Appendix A: Symbol definitions and preliminaries

In this section, we introduce some symbol definitions and the preliminary knowledge [26,27].

.1 Notation and preliminaries for nonconvex analysis

Definition 1. *If A is a set, then define the derived set A' is the set of overall cluster points of A.*

Proposition 1. *A bounded closed set in finite-dimensional space is a compact set.*

Definition 2. *A function $g : \mathbb{R}^n \to (-\infty, +\infty]$ is said to be proper if* $\mathrm{dom}\ g \neq \emptyset$*, where* $\mathrm{dom}\ g = \{x \in \mathbb{R} :\ g(x) < \infty\}$.

Definition 3. *If f satisfies*

$$\lim_{k\to\infty} x_k = x,\ f(x) \leq \liminf_{k\to\infty} f(x_k)(\forall x_k \in \mathrm{dom}\ f),$$

then f is called lower semicontinuous at $\mathrm{dom}\ f$.

Definition 4. *If f is coercive, then $\{x|x \in R^n,\ f(x) < a,\ \forall a \in \mathbb{R}\}$ is bounded and* $\inf_x f(x) > -\infty$.

Definition 5. *Let f be a proper lower semicontinuous function, The Fréchet subdifferential of f at x, written $\hat{\partial} f(x)$, is the set of all vectors u which satisfy*

$$\lim_{y\neq x, y\to x} \cdot \frac{f(y)-f(x)-\langle u,\ y-x\rangle}{\|y-x\|} \geq 0,$$

when $x \notin \mathrm{dom}\ f$, then set $\hat{\partial} f(x) = \emptyset$.

Definition 6. *The limiting subdifferential $\partial f(x) := \{u \in \mathbb{R}^n : \exists x^k \to x, f(x^k) \to f(x), u^k \to u, u^k \in \widehat{\partial} f(x^k)\}$.*

Proposition 2. *Let f be a proper lower semicontinuous function. If f has a local minimum at x^*, then $0 \in \partial f(x^*)$.*

Proposition 3. *Let f be a proper lower semicontinuous function, and g be a continuously differential function, then we have*

$$\forall x \in \mathrm{dom}\ f,\ \partial(f+g)(x) = \partial f(x) + \nabla g(x).$$

Definition 7. *Set f: $X \to R,\ X \subseteq R^n$, define the set $\mathcal{C}_L^p(X)$ as a set composed of all functions satisfying the following property.*

$$\|\nabla^p f(x) - \nabla^p f(y)\| \leq L\|x-y\|,$$

where L is the Lipschitz constant of $\nabla^p f$. $\mathcal{C}_L^p(X)$ is also known as a set of functions satisfying the Lipschitz property. in particular, when $p = 1$, ∇_f is Lipschitz continuous.

Proposition 4. *Set f: $R^n \to R$, if $f \in \mathcal{C}_L^1(R^n)$, then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{L}{2}\|y - x\|^2.$$

Definition 8. *Let f be a proper lower semicontinuous function, f is said to have the KŁ property on $\bar{u} \in dom(\partial f)$, if there exists $\eta \in (0, +\infty]$, and U is the neighborhood of $\bar{u}$, any u in U that satisfies the condition $f(\bar{u}) < f(u) < f(\bar{u}) + \eta$ has the following inequalities established.*

$$\phi'(f(u) - f(\bar{u}))dist(0, \partial f(u)) > 1,$$

where ϕ is the desingularization function, i.e. $\phi \in C^1([0, \eta)),\ \phi' > 0,\ \phi(0) = 0$, and ϕ is a concave function from in $(0, \eta)$. If f has the KŁ property at each point of dom ∂f*, then f is a KŁ function.*

Kurdyka–Łojasiewicz (KŁ) property [28] play a very important role for global convergence analysis in the nonconvex optimization.

Appendix B: Proof for convergence analysis

In this section, we prove the theorems in this paper. Therefore, it is essential to verify that Eq. (3) satisfies the following properties.

Assumption 1. *Define $J(x_{(n)}) = H(x_{(n)}) + \sum_{i=1}^{n} F_i(x_i)$, where*

(i) $\mathbb{D}_H \subseteq \prod_i^n \mathbb{R}^{d_i}$, $H : \mathbb{D}_H \to \mathbb{R}$, $H \in \mathcal{C}_L^1(\mathbb{D}_H)$ and H is continuously differentiable.

