CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY IN PRAGUE

RESEARCH REPORT

ISSN 1213-2365

On Coordinate Minimization of Convex Piecewise-Affine Functions

Tomáš Werner

CTU–CMP–2017–05

September 14, 2017

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/werner/Werner-TR-2017-05.pdf

This work has been supported by the Czech Science Foundation grant 16-05872S.

Research Reports of CMP, Czech Technical University in Prague, No. 5, 2017

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

On Coordinate Minimization of Convex Piecewise-Affine Functions

Tomáš Werner

Dept. of Cybernetics, Faculty of Electrical Engineering, Czech Technical Univ. in Prague

Research Report CTU–CMP–2017–05

September 14, 2017

Abstract

A popular class of algorithms to optimize the dual LP relaxation of the discrete energy minimization problem (a.k.a. MAP inference in graphical models or valued constraint satisfaction) are convergent message-passing algorithms, such as max-sum diffusion, TRW-S, MPLP and SRMP. These algorithms are successful in practice, despite the fact that they are a version of coordinate minimization applied to a convex piecewise-affine function, which is not guaranteed to converge to a global minimizer. These algorithms converge only to a local minimizer, characterized by local consistency known from constraint programming. We generalize max-sum diffusion to a version of coordinate minimization applicable to an arbitrary convex piecewise-affine function, which converges to a local consistency condition. This condition can be seen as the sign relaxation of the global optimality condition.

1 Introduction

Coordinate minimization is an iterative method for unconstrained optimization, which in every iteration minimizes the objective function over a single variable while keeping the remaining variables fixed. For some functions, this simple method converges to a global minimum. This class includes differentiable convex functions [1, §2.7] and convex functions whose non-differentiable part is separable [25]. For general non-differentiable convex functions, the method need not converge to a global minimum.

For large-scale non-differentiable convex optimization, coordinate minimization may be an acceptable option despite its inability to find a global minimum. An example is dual LP relaxations of some NP-hard combinatorial optimization problems, such as *discrete energy minimization* [24, 8] (also known as *MAP inference in graphical models* [27] or *valued constraint satisfaction* [17, 26]) and also some other problems [23]. This dual LP relaxation leads to the unconstrained minimization of a convex piecewise-affine function. A number of algorithms for solving this dual LP relaxation have been proposed. One class of algorithms, sometimes referred to as *convergent message passing* [27, 5, 6, 16], consists of various forms of (block-)coordinate minimization applied to various forms of the dual. Examples are max-sum diffusion [13, 20, 28, 29], TRW-S [9], MPLP [5, 22], and SRMP [10]. Besides these, several large-scale convex optimization methods converging to a global minimum have been applied to the problem, such as subgradient methods [21, 11], bundle methods [18], alternating direction method of multipliers [15], and adaptive diminishing smoothing [19]. For practical problems from computer vision, it was observed [8] that convergent message-passing methods converge faster than these global methods and their fixed points are often not far from global minima, especially for sparse instances.

We ask whether convergent message-passing algorithms can be reformulated to become applicable to a wider class of non-differentiable convex functions than those arising from the above dual LP relaxations. In other words, whether these algorithms can be studied independently of any combinatorial optimization problems. In this paper, we make a step in this direction and generalize max-sum diffusion to an algorithm applicable to an arbitrary convex piecewise-affine function.

Consider an objective function in the form of a pointwise maximum of affine functions. In this case, univariate minimizers in each iteration of coordinate minimization need not be unique and therefore some rule must be designed to choose a unique minimizer. We show that for a certain natural deterministic rule, fixed points of coordinate minimization can be poor. Imitating max-sum diffusion, we propose a better rule which we call the *unique rule*: in each iteration, minimize the maximum of only those affine functions that depend on the current variable. With this rule, univariate minimizers are unique and fixed points of the algorithm satisfy a natural condition, the *sign relaxation* of the true optimality condition. This can be seen as a *local consistency* as used in constraint programming [2], which is known to characterize fixed points of many algorithms to solve the dual LP relaxation, namely message-passing algorithms [28, 29, 9, 10] and the algorithm [12, 3].

