Chapter 1: Background

We first list some basic notations and concepts. Let $\mathbb{N} := \{0, 1, 2, \ldots\}$, $\mathbb{N}^* := \{1, 2, \ldots\}$, $\mathbb{R} := (-\infty, \infty)$, $\mathbb{R}_+ := [0, \infty)$, $\mathbb{R}_{++} := (0, \infty)$, and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. Let $\|\cdot\|$ be the induced norm of an inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$. Given $S \subset \mathbb{R}^n$, let $\mathring{S}$ and $\overline{S}$ denote the interior and closure of S in $\mathbb{R}^n$ respectively. Let $B(a,r)$ and $\mathring{B}(a,r)$ respectively denote the closed and open balls of center $a \in \mathbb{R}^n$ and radius $r \geqslant 0$. Given $x \in \mathbb{R}^n$, consider the distance of x to S defined by $d(x,S) := \inf\{\|x-y\| : y \in S\}$. Given a set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and $y \in \mathbb{R}^m$, let $F^{-1}(y) := \{x \in \mathbb{R}^n : F(x) \ni y\}$. Given $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, the domain, graph, and epigraph are respectively given by $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < \infty\}$, $\operatorname{graph} f := \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : \Phi(x) = t\}$, and $\operatorname{epi} f := \{(x,t) \in \mathbb{R}^{n+1} : f(x) \leqslant t\}$. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex (respectively lower semicontinuous) if $\operatorname{epi} f$ is convex (respectively closed).

1.1 Clarke subdifferential

In this section, we will review some concepts and results on generalized derivative in the sense of Clarke [38, p. 336], since we would like to also consider nonsmooth functions. We begin by recalling the definition of several normal cones [38, Definition 6.3].

Definition 1.1. Given $S \subset \mathbb{R}^n$ and $\bar{x} \in S$, the regular normal, normal, and convexified normal cones are given respectively by

$$\widehat{N}_S(\bar{x}) := \left\{ v \in \mathbb{R}^n : \limsup_{x \underset{S}{\to} \bar{x}, x \neq \bar{x}} \frac{\langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \leqslant 0 \right\},$$

$$N_S(\bar{x}) := \left\{ v \in \mathbb{R}^n : \exists (x_k, v_k) \to (\bar{x}, v),\ (x_k, v_k) \in S \times \widehat{N}_S(x_k) \right\},$$

$$\overline{N}_S(\bar{x}) := \overline{\operatorname{conv}}\, N_S(\bar{x}),$$

where $x \underset{S}{\longrightarrow} \bar{x}$ means $x \in S$ converges to $\bar{x}$.

Definition 1.2. Given $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, the Clarke subdifferential is the set-valued mapping $\partial\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined by

$$\partial f(\bar{x}) := \begin{cases} \left\{v \in \mathbb{R}^n : (v, -1) \in \overline{N}_{\mathrm{epi}\, f}((\bar{x}, f(\bar{x}))\right\} & \text{if } |f(\bar{x})| < \infty \\ \emptyset & \text{else.} \end{cases}$$

We say $x \in \mathrm{dom}(f)$ is a Clarke critical point if $0 \in \partial f(x)$ and $v \in \mathbb{R}$ is a Clarke critical value if $f(x) = v$ for some $x \in \mathrm{dom}(f)$ such that $0 \in \partial f(x)$.

Definition 1.3. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *locally Lipschitz* if for all $a \in \mathbb{R}^n$, there exist positive constants r and L such that

$$\forall x, y \in B(a, r), \quad \|f(x) - f(y)\| \leqslant L\|x - y\|.$$

Notice that for a locally Lipschitz function, by [39, Theorem 3.2], the derivative exists almost everywhere. It is also well known that for any locally Lipschitz function f and any $x \in \mathbb{R}^n$, the Clarke subdifferential $\partial f(x)$ is a nonempty, convex, and compact set [40, Proposition 2.1.2(a)].

1.2 Semialgebraic functions

Definition 1.4. [41, 42] A subset S of $\mathbb{R}^n$ is *semialgebraic* if it is a finite union of sets of the form $\{x \in \mathbb{R}^n : p_i(x) = 0,\ i = 1, \dots, k;\ p_i(x) > 0,\ i = k + 1, \dots, m\}$ where $p_1, \dots, p_m$ are polynomials defined from $\mathbb{R}^n$ to $\mathbb{R}$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is semialgebraic if its graph, that is to say $\{(x, t) \in \mathbb{R}^{n+1} : f(x) = t\}$, is a semialgebraic set.

Next we recall several useful properties of semialgebraic functions. They will be frequently used in later chapters.

Lemma 1.1 (semialgebraic Morse-Sard theorem [43])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and semialgebraic. Then f has finitely many critical values.*

Theorem 1.1 (Kurdyka-Łojasiewicz inequality [44, 43]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and semialgebraic. Let X be a bounded subset of $\mathbb{R}^n$ and $v \in \mathbb{R}$ be a critical value of f in $\overline{X}$. There exists $\rho > 0$ and a strictly increasing continuous semialgebraic function $\psi : [0, \rho) \to [0, \infty)$ which belongs to $C^1((0, \rho))$ with $\psi(0) = 0$ such that*

$$\forall x \in X, \quad |f(x) - v| \in (0, \rho) \quad \Longrightarrow \quad d(0, \partial(\psi \circ |f - v|)(x)) \geqslant 1. \tag{1.1}$$

Proposition 1.1 (Uniform Kurdyka-Łojasiewicz inequality [45, Proposition 5]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and semialgebraic. Let X be a bounded subset of $\mathbb{R}^n$ and V be the set of critical values of f in $\overline{X}$ if it is non-empty, otherwise $V := \{0\}$. There exists a concave semialgebraic diffeomorphism $\psi : [0, \infty) \to [0, \infty)$ such that*

$$\forall x \in X \setminus (\partial \tilde{f})^{-1}(0), \quad d(0, \partial(\psi \circ \tilde{f})(x)) \geqslant 1, \tag{1.2}$$

where $\tilde{f}(x) := d(f(x), V)$ for all $x \in \mathbb{R}^n$.

1.3 Subgradient trajectories

In this section, we will introduce some basic concepts and fundamental properties related to subgradient trajectories.

Definition 1.5. [46, Definition 1 p. 12] Given two real numbers $a < b$, a function $x : [a, b] \to \mathbb{R}^n$ is *absolutely continuous* if for all $\epsilon > 0$, there exists $\delta > 0$ such that, for any finite collection of disjoint subintervals $[a_1, b_1], \ldots, [a_m, b_m]$ of $[a, b]$ such that $\sum_{i=1}^m (b_i - a_i) \leqslant \delta$, we have $\sum_{i=1}^m \|x(b_i) - x(a_i)\| \leqslant \epsilon$.

By virtue of [47, Theorem 20.8], $x : [a, b] \to \mathbb{R}^n$ is absolutely continuous if and only if it is differentiable almost everywhere on (a, b), its derivative x' is Lebesgue integrable, and $x(t) - x(a) = \int_a^t x'(\tau) d\tau$ for all $t \in [a, b]$. Given a non-compact interval I of $\mathbb{R}$, $x : I \to \mathbb{R}^n$ is absolutely continuous if it is absolutely continuous on all compact subintervals of I.

Definition 1.6. An absolutely continuous function $x : [0, \infty) \to \mathbb{R}^n$ is called a *subgradient trajectory* of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ starting at $x_0 \in \operatorname{dom}(\partial f)$ if it satisfies the following differential inclusion with initial condition:

$$x'(t) \in -\partial f(x(t)), \quad \text{for almost every } t \geqslant 0, \quad x(0) = x_0, \tag{1.3}$$

where "almost every" means all elements except for those in a set of zero measure.

However, a subgradient trajectory may not always exist for arbitrary f, even if f is a smooth function. Let $f(x) = -\frac{1}{3}x^3$ and $x_0 = 1$, then it is easy to see $x(t) = \frac{1}{1-t}$ is the unique solution for $t \in [0, 1)$ and it cannot be extended to an absolutely continuous function on $[0, \infty)$ due to the singularity at $t = 1$. In this case, one would seek a family of functions including many loss functions arising in applications that guarantee the existence of a subgradient trajectory.

We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is *bounded below* if $\inf_{\mathbb{R}^n} f = c > -\infty$. It was shown in [48, Theorem 3.2] that a primal lower nice function bounded below by a linear function suffices. However, in general it is not easy to check whether those nonconvex functions in statistical learning problems are primal lower nice. For easily checkable conditions, the following result generalized from [49, Proposition 2.3] for differentiable functions tells us that a locally Lipschitz function bounded below also suffices.

Proposition 1.2. *If $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz and bounded below, then there exists a subgradient trajectory of f starting at arbitrary $x_0 \in \mathbb{R}^n$.*

Proof. For a fixed real number $\tau > 0$, define a sequence x_k^τ recurrently by letting $x_0^\tau := x_0$ and

$$x_{k+1}^\tau \in \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{\|x - x_k^\tau\|^2}{2\tau} \right\}, \quad \forall k \in \mathbb{N}.$$

A solution exists because f is bounded below and the objective function is coercive. Any solution satisfies

$$v_{k+1}^\tau := \frac{x_{k+1}^\tau - x_k^\tau}{\tau} \in -\partial f(x_{k+1}^\tau), \quad \forall k \in \mathbb{N}.$$

Define two functions $x^\tau, \tilde{x}^\tau : \mathbb{R}_+ \to \mathbb{R}^n$ where $\mathbb{R}_+ := [0, \infty)$ by

$$x^\tau(t) := x_{k+1}^\tau, \quad \tilde{x}^\tau(t) := x_k^\tau + (t - k\tau)v_{k+1}^\tau, \quad \forall t \in (k\tau, (k+1)\tau]$$

for all $k \in \mathbb{N}$, with initial condition $x^\tau(0) = \tilde{x}^\tau(0) = x_0$. Note that $\tilde{x}^\tau$ is absolutely continuous because it is piecewise affine. On the contrary, x^τ is not continuous. Also, define $v^\tau : \mathbb{R}_+ \to \mathbb{R}^n$ by

$$v^\tau(t) := v_{k+1}^\tau, \quad \forall t \in (k\tau, (k+1)\tau], \quad \forall k \in \mathbb{N},$$

and choose $v^\tau(0) \in -\partial f(x_0)$. Since $(\tilde{x}^\tau)' = v^\tau$ on $(k\tau, (k+1)\tau)$ for all $k \in \mathbb{N}$, and $v^\tau(t) \in -\partial f(x^\tau(t))$ for all $t \geqslant 0$, we conclude that $(\tilde{x}^\tau)'(t) \in -\partial f(x^\tau(t))$ for almost every $t \in \mathbb{R}_+$. By optimality of x_{k+1}^τ, we have

$$f(x_{k+1}^\tau) + \frac{\|x_{k+1}^\tau - x_k^\tau\|^2}{2\tau} \leqslant f(x_k^\tau), \quad \forall k \in \mathbb{N}.$$

For any $l \in \mathbb{N}$, we have

$$\sum_{k=0}^{l} \frac{\|x_{k+1}^\tau - x_k^\tau\|^2}{2\tau} \leqslant f(x_0^\tau) - f(x_{l+1}^\tau) \leqslant f(x_0) - \inf_{\mathbb{R}^n} f =: C < \infty$$

since f is bounded below. Observe that

$$\sum_{k=0}^{l} \frac{\|x_{k+1}^\tau - x_k^\tau\|^2}{2\tau} = \sum_{k=0}^{l} \frac{\tau}{2}\|v_{k+1}^\tau\|^2 = \frac{1}{2}\sum_{k=0}^{l} \int_{k\tau}^{(k+1)\tau} \|(\tilde{x}^\tau)'(t)\|^2 \, dt.$$

Fix $T \geqslant 0$ from now on. From the above, we have

$$\int_0^T \|(\tilde{x}^\tau)'(t)\|^2 \, dt \leqslant \sum_{k=0}^{\lfloor T/\tau \rfloor} \int_{k\tau}^{(k+1)\tau} \|(\tilde{x}^\tau)'(t)\|^2 \, dt \leqslant 2C. \tag{1.4}$$