(ii) $\mathbb{D}_{F_i} \subseteq \prod_i^n \mathbb{R}^{d_i}$, $F_i : \mathbb{D}_{F_i} \to \mathbb{R}$ is a lower semicontinuous function.

(iii) The objective function J is a Kurdyka-Łojasiewicz (KŁ) function (see Definition 8), and J is also a proper coercive function with a lower bound.

This is because the properties outlined in Assumption 1 regarding Eq. (3) play a crucial role in ensuring the monotonicity and global convergence of Algorithm 1 (we will demonstrate this later).

Therefore, we verify that Eq. (3) satisfies the properties outlined in Assumption 1 one by one. First, it is easy to see from Theorem 1 that $H(\mathcal{W}, b) \in \mathcal{C}_L^1(\prod_i^{p+1} \mathbb{R}^{d_i})$, and F_i is lower semi-continuous (Definition 3), thus Eq. (3) satisfies Assumption 1(i) and (ii). For Assumption 1(iii), since both F_i and $H(\mathcal{W}, b)$ satisfy the KŁ property [26,19], and since the non-negativity of $H(\mathcal{W}, b)$ and F_i, we can conclude that J has a lower bound and satisfies the KŁ property (the sum of a finite number of KŁ functions remains a KŁ function [26]). Therefore, Eq. (3) satisfies Assumption 1.

Since Eq. (3) satisfies Assumption 1, for the sake of simplifying the proof process, we also provide the equivalent expression of Eq. (3) as follows.

$$J(\{x_i\}_{i=1}^{p+1}) = H(\{x_i\}_{i=1}^{p+1}) + \sum_{i=1}^{p+1} F_i(x_i), \tag{9}$$

where J satisfies Assumption 1. Obviously, when $\{x_i\}_{i=1}^{p} = \{w_i\}_{i=1}^{p}$, $x_{p+1} = b$, Eq. (9) degenerates into Eq. (3).

Similarly, let S be the domain of Eq. (9), the proximal operator of Eq. (9) is defined as

$$\begin{aligned} x_j^{k+1} &\in prox_{\sigma_j^k F_j}(y_j^k - \sigma_j^k \nabla_{x_j} h(y_j^k)) \\ &\in \arg\min_{x \in S}(F_j(x) + \frac{1}{2\sigma_j^k}\|x - y_j^k\|^2 \\ &+ \langle \nabla_{x_j} h(y_j^k), x - y_j^k \rangle), \end{aligned} \tag{10}$$

where $h(y_j^k) = H(\{x_i^{k+1}\}_{i=1}^{j-1}, y_j^k, \{y_i^k\}_{i=j+1}^{p+1})$. It is not difficult to see when $\{x_i\}_{i=1}^p = \{w_i\}_{i=1}^p$, $x_{p+1} = b$, the Eq. (10) degenerates into Eq. (5). Therefore, we use Eq. (9) as the objective function for the convergence proof of Algorithm 1 in this section.

Proof of Theorem 1

Proof. For $(\mathcal{W}_{\setminus i}, w_i, b)$ and $(\mathcal{W}_{\setminus i}, \bar{w}_i, \bar{b})$, from Theorem 3.1 of [19], we know that

$$\begin{aligned} &\|\sum_{s=1}^n \frac{y_s \nabla_{w_i} f_{(\mathcal{W},b)}(\mathcal{X}_s)}{1 + \exp(-y_s f_{(\mathcal{W}_{\setminus i}, w_i, b)}(\mathcal{X}_s))} - \sum_{s=1}^n \frac{y_s \nabla_{w_i} f_{(\mathcal{W},b)}(\mathcal{X}_s)}{1 + \exp(-y_s f_{(\mathcal{W}_{\setminus i}, \bar{w}_i, \bar{b})}(\mathcal{X}_s))}\|_2 \\ &\le \sqrt{2}\sum_{s=1}^n (\|\nabla_{w_i} f_{(\mathcal{W},b)}(\mathcal{X}_s)\|_2 + 1)^2 \|(w_i, b) - (\bar{w}_i, \bar{b})\|_2. \end{aligned}$$

Then we infer

$$\begin{aligned} \frac{\|\nabla_{w_i} H(\mathcal{W}_{\setminus i}, w_i, b) - \nabla_{w_i} H(\mathcal{W}_{\setminus i}, \bar{w}_i, b)\|_2}{\|w_i - \bar{w}_i\|_2} &\le \frac{\sqrt{2}}{n}\sum_{s=1}^n (\|\nabla_{w_i} f_{(\mathcal{W},b)}(\mathcal{X}_s)\|_2 + 1)^2 + \lambda_i \\ &= \tau_i. \end{aligned}$$