Little is known theoretically about convergence properties of message-passing algorithms. Although max-sum diffusion and TRW-S were always observed to converge to a fixed point, this was never proved. For finite-valued energy terms, it has been proved that every accumulation point of TRW-S and SRMP satisfy local consistency [9, 10]. Under the same assumption, a stronger result has been proved for max-sum diffusion [20]: a quantity that measures how much the local consistency condition is violated converges to zero (note, this is still a weaker result that convergence to a fixed point). Nothing is known theoretically about convergence rates of max-sum diffusion, TRW-S, MPLP and SRMP.

We revisit the proof from [20] to show that, under a certain assumption, during coordinate minimization with the unique rule the above quantity converges to zero. In contrast to max-sum diffusion, we show there are objective functions for which the algorithm has no fixed point or even no point satisfying sign consistency.

2 Minimizing a Pointwise Maximum of Affine Functions

Consider a function $f\colon \mathbb{R}^n \to \mathbb{R}$ that is a pointwise maximum of affine functions, i.e.,

$$f(x) = \max_{i\in[m]}(a_i^T x + b_i) \tag{1}$$

where $a_1, \dots, a_m \in \mathbb{R}^n$ and $b_1, \dots, b_m \in \mathbb{R}$, and we denote $[m] = \{1, \dots, m\}$. For brevity, we will further on abuse symbols 'max' and 'argmax' and denote, for $y \in \mathbb{R}^m$,

$$\max y = \max_{i\in[m]} y_i, \tag{2a}$$

$$\operatorname{argmax} y = \operatorname*{argmax}_{i\in[m]} y_i = \{\, i \in [m] \mid y_i = \max y \,\}. \tag{2b}$$

Now function (1) can be written simply as

$$f(x) = \max(Ax + b), \tag{3}$$

where matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$ has rows $a_1^T, \dots, a_m^T$ and $b = (b_1, \dots, b_m) \in \mathbb{R}^m$.

The well-known condition for a global minimum of a convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ is that $0 \in \partial f(x)$. For function (1), the subdifferential reads

$$\partial f(x) = \operatorname{conv}\{\, a_i \mid i \in \operatorname{argmax}(Ax + b) \,\}. \tag{4}$$

In more detail, the following holds:

Proposition 1. For function f given by (1) it holds:

- f is bounded from below iff $0 \in \operatorname{conv}\{a_1, \dots, a_m\}$.
- x is a minimizer of f iff $0 \in \operatorname{conv}\{\, a_i \mid i \in \operatorname{argmax}(Ax + b) \,\}$.
- If f is bounded from below, it has a minimizer x such that $0 \in \operatorname{rint}\operatorname{conv}\{\, a_i \mid i \in \operatorname{argmax}(Ax + b) \,\}$.

Proof. The minimization of function (1) can be written as the linear program

$$\min\{\, z \mid Ax + b \le z,\ x \in \mathbb{R}^n,\ z \in \mathbb{R} \,\}. \tag{5}$$

The dual to this linear program reads

$$\max\{\, b^T\lambda \mid \lambda \in \mathbb{R}^m,\ \lambda \ge 0,\ 1^T\lambda = 1,\ A^T\lambda = 0 \,\}. \tag{6}$$

The primal (5) is always feasible. By strong duality, f is bounded from below iff the dual (6) is feasible. By complementary slackness, x and λ are both optimal iff $\lambda_i = 0$ for every $i \notin \operatorname{argmax}(Ax + b)$. If f is bounded from below, by strict complementary slackness there exist optimal x and λ such that $\lambda_i > 0$ iff $i \in \operatorname{argmax}(Ax + b)$. Note that $A^T\lambda = \sum_{i\in[m]} \lambda_i a_i$. Note that, for any $k \in \mathbb{N}$ and $a_1, \ldots, a_k \in \mathbb{R}^n$,