Since $\tilde{x}^\tau$ is absolutely continuous, for any $s, t \in [0, T]$ we have

$$\|\tilde{x}^\tau(t) - \tilde{x}^\tau(s)\| = \left\| \int_s^t (\tilde{x}^\tau)'(u)\ du \right\| \tag{1.5a}$$

$$\leqslant \left(\int_0^T \|(\tilde{x}^\tau)'(t)\|^2\ dt \right)^{1/2} |t-s|^{1/2} \leqslant \sqrt{2C}|t-s|^{1/2} \tag{1.5b}$$

where we use the Cauchy-Schwarz inequality. Now one can see $(\tilde{x}^\tau)_{\tau>0}$ is a family of uniformly bounded and equicontinuous functions on the compact interval $[0, T]$. Therefore, by Arzelà-Ascoli theorem [50, Theorem 7.25], there exists a sequence of positive reals $(\tau_k)_{k\in\mathbb{N}}$ such that $\tau_k \to 0$ and $\tilde{x}^{\tau_k} \to x^*$ uniformly on $[0, T]$ as $k \to \infty$. For all $k \in \mathbb{N}$ and $t \in (k\tau, (k+1)\tau]$, we have $\tilde{x}^\tau((k+1)\tau) = x_k^\tau + \tau v_{k+1}^\tau = x_{k+1}^\tau = x^\tau(t)$. Thus $\|\tilde{x}^\tau(t) - x^\tau(t)\| = \|\tilde{x}^\tau(t) - \tilde{x}^\tau((k+1)\tau)\| \leqslant \sqrt{2C}\tau^{1/2}$ for all $t \in [0, T]$ where the inequality is due to (1.5) (take $s := (k+1)\tau$). Combined with the fact that $\tilde{x}^{\tau_k} \to x^*$ uniformly on $[0, T]$, one can see that $x^{\tau_k} \to x^*$ uniformly on $[0, T]$. Since (1.4) implies that $((\tilde{x}^{\tau_k})')_{k\in\mathbb{N}}$ is a bounded sequence in $L^2([0, T], \mathbb{R}^n)$, there exists a subsequence $(\tau_{k_j})_{j\in\mathbb{N}}$ such that $(\tilde{x}^{\tau_{k_j}})' \to v^*$ weakly in $L^1([0, T], \mathbb{R}^n)$ as $j \to \infty$ by [51, Corollary 14 p. 413]. Since $\tilde{x}^{\tau_{k_j}}$ is absolutely continuous, for all $t \in [0, T]$, we have

$$\tilde{x}^{\tau_{k_j}}(t) - \tilde{x}^{\tau_{k_j}}(0) = \int_0^t (\tilde{x}^{\tau_{k_j}})'(u)\ du.$$

Take $j \to \infty$ on both sides, we have

$$x^*(t) - x^*(0) = \int_0^t v^*(u)\ du,$$

where the convergence of the integral relies on the fact that the constant functions equal to the canonical basis of $\mathbb{R}^n$ lie in $L^\infty([0, T], \mathbb{R}^n)$. Thus, x^* is absolutely continuous and $(x^*)'(t) = v^*(t)$ for almost every $t \in [0, T]$. Recall that for all $k \in \mathbb{N}$, it holds for almost every $t \in [0, T]$ that

$$(\tilde{x}^{\tau_k})'(t) = v^{\tau_k}(t) \in -\partial f(x^{\tau_k}(t)).$$

Since f is locally Lipschitz, the set-valued function $-\partial f$ is upper semicontinuous [40, 2.1.5 Proposition (d) p. 29] with nonempty compact values [40, 2.1.2 Proposition (a) p. 27], hence proper upper hemicontinuous [46, Proposition 1 p. 60]. In addition, $x^{\tau_k} \to x^*$ uniformly on $[0,T]$ and $(\tilde{x}^{\tau_k})' \to (x^*)'$ weakly in $L^1([0,T],\mathbb{R}^n)$. Therefore, $(x^*)'(t) \in -\partial f(x^*(t))$ for almost all $t \in [0,T]$ by [46, Theorem 1 p. 60][1]. The initial condition also holds since $\tilde{x}^{\tau}(0) = x_0$ for all $\tau > 0$.

We have proved that for any initial point x_0, there exists $x^* : [0,T] \to \mathbb{R}^n$ such that $(x^*)'(t) = -\partial f(x^*(t))$ holds for almost every $t \in [0,T]$ with any $T > 0$. Since T is independent of x_0, by setting $T = 1$, there exists a sequence of absolutely continuous functions $(x_k)_{k\in\mathbb{N}}$ such that

$$x_k'(t) \in -\partial f(x_k(t)), \quad \text{for a.e. } t \in [0,1], \quad x_k(0) = x_{k-1}(1),$$

for all $k \in \mathbb{N}$ where $x_{-1}(0) = x_0$. Therefore, the desired function $x : [0,\infty) \to \mathbb{R}^n$ can be defined in a piecewise fashion by

$$x(t) := x_k(t-k), \quad t \in [k, k+1), \quad \forall k \in \mathbb{N}.$$

By construction, x is absolutely continuous on any compact interval $[a,b] \subset [0,\infty)$. □

We remark here that with Proposition 1.2, one can recover Ekeland's variational principle [52, Corollary 2.3] [53, Corollary] for locally Lipschitz lower bounded functions with a chain rule (see [54, Theorem 3.1] for an extension to lower semi-continuous lower bounded functions). Indeed, Proposition 1.2 implies that for all $\epsilon > 0$, there exists $(x,s) \in \operatorname{graph} \partial f$ such that $f(x) \leqslant \inf f + \epsilon$ and $\|s\| \leqslant \epsilon$.[2] Note that Proposition 1.2 only guarantees the existence of a solution to (1.3) for all $t \geqslant 0$, but the solution $x(t)$ could go to infinity as $t \to \infty$. This motivates the following definition.

Definition 1.7. A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ has *bounded subgradient trajectories* if for any $x_0 \in \operatorname{dom}(f)$, there exists a constant $r > 0$, such that for any subgradient trajectory x of f starting at x_0,

[1] In the theorem we take $F := -\partial f$, $X = Y := \mathbb{R}^n$, and $I := [0,T]$.

[2] This follows from the formula $f(x(t)) - \inf f \geqslant \int_t^\infty d(0, \partial f(x(\tau)))^2 d\tau$ where $d(x,X) := \inf_{y\in X} \|x-y\|$ (see [22, Lemma 5.2] and [55, Proposition 4.10]).

we have $\|x(t)\| \leqslant r$ for all $t \geqslant 0$.

Finally, notice that when f is continuously differentiable, by [40, Proposition 2.2.4], (1.3) reduces to the classical Cauchy problem of differential equation

$$x'(t) = -\nabla f(x(t)), \quad \text{for all } t \geqslant 0, \quad x(0) = x_0.$$

and subgradient trajectory reduces to gradient trajectory by imposing x to be continuously differentiable. Recall the descent property of gradient trajectories [56, Proposition 17.1.1], i.e., $f \circ x$ is a decreasing function for any gradient trajectory x of f. We want this nice property to hold even in a more general case. We adopt the notion of chain rule in [22, Definition 5.1]. Note that functions admitting a chain rule are also referred to as path differentiable [57, Definition 3].

Definition 1.8. Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be locally Lipschitz over $\operatorname{dom}(f)$. We say f *admits a chain rule* if for any absolutely continuous function $x : [0, \infty) \to \mathbb{R}^n$, we have

$$(f \circ x)'(t) = \langle v, x'(t) \rangle, \quad \forall\, v \in \partial f(x(t)),$$

for almost every $t \in [0, \infty)$.

Thus, for any locally Lipschitz function that admits a chain rule, by [22, Lemma 5.2], the function value is always decreasing in time along the subgradient trajectory. A detailed discussion on what class of functions admits a chain rule can be found in [57]. Note that general Lipschitz functions are far from admitting a chain rule since they generically have a maximal Clarke subdifferential [34, 58, 59].

Chapter 2: Typical models with bounded subgradient trajectories

In this chapter, we focus on the main condition of this thesis: the boundedness of subgradient trajectories. We formalize this notion and show that it holds in a wide range of models commonly encountered in data science. The results on bounded subgradient trajectories form a foundational tool for the convergence and landscape analyses developed in the subsequent chapters. To help with exposition, we proceed in order of increasing complexity, starting with convex models, followed by nonconvex coercive problems, and concluding with nonconvex noncoercive settings.

2.1 Convex model

Theorem 2.1. *For a convex proper l.s.c. function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, suppose $x^* \in \mathrm{dom}(f)$ is a minimizer, then f has bounded subgradient trajectories.*

Proof. By [56, Theorem 17.2.2], for any initial point $x_0 \in \mathrm{dom}(f)$, there exists a unique subgradient trajectory $x : \mathbb{R}_+ \to \mathbb{R}^n$. It is easy to see $t \mapsto \|x(t) - x^*\|^2$ is absolutely continuous and the chain rule can be applied so that for a.e. $t \in \mathbb{R}_+$,

$$\frac{d}{dt}\|x(t) - x^*\|^2 = 2\langle x'(t), x(t) - x^*\rangle = -2\langle g_t, x(t) - x^*\rangle \le -2(f(x(t)) - f(x^*)) \leqslant 0,$$

where $g_t \in \partial f(x(t))$. Thus, $\|x(t) - x^*\| \leqslant \|x_0 - x^*\|$ for all $t \in \mathbb{R}_+$, which implies $\|x(t)\| \leqslant \|x_0 - x^*\| + \|x^*\|$. Therefore, f has bounded subgradient trajectories. $\square$

Example 2.1. One of the most fundamental and widely used convex models in data science is linear regression. A general form of linear regression can be written as a special case of a linear feedforward neural network without hidden layers [60], given by

$$f(W) := \sum_{i=1}^{m} \|Wx_i - y_i\|^2, \tag{2.1}$$

where $W \in \mathbb{R}^{n\times r}$ is the parameter matrix, $x_i \in \mathbb{R}^r$ are the input vectors, and $y_i \in \mathbb{R}^n$ are the target outputs for $i = 1, \ldots, m$. The function f is convex, as it is a sum of convex quadratic functions. More importantly, f always admits a minimizer for any given dataset $\{(x_i, y_i)\}_{i=1}^m$, since the associated optimality condition is the normal equation, whose solution exists even if the original system $Wx_i = y_i$ is inconsistent. Consequently, Theorem 2.1 is applicable in this setting.

2.2 Nonconvex coercive model

Recall that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be *coercive* if $f(x) \to \infty$ as $\|x\| \to \infty$. Coercive functions arise naturally in data science, particularly when regularization is used to control model complexity. Intuitively, the coercive shape of the function prevents subgradient trajectories from escaping to infinity, as long as the function value decreases along the trajectory.

Proposition 2.1. *For a locally Lipschitz continuous function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, if it is coercive and satisfies chain rule, then f has bounded subgradient trajectories.*

Proof. The existence of subgradient trajectories is ensured by Proposition 1.2. For any absolutely continuous x, $f \circ x$ is also absolutely continuous and $(f \circ x)'(t) = \langle v, x'(t)\rangle$, for all $v \in \partial f(x(t))$. Since $x'(t) \in -\partial f(x(t))$, by taking $v = -x'(t)$, we obtain $(f \circ x)'(t) = -\|x'(t)\|^2 < 0$. Thus, $f \circ x$ is decreasing and $f(x(t)) \le f(x_0)$ for all $t \in \mathbb{R}_+$. By coercivity of f, x is bounded and f has bounded subgradient trajectories. □

Example 2.2. Coercive functions frequently appear in data science models through regularization. One classical example is a ReLU neural network [19] with ℓ_2-regularization:

$$\mathcal{L}(W) := \sum_{i=1}^{m} \|W_L\sigma(W_{L-1}\cdots\sigma(W_1x_i)\cdots) - y_i\|^2 + \lambda \sum_{\ell=1}^{L} \|W_\ell\|_F^2, \tag{2.2}$$

where $\sigma(z) = \max\{0, z\}$ is the ReLU activation and $\lambda > 0$. The regularization renders the function coercive, ensuring the boundedness of gradient or subgradient trajectories during training.

