A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization

C. Josz
EECS, UC Berkeley
cedric.josz@gmail.com

Y. Ouyang
IEOR, UC Berkeley
ouyangyi@berkeley.edu

R. Y. Zhang
IEOR, UC Berkeley
ryz@berkeley.edu

J. Lavaei
IEOR, UC Berkeley
lavaei@berkeley.edu

S. Sojoudi
EECS, UC Berkeley
sojoudi@berkeley.edu

Abstract

We study the set of continuous functions that admit no spurious local optima (i.e. local minima that are not global minima) which we term *global functions*. They satisfy various powerful properties for analyzing nonconvex and nonsmooth optimization problems. For instance, they satisfy a theorem akin to the fundamental uniform limit theorem in the analysis regarding continuous functions. Global functions are also endowed with useful properties regarding the composition of functions and change of variables. Using these new results, we show that a class of nonconvex and nonsmooth optimization problems arising in tensor decomposition applications are global functions. This is the first result concerning nonconvex methods for nonsmooth objective functions. Our result provides a theoretical guarantee for the widely-used ℓ_1 norm to avoid outliers in nonconvex optimization.

1 Introduction

A recent branch of research in optimization and machine learning consists in proving that simple and practical algorithms can solve nonconvex optimization problems. Applications include, but are not limited to, neural networks [40, 44], dictionary learning [1, 2], deep learning [39, 50], mixed linear regression [49, 43], and phase retrieval [46, 21]. In this paper, we focus our attention on matrix completion/sensing [30, 24, 38] and tensor recovery/decomposition [5, 4, 31, 35]. Matrix completion/sensing aims to recover an unknown positive semidefinite matrix M of known size n and rank r from a finite number of linear measurements modeled by the expression $\langle A_i, M\rangle := \text{trace}(A_i M),\ i = 1, \dots, m$, where the symmetric matrices $A_1, \dots, A_m$ of size n are known. It is assumed that the measurements contain noise which can modeled as $b_i := \langle A_i, M\rangle + \epsilon_i$ where ϵ_i is a realization of a random variable. When the noise is Gaussian, the maximum likelihood estimate of M can be recast as the nonconvex optimization problem

$$\inf_{M \succcurlyeq 0} \quad \sum_{i=1}^{m} \left(\langle A_i, M\rangle - b_i\right)^2 \quad \text{subject to} \quad \text{rank}(M) = r \tag{1}$$

where $M \succcurlyeq 0$ stands for positive semidefinite. One can remove the rank constraint and obtain a convex relaxation. It can then be solved via semidefinite programming after the reformulation of the objective function in a linear way. However, the computational complexity of the resulting problem is high, which makes it impractical for large-scale problems. A popular alternative is due to Burer and Monteiro [18, 12]:

$$\inf_{X \in \mathbb{R}^{n\times r}} \quad \sum_{i=1}^{m} \left(\langle A_i, XX^T\rangle - b_i\right)^2 \tag{2}$$

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

This nonlinear *Least-Squares* (LS) problem can be solved efficiently and on a large-scale with the Gauss-Newton method for instance. It has received a lot of attention recently due to the discovery in [30, 10] stating that the problem admits no spurious local minima (i.e. local minima that are not global minima) under certain conditions. These require adding a regularizer and satisfying the restricted isometry property (RIP) [20]. We raise the question of whether this also holds in the case of Laplacian noise, which is a better model to account for outliers in the data. The maximum likelihood estimate of M can be converted to the *Least-Absolute Value* (LAV) optimization problem

$$\inf_{X \in \mathbb{R}^{n \times r}} \quad \sum_{i=1}^{m} \left| \langle A_i, XX^T \rangle - b_i \right|. \tag{3}$$

The nonlinear problem can be solved efficiently using nonconvex methods (for some recent work, see [36]). For example, one may adopt the famous reformulation technique for converting ℓ_1 norms to linear functions subject to linear inequalities to cast the above problem as a smooth nonconvex quadratically-constrained quadratic program [13]. However, the analysis of this result has not been addressed in the literature - all ensuing papers (e.g. [29, 52, 8]) on matrix completion since the aforementioned discovery deal with smooth objective functions.

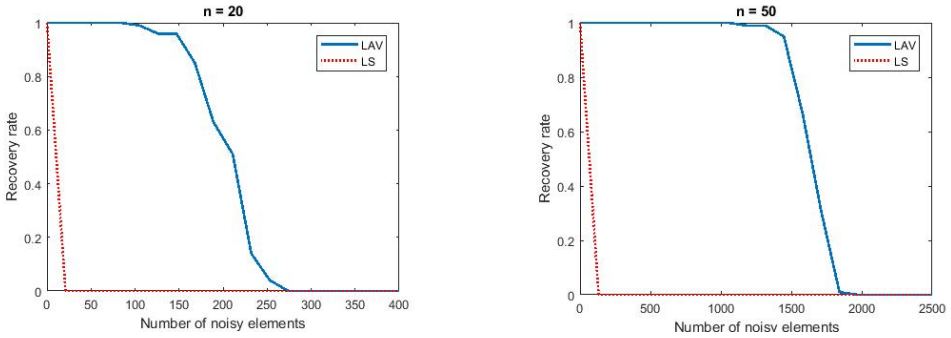
Consider $y \in \mathbb{R}^n$ and assume $r = 1$. On the one hand, in the fully observable case[1] with $M = yy^T$, the above nonconvex LS problem (2) consists in solving

$$\inf_{x \in \mathbb{R}^n} \sum_{i,j=1}^{n} (x_i x_j - y_i y_j - \epsilon_{i,j})^2 \tag{4}$$

for which there are no spurious local minima with high probability when $\epsilon_{i,j}$ are i.i.d. Gaussian variables [30]. On the other hand, in the full observable case, the LAV problem (3) aims to solve

$$\inf_{x \in \mathbb{R}^n} \sum_{i,j=1}^{n} |x_i x_j - y_i y_j - \epsilon_{i,j}|. \tag{5}$$

Although the LS problem has nice properties with Gaussian noise, we observe that stochastic gradient descent (SGD) fails to recover the matrix $M = yy^T$ in the presence of large but sparse noise. In contrast, SGD can perfectly recover the matrix by solving the LAV problem even when the sparse noise $\epsilon_{i,j}$ has a large amplitude. Figures 1a and 1b show our experiments for $n = 20$ and $n = 50$ with the number of noisy elements ranging from 0 to n^2. See Appendix 5.1 for our experiment settings.



(a) $n = 20$ (b) $n = 50$

Figure 1: Experiments with sparse noise

Upon this LAV formulation hinges the potential of nonconvex methods to cope with sparse noise and with Laplacian noise. There is no result on the analysis of the local solutions of this nonsmooth problem in the literature even for the noiseless case. This could be due to the fact that the optimality conditions for the smooth reformulated version of this problem in the form of quadratically-constrained quadratic program are highly nonlinear and lead to an exponential number of scenarios.

[1]This corresponds to the case where the sensing matrices $A_1, \ldots, A_{n^2}$ have all zeros terms apart from one element which is equal to 1.

As such, the goal of this paper is to prove the following proposition, which as the reader will see, is a significant hurdle. It addresses the matrix noiseless case and more generally the case of a tensor of order $d \in \mathbb{N}$.

***Proposition* 1.1.** *The function $f_1 : \mathbb{R}^n \longrightarrow \mathbb{R}$ defined as*

$$f_1(x) := \sum_{i_1,\ldots,i_d=1}^{n} |x_{i_1} \ldots x_{i_d} - y_{i_1} \ldots y_{i_d}| \tag{6}$$

has no spurious local minima.

A direct consequence of Proposition 1.1 is that one can perform the rank-one tensor decomposition by minimizing the function in Proposition 1.1 using a local search algorithm (e.g. [19]). Whenever the algorithm reaches a local minimum, it is a globally optimal solution leading to the desired decomposition. Existing proof techniques, e.g. [29, 30, 24, 38, 5, 4, 31, 35], are not directly useful for the analysis of the nonconvex and nonsmooth optimization problem stated above. In particular, results on the absence of spurious local minima neural networks with a Rectified Linear Unit (ReLU) activation function pertain to smooth objective functions (e.g. [48, 14]). The Clarke derivative [22, 23] provides valuable insight (see Lemma 3.1) but it is not conclusive. In order to pursue the proof (see Lemma 3.2), we propose the new notion of global function. Unlike the previous approaches, it does not require one to exhibit a direction of descent. After some successive transformations, we reduce the problem to a linear program. It is then obvious that there are no spurious local minima. Incidentally, global functions provide a far simpler and shorter proof to a slightly weaker result, that is to say, the absence of spurious strict local minima. It eschews the Clarke derivative all together and instead considers a sequence of converging differentiable functions that have no spurious local minima (see Proposition 3.1). In fact, this technique also applies if we substitute the ℓ_1 norm with the ℓ_∞ norm (see Proposition 3.2).

The paper is organized as follows. Global functions are examined in Section 2 and their application to tensor decomposition is discussed in Section 3. Section 4 concludes our work. The proofs may be found in the supplementary material (Section 5 of the supplementary material).

2 Notion of global function

Given an integer n, consider the Euclidian space $\mathbb{R}^n$ with norm $\|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2}$ along with a subset $S \subset \mathbb{R}^n$. The next two definitions are classical.