In the same way, we infer τ_b is

$$\frac{\|\nabla_b H(\mathcal{W}, b) - \nabla_b H(\mathcal{W}, \bar{b})\|_2}{\|b - \bar{b}\|_2} \le \frac{\sqrt{2}}{n}\sum_{s=1}^n (\|\nabla_b f_{(\mathcal{W},b)}(\mathcal{X}_s)\|_2 + 1)^2 = \tau_b,$$

where $\nabla_b f_{(\mathcal{W},b)}(\mathcal{X}_i) = 1$, thus $\tau_b = 4\sqrt{2}$.

Proof of Theorem 2

Proof. (i) From Proposition 4, when $n = 1$, we know that

$$\begin{aligned} H(x_{(1)}^{k+1}, \{y_i^k\}_{i=2}^P) &\le \langle \nabla_{x_1} H(y_{(P)}^k), x_1^{k+1} - y_1^k \rangle \\ &+ H(y_{(P)}^k) + \frac{L_{\nabla H_{x_1^k}}}{2}\|x_1^{k+1} - y_1^k\|^2. \end{aligned} \tag{11}$$

Since Eq. (10), we obtain

$$F_1(y_1^k) \geq F_1(x_1^{k+1}) + \frac{\sigma_1^k}{2}\|x_1^{k+1} - y_1^k\|^2 + \langle \nabla_{x_1} H(\{y_i^k\}_{i=1}^P), x_1^{k+1} - y_1^k \rangle. \quad (12)$$

Sum of the Eq. (11) and Eq. (12), we have

$$H(x_1^{k+1}, \{y_i^k\}_{i=2}^P) + F_1(x_1^{k+1}) \leq H(y_{(P)}^k) + F_1(y_1^k) - \rho\|x_1^{k+1} - y_1^k\|^2,$$

where $\rho = \frac{1}{2\sigma_1^k} - L_{\nabla H_{x_1^k}}$. Assuming Theorem 2(i) holds when $n = N$, i.e.,

$$H(x_{(N)}^{k+1}, \{y_i^k\}_{i=N+1}^P) \leq H(y_{(P)}^k) + \sum_{i=1}^N F_i(y_i^k) - \rho\|x_{(N)}^{k+1} - y_{(N)}^k\|^2 - \sum_{i=1}^N F_i(x_i^{k+1}), \quad (13)$$

where $\rho = \min(\left\{\frac{1}{2\sigma_i^k} - L_{\nabla H_{x_i^k}}\right\}_{i=1}^N)$. Since Eq. (10), we also have

$$\begin{aligned} F_{N+1}(y_{N+1}^k) \geq{}& F_{N+1}(x_{N+1}^{k+1}) + \frac{\sigma_{N+1}^k}{2}\|x_{N+1}^{k+1} - y_{N+1}^k\|^2 \\ &+ \langle \nabla_{x_{N+1}} H(x_{(N)}^{k+1}, \{y_i^k\}_{i=N+1}^P), x_{N+1}^{k+1} - y_{N+1}^k \rangle. \end{aligned} \quad (14)$$

From Proposition 4, we infer

$$\begin{aligned} H(x_{(N+1)}^{k+1}, \{y_i^k\}_{i=N+2}^P) \leq{}& H(x_{(N)}^{k+1}, \{y_i^k\}_{i=N+1}^P) \frac{L_{\nabla H_{x_{N+1}^k}}}{2}\|x_{N+1}^{k+1} - y_{N+1}^k\|^2 \\ &+ \langle \nabla_{x_{N+1}} H(x_{(N)}^{k+1}, \{y_i^k\}_{i=N+1}^P), x_{N+1}^{k+1} - y_{N+1}^k \rangle. \end{aligned} \quad (15)$$

Sum of the Eq. (13), Eq. (14) and Eq. (15), when $n = P = p + 1$, we have

$$\begin{aligned} J(\{x_i^{k+1}\}_{i=1}^{p+1}) &\leq J(\{y_i^k\}_{i=1}^{p+1}) - \rho\|x_{(p+1)}^{k+1} - y_{(p+1)}^k\|^2 \\ &\leq J(\{x_i^k\}_{i=1}^{p+1}) - \rho\|x_{(p+1)}^{k+1} - y_{(p+1)}^k\|^2, \end{aligned} \quad (16)$$

where $\rho = \min(\left\{\frac{1}{2\sigma_i^k} - L_{\nabla H_{x_i^k}}\right\}_{i=1}^{p+1})$. This means the Theorem 2(i) is true.