2.3 Nonconvex noncoercive model

2.3.1 Phase Retrieval

The problem of solving systems of quadratic equations of the form $y_i = \langle a_i, x^\natural \rangle^2$, $1 \leq i \leq m$, has applications in numerous contexts. One of the most classical applications is the so-called phase retrieval problem. This problem has attracted high interest due to its broad applications in X-ray crystallography [61], microscopy [62], astronomy [63] and optical imaging [64]. Here we consider a slightly more general formulation

$$\min_x \quad f(x) := \sum_{i=1}^m \left(\langle A_i x, x \rangle - y_i \right)^2 . \tag{2.3}$$

Proposition 2.2. *Let $A_i \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite and $y_i \in \mathbb{R}$ for all $i = 1, \ldots, m$. Then* (2.3) *has bounded subgradient trajectories.*

Proof. Since A_i is real symmetric and positive semidefinite for all $i = 1, \ldots, m$, by orthogonal decomposition, we can write $A_i = \sum_{j=1}^r \lambda_{ij} v_{ij} v_{ij}^{\mathrm{T}}$, where $(v_{ij})_{j=1}^r$ are orthonormal, for some $1 \leq r \leq n$ and $\lambda_{ij} > 0$ for all $j = 1, \ldots, r$. Define

$$V := \mathrm{Span}\{v_{ij} : i = 1, \ldots, m,\ j = 1, \ldots, r\}.$$

Notice that

$$\nabla f(x) = 4 \sum_{i=1}^m \left(\langle A_i x, x \rangle - b_i \right) A_i x = 4 \sum_{i=1}^m \sum_{j=1}^r \lambda_{ij} \langle v_{ij}, x \rangle \left(\langle A_i x, x \rangle - b_i \right) v_{ij} \in V.$$

Therefore, $\nabla f(x) \in V$ for all $x \in \mathbb{R}^{2n}$. Denote $V^\perp$ as the orthogonal complement of the subspace V, then for any given initial point x_0, the subgradient trajectory $x(\cdot)$ of f can be decomposed as $x(t) =$

$x_V(t) + x_{V^\perp}(t)$, where $x_V(t) \in V$ and $x_{V^\perp}(t) \in V^\perp$ for all $t \geq 0$. Note that $x'(t) = -\nabla f(x(t)) \in V$, thus $x_{V^\perp}(t) \equiv x_{V^\perp}(0)$ and we write $x(t) = x_V(t) + x_{V^\perp}(0)$ for all $t \geq 0$. Since $f(x)$ is a decreasing function over $t \geq 0$,

$$
\begin{aligned}
\sum_{j=1}^{r} \lambda_{ij} \langle v_{ij}, x(t) \rangle^2 = |\langle A_i x(t), x(t) \rangle| &\leq \sqrt{2(\langle A_i x(t), x(t) \rangle - b_i)^2 + 2b_i^2} \\
&\leq \sqrt{2f(x(t)) + 2b_i^2} \leq \sqrt{2f(x_0) + 2b_i^2}.
\end{aligned}
$$

Recall that $\lambda_{ij} > 0$ for all $j = 1, \ldots, r$, hence $\langle v_{ij}, x(t) \rangle$ is bounded over $t \geq 0$ and so is $\langle v_{ij}, x_V(t) \rangle$. As $\mathrm{Span}\{v_{ij} : i = 1, \ldots, m,\ j = 1, \ldots, r\} = V$, we can extract a basis of vectors v_{ij} to form a basis of V, and denote this basis as $\{u_\ell : \ell = 1, \ldots, d\}$. Then one can write $x_V(t) = \sum_{\ell=1}^{d} \zeta_\ell(t) u_\ell$. Notice that for each $\ell = 1, \ldots, d$, there must exist (i_ℓ, j_ℓ) such that $v_{i_\ell j_\ell} = u_\ell$. Thus,

$$
\|x_V(t)\|^2 = \sum_{\ell=1}^{d} \zeta_\ell(t)^2 = \sum_{\ell=1}^{d} \langle u_\ell, x_V(t) \rangle^2 = \sum_{\ell=1}^{d} \langle v_{i_\ell j_\ell}, x_V(t) \rangle^2
$$

is bounded over $t \geq 0$. Finally, $x(t) = x_V(t) + x_{V^\perp}(0)$ is bounded over $t \geq 0$. □

2.3.2 Asymmetric matrix sensing

Matrix sensing is a widely used model in computer vision and statistics; see for instance [65, 66]. Given $r \geqslant 1$, the goal is to recover an unknown target matrix $M \in \mathbb{R}^{n_1 \times n_2}$ of rank less than or equal to r from a set of linear measurements $b_i = \langle A_i, M \rangle_F$, where $A_i \in \mathbb{R}^{n_1 \times n_2}$ for $i = 1, \ldots, m$ are sensing matrices and $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. In order to do so, we minimize the mean square loss

$$
f(X, Y) := \frac{1}{2m} \sum_{i=1}^{m} (\langle A_i, XY^{\mathrm{T}} \rangle_F - b_i)^2. \tag{2.4}
$$

where $X \in \mathbb{R}^{n_1 \times r}$ and $Y \in \mathbb{R}^{n_2 \times r}$.

A sufficient condition is to require the sensing matrices to be *lower bounded*, i.e., there exists a constant $c > 0$ such that for any matrix $\widetilde{M} \in \mathbb{R}^{n_1 \times n_2}$ with $\mathrm{rank}(\widetilde{M}) \leqslant r$,

$$\frac{1}{m}\sum_{i=1}^{m}\langle A_i, \widetilde{M}\rangle_F^2 \geqslant c\|\widetilde{M}\|_F^2.$$

A special case of lower bounded sensing matrices is to take each A_i be a matrix unit $E_{j,k}$, i.e., a matrix with only one nonzero entry at j-th row and k-th column with value 1. For example, we can let $m = n_1 n_2$, and $A_1 = E_{1,1}, A_2 = E_{1,2}, \ldots, A_{n_1 n_2} = E_{n_1,n_2}$. In this case, $\sum_{i=1}^{m}\langle A_i, \widetilde{M}\rangle_F^2 = \|\widetilde{M}\|_F^2$, so the above condition holds. In fact, under such setting, the objective function is equivalent to simple matrix factorization $f(X,Y) = \|XY^{\mathrm{T}} - M\|_F^2$.

Proposition 2.3. *Matrix sensing with loss function* (2.4) *and lower bounded sensing matrices has bounded gradient trajectories.*

Proof. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a gradient trajectory for any initial point. The gradient trajectories of f satisfy the initial value problem

$$\begin{aligned}
\dot{X} &= -\frac{1}{m}\sum_{i=1}^{m}(\langle A_i, XY^{\mathrm{T}}\rangle_F - b_i)A_i Y,\\
\dot{Y} &= -\frac{1}{m}\sum_{i=1}^{m}(\langle A_i, XY^{\mathrm{T}}\rangle_F - b_i)A_i^{\mathrm{T}} X,\\
X(0) &= X_0, \quad Y(0) = Y_0.
\end{aligned}$$

Notice that $\dot{X}^{\mathrm{T}}X = Y^{\mathrm{T}}\dot{Y}$ and $X^{\mathrm{T}}\dot{X} = \dot{Y}^{\mathrm{T}}Y$, so

$$\frac{d}{dt}(X^{\mathrm{T}}X - Y^{\mathrm{T}}Y) = \dot{X}^{\mathrm{T}}X + X^{\mathrm{T}}\dot{X} - \dot{Y}^{\mathrm{T}}Y - Y^{\mathrm{T}}\dot{Y} = 0.$$

This implies that $X^{\mathrm{T}}X - Y^{\mathrm{T}}Y = C$ where $C \in \mathbb{R}^{r\times r}$ is a constant. Since the function value is decreasing along gradient trajectories [22, Lemma 5.2], there exists a constant c_1 such that $f(X(t), Y(t)) \leqslant c_1$ for all $t \geqslant 0$. Combined with the assumption that sensing matrices are lower bounded, there exist constants c and c_2 such that

$$c\|XY^{\mathrm{T}}\|_F^2 \leqslant \frac{1}{m}\sum_{i=1}^{m}\langle A_i, XY^{\mathrm{T}}\rangle_F^2 \leqslant \frac{1}{m}\sum_{i=1}^{m}[2(\langle A_i, XY^{\mathrm{T}}\rangle_F - b_i)^2 + 2b_i^2]$$
$$= 2f(X,Y) + \frac{2}{m}\sum_{i=1}^{m} b_i^2 \leqslant 2c_1 + \frac{2}{m}\sum_{i=1}^{m} b_i^2 =: c_2.$$

We have $\|XY^{\mathrm{T}}\|_F^2 \leqslant c_3 := c_2/c$. Notice that

$$\|X^{\mathrm{T}}X\|_F^2 + \|Y^{\mathrm{T}}Y\|_F^2 = \|X^{\mathrm{T}}X - Y^{\mathrm{T}}Y\|_F^2 + 2\|XY^{\mathrm{T}}\|_F^2 \leqslant \|C\|_F^2 + 2c_3.$$

Define the constant $c_4 := 2c_3 + \|C\|_F^2$. By the Cauchy-Schwarz inequality,

$$\|X\|_F^4 + \|Y\|_F^4 \leqslant \mathrm{rank}(X)\|X^{\mathrm{T}}X\|_F^2 + \mathrm{rank}(Y)\|Y^{\mathrm{T}}Y\|_F^2 \leqslant (n_1 + n_2 + r)c_4.$$

Thus, X and Y are bounded. □

2.3.3 Nonsmooth matrix factorization

In this subsection, we consider the application of Theorem 3.1 in a nonsmooth setting, namely, the nonsmooth matrix factorization problem. We consider minimizing the loss function

$$f(X,Y) := \|XY^{\mathrm{T}} - M\|_1, \tag{2.5}$$

where $X \in \mathbb{R}^{m\times r}$, $Y \in \mathbb{R}^{n\times r}$ are decision variables and $M \in \mathbb{R}^{m\times n}$ is the given data matrix. Here $\|A\|_1 := \sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|$ for any $A \in \mathbb{R}^{m\times n}$. In robust principal component analysis (PCA) problem with sparse noise, (2.5) is usually used as a surrogate function for the original ℓ_0-norm formulation; see [67, 68].

To verify (2.5) has bounded subgradient trajectories, we discover that the auto-balancing property in [69, Theorem 2.2] also holds for nonsmooth matrix factorization. The result can be summarized in the following proposition.