***Definition* 2.1.** *We say that $x \in S$ is a global minimum of $f : S \longrightarrow \mathbb{R}$ if for all $y \in S \setminus \{x\}$, it holds that $f(x) \leqslant f(y)$.*

***Definition* 2.2.** *We say that $x \in S$ is a local minimum (respectively, strict local minimum) of $f : S \longrightarrow \mathbb{R}$ if there exists $\epsilon > 0$ such that for all $y \in S \setminus \{x\}$ satisfying $\|x - y\|_2 \leqslant \epsilon$, it holds that $f(x) \leqslant f(y)$ (respectively, $f(x) < f(y)$).*

We introduce the notion of global functions below.

***Definition* 2.3.** *We say that $f : S \longrightarrow \mathbb{R}$ is a global function if it is continuous and its local minima are all global minima. Define $\mathcal{G}(S)$ as the set of all global functions on S.*

In the following, we compare global functions with other classes of functions in the literature, particularly those that seek to generalize convex functions.

When the domain S is convex, two important proper subsets of $\mathcal{G}(S)$ are the sets of convex and strict quasiconvex functions. Convex functions (respectively, strict quasiconvex [27, 26]) are such that $f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y)$ (respectively, $f(\lambda x + (1-\lambda)y) < \max\{f(x), f(y)\}$) for all $x, y \in S$ (with $x \neq y$) and $0 < \lambda < 1$. To see why these are proper subsets, notice that the cosine function on $[0, 4\pi]$ is a global function that is neither convex nor strict quasiconvex. In dimension one, global and strict quasiconvex functions are very closely related. Indeed, when the domain is convex and compact (i.e. an interval $[a, b]$ where $a, b \in \mathbb{R}$), it can be shown that a function is strict quasiconvex if and only if it is global and has a unique global minimum. However, this is not true in higher dimensions, as can be seen in Figure 4b in Appendix 5.2, or in the existing literature, i.e. in

[25] or in [9, Figure 1.1.10]. It is also not true in dimension one if we remove the assumption that the domain is compact (consider $f(x) := (x^2 + x^4)/(1 + x^4)$ defined on $\mathbb{R}$ and illustrated in Figure 4a in Appendix 5.2).

When the domain S is not necessarily convex, a proper subset of $\mathcal{G}(S)$ is the set of star-convex functions. For a star-convex function f, there exists $x \in S$ such that $f(\lambda x+(1-\lambda)y) \leqslant \lambda f(x)+(1-\lambda)f(y)$ for all $y \in S \setminus \{x\}$ and $0 < \lambda < 1$. Again, the cosinus function on $[0, 4\pi]$ is a global function that is not star-convex. Another interesting proper subset of $\mathcal{G}(S)$ is the set of functions for which, informally, given any point, there exists a strictly decreasing path from that point to a global minimum. This property is discussed in [47, P.1] (see also [28]) to study the landscape of loss functions of neural networks. Formally, the property is that for all $x \in S$ such that $f(x) > \inf_{y\in S} f(y)$, there exists a continuous function $g : [0, 1] \longrightarrow S$ such that $g(0) = x$, $g(1) \in \operatorname{argmin}\{f(y) \mid y \in S\}$, and $t \in [0, 1] \longmapsto f(g(t))$ is strictly decreasing (i.e. $f(g(t_1)) > f(g(t_2))$ if $0 \leqslant t_1 < t_2 \leqslant 1$). Not all global functions satisfy this property, as illustrated by the function in Figure 4a. For instance, there exists no strictly decreasing path from $x = -3$ to the global minimizer 0. However, in the funtion in Figure 4b in Appendix 5.2, there exists a strictly decreasing path from any point to the unique global minimizer. One could thus think that if S is compact, or if f is coercive, then one should always be able to find a strictly decreasing path. However, there need not exist a strictly decreasing path in general. Consider for example the function defined on $([-1, 1] \setminus \{0\}) \times [-1, 1]$ as follows

$$f(x_1, x_2) := \begin{cases} -4|x_1|^3(1 - x_2)\left(\sin\left(-\frac{1}{|x_1|}\right) + 1\right) & \text{if} \quad 0 \leqslant x_2 \leqslant 1, \\ \left\{12|x_1|^3\left(\sin\left(-\frac{1}{|x_1|}\right) + 1\right) - 2\right\} x_2^3 + & \\ \left\{20|x_1|^3\left(\sin\left(-\frac{1}{|x_1|}\right) + 1\right) - 3\right\} x_2^2 + & \text{if } -1 \leqslant x_2 < 0. \\ 4|x_1|^3\left(\sin\left(-\frac{1}{|x_1|}\right) + 1\right) x_2 - 4|x_1|^3\left(\sin\left(-\frac{1}{|x_1|}\right) + 1\right) & \end{cases}$$

The function and its differential can readily be extended continuously to $[-1, 1] \times [-1, 1]$. This is illustrated in Figure 6a in Appendix 5.2. This yields a smooth[2] global function for which there exists no strictly decreasing path from the point $x = (0, 1/2)$ to a global minimizer (i.e. any point in $[-1, 1] \times \{-1\}$). We find this to be rather counter-intuitive. To the best of our knowledge, no such function has been presented in past literature. Hestenes [32] considered the function defined on $[-1, 1] \times [-1, 1]$ by $f(x_1, x_2) := (x_2 - x_1^2)(x_2 - 4x_1^2)$ (see also [9, Figure 1.1.18]). It is a global function for which the point $x = (0, 0)$ (which is not a global minimizer) admits no direction of descent, i.e. $d \in \mathbb{R}^2$ such that $t \in [0, 1] \longmapsto f(x + td)$ is strictly decreasing. However, it does admit a strictly decreasing path to a global minimizer, i.e. $t \in [0, 1] \longmapsto (\frac{\sqrt{10}}{4}t, t^2)$, along which the objective equals $-\frac{9}{16}t^4$. This is unlike the function exhibited in Figure 6a. As a byproduct, our function shows that the generalization of quasiconvexity to non-convex domains described in [6, Chapter 9] is a proper subset of global functions. This generalization was proposed in [41] and further investigated in [7, 33, 34, 15, 16, 17]. It consists in replacing the segment used to define convexity and quasiconvexity by a continuous path.

Finally, we note that there exists a characterization of functions whose local minima are global, without requiring continuity as in global functions. It is based on a certain notion of continuity of sublevel sets, namely lower-semicontinuity of point-to-set mappings [51, Theorem 3.3]. We will see below that continuity is a key ingredient for obtaining our results. We do not require more regularity precisely because one of our goals is to study nonsmooth functions. Speaking of which, observe that global functions can be nowhere differentiable, contrary to convex functions [11, Theorems 2.1.2 and 2.5.1]. Consider for example the global function defined on $]0, 1[\times]0, 1[$ by $f(x_1, x_2) := |2x_2 - 1| \sum_{n=0}^{+\infty} s(2^n x_1)/2^n$ where $s(x) := \min_{n\in\mathbb{N}} |x - n|$ is the distance to nearest integer. For any fixed $x_2 \neq 0$, the function $x_1 \in [0, 1] \longrightarrow f(x_1, x_2)/|x_2|$ is the Takagi curve [45, 3, 37] which is nowhere differentiable. It can easily be deduced that the bivariate function is nowhere differentiable. This is illustrated in Figure 6b.

In the following, we review some of the properties of global functions. Their proofs can be found in the appendix. We begin by investigating the composition operation.

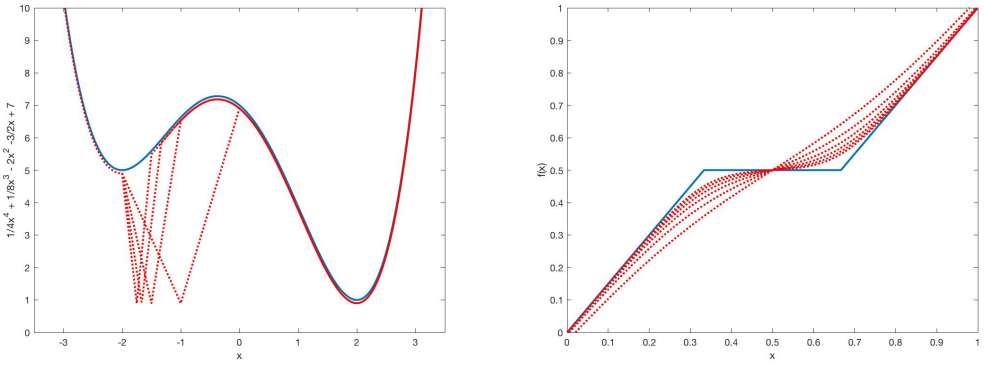
[2] In fact, one could make it infinitely differentiable by using the exponential function in the construction, but it is more cumbersome.

***Proposition* 2.1** (Composition of functions)**.** *Consider $f : S \longrightarrow \mathbb{R}$. Let $\phi : f(S) \longrightarrow \mathbb{R}$ denote a strictly increasing function where $f(S)$ is the range of f. It holds that $f \in \mathcal{G}(S)$ if and only if $\phi \circ f \in \mathcal{G}(S)$.*

However, the set of global functions is not closed under composition of functions in general. For instance, $f(x) := |x|$ and $g(x) := \max(-1, |x| - 2)$ are global functions on $\mathbb{R}$, but $f \circ g$ is not global function on $\mathbb{R}$.