$$0 \in \operatorname{conv}\{a_1, \ldots, a_k\} \iff \exists\lambda \in \mathbb{R}^k\colon \lambda_i \geq 0,\ \textstyle\sum_i \lambda_i = 1,\ \sum_i \lambda_i a_i = 0, \tag{7a}$$

$$0 \in \operatorname{rint}\operatorname{conv}\{a_1, \ldots, a_k\} \iff \exists\lambda \in \mathbb{R}^k\colon \lambda_i > 0,\ \textstyle\sum_i \lambda_i = 1,\ \sum_i \lambda_i a_i = 0. \tag{7b}$$

All three claims are now obvious. □

Proposition 2. A non-empty set $\{\, Ax \mid x \in \mathbb{R}^n,\ Ax \leq b \,\}$ is bounded iff $0 \in \operatorname{rint}\operatorname{conv}\{a_1, \ldots, a_m\}$.

Proof. As the set is non-empty, it is bounded iff the linear program

$$\min\{\, c^T Ax \mid x \in \mathbb{R}^n,\ Ax \leq b \,\}$$

is bounded for every $c \in \mathbb{R}^m$. This holds iff the dual linear program

$$\max\{\, -b^T\lambda \mid A^T(\lambda + c) = 0,\ \lambda \geq 0 \,\}$$

is feasible for every $c \in \mathbb{R}^m$. We show that

$$\forall c \in \mathbb{R}^m\ \exists\lambda \geq 0\colon A^T(\lambda + c) = 0 \iff \exists\lambda' > 0\colon A^T\lambda' = 0. \tag{8}$$

- To prove $\Rightarrow$, take $c = 1$ and let $\lambda \geq 0$ be such that $A^T(\lambda + c) = 0$. Then set $\lambda' = \lambda + c = \lambda + 1$.
- To prove $\Leftarrow$, set $\lambda = \alpha\lambda' - c$ where $\alpha > 0$ is arbitrary such that $\lambda \geq 0$. Such α clearly exists.

Note that λ' can be multiplied by a positive scale to satisfy $1^T\lambda' = 1$. By (7b), the right-hand statement in (8) is thus equivalent to $0 \in \operatorname{rint}\operatorname{conv}\{a_1, \ldots, a_m\}$. □

3 Sign Relaxation of the Optimality Condition

We said that deciding if $x \in \mathbb{R}^n$ is a minimizer of function (1) requires deciding if the convex hull of a finite set of vectors contains the origin. Deciding this condition for large problems may be infeasible. We define the *sign relaxation* of this condition which is cheaper to decide, obtained by considering only the signs of the vectors' components and dropping their magnitudes.

To describe the key idea, suppose that the convex hull of some vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ contains the origin, $0 \in \operatorname{conv}\{a_1, \ldots, a_m\}$. That is, there are numbers $\lambda_1, \ldots, \lambda_m \geq 0$ such that

$$\sum_{i\in[m]} \lambda_i = 1, \tag{9a}$$

$$\sum_{i\in[m]} \lambda_i a_{ij} = 0, \quad j \in [n]. \tag{9b}$$

Let us relax these conditions, considering only the signs $\sigma_i = \operatorname{sgn}\lambda_i \in \{0, 1\}$ and $s_{ij} = \operatorname{sgn} a_{ij} \in \{-1, 0, 1\}$. Equality (9a) implies that at least one of the numbers $\sigma_1, \ldots, \sigma_m$ equals 1. For each $j \in [n]$, equality (9b) implies that the numbers $\sigma_1 s_{1j}, \ldots, \sigma_m s_{mj}$ either are all zero, or some are positive and some negative. One way to write this is as follows:

$$\exists i \in [m]\colon \sigma_i = 1, \tag{10a}$$

$$(\exists i \in [m]\colon \sigma_i s_{ij} = -1) \Leftrightarrow (\exists i' \in [m]\colon \sigma_{i'} s_{i'j} = 1), \quad j \in [n]. \tag{10b}$$

Thus, it is necessary for $0 \in \operatorname{conv}\{a_1, \ldots, a_m\}$ that there exist some $\sigma_1, \ldots, \sigma_m \in \{0, 1\}$ satisfying (10).

