of all nonnegative factor matrices with Frobenius norm bounded by any constant at least $2r\|X\|_F^{1/2}$. Indeed, if (W,H) is any pair of nonnegative factors such that $\|X - WH\|_F \leq \|X\|_F$, then one can rescale the columns of W and the corresponding rows of H suitably so that their norm is at most $2\|X\|_F^{1/2}$ and the rescaled factors have the same objective value $\|X - WH\|_F^2$ (see [28, sect. 6]).

To justify Corollary 6.1, we first explain why MUR (6.3) can be viewed as a special instance of the BMM algorithm (algorithm (1.6)). Consider the following convex subproblem of (6.1): $\min_{H\geq 0}(f_n(H) := \frac{1}{2}\|X - W_n H\|_F^2)$. For any matrix A and positive integer j at most the number of columns in A, denote A^j to be its jth column. Then $f_n(H) = \sum_{j=1}^N f_n^j(H^j)$, where $f_n^j(\mathbf{h}) := \frac{1}{2}\|X^j - W_n\mathbf{h}\|_F^2$.

Fix parameters $\rho, \delta \geq 0$. Define a function $g_n : \mathbb{R}^{r\times N} \to \mathbb{R}$ by $g_n(H) := \sum_{j=1}^N g_n^j (H^j)$, where

$$g_n^j(\mathbf{h}) := f_n^j(\mathbf{h}_n) + (\mathbf{h} - \mathbf{h}_n)^T \nabla_{\mathbf{h}} f_n^j(\mathbf{h}_n) + \frac{1}{2}(\mathbf{h} - \mathbf{h}_n)^T (\mathbf{D} + \rho\mathbf{I})(\mathbf{h} - \mathbf{h}_n),$$

where $\mathbf{h}_n := H_n^j$ and $\mathbf{D} = \mathbf{D}_n^j(\delta, \rho)$ is the $r \times r$ diagonal matrix given by

$$\mathbf{D}_{a,b} := \mathbf{1}(a = b)\frac{[W_n^T W_n(\mathbf{h}_n \vee \delta)]_a}{[\mathbf{h}_n \vee \delta]_a}. \tag{6.4}$$

Note that taking $\delta = 0$ in (6.4), g_n^j defined above becomes the majorizer of f_n^j used in [23]. The thresholding parameter $\delta \geq 0$ prevents the denominator above from vanishing when H_n^j has zero on some coordinates. Since the thresholding is applied to $\mathbf{h}_n$ in both the numerator and the denominator in (6.4) and since we do not threshold $W_n^T W_n$, it turns out that g_n^j is still a majorizer of f_n^j for $\delta > 0$. We claim that the following properties hold:

- [a] ∇f is L-Lipschitz continuous for some $L > 0$ when restricted onto a compact set.
- [b] $g_n \geq f_n$ and $g_n(H_n) = f_n(H_n)$; g_n is a quadratic function.
- [c] H_{n+1} in (6.3) is an exact minimizer of g_n over $\mathbb{R}^{p\times r}_{\geq 0}$.
- [d] g_n is ρ-strongly convex.
- [e] If $\delta > 0$, then there exists a constant $L_h > 0$ such that $h_n := g_n - f_n$ has L_h-Lipschitz continuous gradient for some constant $L_h > 0$.

A similar construction of surrogate function and claim will hold for W_{n+1} by symmetry. Points [b]–[c] verify that (6.3) is a particular instance of BMM (algorithm (1.6)). Given this, we can easily verify that the hypothesis of Theorem 2.1 holds for the NMF problem (6.1) and the algorithm MUR (6.3). Indeed, assumptions A1 and A2 hold trivially, except the sublevel set compactness for NMF holds if and only if $\boldsymbol{\Theta}$ is compact, which holds by the hypothesis. The smoothness property of the majorizer in A3 holds by point [e]. A3(b) holds for $\rho > 0$ due to point [d] (this is why we should require $\rho > 0$ in Corollary 6.1). Last, A4 holds by point [c]. This is enough to deduce Corollary 6.1 from Theorem 2.1.

Proof of points [a]–[e]. We first justify [a]. Note that

$$\nabla f(W,H) - \nabla f(W',H') = \left[W^T W(H-H'), (W-W')HH^T\right].$$

Thus, if we restrict (W,H) on a compact subset of the parameter space, then ∇f is L-Lipschitz continuous for some constant (depending on the compact subset) $L > 0$. This verifies [a].