***Proposition* 2.2** (Change of variables)**.** *Consider $f : S \longrightarrow \mathbb{R}$, a subset $S' \subset \mathbb{R}^n$, and a homeomorphism $\varphi : S \longrightarrow S'$ (i.e. continuous bijection with continuous inverse). It holds that $f \in \mathcal{G}(S)$ if and only if $f \circ \varphi^{-1} \in \mathcal{G}(S')$.*

Next, we consider what happens if we have a sequence of global functions. Figure 2a shows that the sequence of global functions (red dotted curves) pointwise converges to a function with a spurious local minimum (blue curve). Figure 2b shows that uniform convergence also does not preserve the property of being a global function: all points on the middle part of the limit function (blue curve) are spurious local minima. However, it suggests that uniform convergence preserves a slightly weaker property than being a global function. Intuitively, the limit should behave like a global function except that it may have "flat" parts. We next formalize this intuition. To do so, we consider the notions of global minimum, local minimum, and strict local minimum (Definition 2.1 and Definition 2.2), which apply to points in $\mathbb{R}^n$, and generalize them to subsets of $\mathbb{R}^n$. We will borrow the notion of neighborhood of a set (*uniform neighborhood* to be precise, see Definition 2.5).



(a) Pointwise convergence

(b) Uniform convergence

Figure 2: Convergence of a sequence of global functions

***Definition* 2.4.** *We say that a subset $X \subset S$ is a global minimum of $f : S \longrightarrow \mathbb{R}$ if $\inf_X f \leqslant \inf_{S \setminus X} f$.*

We note in passing the following two propositions. We will use them repeatedly in the next section. The proofs are omitted as they follow directly from the definitions.

***Proposition* 2.3.** *Assume that the following statements are true:*

1. *$X \subset S$ is a global minimum of f;*
2. *$f \in \mathcal{G}(X)$;*
3. *f does not have any local minima on $S \setminus X$.*

Then, $f \in \mathcal{G}(S)$.

Note that the first assumption is needed; otherwise the function may not be global because it could take a smaller value at a non local min outside X (possible when S is unbounded).

***Proposition* 2.4.** *If $f : S \longrightarrow \mathbb{R}$ is a global function on global minima $(X_\alpha)_{\alpha \in A}$ for some index set A, then it is a global function on $\bigcup_{\alpha \in A} X_\alpha$.*

We proceed to generalize the definition of local minimum.

***Definition* 2.5.** *We say that a compact subset $X \subset S$ is local minimum (respectively, strict local minimum) of $f : S \longrightarrow \mathbb{R}$ if there exists $\epsilon > 0$ such that for all $x \in X$ and for all $y \in S \setminus X$ satisfying $\|x - y\|_2 \leqslant \epsilon$, it holds that $f(x) \leqslant f(y)$ (respectively, $f(x) < f(y)$).*[3]

The above definitions are distinct from the notion of valley proposed in [47, Definition 1]. The latter is defined as a connected component[4] of a sublevel set (i.e. $\{x \in S \mid f(x) \leqslant \alpha\}$ for some $\alpha \in \mathbb{R}$). Local minima and strict local minima need not be valleys, and vice-versa. One may easily check that when the set is a point, i.e. $X = \{x\}$ with $x \in S$, the two definitions above are the same as the previous definitions of minimum (Definition 2.1 and Definition 2.2). They are therefore consistent. It turns out that the notion of global function (Definition 2.3) does not change when we interpret it in the sense of sets. We next verify this claim.

***Proposition* 2.5** (Consistency of Definition 2.3)**.** *Let $f : S \longrightarrow \mathbb{R}$ denote a continuous function. All local minima are global minima in the sense of points if only if all local minima are global minima in the sense of sets.*

We are ready to define a slightly weaker notion than being a global function.

***Definition* 2.6.** *We say that $f : S \longrightarrow \mathbb{R}$ is a weakly global function if it is continuous and if all strict local minima are global minima in the sense of sets.*

The generalization from points to sets in the definition of a minimum is justified here, as can be seen in Figure 3. All strict local minima are global minima in the sense of points. However, $X = [a, b]$ with $a \approx -2.6$ and $b = -1$ is a strict local minimum that is not a global minimum. Indeed, $\inf_X f = 6 > 1 = \inf_{\mathbb{R} \setminus X} f$. Hence, the function is not weakly global.

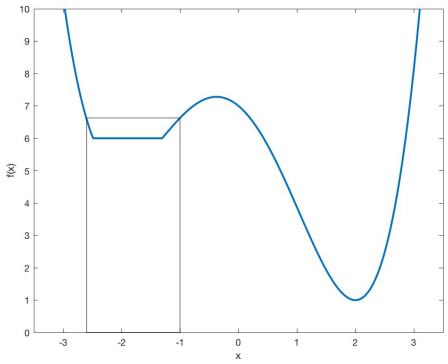


Figure 3: All strict local minima are global minima in the sense of points but not in the sense of sets.

We next make the link with the intuition regarding the flat part in Figure 2b.

***Proposition* 2.6.** *If $f : S \longrightarrow \mathbb{R}$ is a weakly global function, then it is constant on all local minima that are not global minima.*

We are interested in functions that are potentially defined on all of $\mathbb{R}^n$ (i.e. unconstrained optimization) or on subsets $S \subset \mathbb{R}^n$ that are not necessarily compact (i.e. general constrained optimization). We therefore need to borrow a slightly more general notion than uniform convergence [42, page 95, Section 3].

***Definition* 2.7.** *We say that a sequence of continuous functions $f_k : S \longrightarrow \mathbb{R}, k = 1, 2, \ldots,$ converges compactly towards $f : S \longrightarrow \mathbb{R}$ if for all compact subsets $K \subset S$, the restrictions of f_k to K converge uniformly towards the restriction of f to K.*

We are now ready to state a result regarding the convergence of a sequence of global functions and an important property that is preserved in the process.

[3] Note that the neighborhood of a compact set is always uniform.

[4] A subset $C \subset S$ is connected if it is not equal to the union of two disjoint nonempty closed subsets of S. A maximal connected subset (ordered by inclusion) of S is called a connected component.

***Proposition* 2.7** (Compact convergence)**.** *Consider a sequence of functions $(f_k)_{k\in\mathbb{N}}$ and a function f, all from $S \subset \mathbb{R}^n$ to $\mathbb{R}$. If*

$$f_k \longrightarrow f \quad \textit{compactly} \tag{7}$$

and if f_k are global functions on S, then f is a weakly global function on S.

Note that the proofs in this section are not valid if we replace the Euclidian space by an infinite-dimensional metric space. Indeed, we have implicitely used the fact that the unit ball is compact in order for the uniform neighborhood of a minimum to be compact.

3 Application to tensor decomposition

Global functions can be used to prove the following two significant results on nonlinear functions involving ℓ_1 norm and ℓ_∞ norm, as explained below.

***Proposition* 3.1.** *The function $f_1 : \mathbb{R}^n \longrightarrow \mathbb{R}$ defined as*

$$f_1(x) := \sum_{i_1,\dots,i_d=1}^{n} |x_{i_1} \dots x_{i_d} - y_{i_1} \dots y_{i_d}| \tag{8}$$

is a weakly global function; in particular, it has no spurious strict local minima.

Proof. The functions

$$f_p(x) := \sum_{i_1,\dots,i_d=1}^{n} |x_{i_1} \dots x_{i_d} - y_{i_1} \dots y_{i_d}|^p \tag{9}$$

for $p \longrightarrow 1$ with $p > 1$ form a set of global functions that converge compactly towards the function f_1. This is illustrated in Figure 5 in Appendix 5.2 for $n = d = 2$ and $y = (1, -3/4)$. The desired result then follows from Proposition 2.7. To see why each f_p is a global function, observe that f_p is differentiable with the first-order optimality condition as follows:

$$\sum_{i_1,\dots,i_{d-1}=1}^{n} x_{i_1} \dots x_{i_{d-1}} (x_{i_1} \dots x_{i_{d-1}} x_i - y_{i_1} \dots y_{i_{d-1}} y_i) |x_{i_1} \dots x_{i_{d-1}} x_i - y_{i_1} \dots y_{i_{d-1}} y_i|^{p-2} = 0$$

for all $i \in \{1, \dots, n\}$. Note that each term in the sum converges towards zero if the expression inside the absolute value converges towards zero, so that the equation is well-defined. Consider a local minimum $x \in \mathbb{R}^n$; then, x must satisfy the above first-order optimality condition. If $y_i = 0$, then the above equation readily yields $x_i = 0$. This reduces the problem dimension from n variables to $n - 1$ variables, so without loss of generality we may assume that $y_i \neq 0,\ i = 1, \dots, m$. After a division, observe that the following equation is satisfied

$$\sum_{i_1,\dots,i_{d-1}=1}^{n} |y_{i_1} \dots y_{i_{d-1}}|^p \, \frac{x_{i_1} \dots x_{i_{d-1}}}{y_{i_1} \dots y_{i_{d-1}}} \left(\frac{x_{i_1} \dots x_{i_{d-1}}}{y_{i_1} \dots y_{i_{d-1}}} t - 1 \right) \left| \frac{x_{i_1} \dots x_{i_{d-1}}}{y_{i_1} \dots y_{i_{d-1}}} t - 1 \right|^{p-2} = 0$$

for all $t \in \{x_1/y_1, \dots, x_n/y_n\}$. Each term with $x_{i_1} \dots x_{i_{d-1}} \neq 0$ in the above sum is a strictly increasing function of $t \in \mathbb{R}$ since it is the derivative of the strictly convex function

$$g(t) = |x_{i_1} \dots x_{i_{d-1}} t - y_{i_1} \dots y_{i_{d-1}}|^p. \tag{10}$$

The point $x = 0$ is not a local minimum (y is a direction of descent of f_p at 0), and thus $x \neq 0$. As a result, the above sum is a strictly increasing function of $t \in \mathbb{R}$. Hence, it has at most one root, that is to say $t = x_1/y_1 = \dots = x_n/y_n$. Plugging in, we find that $t^d = 1$. If d is odd, then $x = y$ and if d is even, then $x = \pm y$. To conclude, any local minimum x is a global minimum of f_p. □

***Proposition* 3.2.** *$f_\infty : \mathbb{R}^n \longrightarrow \mathbb{R}$ defined as*

$$f_\infty(x) := \max_{1 \leqslant i_1,\dots,i_d \leqslant n} |x_{i_1} \dots x_{i_d} - y_{i_1} \dots y_{i_d}| \tag{11}$$

is a weakly global function; in particular, it has no spurious strict local minima.