(ii) From Eq. (16), we have

$$\rho\|x_{(P)}^{k+1} - y_{(P)}^k\|^2 \leq J(\{x_i^k\}_{i=1}^P) - J(\{x_i^{k+1}\}_{i=1}^P).$$

Sum of both side, since J is not negative, thus we have

$$\begin{aligned} \rho\sum_{k=1}^{\infty}\|x_{(P)}^{k+1} - y_{(P)}^k\|^2 &= \sum_{k=1}^{\infty}(J(x_{(P)}^k) - J(x_{(P)}^{k+1})) \\ &= J(x_{(P)}^1) - \inf J \\ &< \infty. \end{aligned}$$

It follows that

$$\lim_{k\to\infty}\|x_{(P)}^{k+1} - y_{(P)}^k\|^2 = 0,$$

when $P = p + 1$, $\lim_{k\to\infty}\|x_{(p+1)}^{k+1} - y_{(p+1)}^k\| = 0$.

Proof of Theorem 3

Proof. Let

$$h(y_j^k) = H(\{x_i{}^{k+1}\}_{i=1}^{j-1}, y_j^k, \{y_i{}^k\}_{i=j+1}^{P}).$$

By Proposition 2 and Proposition 3, Eq. (10) follows that

$$0 \in \nabla_{x_j} h(y_j^k) - \frac{1}{\sigma_j^k}(y_j^k - x_j^{k+1}) + \partial_{x_j}(\sum_{i=1}^{j} F_i(x_i^{k+1}) + \sum_{i=j+1}^{P} F_i(y_i^k)).$$

Thus we infer

$$\begin{aligned} &\nabla_{x_j} h(x_j^{k+1}) - \nabla_{x_j} h(y_j^k) + \frac{1}{\sigma_j^k}(y_j^k - x_j^{k+1}) \\ &\in \nabla_{x_j} h(x_j^{k+1}) + \partial_{x_j}(\sum_{i=1}^{j} F_i(x_i^{k+1}) + \sum_{i=j+1}^{P} F_i(y_i^k)) \\ &= \partial_{x_j} J(\{x_i^{k+1}\}_{i=1}^{j-1}, x_j^{k+1}, \{y_i^k\}_{i=j+1}^{P}). \end{aligned} \tag{17}$$

Define

$$g_{x_j^{k+1}} = \nabla_{x_j} h(x_j^{k+1}) - \nabla_{x_j} h(y_j^k) + \frac{1}{\sigma_j^k}(y_j^k - x_j^{k+1}). \tag{18}$$

Since Definition 7, we have

$$\begin{aligned} \|g_{x_j^{k+1}}\| &= \|\nabla_{x_j} h(x_j^{k+1}) - \nabla_{x_j} h(y_j^k) + \frac{1}{\sigma_j^k}(y_j^k - x_j^{k+1})\| \\ &\le \|\nabla_{x_j} h(x_j^{k+1}) - \nabla_{x_j} h(y_j^k)\| + \|\frac{1}{\sigma_j^k}(y_j^k - x_j^{k+1})\| \\ &\le L_{\nabla H_{x_j^k}} \|x_j^{k+1} - y_j^k\| + \frac{1}{\sigma_j^k}\|x_j^{k+1} - y_j^k\| \\ &= (L_{\nabla H_{x_j^k}} + \frac{1}{\sigma_j^k})\|x_j^{k+1} - y_j^k\|. \end{aligned} \tag{19}$$

Thus we infer

$$\begin{aligned} \| \left\{ g_{x_i^{k+1}} \right\}_{i=1}^{P} \| &= \sum_{i=1}^{P} \|g_{x_j^{k+1}}\| \\ &\le \rho_b \| \{x_i^{k+1} - y_i^k\}_{i=1}^{P} \|, \end{aligned} \tag{20}$$

where $\rho_b = max(\left\{ \frac{1}{\sigma_i^k} + L_{\nabla H_{x_i^k}} \right\}_{i=1}^{P})$, when $P = p + 1$, Theorem 3 is true.

The definition of z^k and c^k is the same as the main body text, i.e.,

$$z^k = \{x_i^k\}_{i=1}^{P},\ c^k = \{y_i^k\}_{i=1}^{P}.$$

Proof of Theorem 4

Proof. (i) Observe that z' can be viewed as an intersection of compact sets

$$z' = \cap_{q\in\mathbb{N}}\overline{\cup_{k\geq q}\{z^k\}}.$$

Thus z' is compact.