Next, we show point [b]. Clearly g_n is a quadratic function. Next, fix $j \in \{1,\dots,N\}$. Note that we can expand $f_n^j(\mathbf{h})$ at $\mathbf{h} := H_n^j$ as a quadratic function. Subtracting this from g_n^j,

$$g_n^j(\mathbf{h}) - f_n^j(\mathbf{h}) = \frac{1}{2}(\mathbf{h}-\mathbf{h}_n)^T\left(\mathbf{D}_n^j - W_n^T W_n + \rho\mathbf{I}\right)(\mathbf{h}-\mathbf{h}_n). \tag{6.5}$$

We claim that the matrix $\mathbf{D}_n^j - W_n^T W_n$ is positive semidefinite. To justify the claim, it is enough to show that the following rescaled matrix is positive semidefinite:

$$Q := \operatorname{diag}(\tilde{H}_n^j)\,(\mathbf{D}_n^j - W_n^T W_n)\,\operatorname{diag}(\tilde{H}_n^j).$$

Indeed, this can be shown from the following observation (a similar computation was used in the proof of [23, Lem. 2] with $\tilde{H}_n^j$ replaced with H_n^j): For all $\mathbf{x} \in \mathbb{R}^r$,

$$\mathbf{x}^T Q\mathbf{x} = \sum_{a,b}[W_n^T W_n]_{a,b}[\tilde{H}_n^j]_a[\tilde{H}_n^j]_b\,(\mathbf{x}_a - \mathbf{x}_b)^2/2 \geq 0.$$

Now, since Q is positive semidefinite, the identity (6.5) implies point [b]. Point [d] follows by noting that $\mathbf{D}_n^j$ is a nonnegative diagonal matrix.

Next, we justify point [c]. First note that minimizing $g_n(H)$ over $H \in \mathbb{R}_{\geq 0}^{r\times N}$ can be done separately over the columns of H. Note that

$$\nabla g_n^j(H^j) = -W_n^T X^j + (\mathbf{D}_n^j + \rho\mathbf{I})H^j - \rho\tilde{H}_n^j.$$

The global minimizer of g_n^j is given by the solution to $\nabla g_n^j = \mathbf{0}$, which is

$$H_{n+1}^j := (\mathbf{D}_n^j + \rho\mathbf{I})^{-1}\left(W_n^T X^j + \rho\tilde{H}_n^j\right).$$

Assuming H_n and W_n are entrywise nonnegative, recursively, H_{n+1}^j is also entrywise nonnegative. Hence H_{n+1}^j above is the global minimizer of g_n^j on $\mathbb{R}_{\geq 0}^r$. By collecting the columns $j = 1,\dots,N$, it follows that H_{n+1} in (6.3) is the global minimizer of g_n over $\mathbb{R}_{\geq 0}^{r\times N}$.

Last, we justify point [e]. From (6.5) it suffices to show that $\mathbf{D}_n^j - W_n^T W_n$ has bounded maximum eigenvalue. A key difficulty in analyzing the original MU algorithm in [23] is that even if one assumes that the factors (W,H) live in a compact set, the definition of $\mathbf{D}_n^j$ in (6.4) involves the denominator of the coordinates of $\mathbf{h}_n$, which could vanish while the numerator $[W_n^T W_n \mathbf{h}_n]_a$ remains bounded. Our up-thresholding of $\mathbf{h}_n$ by $\delta > 0$ exactly prevents this issue while maintaining majorization. Indeed, by the hypothesis, the iterates $(W_n, H_n)_{n\geq 1}$ (and hence $\tilde{H}_n$s) are contained in a bounded set. Take L to be

$$L := \sup_{n\geq 1} \max_{1\leq j\leq N} \left(\|\mathbf{D}_n^j\|_\infty + \rho\right) \leq \sup_{n\geq 1} \delta^{-1} \left(\|W_n^T W_n \tilde{H}_n\|_\infty + \rho\right) < \infty.$$

Then L uniformly upper bounds the largest eigenvalue of $\mathbf{D}_n^j - W_n^T W_n + \rho \mathbf{I}$ in (6.5). This shows [e]. □

In section 7.3, we provide numerical experiments showing the advantage of MUR when dealing with sparse data.

6.2. Applications to constrained tensor factorization. As matrix factorization is for unimodal data, *nonnegative tensor factorization* (NTF) provides a powerful and versatile tool that can extract useful latent information out of multimodel data tensors. As a result, tensor factorization methods have witnessed increasing popularity and adoption in modern data science [35].