Proof. The functions $h_p(x) := \left(\sum_{i_1,\ldots,i_d=1}^{n} |x_{i_1} \ldots x_{i_d} - y_{i_1} \ldots y_{i_d}|^p \right)^{\frac{1}{p}}$ for $p \longrightarrow +\infty$ form a set of global functions that converge compactly towards the function f_∞. We know that each h_p is a global function by applying Proposition 2.1 to (9) with the fact that $(\cdot)^{\frac{1}{p}}$ is increasing for nonnegative arguments. □

Note that the functions in Proposition 3.1 and Proposition 3.2 are *a priori* utterly different, yet both proofs are essentially the same. This highlights the usefulness of the new notion of global functions.

***Remark* 3.1.** *The notion of weakly global functions explains that one can perform tensor decomposition by minimizing the nonconvex and nonsmooth functions in Proposition 3.1 and Proposition 3.2 with a local search algorithm. Whenever the algorithm reports a strict local minimum, it is a globally optimal solution.*

In order to strengthen the conclusion in Proposition 3.1 and to establish the absence of spurious local minima, we propose the following two lemmas. Using Proposition 2.3 and these two lemmas, we arrive at the stronger result stated in Proposition 1.1.

***Lemma* 3.1.** *If $x \in \mathbb{R}^n$ is a first-order stationary point of f_1 in the sense of the Clarke derivative, then the following statements hold:*

1. *If $y_i = 0$ for some $i \in \{1, \ldots, n\}$, then $x_i = 0$;*
2. *For all $i_1, \ldots, i_d \in \{1, \ldots, n\}$, it holds that $\frac{x_{i_1} \ldots x_{i_d}}{y_{i_1} \ldots y_{i_d}} \leqslant 1$.*

Proof. Similar in spirit to the proof of Proposition 3.1, the ratios $t \in \{x_1/y_1, \ldots, x_n/y_n\}$ for a first-order stationary point must all be the roots of an increasing (set-valued) "staircase function". We then obtain the results by analyzing the relation between the roots and the jump points of the staircase function. See Appendix 5.8 for the complete proof. □

Note that the above lemma only uses the first-order optimality condition (in the sense of Clarke derivative) without any direction of decent.

***Remark* 3.2.** *One cannot show that there are no spurious local minima with only the first-order optimality condition (in the Clarke derivative sense). In fact, any $x \in \mathbb{R}^n$ satisfying $\sum_{i=1}^{n} |y_i| \frac{x_i}{y_i} = 0$ and $\frac{x_{i_1} \ldots x_{i_d}}{y_{i_1} \ldots y_{i_d}} \leqslant 1$ for all $i_1, \ldots, i_d \in \{1, \ldots, n\}$, is a first-order stationary point, but is not a local minimum.*

***Lemma* 3.2.** *If $y_1 \ldots y_n \neq 0$, define the set*

$$S := \left\{ x \in \mathbb{R}^n \;\middle|\; \frac{x_{i_1} \ldots x_{i_d}}{y_{i_1} \ldots y_{i_d}} \leqslant 1 \,, \quad \forall\, i_1, \ldots, i_d \in \{1, \ldots, n\} \right\}. \tag{12}$$

Then, $f_1 \in \mathcal{G}(S)$.

Proof. We provide a sketch here, and the complete proof is deferred to Appendix 5.9. The objective function on S is equal to $f_1(x) = \left(\sum_{i=1}^{n} |y_i| \right)^d - \left(\sum_{i=1}^{n} |y_i| \frac{x_i}{y_i} \right)^d$. Define the set $S' := \{\, x \in \mathbb{R}^n \mid x_{i_1} \ldots x_{i_d} \leqslant 1 \,, \quad \forall\, i_1, \ldots, i_d \in \{1, \ldots, n\} \,\}$. When d is an odd number, the composition and change of variables properties of global functions (Propositions 2.1 and 2.2) imply that f_1 is a global function on S if and only if $f_{\text{odd}}(x) = -\sum_{i=1}^{n} |y_i| x_i \in \mathcal{G}(S')$. Similarly, when d is an even number, f is a global function if and only if $f_{\text{even}}(x) = -\left(\sum_{i=1}^{n} |y_i| x_i\right)^2 \in \mathcal{G}(S')$. For the case when d is odd, we apply the Karush-Kuhn-Tucker conditions to restrict attention to the positive orthant and conclude by showing its association with a linear program. For the case when d is even, we divide the set S' into two subsets: $S' \cap \{x | \sum_{i=1}^{n} |y_i| x_i \geq 0\}$ and $S' \cap \{x | \sum_{i=1}^{n} |y_i| x_i \leq 0\}$. Observe that $f_{\text{even}}(x)$ is a global function on each of the subset by associating each subset with a linear program. Then, Proposition 2.3 establishes the result. □

The two previous lemmas prove Proposition 1.1; the notion of global function is used to the prove the latter.

4 Conclusion

Nonconvex optimization appears in many applications, such as matrix completion/sensing, tensor recovery/decomposition, and training of neural networks. For a general nonconvex function, a local search algorithm may become stuck at a local minimum that is arbitrarily worse than a global minimum. We develop a new notion of global functions for which all local minima are global minima. Using certain properties of global functions, we show that the set of these functions include a class of nonconvex and nonsmooth functions that arise in matrix completion/sensing and tensor recovery/decomposition with Laplacian noise. This paper offers a new mathematical technique for the analysis of nonconvex and nonsmooth functions such as those involving ℓ_1 norm and ℓ_∞ norm.

Acknowledgments

This work was supported by the ONR Awards N00014-17-1-2933 ONR and N00014-18-1-2526, NSF Award 1808859, DARPA Award D16AP00002, and AFOSR Award FA9550- 17-1-0163. We wish to thank the anonymous reviewers for their valuable feedback, as well as Chris Dock for fruitful discussions.

References

[1] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *https://arxiv.org/abs/1310.7991*, 2014.

[2] A. Agarwal, A. Anandkumar, and P. Netrapalli. A Clustering Approach to Learn Sparsely-Used Overcomplete Dictionaries. *https://arxiv.org/abs/1309.1952*, 2014.

[3] P. C. Allaart and K. Kawamura. The Takagi function: a survey. *https://arxiv.org/pdf/1110.1691.pdf*, 2011.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *https://arxiv.org/pdf/1210.7559.pdf*, 2014.

[5] A. Anandkumar, R. Ge, and M. Janzamin. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *https://arxiv.org/pdf/1402.5180.pdf*, 2015.

[6] M. Avriel, Walter E. Diewert, S. Schaible, and I. Zang. *Generalized Concavity*. Plenum Press, New York, 1988.

[7] M. Avriel and I. Zang. Generalized arcwise connected function and characterizations of local-global minimum properties. *J. Opt. Theory and Appl.*, 32:407–425, 1980.

[8] M.-F. Balcan, Y. Liang, D. P. Woodruff, and H. Zhang. Matrix Completion and Related Problems via Strong Duality. *https://arxiv.org/abs/1704.08683*, 2017.

[9] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2016.

[10] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global Optimality of Local Search for Low Rank Matrix Recovery. *https://arxiv.org/abs/1605.07221*, 2016.

[11] J. M. Borwein and J. D. Vanderwerff. *Convex functions: constructions, characterizations and counterexamples*. Cambridge University Press, 2010.

[12] N. Boumal, V. Voroninski, and A. S. Bandeira. Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs. *https://arxiv.org/abs/1804.02008*, 2018.

[13] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2009.

[14] A. Brutzkus and A. Globerson. Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs. *https://arxiv.org/pdf/1702.07966.pdf*, 2017.

[15] P. Burai. Local-global minimum property in unconstrained minimization problems. *J. Opt. Theory and Appl.*, 162:34–46, 2013.

[16] P. Burai. Necessary and sufficient condition on global optimality without convexity and second order differentiability. *Springer Opt. Letters*, 7:903–911, 2013.

[17] P. Burai. Convexity with respect to families of sections and lines and their application in optimization. *J. of Global Opt.*, 64:649–662, 2016.

[18] S. Burer and R. D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95:329–357, 2003.

[19] J. V. Burke, A. S. Lewis, and M. L. Overton. A Robust Gradient Sampling Algorithm for Nonsmooth, Nonconvex Optimization. *SIAM J. Optim.*, 15:751–779, 2005.

[20] E. J. Candes and T. Tao. Decoding by Linear Programming. *IEEE Trans. Inf. Th.*, 51:4203–4215, 2005.

[21] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval. *https://arxiv.org/abs/1803.07726*, 2018.

[22] F. H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 1975.

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

[24] L. Ding and Y. Chen. The Leave-one-out Approach for Matrix Completion: Primal and Dual Analysis. *https://arxiv.org/abs/1803.07554*, 2018.

[25] J. C. Dunn. On the convergence of projected gradient processes to singular critical points. *J. Opt. Theory and Appl.*, 55:203–216, 1987.

[26] W. Fenchel. Convex cones, sets and functions. *Mimeographed lecture notes, Princeton University, New Jersey*, 1951.

[27] J. A. De Finetti. Sulle stratificazioni convesse. *Ann. Math. Pura Appl.*, 30:173–183, 1949.