(ii) $\forall \overline{z} \in z'$, there exists a subsequence z^{k_j} such that:

$$\lim_{j\to\infty} z^{k_j} = \overline{z},$$

where $\overline{z} = \{\overline{x_i}\}_{i=1}^P$. Since H and F_i are lower semicontinuous (Assumption 1), we obtain that

$$\begin{aligned}\lim_{j\to\infty}\inf H(z^{k_j}) &\geq H(\overline{z}),\\ \lim_{j\to\infty}\inf F(z^{k_j}) &\geq F(\overline{z}),\end{aligned} \tag{21}$$

where $F(z^{k_j}) = \sum_{i=1}^P F_i(x_i^{k_j})$. Let

$$h(y_j^k) = H(\{{x_i}^{k+1}\}_{i=1}^{j-1}, y_j^k, \{{y_i}^k\}_{i=j+1}^P).$$

From Eq. (10), we have

$$\begin{aligned}&F_i(x_i^{k+1}) + \frac{1}{2\sigma_i^k}\|x_i^{k+1} - y_i^k\|^2 + \langle\nabla_{x_i}h(y_i^k), x_i^{k+1} - y_i^k\rangle.\\ &\leq F_i(\overline{x_i}) + \frac{1}{2\sigma_i^k}\|\overline{x_i} - y_i^k\|^2 + \langle\nabla_{x_i}h(y_i^k), \overline{x_i} - y_i^k\rangle\end{aligned} \tag{22}$$

Choosing $k = k_j - 1$, from Theorem 2(ii) and Eq. (22), we infer

$$\lim_{j\to\infty}\sup F_i(x_i^{k_j}) \leq \lim_{j\to\infty}\sup(F_i(\overline{x_i}) + \frac{1}{2\sigma_i^k}\|\overline{x_i} - y_i^k\|^2 + \langle\nabla_{x_i}h(y_i^k), \overline{x_i} - y_i^k\rangle). \tag{23}$$

Since $\lim_{j\to\infty} z^{k_j} = \overline{z}$ and $\lim_{j\to\infty}\|z^k - c^{k-1}\| = 0$ (Theorem 2(ii)), we can get that

$$\begin{aligned}\lim_{j\to\infty}\|\overline{x_i} - y_i^k\| &\leq \lim_{j\to\infty}\|\overline{x_i} - x_i^{k+1}\| + \lim_{j\to\infty}\|x_i^{k+1} - y_i^k\|.\\ &= 0.\end{aligned} \tag{24}$$

From Eq. (23) and Eq. (24), we infer

$$\lim_{j\to\infty}\sup F_i(x_i^{k_j}) \leq \lim_{j\to\infty}\sup F_i(\overline{x_i}).$$

Thus we have

$$\lim_{j\to\infty}\sup F(z^{k_j}) \leq \lim_{j\to\infty}\sup F(\overline{z}). \tag{25}$$

From Eq. (25) and Eq. (21), we infer

$$\lim_{j\to\infty} F(z^{k_j}) = \lim_{j\to\infty} F(\overline{z}). \tag{26}$$

Since Theorem 2, we have

$$\lim_{j\to\infty} H(z^{k_j}) = H(\overline{z}). \tag{27}$$

Thus from Eq. (26) and Eq. (27), we infer

$$\begin{aligned}\lim_{j\to\infty} H(z^{k_j}) + \lim_{j\to\infty} F(z^{k_j}) &= \lim_{j\to\infty} (H(z^{k_j}) + F(z^{k_j}))\\ &= \lim_{j\to\infty} J(z^{k_j})\\ &= J(\overline{z})\\ &= J^*,\end{aligned}$$

which means J is constant on z'.