Proposition 2.4. *Nonsmooth matrix factorization with loss function* (2.5) *has bounded subgradient trajectories.*

Proof. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a subgradient trajectory for any initial point. Let $(X_0, Y_0) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$. Consider an absolutely continuous function $Z : [0,\infty) \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$ such that

$$Z'(t) \in -\partial f(Z(t)), \quad \text{for almost every } t \geqslant 0, \quad \text{and } Z(0) = (X_0, Y_0).$$

By [40, Theorem 2.3.10],

$$\partial f(X,Y) = \left\{ \left. \begin{pmatrix} \Lambda Y \\ \Lambda^T X \end{pmatrix} \right| \Lambda \in \operatorname{sign}(XY^T - M) \right\}$$

where sign is an element-wise operation mapping each entry of a matrix to a real number in $[-1, 1]$ such that

$$\operatorname{sign}(x) := \begin{cases} -1 & \text{if } x < 0, \\ \big[-1, 1\big] & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Hence, with $Z =: (X, Y)$, for almost every $t \geqslant 0$ we have

$$X'(t) = -\Lambda(t)Y(t), \quad Y'(t) = -\Lambda(t)^T X(t), \tag{2.6a}$$

$$\Lambda(t) \in \operatorname{sign}(X(t)Y(t)^T - M). \tag{2.6b}$$

Consider $\phi : [0,\infty) \to \mathbb{R}$ defined by $\phi(t) := X(t)^T X(t) - Y(t)^T Y(t)$. By taking derivative, we have

$$\phi'(t) = X'(t)^T X(t) + X(t)^T X'(t) - Y'(t)^T Y(t) - Y(t)^T Y'(t). \tag{2.7}$$

Combining (2.6a) and (2.7), we have

$$\begin{aligned}\phi'(t) = &-Y(t)^T\Lambda(t)^T X(t) - X(t)^T\Lambda(t)Y(t)\\ &+ X(t)^T\Lambda(t)Y(t) + Y(t)^T\Lambda(t)^T X(t) = 0.\end{aligned}$$

Hence the continuous function ϕ is constant on $[0, \infty)$. Also, we have

$$\begin{aligned}\|X^TX - Y^TY\|_F^2 &= \|X^TX\|_F^2 + \|Y^TY\|_F^2 - 2\langle X^TX, Y^TY\rangle_F\\ &= \|X^TX\|_F^2 + \|Y^TY\|_F^2 - 2\|XY^T\|_F^2\\ &\geqslant \|X^TX\|_2^2 + \|Y^TY\|_2^2 - 2\|XY^T\|_F^2\\ &= \|X\|_2^4 + \|Y\|_2^4 - 2\|XY^T\|_F^2\\ &\geqslant \|X\|_2^4 + \|Y\|_2^4 - 2mn\|XY^T\|_1^2\\ &\geqslant \|X\|_2^4 + \|Y\|_2^4 - 2mn(\|XY^T - M\|_1 + \|M\|_1)^2.\end{aligned}$$

Here $\|\cdot\|_2$ denotes the spectral norm. Therefore, for all $t \geqslant 0$, we have

$$\begin{aligned}\|X(t)\|_2^4 + \|Y(t)\|_2^4 \leqslant &\|X(t)^TX(t) - Y(t)^TY(t)\|_F^2\\ &+ 2mn(\|X(t)Y(t)^T - M\|_1 + \|M\|_1)^2\\ \leqslant &\|X_0^TX_0 - Y_0^TY_0\|_F^2\\ &+ 2mn(\|X_0Y_0^T - M\|_1 + \|M\|_1)^2.\end{aligned}$$

$\square$

2.3.4 Nonnegative matrix factorization

Let $A \in \mathbb{R}^{m\times n}$ and $p \geqslant 1$, we define the p-norm of A by

$$\|A\|_p := \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|^p\right)^{1/p}.$$

Let $M \in \mathbb{R}^{m\times n}$, in nonnegative ℓ_p matrix factorization, we aim to minimize

$$\begin{array}{rccc} f: & \mathbb{R}^{m\times r}\times\mathbb{R}^{n\times r} & \longrightarrow & \mathbb{R} \\ & (X,Y) & \longmapsto & \|XY^T - M\|_p^p, \end{array}$$

subject to $(X,Y) \in \mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{n\times r} =: C$. Without the nonnegativity constraints, ℓ_p matrix factorization was studied in [70], and was shown to be robust against outlines when $p < 2$. Note that when $p = 2$, the above example reduces to the problem of nonnegative matrix factorization (NMF) [71, 72, 18, 73].

Next we verify the boundedness of the subgradient trajectories. We begin with some notations. Let $A, B \in \mathbb{R}^{m\times n}$, we denote by $A \odot B \in \mathbb{R}^{m\times n}$ their Hadamard product, whose (i, j)-entry is given by $(A \odot B)_{ij} := A_{ij}B_{ij}$. Let $p \geqslant 0$, we denote by $|A|^{\circ p} \in \mathbb{R}^{m\times n}$ the matrix obtained by taking absolute value and then raising to pth power for each element in A, namely, $(|A|^{\circ p})_{ij} := |A_{ij}|^p$. We use the convention that $0^0 = 1$. Let sign denote the element-wise operation that maps each entry of a matrix to a subset of $[-1, 1]$ such that

$$\operatorname{sign}(t) := \begin{cases} -1 & \text{if } t < 0, \\ \big[-1, 1\big] & \text{if } t = 0, \\ 1 & \text{if } t > 0. \end{cases}$$

By [40, 2.3.10 Theorem (Chain Rule II)], we have

$$\partial f(X,Y) = \left\{ \begin{pmatrix} \left(\Lambda \odot |XY^T - M|^{\circ(p-1)}\right) Y \\ \left(\Lambda \odot |XY^T - M|^{\circ(p-1)}\right)^T X \end{pmatrix} : \Lambda \in \operatorname{sign}(XY^T - M) \right\}.$$

We next study the solutions to (2.8), which is an equivalent characterization of the subgradient trajectories of Φ by [74, Theorem 2.3(b)].

Lemma 2.1. *Given $X_0 \in \mathbb{R}_+^{m\times r}$, $Y_0 \in \mathbb{R}_+^{n\times r}$, and $M \in \mathbb{R}^{m\times n}$, there exist $c_1, \ldots, c_r \in \mathbb{R}$ such that any solution $(X,Y,\Lambda) : \mathbb{R}_+ \to \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times n}$ to*

$$\begin{cases} X' = P_{T_{\mathbb{R}_+^{m\times r}}(X)} \left(-\left(\Lambda \odot |XY^T - M|^{\circ(p-1)}\right) Y \right) \\ Y' = P_{T_{\mathbb{R}_+^{n\times r}}(Y)} \left(-\left(\Lambda \odot |XY^T - M|^{\circ(p-1)}\right)^T X \right) \\ \Lambda \in \operatorname{sign}(XY^T - M), \;\; X(0) = X_0, \;\; Y(0) = Y_0 \end{cases} \tag{2.8}$$

satisfies that

$$\sum_{i=1}^m X_{ik}(t)^2 - \sum_{j=1}^n Y_{jk}(t)^2 = c_k, \quad \forall t \in \mathbb{R}_+, \quad \forall k \in [\![1,r]\!].$$

Proof. For all $t \in \mathbb{R}_+$, let

$$L(t) := X'(t)^T X(t) + X(t)^T X'(t), \quad R(t) := Y'(t)^T Y(t) + Y^T(t) Y'(t)$$
$$\text{and } E(t) := -\Lambda(t) \odot |X(t)Y(t)^T - M|^{\circ(p-1)}.$$

For $k \in [\![1,r]\!]$ and $t \in \mathbb{R}_+$, define the following index sets

$$I_k^X(t) := \left\{ i \in [\![1,m]\!] : X'_{ik}(t) \neq \sum_{j=1}^n E_{ij}(t) Y_{jk}(t) \right\}$$

and

$$I_k^Y(t) := \left\{ j \in [\![1,n]\!] : Y'_{jk}(t) \neq \sum_{i=1}^m E_{ij}(t) X_{ik}(t) \right\}.$$

Consider the k-th diagonal element of $L(t)$ for $k \in [\![1, r]\!]$, we have that

$$L_{kk}(t) = 2\sum_{i=1}^{m} X_{ik}(t)X'_{ik}(t) = 2\sum_{i \in I_k^X(t)} X_{ik}(t)X'_{ik}(t) + 2\sum_{i \notin I_k^X(t)} X_{ik}(t)X'_{ik}(t).$$

Notice that if $i \in I_k^X(t)$, then $X'_{ik}(t) = 0$. Thus

$$L_{kk}(t) = 2\sum_{i \notin I_k^X(t)} X_{ik}(t)\sum_{j=1}^{n} E_{ij}(t)Y_{jk}(t) = 2\sum_{i \notin I_k^X(t)}\sum_{j=1}^{n} E_{ij}(t)X_{ik}(t)Y_{jk}(t).$$

Furthermore, notice that if $j \in I_k^Y(t)$, then $Y_{jk}(t) = 0$ and

$$L_{kk}(t) = 2\sum_{i \notin I_k^X(t)}\sum_{j \notin I_k^Y(t)} E_{ij}(t)X_{ik}(t)Y_{jk}(t).$$

Similarly, consider the k-th diagonal element of $R(t)$ for $k \in [\![1, r]\!]$, one has

$$R_{kk}(t) = 2\sum_{j=1}^{n} Y_{jk}(t)Y'_{jk}(t) = 2\sum_{j \in I_k^Y(t)} Y_{jk}(t)Y'_{jk}(t) + 2\sum_{j \notin I_k^Y(t)} Y_{jk}(t)Y'_{jk}(t).$$

Notice that if $j \in I_k^Y(t)$, then $Y'_{jk}(t) = 0$. Thus,

$$R_{kk}(t) = 2\sum_{j \notin I_k^Y(t)} Y_{jk}(t)\sum_{i=1}^{m} E_{ij}(t)X_{ik}(t) = 2\sum_{j \notin I_k^Y(t)}\sum_{i=1}^{m} E_{ij}(t)X_{ik}(t)Y_{jk}(t).$$

Moreover, if $i \in I_k^X(t)$, then $X_{ik}(t) = 0$, thus

$$R_{kk}(t) = 2\sum_{j \notin I_k^Y(t)}\sum_{i \notin I_k^X(t)} E_{ij}(t)X_{ik}(t)Y_{jk}(t) = 2\sum_{i \notin I_k^X(t)}\sum_{j \notin I_k^Y(t)} E_{ij}(t)X_{ik}(t)Y_{jk}(t) = L_{kk}(t)$$

by exchanging the order of two finite sums. Finally, for all $k \in [\![1, r]\!]$ and $t \in \mathbb{R}_+$,

$$\frac{d}{dt}\left(\sum_{i=1}^{m} X_{ik}(t)^2 - \sum_{j=1}^{n} Y_{jk}(t)^2\right) = L_{kk}(t) - R_{kk}(t) = 0.$$

Therefore, for all $k \in [\![1, r]\!]$ and $t \in \mathbb{R}_+$,

$$\sum_{i=1}^{m} X_{ik}(t)^2 - \sum_{j=1}^{n} Y_{jk}(t)^2 = c_k := \sum_{i=1}^{m} X_{ik}(0)^2 - \sum_{j=1}^{n} Y_{jk}(0)^2,$$

where c_k's are constants independent of t. □

Using lemma 2.1, we next prove that the products of the entries of X and Y in (2.8) remain bounded throughout time.

Lemma 2.2. *Given $X_0 \in \mathbb{R}_+^{m\times r}$, $Y_0 \in \mathbb{R}_+^{n\times r}$, and $M \in \mathbb{R}^{m\times n}$, there exists $d > 0$ such that any solution $(X, Y) : \mathbb{R}_+ \to \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{n\times r}$ to* (2.8) *satisfies that for every $i \in [\![1, m]\!]$ and $j \in [\![1, n]\!]$,*

$$|X_{ik}(t)Y_{jk}(t)| \leqslant d, \quad \forall t \in \mathbb{R}_+, \quad \forall k \in [\![1, r]\!].$$

Proof. Note that any solution to (2.8) is a subgradient trajectory of $f + \delta_{\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{n\times r}}$ [74, Theorem 2.3(b)]. By [55, Corollary 5.4] and [22, Lemma 6.3], we have that $t \mapsto f(X(t), Y(t))$ is a decreasing function over $\mathbb{R}_+$. Thus, we have that for all $t \in \mathbb{R}_+$,

$$\begin{aligned} \|X(t)Y(t)^T\|_p &\leqslant \|X(t)Y(t)^T - M\|_p + \|M\|_p \\ &\leqslant \|X_0Y_0^T - M\|_p + \|M\|_p =: \hat{d} \end{aligned}$$

where $\hat{d} > 0$ is a constant. By the equivalence of norms, there is a constant $\hat{c}_p > 0$ such that

$$\|X(t)Y(t)^T\|_1 \leqslant \hat{c}_p\|X(t)Y(t)^T\|_p \leqslant \hat{c}_p\hat{d}$$

This implies that for all $t \in \mathbb{R}_+$,

$$\left|\sum_{k=1}^{r} X_{ik}(t)Y_{jk}(t)\right| = \left|[X(t)Y(t)^T]_{ij}\right| \leqslant \|X(t)Y(t)^T\|_1 \leqslant \hat{c}_p\hat{d}, \quad \forall i \in [\![1, m]\!], j \in [\![1, n]\!].$$

Notice that $X \in \mathbb{R}_+^{m\times r}$ and $Y \in \mathbb{R}_+^{n\times r}$, Thus, we have for all $t \in \mathbb{R}_+$,

$$\sum_{k=1}^{r} \left| X_{ik}(t) Y_{jk}(t) \right| = \left| \sum_{k=1}^{r} X_{ik}(t) Y_{jk}(t) \right| \leqslant \hat{c}_p \hat{d},$$

and the desired result follows immediately by setting $d := \hat{c}_p \hat{d}$. □

Finally, by combining Lemma 2.1 and Lemma 2.2, we can obtain the desired result that ℓ_p-nonnegative matrix factorization has bounded subgradient trajectories.