[28] C. D. Freeman and J. Bruna. Topology and geometry of half-rectified network optimization. *ICLR 2017*, 2017.

[29] R. Ge, C. Jin, and Y. Zheng. No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis. *https://arxiv.org/pdf/1704.00708.pdf*, 2017.

[30] R. Ge, J. D. Lee, and T. Ma. Matrix Completion has No Spurious Local Minimum. *NIPS*, 2016.

[31] R. Ge and T. Ma. On the Optimization Landscape of Tensor Decompositions. *NIPS*, 2016.

[32] M. R. Hestenes. *Optimization Theory: The Finite Dimensional Case*. Wiley, N.Y., 1975.

[33] R. Horst. A note on functions whose local minima are global. *J. Opt. Theory and Appl.*, 36:457–463, 1982.

[34] R. Horst. A topolical property of limes-arciwise strictly quasiconvex functions. *J. Math. Anal. and Appl.*, 134:426–430, 1988.

[35] P. Jain and S. Oh. Provable Tensor Factorization with Missing Data. *NIPS*, 2014.

[36] K. Khamaru and M. J. Wainwright. Convergence guarantees for a class of non-convex and non-smooth optimization problems . *https://arxiv.org/pdf/1804.09629.pdf*, 2018.

[37] J. C. Lagarias. The Takagi function and its properties. *RIMS Kokyuroku Bessatsu*, pages 153–189, 2012.

[38] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex Matrix Factorization from Rank-One Measurements. *https://arxiv.org/abs/1802.06286*, 2018.

[39] H. Lu and K. Kawaguchi. Depth Creates No Bad Local Minima. *https://arxiv.org/pdf/1702.08580.pdf*, 2017.

[40] S. Mei, Y. Bai, and A. Montanari. The Landscape of Empirical Risk for Non-convex Losses. *https://arxiv.org/abs/1607.06534*, 2016.

[41] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, 1970.

[42] R. Remmert. *Theory of Complex Functions*. Graduate Texts in Mathematics, Springer, 1991.

[43] H. Sedghi, M. Janzamin, and A. Anandkumar. Provable Tensor Methods for Learning Mixtures of Generalized Linear Models. *https://arxiv.org/abs/1412.3046*, 2016.

[44] M. Soltanolkotabi. Learning ReLUs via Gradient Descent. *http://arxiv.org/abs/1705.04591*, 2017.

[45] T. Takagi. A simple example of the continuous function without derivate. *Tokyo Sugaku-Butsurigakkwai Hokoku*, 1:176–177, 1901.

[46] N. Vaswani, S. Nayer, and Y. C. Eldar. Low Rank Phase Retrieval. *https://arxiv.org/abs/1608.04141*, 2017.

[47] L. Venturi, A. S. Bandeira, and J. Bruna. Neural Networks with Finite Intrinsic Dimension have no Spurious Valleys. *https://arxiv.org/pdf/1802.06384.pdf*, 2018.

[48] C. Wu, J. Luo, and J. D. Lee. No Spurious Local Minima in a Two Hidden Unit ReLU Network. *ICLR*, 2018.

[49] X. Yi, C. Caramanis, and S. Sanghavi. Solving a Mixture of Many Random Linear Equations by Tensor Decomposition and Alternating Minimization. *https://arxiv.org/abs/1608.05749*, 2016.

[50] C. Yun, S. Sra, and A. Jadbabaie. Global optimality conditions for deep neural networks. *https://arxiv.org/pdf/1707.02444.pdf*, 2018.

[51] I. Zang and M. Avriel. On functions whose local minima are global. *J. Opt. Theory and Appl.*, 16:183–190, 1975.

[52] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. The Global Optimization Geometry of Low-Rank Matrix Optimization. *https://arxiv.org/pdf/1703.01256.pdf*, 2018.

5 Appendix

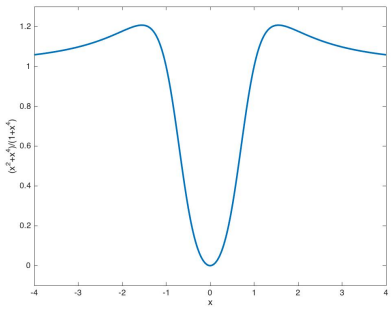
5.1 Experiment settings

We use SGD to solve the problems (4) and (5) for randomly generated rank-one matrices. In the experiments, each y is generated according to the n-dimensional i.i.d. standard Gaussian distribution. The positions of the sparse noise are uniformly selected from all the n^2 elements, and each noisy element is replaced by a Gaussian random variable with standard deviation 10. With regard to SGD, we set the learning rate to 0.001 and momentum to 0.9. The initial point is a Gaussian random vector.

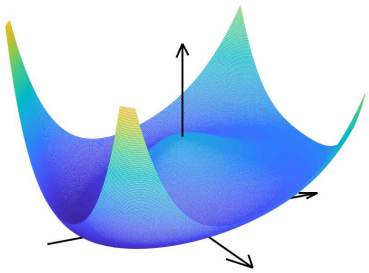
In our experiments, a successful recovery means that the solution x has a relative error less than 0.1 compared with the optimal solution y. We consider $n = 20$ and $n = 50$ and vary the number of noisy elements from 0 to n^2. For each case, we run 100 experiments and report the successful recovery rate. As shown in Figures (1a) and (1b), the LS problem (4) fails to recover the matrix except for the noiseless case. On the other hand, the LAV problem (4) provides perfect recovery in the presence of sparse noise.

5.2 Illustrations

This section is composed of Figure 4, Figure 5, and Figure 6.



(a) Global function on $\mathbb{R}$



(b) Global function on $\mathbb{R}^2$

Figure 4: Examples of global functions

Concerning Figure 6b, first note that the path along $x_1 = 0$ is not a strictly decreasing path. Any other path from $(0, 0.5)$ must include a segment of $(0, 0.5)$ to $(\epsilon_1, 0.5 + \epsilon_2)$ for some small ϵ_1, ϵ_2 with $\epsilon_1 \neq 0$. Then this segment is not strictly decreasing due to the alternating behavior of $\sin(-\frac{1}{|x_1|})$ when $x_1 = \epsilon_1 \neq 0$.

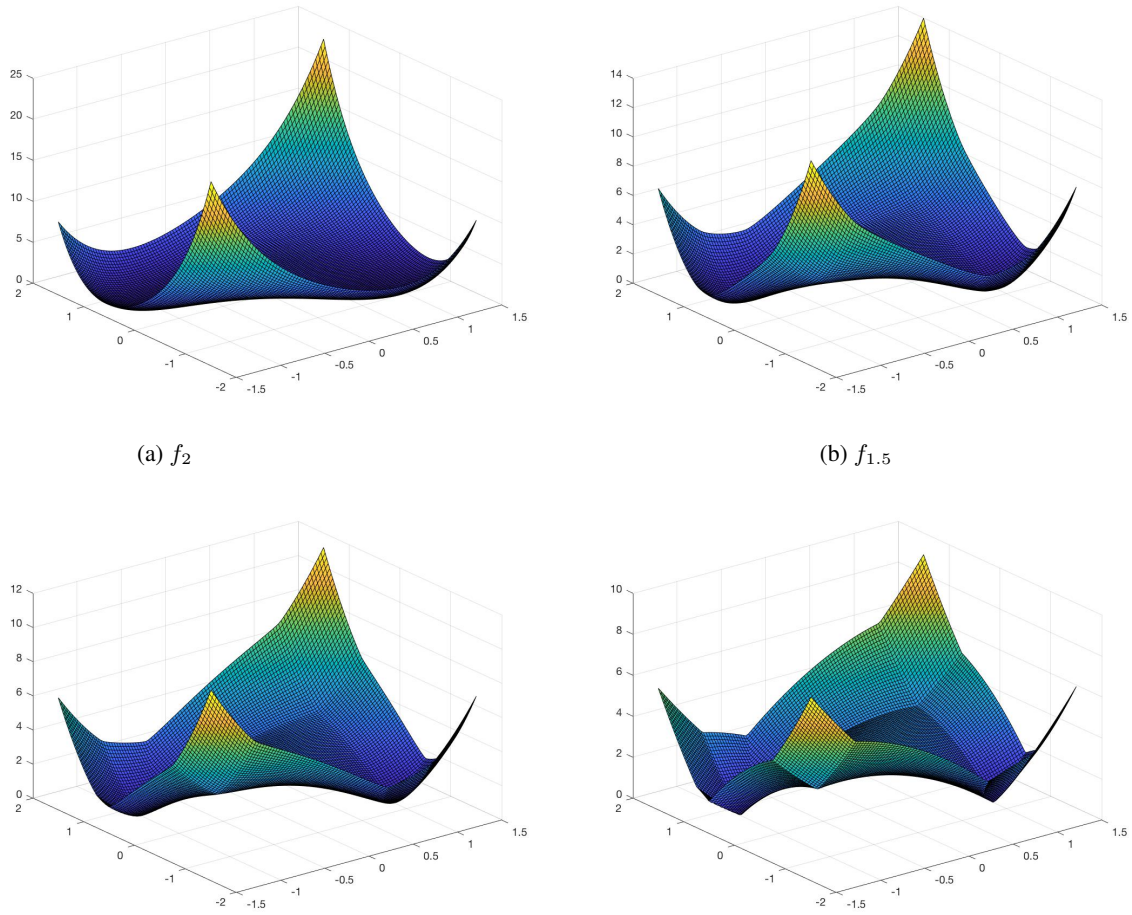
5.3 Proof of Proposition 2.1

($\Longrightarrow$) Let $x \in S$ denote a local minimum of $\phi \circ f$. There exists $\epsilon > 0$ such that $\phi(f(x)) \leqslant \phi(f(y))$ for all $y \in S \setminus \{x\}$ with $\|x - y\|_2 \leqslant \epsilon$. Since ϕ is increasing, it holds that $f(x) \leqslant f(y)$. Since f is global, we deduce that x is a global minimum of f, that is to say $f(x) \leqslant f(y)$ for all $y \in S \setminus \{x\}$. Since ϕ is increasing, it holds that $\phi(f(x)) \leqslant \phi(f(y))$ for all $y \in S \setminus \{x\}$. We conclude that x is a global minimum of $\phi \circ f$.