Suppose a data tensor $\mathbf{X} \in \mathbb{R}^{I_1\times\cdots\times I_m}$ is given and fix an integer $R \geq 1$. In the CP decomposition of $\mathbf{X}$ [20], we would like to find *loading matrices* $U^{(i)} \in \mathbb{R}^{I_i\times R}$ for $i = 1, \ldots, m$ such that the sum of the outer products of their respective columns approximates $\mathbf{X}$: $\mathbf{X} \approx \sum_{k=1}^R \bigotimes_{i=1}^m U^{(i)}[:, k]$, where $U^{(i)}[:, k]$ denotes the kth column of the $I_i \times R$ loading matrix $U^{(i)}$ and $\bigotimes$ denotes the outer product. As an optimization problem, the above CP decomposition model can be formulated as the following *constrained tensor factorization* problem:

$$\text{(6.6)} \qquad \underset{U^{(i)}\in\Theta^{(i)};i=1,\ldots,m}{\arg\min} F(U^{(1)}, \ldots, U^{(m)}) := \left\| \mathbf{X} - \sum_{k=1}^R \bigotimes_{i=1}^m U^{(i)}[:, k] \right\|_F^2 + \sum_{i=1}^m \lambda_i \|U^{(i)}\|_1,$$

where $\Theta^{(i)} \subseteq \mathbb{R}^{I_i\times R}$ denotes a closed and convex constraint set and $\lambda_i \geq 0$ is an ℓ_1-regularizer for the ith loading matrix $U^{(i)}$ for $i = 1, \ldots, m$. In particular, by taking $\lambda_i = 0$ and $\Theta^{(i)}$ to be the set of nonnegative $I_i \times R$ matrices for $i = 1, \ldots, m$, (6.6) reduces to the NCPD [35].

The constrained tensor factorization problem (6.6) falls under the framework of BCD in (1.3), since the objective function F in (6.6) is convex in each loading matrix $U^{(i)}$ for $i = 1, \ldots, m$. Indeed, BCD is a popular approach for both NMF and NTF problems [19]. Namely, when we apply BCD (1.3) for (6.6), each block update amounts to solving a quadratic problem under convex constraint. BCD for (6.6) is known as the form of ALS, which can be implemented by using a few steps of projected gradient descent for solving each subproblem. For NMF, ALS (or vanilla BCD in (1.3)) is known to converge to stationary points [16]. However, for NTF with $m \geq 3$ modes, global convergence to stationary points of ALS is not guaranteed in general [20] and requires some additional regularity conditions [7, 16]. Using our general framework of BMM-DR, we propose the following iterative algorithms for constrained tensor factorization: For each $n \geq 1$ and $i = 1, \ldots, m$,

(6.7)

BMM-DR for CTF

$$\begin{cases} \mathbf{A} \leftarrow \texttt{Out}(U_{n-1}^{(1)}, \ldots, U_{n-1}^{(i-1)}, U_{n-1}^{(i+1)}, \ldots, U_{n-1}^{(m-1)}) \in \mathbb{R}^{(I_1\times\cdots\times I_{i-1}\times I_{i+1}\times\cdots\times I_m)\times R} \\ B \leftarrow \text{unfold}(\mathbf{A}, m) \in \mathbb{R}^{(I_1\cdots I_{i-1} I_{i+1}\cdots I_m)\times R} \\ g_n^{(i)}(U) \leftarrow \text{Majorizing surrogate of } f_n^{(i)}(U) := \|\text{unfold}(\mathbf{X}, i) - BU^T\|_F^2 \\ U_n^{(i)} \in \arg\min_{U\in\Theta^{(i)},\, \|U - U_{n-1}^{(i)}\|_F \leq r_n} G_n^{(i)}(U) := g_n^{(i)}(U) + \lambda_i \|U\|_1, \end{cases}$$

where unfold$(\cdot, i)$ denotes the mode-i tensor unfolding (see [20]) and $r_n \in [0,\infty]$ for $n \geq 1$ denotes the radius of the trust-region. The iterate (6.7) specializes in various tensor factorization algorithms depending on the choice of the surrogate $g_n^{(i)}$. Namely, first, there are four ALS-type algorithms:

(a) (ALS) $g_n^{(i)} = f_n^{(i)}$, $r_n \equiv \infty$.
(b) (ALS with DR) $g_n^{(i)} = f_n^{(i)}$, $\sum_{n=1}^{\infty} r_n = \infty, \sum_{n=1}^{\infty} r_n^2 < \infty$.
(c) (ALS with PR) $g_n^{(i)}(U) = f_n^{(i)}(U) + \frac{\lambda}{2}\|U - U_{n-1}^{(i)}\|_F^2$, $\lambda > 0$, and $r_n \equiv \infty$.
(d) (ALS with DR + PR) $g_n^{(i)}(U) = f_n^{(i)}(U) + \frac{\lambda}{2}\|U - U_{n-1}^{(i)}\|_F^2$, $\lambda > 0$, $\sum_{n=1}^{\infty} r_n = \infty$, and $\sum_{n=1}^{\infty} r_n^2 < \infty$.

Next, specialize (6.6) into NCPD, where $\Theta^{(i)} = \mathbb{R}_{\geq 0}^{I_i \times R}$ and $\lambda_i = 0$ for $i = 1, \ldots m$. There are two MU-type algorithms for NCPD, which can be derived similarly as in the NMF case (see sect. 6.1):

(e) (MU) $U_{n+1}^{(i)} \leftarrow U_n^{(i)} \odot (B^T \mathrm{unfold}(\mathbf{X}, i)) \oslash (B^T B U_n^{(i)})$.
(f) (MUR) $\left\{ \tilde{U} \leftarrow U_n^{(i)} \vee \delta U_{n+1}^{(i)} \leftarrow \tilde{U} \odot \left((B^T \mathrm{unfold}(\mathbf{X}, i)) + \rho \tilde{U} \right) \oslash ((B^T B + \rho \mathbf{I}) \tilde{U}) \right.$.

For many instances of BMM-DR for CTF listed above, we can apply our general result (Theorem 2.1) to deduce their convergence and complexity. In order to guarantee asymptotic convergence to stationary points and iteration complexity as stated in Theorem 2.1, we need to assume that either diminishing radius or strongly convex surrogate is used. This rules out the vanilla ALS (a) for general CTF and MU (e) for NCPD. The following corollary holds for all the other options listed above.

Corollary 6.2 (convergence of BMM-DR for CTF). *Let $\boldsymbol{\theta}_n = [U_n^{(1)}, \ldots, U_n^{(m)}]$, $n \geq 0$, be generated by* (6.7) *with one of the options (*b*), (*c*), (*d*) (for general convex constraints on factor matrices) or (*f*) (for NCPD). Then* (i)–(iii) *of Theorem* 2.1 *hold for $(\boldsymbol{\theta}_n)_{n \geq 0}$.*

6.3. Block projected gradient descent. In the introduction, we discussed that BPGD (1.5) is a special instance of BMM with smooth objectives where prox-linear surrogates are exactly minimized over convex constraint sets. Therefore, our general result (Theorem 2.1) implies that the BPGD algorithm (1.5) converges asymptotically to the stationary points (not only Nash equilibrium) and also has iteration complexity of $\widetilde{O}((1 + \rho + \rho^{-1})\varepsilon^{-2})$, under the hypothesis of Theorem 2.1. See Corollary 6.3.

Corollary 6.3 (complexity of BPGD). *Let $(\boldsymbol{\theta}_n)_{n \geq 0}$ be generated by the BPGD updates* (1.5) *with stepsize* $1/\rho$. *Suppose* A1 *and* A2 *hold, where the convex constraint sets $\Theta^{(i)}$ are not necessarily bounded. If $\rho > L$, then* (i)–(iii) *of Theorem* 2.1 *hold for $(\boldsymbol{\theta}_n)_{n \geq 0}$. In particular, the iteration complexity for smooth nonconvex objectives with convex constraints is $\tilde{O}((1 + \rho + \rho^{-1})\varepsilon^{-2})$.*

Note that though we only state the convergence results for smooth objectives in Corollary 6.3, our general convergence and complexity results hold for nonsmooth nonconvex objectives with prox-linear updates as in (1.5) (when $p \neq 0$).

7. Experimental validation. In this section, we compare the performance of BMM-DR with other classical algorithms on different problems. The performance of the algorithms is evaluated by the *relative reconstruction error* defined as $\mathrm{Error} = \|\mathbf{X} - \hat{\mathbf{X}}\|/\|\mathbf{X}\|$ where $\mathbf{X}$ is the given data matrix (tensor) and $\hat{\mathbf{X}}$ is the reconstructed data matrix (tensor). For algorithms using diminishing radius, we take $r_n = c' n^{-\beta}/\log n$ for $n \geq 1$, where $c' > 0$ is a constant.