Proposition 2.5. *Let $p \geqslant 1$ and let $f : \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \to \mathbb{R}$ be defined by $f(X,Y) := \frac{1}{p}\|XY^T - M\|_p^p$. Let $\Phi := f + \delta_C$ where $C := \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{n\times r}$. Then for any $(X_0, Y_0) \in \operatorname{dom}\Phi$, there exists a unique subgradient trajectory of Φ initialized at (X_0, Y_0), and this subgradient trajectory is bounded.*

Proof. For any $(X_0, Y_0) \in \operatorname{dom}\Phi$, existence of a subgradient trajectory initialized at (X_0, Y_0) is a result of [74, Theorem 3.1]. As Φ is primal lower nice [75, Definition 1.1] at every point in $\operatorname{dom}\Phi$, such a subgradient trajectory must be unique [48, Theorem 2.9]. We next show that this subgradient trajectory is also bounded. Recall that every subgradient trajectory is a solution to (2.8) by [74, Theorem 2.3(b)]. Thus, it suffices to show that every solution to (2.8) is bounded. From lemmas 2.1 and 2.2, there exist constants c_k's for any $k \in [\![1, r]\!]$ and $d > 0$ such that for any solution $(X(\cdot), Y(\cdot))$ to (2.8), we have

$$\begin{aligned}\left(\sum_{i=1}^{m} X_{ik}(t)^2 + \sum_{j=1}^{n} Y_{jk}(t)^2\right)^2 &= \left(\sum_{i=1}^{m} X_{ik}(t)^2 - \sum_{j=1}^{n} Y_{jk}(t)^2\right)^2 + 4\sum_{i=1}^{m} X_{ik}(t)^2 \sum_{j=1}^{n} Y_{jk}(t)^2 \\ &\leqslant c_k^2 + 4mnd^2.\end{aligned}$$

for any $t \in \mathbb{R}_+$. Thus, for all $k \in [\![1, r]\!]$,

$$\sum_{i=1}^{m} X_{ik}(t)^2 + \sum_{j=1}^{n} Y_{jk}(t)^2 \leqslant \sqrt{4mnd^2 + c_k^2}.$$

Summing both sides up over k yields

$$\|X(t)\|_2^2 + \|Y(t)\|_2^2 \leqslant \sum_{k=1}^{r} \sqrt{4mnd^2 + c_k^2}.$$

Hence, every solution to (2.8) is bounded. $\square$

2.3.5 Linear neural network

Consider minimizing the loss function of linear neural network without bias term

$$f(W_1, \ldots, W_L) := \frac{1}{2}\|W_L \cdots W_1 X - Y\|_F^2, \tag{2.11}$$

where $X \in \mathbb{R}^{d_0 \times d_x}$, $Y \in \mathbb{R}^{d_L \times d_x}$, and $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ for $i = 1, \ldots, L$. Here $\|\cdot\|_F$ denotes the Frobenius norm.

Proposition 2.6. *Linear neural network with loss function* (2.11) *has bounded gradient trajectories.*

An existing proof of Proposition 2.6 under additional assumptions on network structure, initialization, input data, or target data can be found, for instance, in [76, 77, 78]. To the best of our knowledge, the closest result to Proposition 2.6 is [76, Theorem 3.2], which shows that gradient trajectories are bounded if XX^T is of full rank. In the proof of Proposition 2.6, we show that this rank assumption on X can be removed and hence Proposition 2.6 applies to any linear neural network.

Proof of Proposition 2.6. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a gradient trajectory for any initial point. By [76, Lemma 2.1], the gradient trajectories of f satisfy the initial value problem

$$\dot{W}_i = -(W_L \cdots W_{i+1})^\mathrm{T}(W_L \cdots W_1 X - Y)(W_{i-1} \cdots W_1 X)^\mathrm{T}, \tag{2.12a}$$

$$W_i(0) = W_i^0, \quad W_i^0 \in \mathbb{R}^{d_i \times d_{i-1}} \text{ is a given constant matrix,} \tag{2.12b}$$

for all $i = 1, \ldots, L$. Note that if $i = L$, (2.12a) reduces to

$$\dot{W}_L = -(W_L \cdots W_1 X - Y)(W_{L-1} \cdots W_1 X)^{\mathrm{T}},$$

and if $i = 1$, (2.12a) reduces to

$$\dot{W}_1 = -(W_L \cdots W_2)^{\mathrm{T}}(W_L \cdots W_1 X - Y)X^{\mathrm{T}}.$$

Note that [76, Theorem 3.2] proved the boundedness of gradient trajectories of f when XX^{T} is invertible. Thus, we only need to show we can always reduce the boundedness of gradient trajectories of f for general X to the boundedness of gradient trajectories of another function g in the same form as f but with invertible XX^{T}. Let $X = U\Sigma V^{\mathrm{T}}$ be a singular value decomposition, where $U \in \mathbb{R}^{d_0 \times d_0}$ and $V \in \mathbb{R}^{d_x \times d_x}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{d_0 \times d_x}$ is a rectangular matrix satisfying

$$\Sigma = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r) \succ 0,$$

where $r \leqslant \min\{d_0, d_x\}$. Eliminating X in (2.12a), it reduces to

$$\begin{aligned} \dot{W}_i &= -(W_L \cdots W_{i+1})^{\mathrm{T}}(W_L \cdots W_1 U\Sigma V^{\mathrm{T}} - Y)(W_{i-1} \cdots W_1 U\Sigma V^{\mathrm{T}})^{\mathrm{T}} \\ &= -(W_L \cdots W_{i+1})^{\mathrm{T}}(W_L \cdots W_1 U\Sigma - YV)(W_{i-1} \cdots W_1 U\Sigma)^{\mathrm{T}}. \end{aligned}$$

Define $Z := YV \in \mathbb{R}^{d_L \times d_x}$, and (2.12) reduces to

$$\dot{W}_i = -(W_L \cdots W_{i+1})^{\mathrm{T}}(W_L \cdots W_1 U\Sigma - Z)(W_{i-1} \cdots W_1 U\Sigma)^{\mathrm{T}}, \tag{2.13a}$$

$$W_i(0) = W_i^0, \quad \forall\, i = 1, \ldots L. \tag{2.13b}$$

Denote $\overline{W}_1 := W_1 U \in \mathbb{R}^{d_1 \times d_0}$ and $\overline{W}_1^0 := W_1^0 U \in \mathbb{R}^{d_1 \times d_0}$. To keep the notation consistent, also let

$\overline{W}_i := W_i$ and $\overline{W}_i^0 := W_i^0$ for $i = 2, \ldots, L$. Thus, (2.13) reduces to

$$\dot{\overline{W}}_i = -(\overline{W}_L \cdots \overline{W}_{i+1})^{\mathrm{T}}(\overline{W}_L \cdots \overline{W}_1 \Sigma - Z)(\overline{W}_{i-1} \cdots \overline{W}_1 \Sigma)^{\mathrm{T}}, \tag{2.14a}$$

$$\overline{W}_i(0) = \overline{W}_i^0, \quad \forall\, i = 1, \ldots L. \tag{2.14b}$$

Partition the matrices $\overline{W}_1$, $\overline{W}_1^0$, and Z into two column blocks:

$$\overline{W}_1 = \begin{bmatrix} \overline{W}_{11} & \overline{W}_{12} \end{bmatrix}, \quad \overline{W}_1^0 = \begin{bmatrix} \overline{W}_{11}^0 & \overline{W}_{12}^0 \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix},$$

where $\overline{W}_{11}$, $\overline{W}_{11}^0$, and Z_1 consist of the first r columns of $\overline{W}_1$, $\overline{W}_1^0$ and Z respectively. Thus, when $i = 1$, (2.14) can be reduced into

$$\dot{\overline{W}}_{11} = -(\overline{W}_L \cdots \overline{W}_2)^{\mathrm{T}}(\overline{W}_L \cdots \overline{W}_2 \overline{W}_{11} \Lambda - Z_1)\Lambda^{\mathrm{T}}, \quad \dot{\overline{W}}_{12} = 0,$$

$$\overline{W}_{11}(0) = \overline{W}_{11}^0, \quad \overline{W}_{12}(0) = \overline{W}_{12}^0.$$

When $i = 2, \ldots, L$, (2.14) can be reduced into

$$\dot{\overline{W}}_i = -(\overline{W}_L \cdots \overline{W}_{i+1})^{\mathrm{T}}(\overline{W}_L \cdots \overline{W}_2 \overline{W}_{11} \Lambda - Z_1)(\overline{W}_{i-1} \cdots \overline{W}_2 \overline{W}_{11} \Lambda)^{\mathrm{T}},$$

$$\overline{W}_i(0) = \overline{W}_i^0.$$

It indicates that $\overline{W}_{12}(t) = \overline{W}_{12}^0$ for all $t \geqslant 0$. Denote $\widetilde{W}_1 := \overline{W}_{11}$ and $\widetilde{W}_1^0 := \overline{W}_{11}^0$. To keep the notation consistent, also let $\widetilde{W}_i := \overline{W}_i$ and $\widetilde{W}_i^0 := \overline{W}_i^0$ for $i = 2, \ldots, L$. Therefore, (2.14) reduces to

$$\dot{\widetilde{W}}_i = -(\widetilde{W}_L \cdots \widetilde{W}_{i+1})^{\mathrm{T}}(\widetilde{W}_L \cdots \widetilde{W}_1 \Lambda - Z_1)(\widetilde{W}_{i-1} \cdots \widetilde{W}_1 \Lambda)^{\mathrm{T}}, \tag{2.15a}$$

$$\widetilde{W}_i(0) = \widetilde{W}_i^0, \quad \forall\, i = 1, \ldots L. \tag{2.15b}$$

Define the new function g as

$$g(\widetilde{W}_1, \ldots, \widetilde{W}_L) := \frac{1}{2}\|\widetilde{W}_L \cdots \widetilde{W}_1 \Lambda - Z_1\|_F^2.$$

Notice that the gradient trajectories of g satisfy (2.15). To prove f has bounded gradient trajectories, it is equivalent to prove g has bounded gradient trajectories, because $\|W_1\|_F = \|W_1 U\|_F = \|\overline{W}_1\|_F$ and $\|\overline{W}_1(t)\|_F^2 = \|\widetilde{W}_1(t)\|_F^2 + \|\overline{W}_{12}^0\|_F^2$ for all $t \geqslant 0$. Since $\Lambda\Lambda^{\mathrm{T}}$ is invertible, by [76, Theorem 3.2], g has bounded gradient trajectories, and so does f. □

2.3.6 One dimensional deep sigmoid neural network

Though famous for its benign theoretical properties, linear neural network is rarely used in practice because of its low representation power. We want to take a step further in the case of nonlinear deep neural network. In this subsection, we focus on neural network with sigmoid activation function in one dimensional case.

Consider minimizing the following loss function of sigmoid neural network

$$f(w_1, \ldots, w_L) := \frac{1}{2}(w_L \sigma(w_{L-1} \cdots \sigma(w_1 x)) - y)^2, \tag{2.16}$$

where $\sigma(z) := (1 + e^{-z})^{-1}$ is the sigmoid function and $w_i, x, y \in \mathbb{R}$ for all $i = 1, \ldots, L$.