($\Longleftarrow$) Simply apply the previous argument to $\phi^{-1} \circ (\phi \circ f)$, where ϕ^{-1} denotes the inverse of $\phi : f(S) \longrightarrow \phi \circ f(S)$.

5.4 Proof of Proposition 2.2

($\Longrightarrow$) Let $x' \in S'$ denote a local minimum of $f \circ \varphi^{-1}$. There exists $\epsilon' > 0$ such that $f(\varphi^{-1}(x')) \leqslant f(\varphi^{-1}(y'))$ for all $y' \in S' \setminus \{x'\}$ with $\|x' - y'\|_2 \leqslant \epsilon'$. Since φ is continuous, there exists $\epsilon > 0$ such that $f(\varphi^{-1}(x')) \leqslant f(y)$ for all $y \in S \setminus \{\varphi^{-1}(x')\}$ with $\|\varphi^{-1}(x') - y\|_2 \leqslant \epsilon$. Hence, $\varphi^{-1}(x')$



(a) f_2 (b) $f_{1.5}$

(c) $f_{1.25}$ (d) f_1

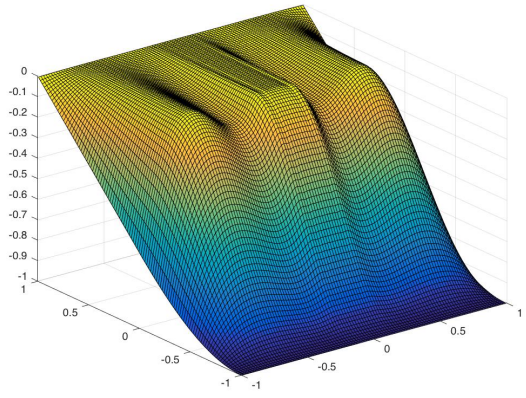
Figure 5: Compact convergence of global functions implies that strict local minima are global

is a local minimum of f. Since f is global, it holds that $f(\varphi^{-1}(x')) \leqslant f(y)$ for all $y \in S$. Since φ is a bijection, $f(\varphi^{-1}(x')) \leqslant f(\varphi^{-1}(y'))$ for all $y' \in S'$, implying that x' is a global minimum of $f \circ \varphi^{-1}$.

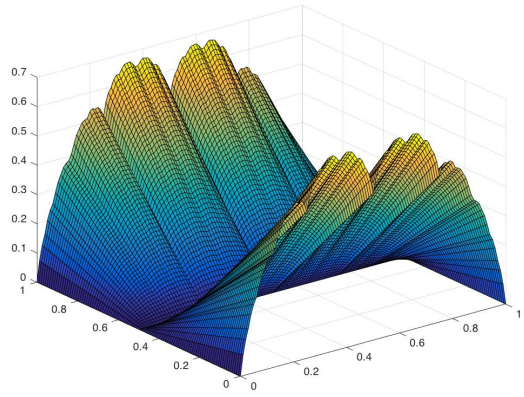
($\Longleftarrow$) Simply apply the previous argument to $(f \circ \varphi^{-1}) \circ \varphi$.

5.5 Proof of Proposition 2.5

One direction is obvious. For the other direction, we propose a proof by contrapositive. Let $X \subset S$ denote a local minimum that is not a global minimum. There exists $\epsilon > 0$ such that the uniform neighborhood $V := \{y \in S \mid \exists\, x \in X : \|x - y\|_2 \leqslant \epsilon\}$ satisfies $f(x) \leqslant f(y)$ for all $x \in X$ and for all $y \in V \setminus X$. Also, there exists $z \in S \setminus V$ such that $f(z) < f(x)$ for all $x \in V$. Since f is continuous on the compact set V, it attains a minimum $x' \in V$ such that $f(z) < f(x')$. If $x' \in X$, then for all $y \in S$ such that $\|x' - y\|_2 \leqslant \epsilon$, it holds that $f(z) < f(x') \leqslant f(y)$. Thus, x' is local minimum that is not a global minimum. If $x' \in V \setminus X$, then $f(x') \leqslant f(x) \leqslant f(x')$ for all $x \in X$. Consider a point $x \in X$. For all $y \in S$ such that $\|x - y\|_2 \leqslant \epsilon$, it holds that $f(x) = f(y)$ if $y \in X$ and $f(x) \leqslant f(y)$ if $y \notin X$. Together with the fact that $f(z) < f(x') = f(x)$, we deduce that x is a local minimum that is not a global minimum.



(a) Global function devoid of a strictly decreasing path from $(0, 1/2)$ to a global minimizer



(b) Global function that is nowhere differentiable

Figure 6: Notable examples (with x_1-axis on the right and x_2-axis on the left)

5.6 Proof of Proposition 2.6

We propose a proof by contrapositive. Assume that f is not constant on a local minimum $X \subset S$ that is not a global minimum. The minimum X admits a uniform neighborhood V such that $f(x) \leqslant f(y)$ for all $x \in X$ and for all $y \in V \setminus X$. Since f is continuous on the compact set V, there exists $x' \in V$ such that $f(x') \leqslant f(x)$ for all $x \in V$. If $x' \in V \setminus \text{int}(X)$ where "int" stands for interior, then f is constant on X because X is a local minimum. Therefore, $x' \in \text{int}(X)$ and $f(x') < f(x)$ for all $x \in \partial X := X \setminus \text{int}(X)$. Consider the compact set defined by $X' := \{x \in X \mid f(x') = f(x)\}$. The set V satisfies $f(x) < f(y)$ for all $x \in X'$ and $y \in V \setminus X'$. Since $X' \subset X$, there exists a uniform neighborhood V' of X' satisfying $f(x) < f(y)$ for all $x \in X'$ and for all $y \in V' \setminus X'$. Hence, X' is a strict local minimum that is not global. To conclude, f is not a weakly global function.

5.7 Proof of Proposition 2.7

Consider a sequence of global functions f_k that converge compactly towards f. Since $S \subset \mathbb{R}^n$ and $\mathbb{R}^n$ is a compactly generated space, it follows that f is continuous. We proceed to prove that f is a weakly global function by contradiction. Suppose $X \subset S$ is a strict local minimum that is not global minimum. There exists $\epsilon > 0$ such that the uniform neighborhood $V := \{y \in S \mid \exists\, x \in X : \|x - y\|_2 \leqslant \epsilon\}$ satisfies $f(x) < f(y)$ for all $x \in X$ and for all $y \in V \setminus X$. Since f is continuous on the compact set X, it attains a minimal value on it, say $\inf_X f := \alpha + \inf_S f$ where $\alpha > 0$ since X is not a global minimum. Consider a compact set $V \subset K \subset S$ such that $\inf_K f \leqslant \alpha/2 + \inf_S f$. Since f is continuous on the compact set ∂V, it attains a minimal value on it, say $\inf_{\partial V} f := \beta + \inf_X f$ where $\beta > 0$ by strict optimality. Let $\gamma := \min\{\alpha/2, \beta\}$. For a sufficiently large value of k, compact convergence implies that $|f_k(y) - f(y)| \leqslant \gamma/3$ for all $y \in K$. Since the function f_k is compact on V, it attains a minimum, say $z \in V$. Therefore,

$$f_k(z) \;\leqslant\; \gamma/3 + \inf_V f \;\leqslant\; \beta/3 + \inf_V f \;<\; 2\beta/3 + \inf_V f \tag{13}$$

$$\leqslant\; -\gamma/3 + \beta + \inf_V f \;\leqslant\; -\gamma/3 + \inf_{\partial V} f \;\leqslant\; \inf_{\partial V} f_k. \tag{14}$$

Thus, $z \in \text{int}(V)$. We now proceed to show by contradiction that z is a local minimum of f_k. Assume that for all $\epsilon' > 0$, there exists $y' \in S \setminus \{x\}$ satisfying $\|x - y'\|_2 \leqslant \epsilon'$ such that $f_k(z) > f_k(y')$. We can choose ϵ' small enough to guarantee that y' belongs to V since $z \in \text{int}(V)$. The point y' then contradicts the minimality of z on V. This means that $z \in V$ is a local minimum of f_k. Now, observe that

$$\inf_K f_k \;\leqslant\; \gamma/3 + \inf_K f \;\leqslant\; \gamma/3 + \alpha/2 + \inf_S f \;\leqslant\; 2\alpha/3 + \inf_S f \;<\; 5\alpha/6 + \inf_S f \tag{15}$$

$$\leqslant \alpha - \gamma/3 + \inf_S f \;=\; -\gamma/3 + \inf_X f \;=\; -\gamma/3 + \inf_V f \;\leqslant\; \inf_V f_k \;\leqslant\; f_k(z). \tag{16}$$

Thus, z is not a global minimum of f_k. This contradicts the fact that f_k is a global function.