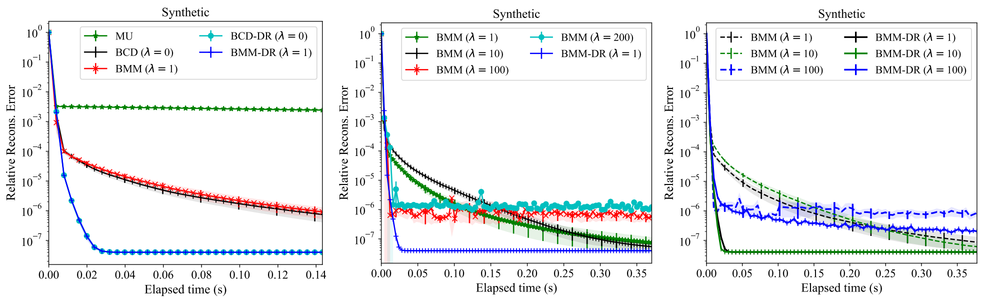

FIG. 2. *Comparison of BMM-DR with BMM on NMF.* $\beta = 0.5$ *is the diminishing radius parameter used for all algorithms with diminishing radius and* λ *is the proximal regularization parameter. The average relative reconstruction error with standard deviation is shown by the lines and shaded regions of respective colors. (Color figures are available online.)*

7.1. BMM-DR for NMF. We compare the performance of MU, BCD, BMM(c), BCD-DR, and BMM-DR(d) on the NMF task. The advantage of implementing diminishing radius becomes significant when matrices are ill-conditioned in NMF. In the experiments, the synthetic data matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{100\times 50}$ is generated by the product of $W \in \mathbb{R}_{\geq 0}^{100\times 7}$ and $H \in \mathbb{R}_{\geq 0}^{7\times 50}$ where W is set to have a large condition number of order 10^7. This makes the block-marginal objective functions have very small strong convexity parameters, if not zero. Each algorithm is run 10 times with random initialization, and the averaged reconstruction error is computed. The first two plots in Figure 2 show the practicality of BMM-DR with square-summable radii. Namely, in the left panel, BCD-DR and BMM-DR converge significantly faster to a more accurate solution than other standard algorithms without trust-region (MU, BCD, and BMM). In the middle panel, we compare a single instance of BMM-DR with various instances of BMM depending on the proximal regularization parameter λ. While excessively large proximal regularization ($\lambda = 100, 200$) seems to compromise the performance of BMM, all four instances of BMM are outperformed by BMM-DR with $\lambda = 1$.

Last, we discuss the experiments in the rightmost panel, where we make a direct comparison between BMM (dashed lines) and BMM-DR (solid lines) with the same surrogates presented. Notice that both the surrogate's strong convexity and smoothness parameters, ρ and L_g, increase additively by increasing λ. Hence from our complexity bound in Theorem 2.1, it is expected to see worse performance using steep surrogates corresponding to excessively large λ (e.g., $\lambda = 100$) for both BMM and BMM-DR, but for flat surrogates with small λ (e.g., $\lambda = 1, 10$), only the performance of BMM should be hindered, since the complexity bound involves the inverse of the strong convexity parameter. The experiments in the right panel indeed confirm this theoretically expected behavior.

7.2. Nonnegative CP-decomposition. In this section, we compare the performance of our proposed BMM-DR algorithm (b)–(d) and (f) for the task of NCPD (6.6) (with no L_1-regularization) against the two most popular approaches in practice: (1) ALS(a), and (2) MU(e). We consider a synthetic tensor data $\mathbf{X}_{\text{synth}}$ and a real-world tensor data $\mathbf{X}_{\text{Cifar10}}$.

For each data tensor $\mathbf{X}$ of shape $I_1 \times I_2 \times I_3$, we used the aforementioned algorithms to learn three loading matrices $U_i \in \mathbb{R}^{I_i \times R}$ for $i = 1, 2, 3$. We set the number of columns $R = 2$ for synthetic data and $R = 10$ for the `Cifar10` dataset. Each algorithm is run 10 times with independent random initial data in each case, and the plot shows the average relative reconstruction error with the standard deviation in shades.