Notice that the techniques in the proof of Proposition 2.6 cannot be adapted to this case because the auto-balancing property in [69, Theorem 2.1] does not hold. Surprisingly, it is still true that (2.16) has bounded gradient trajectories.

Proposition 2.7. *One dimensional sigmoid neural network with loss function* (2.16) *has bounded gradient trajectories.*

Proof. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a gradient trajectory for any initial point. For simplicity, define p_i for $i = 0, \ldots, L-2$ recursively by $p_0 := x$,

$p_1 := \sigma(w_1 x)$ and $p_{i+1} := \sigma(w_{i+1} p_i)$. The gradient trajectories of f satisfy

$$\dot{w}_L = -(w_L \sigma(w_{L-1} p_{L-2}) - y)\sigma(w_{L-1} p_{L-2}), \tag{2.17a}$$

$$\dot{w}_i = \frac{p_{i-1}}{1 + e^{w_i p_{i-1}}} \dot{w}_{i+1} w_{i+1}, \quad i = L-1, \ldots, 1. \tag{2.17b}$$

We will prove each w_i is bounded inductively from the last layer to the first layer. The relation between the last two layers w_L and w_{L-1}, and the relation between the first two layers can be regarded as the base cases.

We claim that there exists a time T such that $\dot{w}_i$ and w_i does not change sign for all $t \geqslant T$ and for all i. To verify this, first notice that the claim is true for the last layer, i.e., $\dot{w}_L$ and w_L will not change sign for all $t \geqslant T$. Suppose $\dot{w}_L$ changes sign, by continuity and mean value theorem, there exists $t^* > 0$ such that $\dot{w}_L(t^*) = 0$. However, $\dot{w}_L(t^*) = 0$ implies $\dot{w}_i(t^*) = 0$ for all i, meaning that a critical point is achieved and the gradient trajectory is stopped for all $t \geqslant t^*$. In this case, all w_i's are trivially bounded. Thus, we assume the trajectory will never stop at a finite time. In this case, either $\dot{w}_L(t) > 0$ or $\dot{w}_L(t) < 0$ for all $t \geqslant 0$. Since w_L is monotonic, it either keeps the sign unchanged or changes the sign only once. Thus, there exists $T_L > 0$ such that w_L does not change sign on $[T_L, \infty)$. Notice that for all $i \geqslant 2$, $p_{i-1}(t) \in (0, 1)$ for all $t \geqslant 0$. Since $\dot{w}_L w_L$ does not change sign on $[T_L, \infty)$, (2.17b) implies that $\dot{w}_{L-1}$ does not change sign on $[T_L, \infty)$ either. Therefore, we conclude that w_{L-1} is monotonic. Similarly, there exists $T_{L-1} > T_L$ such that $\dot{w}_{L-1}$ and w_{L-1} does not change sign on $[T_{L-1}, \infty)$. Recursively using the above argument, we can show the claim is true for all $i \geqslant 2$ on $[T_2, \infty)$. For $i = 1$, although $p_0 = x$ may not be in $(0, 1)$, since x is a constant, the fact that $\dot{w}_2$ and w_2 do not change sign still implies that $\dot{w}_1$ does not change sign and hence there exists $T_1 > T_2$ such that w_1 does not change sign on $[T_1, \infty)$. Therefore, the claim holds for $i = 1, \ldots, L$ by choosing $T = T_1$.

By the claim proved in the last paragraph, for $i = 1, \ldots, L$, either $\dot{w}_i w_i$ is nonnegative or $\dot{w}_i w_i$ is negative on $[T, \infty)$. Now we are going to prove each w_i is bounded. The first step is to prove the last two layers w_L and w_{L-1} are bounded. Consider the case where $\dot{w}_L w_L$ is nonnegative on

$[T, \infty)$. (2.17b) implies that $\dot{w}_{L-1} \geqslant 0$ and w_{L-1} is increasing over $[T, \infty)$, so there exists a constant c_{L-1} such that $w_{L-1}(t) \geqslant c_{L-1}$ for all $t \geqslant 0$. Since $p_{L-2} \in (0,1)$, we have $\sigma(w_{L-1}p_{L-2}) \geqslant \sigma(-|c_{L-1}|) > 0$. Again, by [22, Lemma 5.2], $\frac{d}{dt} f(w_1, \ldots, w_L) \leqslant 0$ and $f(w_1, \ldots, w_L) \leqslant C$ for some constant C on $[0, \infty)$. Thus, it is easy to see $|w_L|\sigma(w_{L-1}p_{L-2}) \leqslant C_1$ for some constant C_1 on $[0, \infty)$. Since $\sigma(w_{L-1}p_{L-2}) \in [\sigma(-|c_{L-1}|), 1)$, we conclude $|w_L|$ is bounded. Suppose w_{L-1} is unbounded. Since it is increasing and does not change sign, $w_{L-1}(t) > 0$ for all $t \geqslant T$ and $w_{L-1}(t) \to \infty$ as $t \to \infty$. By (2.17b),

$$\dot{w}_{L-1} = \frac{p_{L-2}}{1 + e^{w_{L-1}p_{L-2}}} \dot{w}_L w_L \leqslant \dot{w}_L w_L, \tag{2.18}$$

because $p_{L-2} \in (0,1)$ and $1 + e^{w_L p_{L-2}} > 1$. By (2.18), $w_{L-1} - \frac{1}{2}w_L^2$ is a decreasing function on $[T, \infty)$. Hence, $w_{L-1} - \frac{1}{2}w_L^2 \leqslant C_2$ for some constant C_2. Notice that w_L is bounded but $w_{L-1}(t) \to \infty$ as $t \to \infty$, so a contradiction occurs. Therefore, w_{L-1} is bounded.

Now we consider the case where $\dot{w}_L w_L$ is negative on $[T, \infty)$. In this case, (2.17b) implies $\dot{w}_{L-1} \leqslant 0$, so w_{L-1} is decreasing on $[T, \infty)$ and there exists a constant d_{L-1} such that $w_{L-1} \leqslant d_{L-1}$. Since $p_{L-2}/(1 + e^{w_{L-1}p_{L-2}}) \in (0,1)$ and $\dot{w}_L w_L \leqslant 0$ on $[T, \infty)$, we have $\dot{w}_{L-1} \geqslant \dot{w}_L w_L$. This shows $w_{L-1} - \frac{1}{2}w_L^2$ is increasing on $[T, \infty)$, and hence $w_{L-1} \geqslant \tilde{d}_{L-1}$ for some constant $\tilde{d}_{L-1}$. Therefore, $w_{L-1} \in [\tilde{d}_{L-1}, d_{L-1}]$ is bounded. By exactly the same argument as in the case when $\dot{w}_L w_L$ is nonnegative, we know $\sigma(w_{L-1}p_{L-2}) \in [\sigma(-|\tilde{d}_{L-1}|), 1)$ and w_L is bounded by using the boundedness of objective function f.

Up to now, we have proved boundedness for the last two layers w_L and w_{L-1}. For $i = 2, \ldots, L-2$, by discussing two cases $\dot{w}_{i+1}w_{i+1} \geqslant 0$ and $\dot{w}_{i+1}w_{i+1} \leqslant 0$, together with the boundedness of w_{i+1}, we can prove that w_i is bounded by exactly the same argument as we did in the last two paragraphs. The induction starts with proving w_{L-2} is bounded and ends with proving w_2 is bounded. Once we prove w_2 is bounded, consider the relation between w_1 and w_2,

$$(1 + e^{w_1 x})\dot{w}_1 = x\dot{w}_2 w_2.$$

If $x = 0$, then $\dot{w}_1 = 0$ implies w_1 is a constant over $[0, \infty)$, so it must be bounded. Suppose $x \neq 0$, by taking integration with respect to t and multiplying x on both sides, we have

$$w_1 x + e^{w_1 x} = \frac{x^2}{2} w_2^2 + C_3.$$

Let $z = w_1 x$, then $z + e^z \to \pm\infty$ as $z \to \pm\infty$. Thus, the boundedness of w_2 implies the boundedness of $z = w_1 x$. Since $x \neq 0$ is a constant, w_1 is bounded. Therefore, we proved that w_i is bounded for all $i = 1, \ldots, L$. □

References

[1] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[2] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning based recommender system: A survey and new perspectives,” *ACM computing surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.

[3] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, “Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer,” in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.

[4] M Turk and A Pentland, “Eigenfaces for recognition,” *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[6] S. Balaban, “Deep learning and face recognition: The state of the art,” *Biometric and surveillance technology for human and activity identification XII*, vol. 9457, pp. 68–75, 2015.

[7] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, “End to end learning for self-driving cars,” *arXiv preprint arXiv:1604.07316*, 2016.

[8] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” *Journal of field robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[11] V. Cevher, S. Becker, and M. Schmidt, "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 32–43, 2014.

[12] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.

[13] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 13 693–13 696.

[14] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From words to watts: Benchmarking the energy costs of large language model inference," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, IEEE, 2023, pp. 1–9.

[15] D. Bertsekas, *Nonlinear Programming* (Athena Scientific optimization and computation series). Athena Scientific, 1995, ISBN: 9781886529144.

[16] Y. Nesterov, *Lectures on Convex Optimization*. Springer Science & Business Media, 2018, vol. 137.

[17] A. Beck, *First-order methods in optimization*. SIAM, 2017.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[20] P.-A. Absil, R. Mahony, and B. Andrews, "Convergence of the iterates of descent methods for analytic cost functions," *SIAM Journal on Optimization*, vol. 16, no. 2, pp. 531–547, 2005.

[21] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Mathematical Programming*, vol. 137, no. 1, pp. 91–129, 2013.

[22] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.

[23] C. Josz and L. Lai, "Global stability of first-order methods for coercive tame functions," *Mathematical Programming*, pp. 1–26, 2023.

[24] L. Ljung, “Analysis of recursive stochastic algorithms,” *IEEE transactions on automatic control*, vol. 22, no. 4, pp. 551–575, 1977.

[25] H. J. Kushner, “General convergence results for stochastic approximations via weak convergence theory,” *Journal of mathematical analysis and applications*, vol. 61, no. 2, pp. 490–503, 1977.

[26] M. Benaïm, “Dynamics of stochastic approximation algorithms,” in *Seminaire de probabilites XXXIII*, Springer, 2006, pp. 1–68.

[27] J. C. Duchi and F. Ruan, “Stochastic methods for composite and weakly convex optimization problems,” *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3229–3259, 2018.

[28] J. Bolte and E. Pauwels, “Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning,” *Mathematical Programming*, vol. 188, pp. 19–51, 2021.

[29] A. Salim, “Random monotone operators and application to stochastic optimization,” Ph.D. dissertation, Université Paris-Saclay (ComUE), 2018.

[30] E. Pauwels, “Incremental without replacement sampling in nonconvex optimization,” *Journal of Optimization Theory and Applications*, pp. 1–26, 2021.

[31] P. Bianchi, W. Hachem, and S. Schechtman, “Convergence of constant step stochastic gradient descent for non-smooth non-convex functions,” *Set-Valued and Variational Analysis*, vol. 30, no. 3, pp. 1117–1147, 2022.

[32] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, “Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality,” *Mathematics of operations research*, vol. 35, no. 2, pp. 438–457, 2010.

[33] M. Korda, “Stability and performance verification of dynamical systems controlled by neural networks: Algorithms and complexity,” *IEEE Control Systems Letters*, vol. 6, pp. 3265–3270, 2022.

[34] A. Daniilidis and D. Drusvyatskiy, “Pathological subgradient dynamics,” *SIAM Journal on Optimization*, vol. 30, no. 2, pp. 1327–1338, 2020.

[35] J. Bolte and E. Pauwels, “Curiosities and counterexamples in smooth convex optimization,” *Mathematical Programming*, vol. 195, no. 1, pp. 553–603, 2022.

[36] L. Van den Dries, *Tame topology and o-minimal structures*. Cambridge university press, 1998, vol. 248.

[37] M. Coste, *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000.