5.8 Proof of Lemma 3.1

Based on the Clarke derivative [22, 23] for locally Lipschitz functions[5], the first-order optimality condition reads

$$0 \in \sum_{i_1,\ldots,i_{d-1}=1}^{n} x_{i_1}\ldots x_{i_{d-1}}\,\text{sign}(x_{i_1}\ldots x_{i_{d-1}}x_i - y_{i_1}\ldots y_{i_{d-1}}y_i) \quad , \quad i=1,\ldots,n \tag{17}$$

where

$$\text{sign}(x) := \begin{cases} -1 & \text{if } x<0,\\ [-1,1] & \text{if } x=0,\\ 1 & \text{if } x>0. \end{cases} \tag{18}$$

If $y_i = 0$ for some $i \in \{1,\ldots,n\}$, then the above equations readily yield

$$0 \in \sum_{i_1,\ldots,i_{d-1}=1}^{n} x_{i_1}\ldots x_{i_{d-1}}\,\text{sign}(x_{i_1}\ldots x_{i_{d-1}}x_i) = \text{sign}(x_i)\sum_{i_1,\ldots,i_{d-1}=1}^{n}|x_{i_1}\ldots x_{i_{d-1}}| \tag{19}$$

which implies $x_i = 0$. This reduces the dimension of the problem from n to $n-1$, so without loss of generality we may assume that $y_i \neq 0$ for all $i = 1,\ldots,n$. After a division, observe that the following inclusion is satisfied:

$$0 \in \sum_{i_1,\ldots,i_{d-1}=1}^{n} |y_{i_1}\ldots y_{i_{d-1}}|\frac{x_{i_1}\ldots x_{i_{d-1}}}{y_{i_1}\ldots y_{i_{d-1}}}\text{sign}\left(\frac{x_{i_1}\ldots x_{i_{d-1}}}{y_{i_1}\ldots y_{i_{d-1}}}t-1\right) \tag{20}$$

for all $t \in \{x_1/y_1,\ldots,x_n/y_n\}$. Each term with $x_{i_1}\ldots x_{i_{d-1}} \neq 0$ in the above sum is an increasing step (set-valued) function of $t \in \mathbb{R}$ since it is the Clarke derivative of the convex function

$$g(t) = |x_{i_1}\ldots x_{i_{d-1}}t - y_{i_1}\ldots y_{i_{d-1}}|. \tag{21}$$

The above sum is thus a increasing step function of $t \in \mathbb{R}$. Hence, the roots $x_1/y_1,\ldots,x_n/y_n$ all along belong to the same step. Jumps between the steps occur exactly at the following set of points:

$$\left\{ \frac{y_{i_1}\cdots y_{i_{d-1}}}{x_{i_1}\ldots x_{i_{d-1}}} \;\middle|\; i_1,\ldots,i_{d-1}\in\{1,\ldots,n\} \text{ and } x_{i_1}\ldots x_{i_{d-1}}\neq 0 \right\} \tag{22}$$

This set is empty when $x = 0$; otherwise, none of its elements are equal to zero because $y \neq 0$. Given a jump point $\alpha \neq 0$ in the above set, the roots must therefore be all before or all after, that is to say:

$$\frac{x_1}{y_1},\ldots,\frac{x_n}{y_n} \leqslant \alpha \qquad \text{or} \qquad \alpha \leqslant \frac{x_1}{y_1},\ldots,\frac{x_n}{y_n}. \tag{23}$$

We next prove that

$$\alpha>0 \implies \frac{x_1}{y_1},\ldots,\frac{x_n}{y_n} \leqslant \alpha \qquad \text{and} \qquad \alpha<0 \implies \alpha \leqslant \frac{x_1}{y_1},\ldots,\frac{x_n}{y_n}. \tag{24}$$

Let us prove the first implication by contradiction. Assume that there exists $k \in \{1,\ldots,n\}$ such that $\alpha < x_k/y_k$. Since one root is after the jump point α, all other roots are after the jump point α. In particular, for all $i \in \{1,\ldots,n\}$, we have

$$0 < \alpha := \frac{y_{i_1}\cdots y_{i_{d-1}}}{x_{i_1}\ldots x_{i_{d-1}}} \leqslant \frac{x_i}{y_i} \tag{25}$$

[5]For a locally Lipschitz function $f:\mathbb{R}^n \longrightarrow \mathbb{R}$, the directional derivative evaluated at x in the direction v is defined by

$$f^\circ(x;v) := \limsup_{\substack{y\to x\\ \lambda\downarrow 0}} \frac{f(y+\lambda v)-f(y)}{\lambda}$$

and the Clarke derivative is defined by $\partial f(x) := \{d \in \mathbb{R}^n : f^\circ(x;v) \geqslant \langle v,d\rangle, \quad \forall v \in \mathbb{R}^n\}$.

Therefore, all the roots are positive. By multiplying the above equation by the positive number $\frac{x_{i_1} y_i}{y_{i_1} x_i}$, we obtain

$$\frac{y_i y_{i_2} \dots y_{i_{d-1}}}{x_i x_{i_2} \dots x_{i_{d-1}}} \leqslant \frac{x_{i_1}}{y_{i_1}}. \tag{26}$$

Note that the left-hand side is a jump point, and the right-hand side is a root. Therefore, all the roots are after, and in particular:

$$\frac{y_i y_{i_2} \dots y_{i_{d-1}}}{x_i x_{i_2} \dots x_{i_{d-1}}} \leqslant \frac{x_i}{y_i}. \tag{27}$$

Again, since the roots are positive, by multiplying by $\frac{x_{i_2} y_i}{y_{i_2} x_i}$, we get

$$\frac{y_i^2 y_{i_3} \dots y_{i_{d-1}}}{x_i^2 x_{i_3} \dots x_{i_{d-1}}} \leqslant \frac{x_{i_2}}{y_{i_2}}. \tag{28}$$

Similarly, the left-hand side is a jump point, and the right-hand side is a root. Thus, all the roots are after, and in particular:

$$\frac{y_i^2 y_{i_3} \dots y_{i_{d-1}}}{x_i^2 x_{i_3} \dots x_{i_{d-1}}} \leqslant \frac{x_i}{y_i}. \tag{29}$$

Continuing this process, we ultimately obtain that

$$\frac{y_i^{d-1}}{x_i^{d-1}} \leqslant \frac{x_i}{y_i} \tag{30}$$

that is to say $1 \leqslant x_i/y_i$. If the inequality is an equality for all $i \in \{1, \dots, n\}$, then $\alpha = 1 = x_k/y_k$, which is impossible since $\alpha < x_k/y_k$. Thus, there exists one root t of (20) that is strictly greater than one. But this implies that every term in the sum in (20) is strictly positive, which is impossible. As a result, the first implication in (24) is true.

We next prove the second implication in (24) by contradiction. Assume that there exists $k \in \{1, \dots, n\}$ such that $x_k/y_k < \alpha$. Since one root is before the jump point α, all other roots are before the jump point α. In particular, for all $i \in \{1, \dots, n\}$, we have

$$\frac{x_i}{y_i} \leqslant \frac{y_{i_1} \dots y_{i_{d-1}}}{x_{i_1} \dots x_{i_{d-1}}} := \alpha < 0. \tag{31}$$

Therefore, all the roots are negative. Since $\alpha < 0$ is the product of $d-1$ negative terms, it must be that d is even. Observe that $\frac{x_{i_1} y_i}{y_{i_1} x_i} > 0$ because it is a ratio of two roots. Now, similar to the case $\alpha > 0$, we obtain

$$\frac{x_{i_1}}{y_{i_1}} \leqslant \frac{y_i y_{i_2} \dots y_{i_{d-1}}}{x_i x_{i_2} \dots x_{i_{d-1}}}. \tag{32}$$

The right-hand side is a jump point, and the left-hand side is a root. Thus, all the roots are before, and in particular:

$$\frac{x_i}{y_i} \leqslant \frac{y_i y_{i_2} \dots y_{i_{d-1}}}{x_i x_{i_2} \dots x_{i_{d-1}}}. \tag{33}$$

Continuing this process (as in the case where $\alpha > 0$), we ultimately obtain that

$$\frac{x_i}{y_i} \leqslant \frac{y_i^{d-1}}{x_i^{d-1}}. \tag{34}$$

Since d is even and $x_i/y_i < 0$, this implies that $x_i/y_i \leqslant -1$. If the inequality is an equality for all $i \in \{1, \dots, n\}$, then $\alpha = -1 = x_k/y_k$, which is impossible since $x_k/y_k < \alpha$. Thus, there exists one root t of (20) that is strictly less than -1. But this implies that every term in the sum in (20) is strictly negative, which is impossible. Consequently, the second implication in (24) holds.

Let us apply (24) to a root x_{i_d}/y_{i_d} for some $i_d \in \{1, \dots, n\}$:

$$\frac{y_{i_1} \dots y_{i_{d-1}}}{x_{i_1} \dots x_{i_{d-1}}} > 0 \implies \frac{x_{i_d}}{y_{i_d}} \leqslant \frac{y_{i_1} \dots y_{i_{d-1}}}{x_{i_1} \dots x_{i_{d-1}}} \quad \text{and} \quad \frac{y_{i_1} \dots y_{i_{d-1}}}{x_{i_1} \dots x_{i_{d-1}}} < 0 \implies \frac{y_{i_1} \dots y_{i_{d-1}}}{x_{i_1} \dots x_{i_{d-1}}} \leqslant \frac{x_{i_d}}{y_{i_d}}.$$

In both cases we find that

$$\frac{x_{i_1} \dots x_{i_d}}{y_{i_1} \dots y_{i_d}} \leqslant 1. \tag{35}$$

This inequality holds for all jump points (i.e. for all indices $i_1, \dots, i_{d-1} \in \{1, \dots, n\}$ such that $x_{i_1} \dots x_{i_{d-1}} \neq 0$) and it is trivially true for all indices such that $x_{i_1} \dots x_{i_{d-1}} = 0$. Therefore, (35) is true for all $i_1, \dots, i_d \in \{1, \dots, n\}$, which completes the proof of this lemma.