(iii) Since Eq. (17) and Eq. (18), we know that

$$g_{x_j^{k+1}} \in \partial_{x_j} J(\{x_i^{k+1}\}_{i=1}^{j-1}, x_j^{k+1}, \{y_i^k\}_{i=j+1}^{P}). \tag{28}$$

From Theorem 2(ii) and Theorem 3, we infer

$$\lim_{k\to\infty} \| \left\{ g_{x_i^{k+1}} \right\}_{i=1}^{P} \| \leq \lim_{k\to\infty} \rho_b \| x_{(P)}^{k+1} - y_{(P)}^k \| = 0. \tag{29}$$

Since Theorem 2(i) and Theorem 3, we know that

$$\begin{aligned}\lim_{k\to\infty} J(\{x_i^{k+1}\}_{i=1}^{j-1}, x_j^{k+1}, \{y_i^k\}_{i=j+1}^{P}) &= \lim_{k\to\infty} J(x_{(P)}^k)\\ &= J^*. \end{aligned} \tag{30}$$

Therefore, from Eq. (28), Eq. (29) and Eq. (30), we have

$$\begin{aligned}0 &\in \partial_{x_j} \lim_{k\to\infty} J(\{x_i^{k+1}\}_{i=1}^{j-1}, x_j^{k+1}, \{y_i^k\}_{i=j+1}^{P})\\ &= \partial_{x_j} J(x^*).\end{aligned}$$

This means when $P = p + 1$, $0 \in \partial J(x^*)$.

Proof of Theorem 5

Proof. From Definition 8, there exists a concave function ϕ so that

$$\phi^{'}(J(z^k) - J(\overline{z})) dist(0, \partial J(z^k)) \geq 1. \tag{31}$$

From ϕ is the convex function, we have

$$\begin{aligned}\phi(J(z^{k+1}) - J(\overline{z})) &\leq \phi(J(z^k) - J(\overline{z}))\\ &\quad + \phi^{'}(J(z^k) - J(\overline{z}))(J(z^{k+1}) - J(z^k)).\end{aligned} \tag{32}$$

From Theorem 3, we infer:

$$\begin{aligned} dist(0, \partial J(z^k)) &\leq \| \left\{ p_{x_i^k} \right\}_{i=1}^{P} \| \\ &\leq \rho_b \| \left\{ x_i^k - y_i^{k-1} \right\}_{i=1}^{P} \| \\ &= \rho_b \| z^k - c^{k-1} \|. \end{aligned} \tag{33}$$

Since the Eq. (31) and Eq. (33), we infer

$$\begin{aligned} \phi^{'}(J(z^k) - J(\overline{z})) &\geq \frac{1}{dist(0, \partial J(z^k))} \\ &\geq \frac{1}{\rho_b \| z^k - c^{k-1} \|}. \end{aligned} \tag{34}$$

Let $G(k) = J(z^k) - J(\overline{z})$, from Eq. (32), Eq. (33) and Eq. (34), we have

$$\begin{aligned} \phi(G(k)) - \phi(G(k+1)) &\geq \phi^{'}(G(k))(G(k) - G(k+1)) \\ &\geq \frac{G(k) - G(k+1)}{\rho_b \| z^k - c^{k-1} \|} \\ &\geq \frac{\rho \| z^{k+1} - c^k \|^2}{\rho_b \| z^k - c^{k-1} \|}. \end{aligned}$$

Define C=$\frac{\rho}{\rho_b}$, C is a constant, so we infer

$$\| z^{k+1} - c^k \|^2 \leq C(\phi(G(k)) - \phi(G(k+1))) \| z^k - c^{k-1} \|.$$

Using the fact that $2ab \leq a^2 + b^2$

$$2 \| z^{k+1} - c^k \| \leq C(\phi(G(k)) - \phi(G(k+1))) + \| z^k - c^{k-1} \|.$$

Sum both sides

$$\begin{aligned} 2 \sum_{k=l+1}^{K} \| z^{k+1} - c^k \| &\leq \sum_{k=l+1}^{K} \| z^k - c^{k-1} \| + C(\phi(G(l+1)) - \phi(G(K+1))) \\ &= C(\phi(G(l+1)) - \phi(G(K+1))) + \| z^{l+1} - c^l \| \\ &+ \sum_{k=l+1}^{K} \| z^{k+1} - c^k \|. \end{aligned} \tag{35}$$

From Assumption 1 and Eq. (35), we can get that

$$\begin{aligned} \lim_{K \to \infty} \sum_{k=l+1}^{K} \| z^{k+1} - c^k \| &\leq \| z^{l+1} - c^l \| + C\phi(G(l+1)) - \lim_{K \to \infty} C\phi(G(K+1))) \\ &< \infty. \end{aligned}$$

Thus we have

$$\sum_{k=l+1}^{\infty} \|z^{k+1} - c^k\| < \infty. \tag{36}$$

From Assumption 1, no matter $c^k = z^k + \beta_k(z^k - z^{k-1})$ or $c^k = z^k$, we always have

$$\|z^{k+1} - z^k\| - \beta_k\|z^k - z^{k-1}\| \ \leq \|z^{k+1} - c^k\|. \tag{37}$$