[38] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[39] L. C. Evans and R. F. Garzepy, *Measure theory and fine properties of functions*. Oxfordshire: Routledge, 2018.

[40] F. H. Clarke, *Optimization and Nonsmooth Analysis*. Philadelphia: SIAM Classics in Applied Mathematics, 1990.

[41] J. Bochnak, M. Coste, and M.-F. Roy, *Real algebraic geometry*. Springer Science & Business Media, 2013, vol. 36.

[42] T. S. Pham and H. H. Vui, *Genericity in polynomial optimization*. London: World Scientific, 2016, vol. 3.

[43] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, "Clarke subgradients of stratifiable functions," *SIAM Journal on Optimization*, vol. 18, no. 2, pp. 556–572, 2007.

[44] K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, vol. 48, 1998, pp. 769–783.

[45] C. Josz, "Global convergence of the gradient method for functions definable in o-minimal structures," *Mathematical Programming*, pp. 1–29, 2023.

[46] J.-P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory*. Berlin: Springer-Verlag, 1984, vol. 264.

[47] O. A. Nielsen, *An introduction to integration and measure theory*. New York: Wiley-Interscience, 1997, vol. 17.

[48] S. Marcellin and L. Thibault, "Evolution problems associated with primal lower nice functions," *Journal of convex Analysis*, vol. 13, no. 2, p. 385, 2006.

[49] F. Santambrogio, "{Euclidean, metric, and wasserstein} gradient flows: An overview," *Bulletin of Mathematical Sciences*, vol. 7, no. 1, pp. 87–154, 2017.

[50] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[51] P. M. Fitzpatrick and H. L. Royden, *Real Analysis*, 4th ed. Upper Saddle River, NJ: Pearson, Jan. 2010.

[52] I. Ekeland, "On the variational principle," *Journal of Mathematical Analysis and Applications*, vol. 47, no. 2, pp. 324–353, 1974.

[53] J.-B. Hiriart-Urruty, "A short proof of the variational principle for approximate solutions of a minimization problem," *The American Mathematical Monthly*, vol. 90, no. 3, pp. 206–207, 1983.

[54] T. X. D. Ha, "The ekeland variational principle for set-valued maps involving coderivatives," *Journal of mathematical analysis and applications*, vol. 286, no. 2, pp. 509–523, 2003.

[55] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, "Curves of descent," *SIAM Journal on Control and Optimization*, vol. 53, no. 1, pp. 114–138, 2015.

[56] H. Attouch, G. Buttazzo, and G. Michaille, *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. Philadelphia: SIAM, 2014.

[57] J. Bolte and E. Pauwels, "Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning," *Mathematical Programming*, pp. 1–33, 2020.

[58] J. Borwein and X. Wang, "Lipschitz functions with maximal clarke subdifferentials are generic," *Proceedings of the American Mathematical Society*, vol. 128, no. 11, pp. 3221–3229, 2000.

[59] A. Daniilidis and G. Flores, "Linear structure of functions with maximal clarke subdifferential," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 511–521, 2019.

[60] Z. Luo, "On the convergence of the lms algorithm with adaptive learning rate for linear feedforward networks.," *Neural Computation*, vol. 3, no. 2, pp. 226–245, 1991.

[61] V. Elser, T.-Y. Lan, and T. Bendory, "Benchmark problems for phase retrieval," *SIAM Journal on Imaging Sciences*, vol. 11, no. 4, pp. 2429–2455, 2018.

[62] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest, "Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes," *Annu. Rev. Phys. Chem.*, vol. 59, no. 1, pp. 387–410, 2008.

[63] C Fienup and J Dainty, "Phase retrieval and image reconstruction for astronomy," *Image recovery: theory and application*, vol. 231, p. 275, 1987.

[64] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: A contemporary overview," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[65] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.

[66] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[67] N. Gillis and S. A. Vavasis, "On the complexity of robust pca and ℓ_1-norm low-rank matrix approximation," *Mathematics of Operations Research*, vol. 43, no. 4, pp. 1072–1084, 2018.

[68] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, "Low-rank matrix recovery with composite optimization: Good conditioning and rapid convergence," *Foundations of Computational Mathematics*, pp. 1–89, 2021.

[69] S. S. Du, W. Hu, and J. D. Lee, "Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[70] W.-J. Zeng and H. C. So, "Outlier-robust matrix completion via ℓ_p-minimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1125–1140, 2017.

[71] R. M. Wallace, "Analysis of absorption spectra of multicomponent systems," *The Journal of Physical Chemistry*, vol. 64, no. 7, pp. 899–901, 1960.

[72] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[73] N. Gillis, *Nonnegative matrix factorization*. SIAM, 2020.

[74] B. Cornet, "Existence of slow solutions for a class of differential inclusions," *Journal of mathematical analysis and applications*, vol. 96, no. 1, pp. 130–147, 1983.

[75] R. Poliquin, "Integration of subdifferentials of nonconvex functions," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 17, no. 4, pp. 385–398, 1991.

[76] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg, "Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers," *Information and Inference: A Journal of the IMA*, vol. 11, no. 1, pp. 307–353, 2022.

[77] A. Eftekhari, "Training linear neural networks: Non-local convergence and complexity results," in *International Conference on Machine Learning*, PMLR, 2020, pp. 2836–2847.

[78] K. Chen, D. Lin, and Z. Zhang, "On non-local convergence analysis of deep linear networks," in *International Conference on Machine Learning*, PMLR, 2022, pp. 3417–3443.

[79] C. Josz and X. Li, "Certifying the absence of spurious local minima at infinity," *SIAM Journal on Optimization*, vol. 33, pp. 1416–1439, 3 2023.

[80] K. L. Blackmore, R. C. Williamson, and I. M. Mareels, "Local minima and attractors at infinity for gradient descent learning algorithms," *Journal of Mathematical Systems Estimation and Control*, vol. 6, pp. 231–234, 1996.

[81] S. Liang, R. Sun, J. D. Lee, and R. Srikant, "Adding one neuron can eliminate all bad local minima," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[82] J. Sohl-Dickstein and K. Kawaguchi, "Eliminating all bad local minima from loss landscapes without even adding an extra unit," *arXiv preprint arXiv:1901.03909*, 2019.

[83] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient Descent Only Converges to Minimizers," *COLT*, 2016.

[84] C. Josz, Y. Ouyang, R. Y. Zhang, J. Lavaei, and S. Sojoudi, "A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization," *NeurIPS*, Dec. 2018.

[85] L. Venturi, A. S. Bandeira, and J. Bruna, "Spurious valleys in one-hidden-layer neural network optimization landscapes," *Journal of Machine Learning Research*, vol. 20, p. 133, 2019.

[86] C. D. Freeman and J. Bruna, "Topology and geometry of half-rectified network optimization," in *International Conference on Learning Representations*, 2017.

[87] T. Ding, D. Li, and R. Sun, "Suboptimal local minima exist for wide neural networks with smooth activations," *Mathematics of Operations Research*, 2022.

[88] D. Li, T. Ding, and R. Sun, "On the benefit of width for neural networks: Disappearance of basins," *SIAM Journal on Optimization*, vol. 32, no. 3, pp. 1728–1758, 2022.

[89] J. R. Munkres, "Topology," *Prenctice Hall, US*, 2000.

[90] K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems*, PMLR, vol. 29, 2016.

[91] T. Laurent and J. Brecht, "Deep linear networks with arbitrary loss: All local minima are global," in *International conference on machine learning*, PMLR, 2018, pp. 2902–2907.

[92] L. Zhang, "Depth creates no more spurious local minima," *arXiv preprint arXiv:1901.09827*, 2019.

[93] A. J. Wilkie, "Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function," *Journal of the American Mathematical Society*, vol. 9, no. 4, pp. 1051–1094, 1996.

[94] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3614–3628, 2018.

[95] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the burer-monteiro approach," in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 65–74.

[96] S. Li, Q. Li, Z. Zhu, G. Tang, and M. B. Wakin, "The global geometry of centralized and distributed low-rank matrix recovery without regularization," *IEEE Signal Processing Letters*, vol. 27, pp. 1400–1404, 2020.

[97] C. Josz and L. Lai, "Nonsmooth rank-one matrix factorization landscape," *Optimization Letters*, pp. 1–21, 2021.

[98] I. Safran and O. Shamir, "Spurious local minima are common in two-layer relu neural networks," in *International conference on machine learning*, PMLR, 2018, pp. 4433–4441.

[99] R. Y. Zhang, S. Sojoudi, and J. Lavaei, "Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery," *Journal of Machine Learning Research*, vol. 20, no. 114, pp. 1–34, 2019.

[100] C. Josz, L. Lai, and X. Li, "Convergence of the momentum method for semialgebraic functions with locally lipschitz gradients," *SIAM Journal on Optimization*, vol. 33, no. 4, pp. 3012–3037, 2023.

[101] N. B. Kovachki and A. M. Stuart, "Continuous time analysis of momentum methods," *Journal of Machine Learning Research*, vol. 22, no. 17, pp. 1–40, 2021.

[102] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Workshop on Autodiff*, Long Beach, California, USA, 2017.

[103] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[104] M. Abadi, "Tensorflow: Learning functions at scale," in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, 2016, pp. 1–1.

[105] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[106] P. Ochs, Y. Chen, T. Brox, and T. Pock, "Ipiano: Inertial proximal algorithm for nonconvex optimization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1388–1419, 2014.

[107] P. Ochs, "Local convergence of the heavy-ball method and ipiano for non-convex optimization," *Journal of Optimization Theory and Applications*, vol. 177, no. 1, pp. 153–180, 2018.

[108] S. Łojasiewicz, "Ensembles semi-analytiques," *IHES notes*, 1965.

[109] S. Zavriev and F. Kostyuk, "Heavy-ball method in nonconvex optimization problems," *Computational Mathematics and Modeling*, vol. 4, no. 4, pp. 336–341, 1993.

[110] B. Wen, X. Chen, and T. K. Pong, "Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 124–145, 2017.

[111] Z. Jia, Z. Wu, and X. Dong, "An inexact proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth optimization problems," *Journal of Inequalities and Applications*, vol. 2019, no. 1, pp. 1–16, 2019.

[112] T. Sun, D. Li, Z. Quan, H. Jiang, S. Li, and Y. Dou, "Heavy-ball algorithms always escape saddle points," *IJCAI*, 2019.

[113] M. O'Neill and S. J. Wright, "Behavior of accelerated gradient methods near critical points of nonconvex functions," *Mathematical Programming*, vol. 176, no. 1, pp. 403–427, 2019.

[114] R. Pemantle, "Nonconvergence to unstable points in urn models and stochastic approximations," *The Annals of Probability*, vol. 18, no. 2, pp. 698–712, 1990.

[115] G. Garrigos, "Descent dynamical systems and algorithms for tame optimization, and multiobjective problems," Ph.D. dissertation, Université Montpellier; Universidad técnica Federico Santa María (Valparaiso), 2015.

[116] E. A. Coddington and N. Levinson, *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.

[117] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Mathematical Programming*, vol. 156, no. 1-2, pp. 59–99, 2016.

[118] M. Shub, *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

[119] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[120] G. Li and T. K. Pong, "Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods," *Foundations of computational mathematics*, vol. 18, no. 5, pp. 1199–1232, 2018.

[121] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[122] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid strict saddle points," *Mathematical programming*, vol. 176, no. 1, pp. 311–337, 2019.

[123] V. Zorich and R. Cooke, *Mathematical Analysis I* (Mathematical Analysis). Springer, 2004, ISBN: 9783540403869.

[124] S. P. Ponomarev, "Submersions and preimages of sets of measure zero," *Siberian Mathematical Journal*, vol. 28, no. 1, pp. 153–163, 1987.

[125] J. R. Silvester, "Determinants of block matrices," *The Mathematical Gazette*, vol. 84, no. 501, pp. 460–467, 2000.

[126] C. Josz, L. Lai, and X. Li, "Proximal random reshuffling under local lipschitz continuity," *arXiv preprint arXiv:2408.07182*, 2024.