This can be seen as a *constraint satisfaction problem* (CSP) [14, 4] with m binary variables and $n + 1$ constraints. This particular CSP class can be solved by enforcing *(generalized) arc consistency* [2]. Suppose that for some j, the signs $s_{1j}, \ldots, s_{mj}$ are, say, all non-negative and some of them is positive. Then constraint (10b) enforces that for all i for which $s_{ij} \neq 0$ we have $\sigma_i = 0$. Repeating this for various coordinates j progressively sets some σ_i to zero. If finally $\sigma_i = 0$ for all $i \in [n]$, constraint (10) is violated and the CSP has no solution. Otherwise, the CSP has a solution.

Since we believe that the described concept of sign relaxation might have a wider applicability in non-differential convex optimization, we further develop it in more detail and greater generality than is needed in this paper. Namely, we consider the sign relaxation of the condition $0 \in \operatorname{conv} X$ for any (not necessarily finite) set $X \subseteq \mathbb{R}^n$. This is straightforward because the sign relaxation of this condition depends only on the set $\{ \operatorname{sgn} a \mid a \in X \} \subseteq \{-1, 0, 1\}^n$ which is finite. Here, for a vector $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, we denoted $\operatorname{sgn} a = (\operatorname{sgn} a_1, \ldots, \operatorname{sgn} a_n) \in \{-1, 0, 1\}^n$.

Theorem 3. Let $X \subseteq \mathbb{R}^n$ and $a \in \operatorname{conv} X$. Then $a \in \operatorname{rint} \operatorname{conv}(X \cap F)$ where F is the intersection of all faces of $\operatorname{conv} X$ that contain[1] a.

Proof. It can be shown that for any face F of $\operatorname{conv} X$ it holds that[2] $\operatorname{conv}(X \cap F) = F$. So it suffices to show that $a \in \operatorname{rint} F$. For contradiction, suppose a is a relative boundary point of F. But every relative boundary point of a face is contained in some subface of the face, so F cannot be the intersection of all faces of $\operatorname{conv} X$ containing a. □

Definition 1. A set $S \subseteq \{-1, 0, 1\}^n$ is *consistent in coordinate* $j \in [n]$ if it holds that

$$\exists s \in S\colon s_j = -1 \quad \Longleftrightarrow \quad \exists t \in S\colon t_j = 1. \tag{11}$$

Set S is *consistent* if it is consistent in every coordinate $j \in [n]$.

In particular, note that the sets $\emptyset$ and $\{0\}$ (where 0 denotes the vector with n zeros) are consistent.

Theorem 4. Let $X \subseteq \mathbb{R}^n$. If $0 \in \operatorname{rint} \operatorname{conv} X$, then the set $\{ \operatorname{sgn} a \mid a \in X \}$ is consistent.

Proof. Let $0 \in \operatorname{rint} \operatorname{conv} X$. Then the projection of the set $\operatorname{rint} \operatorname{conv} X$ onto each coordinate axis is either the set $\{0\}$ or an interval containing zero as its interior point. Noting that projections commute with the convex hull operator, this means for every $j \in [n]$ we have that

$$\exists a \in X\colon a_j < 0 \quad \Longleftrightarrow \quad \exists b \in X\colon b_j > 0. \tag{12}$$

This is equivalent to condition (11) for the set $\{ \operatorname{sgn} a \mid a \in X \}$. □

Theorem 5. Let $X \subseteq \mathbb{R}^n$. If $0 \in \operatorname{conv} X$, then the set $\{ \operatorname{sgn} a \mid a \in X \}$ has a non-empty consistent subset.