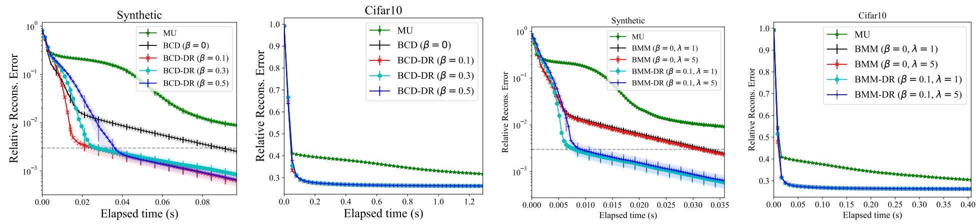

FIG. 3. *Comparison of the performance of BMM-DR (algorithm* (1.6)*) and MUR against BCD and MU for the NCPD problem. BCD (equivalent to ALS) is implmented as* (6.7)(a) *with* $r_n = \infty$ *for* $n \geq 1$. *BCD-DR is implemented as* (6.7)(b) *with* $c' = \|\mathbf{X}\|_F/(1.5 \times 10^5)$ *for synthetic data and* $c' = \|\mathbf{X}\|_F/(3 \times 10^5)$ *for* `Cifar` 10 *data. BMM*(c) *is implemented with a proximal regularizer with parameter* λ. *BMM-DR*(d) *is implemented on top of BMM with diminishing radius parameter* β *and the same* c' *as BCD-DR. The average relative reconstruction error with standard deviation is shown by the solid lines and shaded regions of respective colors.*

For synthetic data, as shown in Figure 3, BCD-DR with proper diminishing radius parameters β and c' is significantly faster than MU and also the standard vanilla BCD in terms of elapsed time. Here we take $c' = \|\mathbf{X}\|_F/(1.5 \times 10^5)$ where $\mathbf{X}$ denotes the synthetic data tensor, and 1.5×10^5 is the number of elements in the tensor. BCD-DR with $\beta = 0.1$ attains its best performance. In the third plot, a direct comparison between BMM and BMM-DR with the same surrogates is shown. One can observe that while the effect of different proximal parameters ρ is negligible, applying diminishing radius improves the performance of BMM with the same surrogates.

For `Cifar 10`, the same experiments are conducted with $c' = \|\mathbf{X}\|_F/(3 \times 10^5)$. All BCD-DR and BMM-DR outperform MU in terms of elapsed time and demonstrate a comparable convergence rate to the vanilla BCD. Diminishing radius does not accelerate the convergence as in the synthetic data case. In fact, the acceleration from diminishing radius becomes significant when the relative reconstruction error is of order 10^{-2}. However, decomposing real-world tensors to loading matrices with such a small relative reconstruction error may not be possible. Hence, the acceleration from diminishing radius is not observed in the `Cifar 10` data set case.

7.3. Comparison between MU and MUR for NMF. In this section, we compare the performance of our MUR (6.3) for the task of NMF against the original MU (6.2). We consider synthetic dense data $\mathbf{X}_{\text{synth}}$, synthetic sparse data $\mathbf{X}_{\text{synth-sp}}$ with 20% nonzero elements, and real-world data $\mathbf{X}_{\text{MNIST}}$.

In numerical experiments, we use MU and MUR to learn nonnegative matrices $W \in \mathbb{R}_{\geq 0}^{100\times 2}$ and $H \in \mathbb{R}_{\geq 0}^{2\times 50}$ with synthetic data, and $W \in \mathbb{R}_{\geq 0}^{28\times 15}$ and $H \in \mathbb{R}_{\geq 0}^{15\times 28}$ with `MNIST` data. MU and MUR with various threshold parameters δ and regularization parameter ρ are run 100 times in each experiment with random initial data. The average relative reconstruction error with standard deviation is computed and shown in Figure 4 with solid lines and shaded regions. As shown in Figure 4, in the synthetic data case without the sparsity feature, MU and MUR show similar convergence speeds. In the sparse data case, for both synthetic and `MNIST` data, MUR with various parameters significantly outperforms MU. In fact, writing MU in gradient descent form [24], the step size of gradient descent updates involves H and W in both the denominator and the numerator, whose elements could possibly be zero especially when the data is sparse. A zero numerator of the step size results in no change during updates, while a zero denominator would lead to blow-up issues. These challenges contribute to the comparatively poorer performance of MU when compared to MUR.