[127] H Kushner, "Convergence of recursive adaptive and identification procedures via weak convergence theory," *IEEE Transactions on Automatic Control*, vol. 22, no. 6, pp. 921–930, 1977.

[128] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.

[129] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions, part ii: Applications," *Mathematics of Operations Research*, vol. 31, no. 4, pp. 673–695, 2006.

[130] H Brézis, "Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de hilbert. number 5 in north holland math," *Studies. North-Holland, Amsterdam*, 1973.

[131] S. Schechtman, "Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation set," *Optimization Letters*, vol. 17, no. 1, pp. 177–190, 2023.

[132] X. Li, A. Milzarek, and J. Qiu, "Convergence of Random Reshuffling under the Kurdyka–Łojasiewicz Inequality," *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 1092–1120, 2023.

[133] Y.-X. Wang and Y.-J. Zhang, “Nonnegative matrix factorization: A comprehensive review,” *IEEE Transactions on knowledge and data engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.

[134] M. V. Solodov, “Convergence properties of proximal (sub) gradient methods without convexity or smoothness of any of the functions,” *SIAM Journal on Optimization*, vol. 35, no. 1, pp. 28–41, 2025.

[135] J. Duchi and Y. Singer, “Efficient online and batch learning using forward backward splitting,” *The Journal of Machine Learning Research*, vol. 10, pp. 2899–2934, 2009.

[136] D. Davis and D. Drusvyatskiy, “Stochastic model-based minimization of weakly convex functions,” *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 207–239, 2019.

[137] J. Bolte, T. Le, É. Moulines, and E. Pauwels, “Inexact subgradient methods for semialgebraic functions,” *arXiv preprint arXiv:2404.19517*, 2024.

[138] K. Ding, N. Xiao, and K.-C. Toh, “Adam-family methods with decoupled weight decay in deep learning,” *arXiv preprint arXiv:2310.08858*, 2023.

[139] N. Xiao, X. Hu, and K.-C. Toh, “Stochastic subgradient methods with guaranteed global stability in nonsmooth nonconvex optimization,” *arXiv preprint arXiv:2307.10053*, 2023.

[140] Y. Nesterov, “Gradient methods for minimizing composite functions,” *Mathematical programming*, vol. 140, no. 1, pp. 125–161, 2013.

[141] C. Kanzow and P. Mehlitz, “Convergence properties of monotone and nonmonotone proximal gradient methods revisited,” *Journal of Optimization Theory and Applications*, vol. 195, no. 2, pp. 624–646, 2022.

[142] P. Frankel, G. Garrigos, and J. Peypouquet, “Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates,” *Journal of Optimization Theory and Applications*, vol. 165, pp. 874–900, 2015.

[143] X. Jia, C. Kanzow, and P. Mehlitz, “Convergence Analysis of the Proximal Gradient Method in the Presence of the Kurdyka-Łojasiewicz Property without Global Lipschitz Assumptions,” *SIOPT*, vol. 33, no. 4, pp. 3038–3056, 2023.

[144] D. P. Bertsekas *et al.*, “Incremental gradient, subgradient, and proximal methods for convex optimization: A survey,” *Optimization for Machine Learning*, vol. 2010, no. 1-38, p. 3, 2011.

[145] K. Mishchenko, A. Khaled, and P. Richtárik, “Random reshuffling: Simple analysis with vast improvements,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 309–17 320, 2020.

[146] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo, "Why random reshuffling beats stochastic gradient descent," *Mathematical Programming*, vol. 186, no. 1, pp. 49–84, 2021.

[147] O. Mangasariany and M. Solodovy, "Serial and parallel backpropagation convergence via nonmonotone perturbed minimization," *Optimization Methods and Software*, 1994.

[148] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

[149] M Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo, "Convergence rate of incremental gradient and incremental newton methods," *SIAM Journal on Optimization*, vol. 29, no. 4, pp. 2542–2565, 2019.

[150] J. Haochen and S. Sra, "Random shuffling beats sgd after finite epochs," in *International Conference on Machine Learning*, PMLR, 2019, pp. 2624–2633.

[151] A. Khaled and P. Richtárik, "Better theory for sgd in the nonconvex world," *arXiv preprint arXiv:2002.03329*, 2020.

[152] L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. van Dijk, "A unified convergence analysis for shuffling-type gradient methods," *Journal of Machine Learning Research*, vol. 22, no. 207, pp. 1–44, 2021.

[153] X. Li, A. Milzarek, and J. Qiu, "A new random reshuffling method for nonsmooth nonconvex finite-sum optimization," *arXiv preprint arXiv:2312.01047*, 2023.

[154] L. Rosasco, S. Villa, and B. C. Vũ, "Convergence of stochastic proximal gradient algorithm," *Applied Mathematics & Optimization*, vol. 82, pp. 891–917, 2020.

[155] S. Majewski, B. Miasojedow, and E. Moulines, "Analysis of nonsmooth stochastic approximation: The differential inclusion approach," *arXiv preprint arXiv:1805.01916*, 2018.

[156] K. Mishchenko, A. Khaled, and P. Richtárik, "Proximal and federated random reshuffling," in *International Conference on Machine Learning*, PMLR, 2022, pp. 15 718–15 749.

[157] V. B. Tadić, "Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema," *Stochastic Processes and their Applications*, vol. 125, no. 5, pp. 1715–1755, 2015.

[158] S. Dereich and S. Kassing, "Convergence of stochastic gradient descent schemes for lojasiewicz-landscapes," *arXiv preprint arXiv:2102.09385*, 2021.

[159] E. A. Nurminskii, "The quasigradient method for the solving of the nonlinear programming problems," *Cybernetics*, vol. 9, no. 1, pp. 145–150, 1973.

[160] J.-P. Vial, “Strong and weak convexity of sets and functions,” *Mathematics of Operations Research*, vol. 8, no. 2, pp. 231–259, 1983.

[161] J.-J. Moreau, “Proximité et dualité dans un espace hilbertien,” *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.

[162] L. Tian and A. M.-C. So, “No dimension-free deterministic algorithm computes approximate stationarities of Lipschitzians,” *Mathematical Programming*, pp. 1–24, 2024.

[163] L. Tian and A. M.-C. So, “On the hardness of computing near-approximate stationary points of clarke regular nonsmooth nonconvex problems and certain dc programs,” in *ICML Workshop on Beyond First-Order Methods in ML Systems*, 2021.

[164] D. Davis and B. Grimmer, “Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems,” *SIAM Journal on Optimization*, vol. 29, no. 3, pp. 1908–1930, 2019.

[165] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie, “Complexity of finding stationary points of nonconvex nonsmooth functions,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 11 173–11 182.

[166] A. Goldstein, “Optimization of Lipschitz continuous functions,” *Mathematical Programming*, vol. 13, pp. 14–22, 1977.

[167] L. Van den Dries and C. Miller, “Geometric categories and o-minimal structures,” *Duke Mathematical Journal*, vol. 84, no. 2, pp. 497–540, 1996.

[168] J. Bolte, T. Le, and E. Pauwels, “Subgradient sampling for nonsmooth nonconvex minimization,” *SIAM Journal on Optimization*, vol. 33, no. 4, pp. 2542–2569, 2023.

[169] F. H. Clarke, “Generalized gradients and applications,” *Transactions of the American Mathematical Society*, vol. 205, pp. 247–262, 1975.

[170] R. Rockafellar, *Convex Analysis* (Princeton mathematical series). Princeton University Press, 1970, ISBN: 9780691080697.

[171] B. Schwartz, “Finite extensions of convex functions,” *Mathematische Operationsforschung und Statistik. Series Optimization*, vol. 10, no. 4, pp. 501–509, 1979.

[172] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition,” *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[173] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, "Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery," *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 1, pp. 10–22, 2014.

[174] D. Davis, D. Drusvyatskiy, and C. Paquette, "The nonsmooth landscape of phase retrieval," *IMA Journal of Numerical Analysis*, vol. 40, no. 4, pp. 2652–2695, 2020.

[175] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[176] B. T. Polyak, "Minimization of unsmooth functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3, pp. 14–29, 1969.

[177] J.-L. Goffin, "On convergence rates of subgradient optimization methods," *Mathematical programming*, vol. 13, pp. 329–347, 1977.

[178] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, "Subgradient methods for sharp weakly convex functions," *Journal of Optimization Theory and Applications*, vol. 179, pp. 962–982, 2018.

[179] X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal, "Nonconvex robust low-rank matrix recovery," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 660–686, 2020.

[180] D. Davis, D. Drusvyatskiy, and V. Charisopoulos, "Stochastic algorithms with geometric step decay converge linearly on sharp functions," *Mathematical Programming*, vol. 207, no. 1, pp. 145–190, 2024.

[181] H. Yu and X. Li, "High probability guarantees for random reshuffling," *arXiv preprint arXiv:2311.11841*, 2023.

[182] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[183] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, "Group sparse regularization for deep neural networks," *Neurocomputing*, vol. 241, pp. 81–89, 2017.

[184] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.

[185] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[186] S. Fattahi and S. Sojoudi, "Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis," *Journal of machine learning research*, 2020.

[187] D. Drusvyatskiy and C. Paquette, "Efficiency of minimizing compositions of convex functions and smooth maps," *Mathematical Programming*, vol. 178, pp. 503–558, 2019.

[188] J. Ma and S. Fattahi, "Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization," *Journal of Machine Learning Research*, vol. 24, no. 96, pp. 1–84, 2023.

[189] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[190] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000.

[191] A. Cichocki, R. Zdunek, and S.-i. Amari, "Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization," in *International conference on independent component analysis and signal separation*, Springer, 2007, pp. 169–176.

[192] J. Kim, Y. He, and H. Park, "Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework," *Journal of Global Optimization*, vol. 58, pp. 285–319, 2014.

[193] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.

[194] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[195] D. Hajinezhad, T.-H. Chang, X. Wang, Q. Shi, and M. Hong, "Nonnegative matrix factorization using admm: Algorithm and convergence analysis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 4742–4746.

[196] M. C. Mukkamala and P. Ochs, "Beyond alternating updates for matrix factorization with inertial bregman proximal gradient algorithms," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[197] M. Teboulle and Y. Vaisbourd, "Novel proximal gradient methods for nonnegative matrix factorization with sparsity constraints," *SIAM Journal on Imaging Sciences*, vol. 13, no. 1, pp. 381–421, 2020.

[198] N. Takahashi and R. Hibi, “Global convergence of modified multiplicative updates for nonnegative matrix factorization,” *Computational Optimization and Applications*, vol. 57, pp. 417–440, 2014.

[199] T. Kimura and N. Takahashi, “Global convergence of a modified HALS algorithm for nonnegative matrix factorization,” in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, 2015, pp. 21–24.

[200] P. H. Calamai and J. J. Moré, “Projected gradient methods for linearly constrained problems,” *Mathematical programming*, vol. 39, no. 1, pp. 93–116, 1987.

[201] A. Rakotomamonjy, “Direct optimization of the dictionary learning problem,” *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5495–5506, 2013.

[202] C. Josz and L. Lai, “Lyapunov stability of the subgradient method with constant step size,” *Mathematical Programming*, pp. 1–10, 2023.

[203] S Cobzas and C Mustata, “Norm preserving extension of convex lipschitz functions,” *J. Approx. theory*, vol. 24, no. 3, pp. 236–244, 1978.

[204] A. Lewis and T. Tian, “Identifiability, the KL property in metric spaces, and subgradient curves,” *arXiv preprint arXiv:2205.02868*, 2022.

[205] P. W. Gwanyama, “The HM-GM-AM-QM inequalities,” *College Mathematics Journal*, pp. 47–50, 2004.

[206] P. Maréchal, “On a functional operation generating convex functions, part 1: Duality,” *Journal of Optimization Theory and Applications*, vol. 126, no. 1, pp. 175–189, 2005.

[207] S. Lojasiewicz, “Sur les trajectoires du gradient d’une fonction analytique,” *Seminari di geometria*, vol. 1983, pp. 115–117, 1982.