From Eq. (36) and Eq. (37), we know that

$$\begin{aligned}\sum_{k=l+1}^{\infty} (\|z^{k+1} - z^k\| - \beta_k\|z^k - z^{k-1}\|) &\leq \sum_{k=l+1}^{\infty} \|z^{k+1} - c^k\| \\ &< \infty.\end{aligned}$$

Since

$$\|z^{k+1} - z^k\| - \beta_{max}\|z^k - z^{k-1}\| \leq \|z^{k+1} - z^k\| - \beta_k\|z^k - z^{k-1}\|,$$

we also have

$$\begin{aligned}\sum_{k=l+1}^{\infty} (\|z^{k+1} - z^k\| - \beta_{max}\|z^k - z^{k-1}\|) &\leq \sum_{k=l+1}^{\infty} (\|z^{k+1} - z^k\| - \beta_k\|z^k - z^{k-1}\|) \\ &< \infty, \end{aligned} \tag{38}$$

and

$$\begin{aligned}&\sum_{k=l+1}^{\infty} \|z^{k+1} - z^k\| - \sum_{k=l+1}^{\infty} \beta_{max}\|z^k - z^{k-1}\| \\ &= \sum_{k=l+1}^{\infty} (\|z^{k+1} - z^k\| - \beta_{max}\|z^{k+1} - z^k\|) + \beta_{max}\|z^{l+1} - z^l\| \\ &= \sum_{k=l+1}^{\infty} (1 - \beta_{max})\|z^{k+1} - z^k\| + \beta_{max}\|z^{l+1} - z^l\|. \end{aligned} \tag{39}$$

Let $s^k = \|z^{k+1} - z^k\|$, from Eq. (38) and Eq. (39), we have

$$\begin{aligned}\sum_{k=l+1}^{\infty} (1 - \beta_{max})s^k - \beta_{max}s^l &= \sum_{k=l}^{\infty}(s^{k+1} - \beta_{max}s^k) \\ &\leq \sum_{k=l+1}^{\infty} \|z^{k+1} - c^k\|. \\ &< \infty.\end{aligned}$$

Thus we infer

$$\sum_{k=l+1}^{\infty} (1 - \beta_{max})s^k < \infty. \tag{40}$$

From $(1-\beta_{max})$ is a positive constant and Eq. (40), we infer

$$\sum_{k=l+1}^{\infty} \|z^{k+1}-z^k\| = \sum_{k=l+1}^{\infty} s^k < \infty.$$

Thus we have

$$\sum_{k=0}^{\infty} \|z^{k+1}-z^k\| = \sum_{k=0}^{\infty} s^k < \infty.$$

This shows that

$$\begin{aligned} \lim_{K\to\infty} \|z^{K+p}-z^K\| &= \lim_{K\to\infty} \|\sum_{k=K+1}^{K+p} (z^{k+1}-z^k)\| \\ &\le \lim_{K\to\infty} \sum_{k=K+1}^{K+p} \|z^{k+1}-z^k\| \\ &= \lim_{K\to\infty} \sum_{k=K+1}^{\infty} s^k \\ &= 0. \end{aligned} \tag{41}$$

This means the Theorem 5 is true.

Proof of Theorem 6

Proof. Combined with Theorem 2, Theorem 4 and Theorem 5, we can follow the same technique of the proof of Theorem 2 in [29] to get the result.

References

1. Lei Wang, Zhu-Hong You, Xing Chen, Yang-Ming Li, Ya-Nan Dong, Li-Ping Li, and Kai Zheng. Lmtrda: Using logistic model tree to predict mirna-disease associations by fusing multi-source information of sequences and similarities. *PLOS Computational Biology*, 15(3):e1006865, 2019.
2. Wenwen Min, Juan Liu, and Shihua Zhang. Network-regularized sparse logistic regression models for clinical risk prediction and biomarker discovery. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(3):944–953, 2018.
3. Gijsbert M Kalisvaart, Thomas Van Den Berghe, Willem Grootjans, Maryse Lejoly, Wouter CJ Huysse, Judith VMG Bovée, David Creytens, Hans Gelderblom, Frank M Speetjens, Lore Lapeire, et al. Evaluation of response to neoadjuvant chemotherapy in osteosarcoma using dynamic contrast-enhanced mri: development and external validation of a model. *Skeletal radiology*, pages 1–10, 2023.
4. Bernd Erber, Nina Hesse, Sophia Goller, Fabian Gilbert, Jens Ricke, Christian Glaser, and Andreas Heuck. Diagnostic performance and interreader agreement of individual and combined non-enhanced and contrast-enhanced mr imaging parameters in adhesive capsulitis of the shoulder. *Skeletal Radiology*, 11153(112):4183–4215263–273, 2024.

5. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 60(6):84–90, 2017.
6. Xiaorou Zheng, Jianxin Jia, Jinsong Chen, Shanxin Guo, Luyi Sun, Chan Zhou, and Yawei Wang. Hyperspectral image classification with imbalanced data based on semi-supervised learning. *Applied Sciences*, 12(8):3943, 2022.
7. Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(Jul):1519–1555, 2007.
8. Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *The Journal of Machine Learning Research*, 14(1):807–841, 2013.
9. Yanou Ramon, David Martens, Theodoros Evgeniou, and Stiene Praet. Can metafeatures help improve explanations of prediction models when using behavioral and textual data? *Machine Learning*, pages 1–40, 2021.
10. Dimitris Bertsimas and Vassilis Digalakis Jr. The backbone method for ultra-high dimensional sparse machine learning. *Machine Learning*, 111(6):2161–2212, 2022.
11. Wenwen Min, Juan Liu, and Shihua Zhang. Group-sparse svd models via l_1- and l_0-norm penalties and their applications in biological data. *IEEE Transactions on Knowledge and Data Engineering*, 33(2):536–550, 2021.
12. Wenwen Min, Xiang Wan, Tsung-Hui Chang, and Shihua Zhang. A novel sparse graph-regularized singular value decomposition model and its application to genomic data analysis. *IEEE Transactions on Neural Networks and Learning Systems.*, 33(8):3842–3856, 2022.
13. Lili Pan, Shenglong Zhou, Naihua Xiu, and Hou-Duo Qi. A convergent iterative hard thresholding for nonnegative sparsity optimization. *Pacific Journal of Optimization*, 13(2):325–353, 2017.
14. Rui Wang, Naihua Xiu, and Chao Zhang. Greedy projected gradient-newton method for sparse logistic regression. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2):527–538, 2019.
15. Ying Wu, Dan Chen, Chaoqian Li, and Niansheng Tang. Bayesian tensor logistic regression with applications to neuroimaging data analysis of alzheimer's disease. *Statistical Methods in Medical Research*, 31(12):2368–2382, 2022.
16. Wenwen Min, Tsung-Hui Chang, Shihua Zhang, and Xiang Wan. TSCCA: a tensor sparse CCA method for detecting microrna-gene patterns from multiple cancers. *PLOS Computational Biology*, 17(6):e1009044, 2021.
17. Xu Tan, Yin Zhang, Siliang Tang, Jian Shao, Fei Wu, and Yueting Zhuang. Logistic tensor regression for classification. In *Intelligent Data Analysis*, pages 573–581. Springer, 2013.
18. Weiwei Guo, Irene Kotsia, and Ioannis Patras. Tensor learning for regression. *IEEE Transactions on Image Processing*, 21(2):816–827, 2011.
19. Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154, 2014.
20. Bilian Chen, Jiewen Guan, Zhening Li, and Zhehao Zhou. Robust feature extraction via ℓ_∞-norm based nonnegative tucker decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
21. Cristian Minoccheri, Reza Soroushmehr, Jonathan Gryak, and Kayvan Najarian. Tensor methods for clinical informatics. In *Artificial Intelligence in Healthcare and Medicine*, pages 261–281. CRC Press, 2022.
22. Brett W Bader and Tamara G Kolda. Algorithm 862: Matlab tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653, 2006.

23. Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on International Conference*, pages 37–45, 2013.
24. Lei Zhang, Fan Yang, Yimin Daniel Zhang, and Ying Julie Zhu. Road crack detection using deep convolutional neural network. In *International Conference on Signal and Information Processing*, pages 3708–3712, 2016.
25. HAMADA AHMED. Br35h :: Brain tumor detection 2020, oct 2022.
26. Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math Program*, 146(1):459–494, 2014.
27. Mark Anthony Armstrong. *Basic topology*. Springer Science & Business Media, 2013.
28. Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.
29. Qunwei Li, Yi Zhou, Yingbin Liang, and Pramod K Varshney. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *International Conference on Machine Learning*, pages 2111–2119. PMLR, 2017.