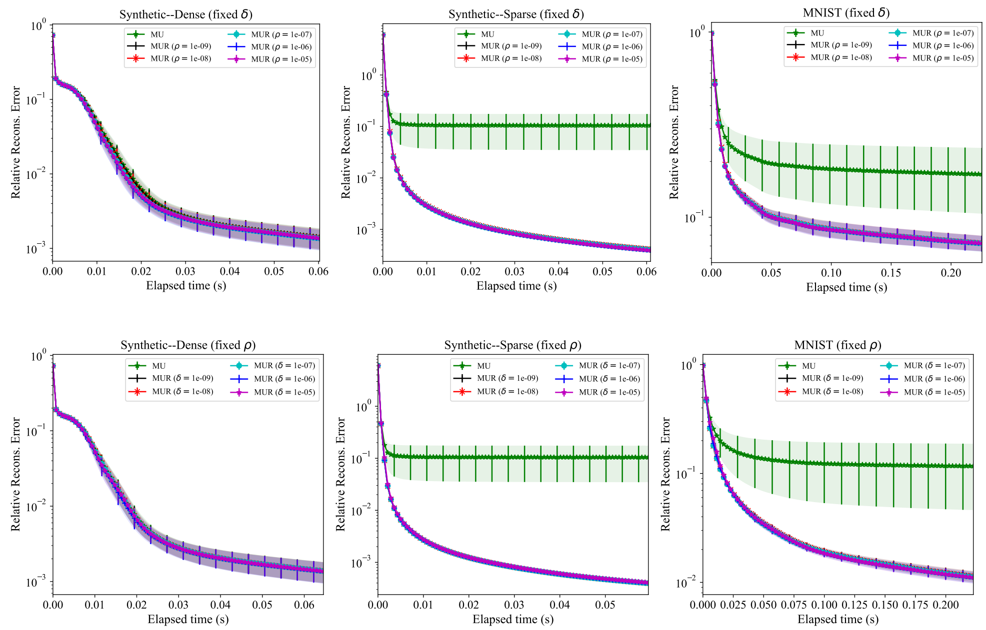

FIG. 4. *Comparison of the performance of MUR for the NMF problem against MU. For MUR in the first (second) row, δ (ρ) is fixed as 10^{-8}. The number r of columns of loading matrices is set to be* 2 *for synthetic data and* 15 *for MNIST data. The average relative reconstruction error with standard deviation is shown by the solid lines and shaded regions of respective colors.*

Acknowledgments. The authors appreciate valuable discussions with Stephen Wright. The authors also thank the anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] A. Alacaoglu and H. Lyu, *Convergence of stochastic first-order methods with dependent data for constrained nonconvex optimization*, in Proceedings of the International Conference on Machine Learning, 2023.

[2] A. Auslender, *Optimisation : Méthodes numériques*, Masson, Paris, 1976.

[3] A. Beck, *First-Order Methods in Optimization*, SIAM, Philadelphia, 2017.

[4] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060, https://doi.org/10.1137/120887679.

[5] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, *Algorithms and applications for approximate nonnegative matrix factorization*, Comput. Statist. Data Anal., 52 (2007), pp. 155–173, https://doi.org/10.1016/j.csda.2006.11.006.

[6] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 2015.

[7] D. P. Bertsekas, *Nonlinear programming*, J. Oper. Res. Soc., 48 (1997), pp. 334–334, https://doi.org/10.1057/palgrave.jors.2600425.

[8] A. Breloy, S. Kumar, Y. Sun, and D. P. Palomar, *Majorization-minimization on the stiefel manifold with application to robust sparse PCA*, IEEE Trans. Signal Process., 69 (2021), pp. 1507–1520, https://doi.org/10.1109/TSP.2021.3058442.

[9] X. Cai, C. Song, S. Wright, and J. Diakonikolas, *Cyclic block coordinate descent with variance reduction for composite nonconvex optimization*, in Proceedings of the International Conference on Machine Learning, PMLR, 2023, pp. 3469–3494.