5.9 Proof of Lemma 3.2

When $x \in S$, notice that

$$\begin{aligned} f_1(x) &= \sum_{i_1,\ldots,i_d=1}^{n} |x_{i_1}\ldots x_{i_d} - y_{i_1}\ldots y_{i_d}| \\ &= \sum_{i_1,\ldots,i_d=1}^{n} |y_{i_1}\ldots y_{i_d}| \left|\tfrac{x_{i_1}\ldots x_{i_d}}{y_{i_1}\ldots y_{i_d}} - 1\right| \\ &= \sum_{i_1,\ldots,i_d=1}^{n} |y_{i_1}\ldots y_{i_d}| - |y_{i_1}\ldots y_{i_d}| \tfrac{x_{i_1}\ldots x_{i_d}}{y_{i_1}\ldots y_{i_d}} \\ &= \left(\sum_{i=1}^{n} |y_i|\right)^d - \left(\sum_{i=1}^{n} |y_i| \tfrac{x_i}{y_i}\right)^d . \end{aligned} \tag{36}$$

Given $\alpha > 0$, consider the function $\phi_\alpha : f_1(S) \longrightarrow \mathbb{R}$ defined by

$$\phi_\alpha(t) = -\left[-t - \left(\sum_{i=1}^{n} |y_i|\right)^d \right]^\alpha . \tag{37}$$

If d is odd, then ϕ_α is increasing when taking $\alpha = 1/d$. If d is even, then ϕ_α is increasing when $-t - (\sum_{i=1}^n |y_i|)^d$ is positive and $\alpha = 2/d$. Next, define the set

$$S' := \{\, x \in \mathbb{R}^n \mid x_{i_1}\ldots x_{i_d} \leqslant 1 \,, \quad \forall\, i_1,\ldots,i_d \in \{1,\ldots,n\} \,\} \tag{38}$$

and consider the homeomorphism $\varphi : S \longrightarrow S'$ defined by

$$\varphi(x) = \left(\frac{x_1}{y_1},\ldots,\frac{x_n}{y_n}\right). \tag{39}$$

According to Proposition 2.1 and Proposition 2.2, f is a global function on S, i.e. $f_1 \in \mathcal{G}(S)$, if and only if $\phi_\alpha \circ f_1 \circ \varphi^{-1}$ is a global function on S'. Thus, when d is odd, $f_1 \in \mathcal{G}(S)$ if and only if $f_{\text{odd}}(x) := \phi_{1/d} \circ f_1 \circ \varphi^{-1}(x) = -\sum_{i=1}^n |y_i| x_i \in \mathcal{G}(S')$. When d is even, $f_1 \in \mathcal{G}(S)$ if and only if $f_{\text{even}}(x) := \phi_{2/d} \circ f_1 \circ \varphi^{-1}(x) = -\left(\sum_{i=1}^n |y_i| x_i\right)^2 \in \mathcal{G}(S')$.

Consider the case when d is odd. For all $i_1,\ldots,i_d \in \{1,\ldots,n\}$, define the constraint function $g_{i_1,\ldots i_d}(x) := x_{i_1}\ldots x_{i_d} - 1$. If $x_1\ldots x_n \neq 0$, then for any $i_1,\ldots,i_d \in \{1,\ldots,n\}$, it satisfies

$$\nabla\, g_{i_1,\ldots i_d}(x) = \begin{pmatrix} N(1,i_1,\ldots i_d)/x_1 \\ \vdots \\ N(n,i_1,\ldots i_d)/x_n \end{pmatrix} x_{i_1}\ldots x_{i_d} \tag{40}$$

where $\nabla\, g_{i_1,\ldots i_d}(x)$ denotes the gradient of $g_{i_1,\ldots i_d}$ at x and $N(i,i_1,\ldots,i_d)$ denotes the number of indices among $i_1,\ldots,i_d$ that are equal to i. If the constraint $g_{i_1,\ldots i_d}(x) \leqslant 0$ is active, then

$$-x^T\, \nabla g_{i_1,\ldots i_d}(x) = -\sum_{k=1}^{n} N(k,i_1,\ldots i_d) < 0. \tag{41}$$

The Mangasarian-Fromovitz constraint qualification thus holds. A local minimum $x \in \mathbb{R}^n$ for the problem $\inf_{x\in S'} f_{\text{odd}}(x)$ must therefore satisfy the Karush-Kuhn-Tucker conditions:

$$\begin{cases} \displaystyle\sum_{\substack{i_1,\ldots,i_d\,=\,1 \\ N(i,i_1,\ldots,i_d)\,\neq\,0}}^{n} \lambda_{i_1,\ldots,i_d} N(i,i_1,\ldots,i_d) \tfrac{x_{i_1}\ldots x_{i_d}}{x_i} = |y_i| \,, \quad \forall\, i \in \{1,\ldots,n\} \,, \\ x_{i_1}\ldots x_{i_d} \leqslant 1 \,, \quad \forall\, i_1,\ldots,i_d \in \{1,\ldots,n\} \,, \\ \lambda_{i_1,\ldots,i_d} \geqslant 0 \,, \\ \lambda_{i_1,\ldots,i_d}(x_{i_1}\ldots x_{i_d} - 1) = 0 \,. \end{cases} \tag{42}$$

Here $\lambda_{i_1,\ldots,i_d} \geq 0, i_1, \ldots, i_d \in \{1, \ldots, n\}$, are the Lagrange multipliers. If $x_i \neq 0$ for some $i \in \{1, \ldots, n\}$, then, by complementarity slackness, the first line yields

$$0 \quad < \quad |y_i| \quad = \frac{1}{x_i} \sum_{\substack{\lambda_{i_1,\ldots,i_d} > 0 \\ N(i,i_1,\ldots,i_d) \neq 0}} \underbrace{\lambda_{i_1,\ldots,i_d} N(i, i_1, \ldots, i_d)}_{>0} \tag{43}$$

which implies that $x_i > 0$. As a result, $x \geqslant 0$. Together with feasibility, it results that $0 \leqslant x_i^d \leqslant 1$, leading to the inequalities $0 \leqslant x_i \leqslant 1$ for all $i \in \{1, \ldots, n\}$. Following Proposition 2.3, f_{odd} is thus global on S' if $f_{\text{odd}} \in \mathcal{G}(S'')$ where

$$S'' := \{\, x \in \mathbb{R}^n \mid 0 \leqslant x_i \leqslant 1\,, \quad \forall\, i \in \{1, \ldots, n\} \,\} \tag{44}$$

which contains the global minimizer 0. From the notion of global functions, $f_{\text{odd}} \in \mathcal{G}(S'')$ if the problem

$$\inf_{x \in \mathbb{R}^n} \quad -\sum_{i=1}^{n} |y_i| x_i \qquad \text{subject to} \qquad 0 \leqslant x_i \leqslant 1\,, \qquad \forall\, i \in \{1, \ldots, n\} \tag{45}$$

has no spurious local minima, which is obvious because the problem is a linear program.

Consider the case when d is even. Since a feasible point $x \in S'$ satisfies $x_i^d \leqslant 1$, it must be that $-1 \leqslant x_i \leqslant 1$. Conversely, any point such that $-1 \leqslant x_i \leqslant 1$ belongs to S'. This implies that

$$S' := \{\, x \in \mathbb{R}^n \mid -1 \leqslant x_i \leqslant 1\,, \quad \forall\, i \in \{1, \ldots, n\} \,\}\,. \tag{46}$$

According to Proposition 2.4, $f_{\text{even}}(x) \in \mathcal{G}(S')$ if $f_{\text{even}}(x)$ is a global function on both sets $S' \cap \{x \in \mathbb{R}^n | \sum_{i=1}^n |y_i| x_i \geq 0\}$ and $S' \cap \{x \in \mathbb{R}^n | \sum_{i=1}^n |y_i| x_i \leq 0\}$, and $f_{\text{even}}(x)$ takes the same optimal value on both sets (the latter is obvious using symmetry). Using Proposition 2.1 again, we find that $f_{\text{even}}(x)$ is a global function on these two sets if and only if

$$-\sum_{i=1}^{n} |y_i| x_i \in \mathcal{G}\left(S' \cap \left\{x \in \mathbb{R}^n | \sum_{i=1}^{n} |y_i| x_i \geq 0\right\}\right) \tag{47}$$

and

$$\sum_{i=1}^{n} |y_i| x_i \in \mathcal{G}\left(S' \cap \left\{x \in \mathbb{R}^n | \sum_{i=1}^{n} |y_i| x_i \leq 0\right\}\right), \tag{48}$$

which are true because they are associated with the following linear programs:

$$\inf_{x \in \mathbb{R}^n} \quad -\sum_{i=1}^{n} |y_i| x_i \qquad \text{subject to} \qquad \begin{cases} -1 \leqslant x_i \leqslant 1\,, \qquad \forall\, i \in \{1, \ldots, n\} \\ \sum\limits_{i=1}^{n} |y_i| x_i \geqslant 0. \end{cases} \tag{49}$$

and

$$\inf_{x \in \mathbb{R}^n} \quad \sum_{i=1}^{n} |y_i| x_i \qquad \text{subject to} \qquad \begin{cases} -1 \leqslant x_i \leqslant 1\,, \qquad \forall\, i \in \{1, \ldots, n\} \\ \sum\limits_{i=1}^{n} |y_i| x_i \leqslant 0. \end{cases} \tag{50}$$