[10] J. D. Carroll and J.-J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition*, Psychometrika, 35 (1970), pp. 283–319, https://doi.org/10.1007/BF02310791.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems*, SIAM J. Optim., 20 (2010), pp. 2833–2852, https://doi.org/10.1137/090774100.
[12] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, New York, 2011, pp. 185–212.
[13] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust Region Methods*, SIAM, Philadelphia, 2000.
[14] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Comm. Pure Appl. Math., 63 (2010), pp. 1–38, https://doi.org/10.1002/cpa.20303.
[15] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM J. Optim., 29 (2019), pp. 207–239, https://doi.org/10.1137/18M1178244.
[16] L. Grippo and M. Sciandrone, *On the convergence of the block nonlinear Gauss–Seidel method under convex constraints*, Oper. Res. Lett., 26 (2000), pp. 127–136, https://doi.org/10.1016/S0167-6377(99)00074-7.
[17] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, *A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing*, IEEE Signal Process. Mag., 33 (2015), pp. 57–77, https://doi.org/10.1109/MSP.2015.2481563.
[18] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo, *Iteration complexity analysis of block coordinate descent methods*, Math. Program., 163 (2017), pp. 85–114, https://doi.org/10.1007/s10107-016-1057-8.
[19] J. Kim, Y. He, and H. Park, *Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework*, J. Global Optim., 58 (2014), pp. 285–319, https://doi.org/10.1007/s10898-013-0035-4.
[20] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500, https://doi.org/10.1137/07070111X.
[21] D. Kwon and H. Lyu, *Complexity of block coordinate descent with proximal regularization and applications to Wasserstein CP-dictionary learning*, in Proceedings of the International Conference on Machine Learning, 2023.
[22] K. Lange, D. R. Hunter, and I. Yang, *Optimization transfer using surrogate objective functions*, J. Comput. Graph. Statist., 9 (2000), pp. 1–20, https://doi.org/10.1080/10618600.2000.10474858.
[23] D. D. Lee and H. S. Seung, *Algorithms for non-negative matrix factorization*, in Advances in Neural Information Processing Systems, 2001, pp. 556–562.
[24] C.-J. Lin, *On the convergence of multiplicative update algorithms for nonnegative matrix factorization*, IEEE Trans. Neural Networks, 18 (2007), pp. 1589–1596, https://doi.org/10.1109/TNN.2007.895831.
[25] C.-J. Lin, *Projected gradient methods for nonnegative matrix factorization*, Neural Comput., 19 (2007), pp. 2756–2779, https://doi.org/10.1162/neco.2007.19.10.2756.
[26] J. Liu, P. Musialski, P. Wonka, and J. Ye, *Tensor completion for estimating missing values in visual data*, IEEE Trans. Pattern Anal. Mach. Intell., 35 (2012), pp. 208–220, https://doi.org/10.1109/TPAMI.2012.39.
[27] Z.-Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35, https://doi.org/10.1007/BF00939948.
[28] H. Lyu, C. Strohmeier, and D. Needell, *Online nonnegative CP-dictionary learning for Markovian data*, J. Mach. Learn. Res., 23 (2022), pp. 6630–6679.
[29] Y. Nesterov, *Introductory Lectures on Convex Programming Volume* I*: Basic Course*, Lecture notes, 1998.
[30] B. A. Olshausen and D. J. Field, *Sparse coding with an overcomplete basis set: A strategy employed by V1?*, Vision Res., 37 (1997), pp. 3311–3325, https://doi.org/10.1016/S0042-6989(97)00169-7.
[31] L. Peng and R. Vidal, *Block Coordinate Descent on Smooth Manifolds*, preprint, arXiv:2305.14744, 2023.
[32] M. J. Powell, *On search directions for minimization algorithms*, Math. Program., 4 (1973), pp. 193–201, https://doi.org/10.1007/BF01584660.
[33] M. Razaviyayn, M. Hong, and Z.-Q. Luo, *A unified convergence analysis of block successive minimization methods for nonsmooth optimization*, SIAM J. Optim., 23 (2013), pp. 1126–1153, https://doi.org/10.1137/120891009.

[34] M. Razaviyayn, M. Hong, Z.-Q. Luo, and J.-S. Pang, *Parallel successive convex approximation for nonsmooth nonconvex optimization*, in Advances in Neural Information Processing Systems, 2014.

[35] A. Shashua and T. Hazan, *Non-negative tensor factorization with applications to statistics and computer vision*, in Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 792–799.

[36] L. Taslaman and B. Nilsson, *A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data*, PloS ONE, 7 (2012), e46331, https://doi.org/10.1371/journal.pone.0046331.

[37] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423, https://doi.org/10.1007/s10107-007-0170-0.

[38] S. J. Wright, *Coordinate descent algorithms*, Math. Program., 151 (2015), pp. 3–34, https://doi.org/10.1007/s10107-015-0892-3.

[39] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci., 6 (2013), pp. 1758–1789, https://doi.org/10.1137/120887795.

[40] A. L. Yuille and A. Rangarajan, *The concave-convex procedure*, Neural Comput., 15 (2003), pp. 915–936, https://doi.org/10.1162/08997660360581958.