Proof. If $0 \in \operatorname{conv} X$, then by Theorem 3 there is $Y \subseteq X$ such that $0 \in \operatorname{rint} \operatorname{conv} Y$. Thus, by Theorem 4, the set $\{ \operatorname{sgn} a \mid a \in Y \}$ is consistent. This set is non-empty and it is a subset of $\{ \operatorname{sgn} a \mid a \in X \}$. □

Theorem 5 gives a necessary condition for a set $X \subseteq \mathbb{R}^n$ to satisfy $0 \in \operatorname{conv} X$. This is what we called the sign relaxation of $0 \in \operatorname{conv} X$. This condition is of course not sufficient, e.g., for $X = \{(1, -2), (-2, 1)\}$ the set $\{ \operatorname{sgn} a \mid a \in X \} = \{(1, -1), (-1, 1)\}$ is consistent but $0 \notin \operatorname{conv} X$.

Next we develop the concept of *consistency closure*, analogous to arc consistency closure [2].

Theorem 6. Let the sets $S, T \subseteq \{-1, 0, 1\}^n$ be consistent. Then the set $S \cup T$ is consistent.

Proof. Let $s \in S \cup T$, so $s \in S$ or $s \in T$. Suppose $s \in S$. Suppose for some $j \in [n]$ we have, say, $s_j < 0$. As S is consistent, there is $t \in S$ such that $s_j > 0$. But $t \in S \cup T$. □

Definition 2. The *consistency closure* of a set $S \subseteq \{-1, 0, 1\}^n$ is the greatest (with respect to partial ordering by inclusion) consistent subset of S, i.e., it is the union of all consistent subsets of S. We will denote it by $\operatorname{cons} S$.

Theorem 7. The cons operator satisfies the axioms of a closure operator, i.e., it is

- intensive ($\operatorname{cons} S \subseteq S$),

[1] More precisely, it can be shown that $Y = X \cap F$ is the greatest (with respect to partial ordering by inclusion) subset of X such that $a \in \operatorname{rint} \operatorname{conv} Y$.

[2] We omit the proof of this claim. The proof is obvious if X is finite and hence $\operatorname{conv} X$ is a convex polytope. For infinite X, recall [7] that a *face* of a convex set C is a convex set $F \subseteq C$ such that every line segment from C whose relative interior has a non-empty intersection with F is contained in F.

- idempotent ($\operatorname{cons}\operatorname{cons} S = \operatorname{cons} S$),
- non-increasing ($S \subseteq T \implies \operatorname{cons} S \subseteq \operatorname{cons} T$).

Proof. Intensivity and idempotency are immediate from Definition 2. To show the non-increasing property, let $S \subseteq T$. By intensivity, we have $\operatorname{cons} S \subseteq T$. But $\operatorname{cons} S$ is consistent, thus by Definition 2 it must be a subset of $\operatorname{cons} T$. $\square$

Definition 3. *Enforcing consistency* of a set $S \subseteq \{-1, 0, 1\}^n$ is the algorithm that repeats the following iteration:

1. choose an arbitrary coordinate $j \in [n]$ in which S is not consistent,
2. remove from S all elements s such that $s_j \neq 0$.

If S is consistent, the algorithm stops.

Theorem 8. Enforcing consistency of any set $S \subseteq \{-1, 0, 1\}^n$ yields $\operatorname{cons} S$.

Proof. The algorithm creates a sequence of sets $S = S_0 \supset S_1 \supset \cdots \supset S_K$, where S_K is consistent and K is the number of iterations. This sequence is given recurrently by $S_k = S_{k-1} \setminus \{\, s \in S_{k-1} \mid s_{j_k} \neq 0 \,\}$ where each $j_k \in [n]$ is such that S_{k-1} is not consistent in coordinate j_k.

Let a set $T \subseteq S$ be consistent. We show by induction that $T \subseteq S_k$ for all k. Assume $T \subseteq S_{k-1}$. Since S_{k-1} is not consistent in coordinate j_k, it follows from Definition 1 that $s_{j_k} = 0$ for every $s \in T$. Therefore $T \subseteq S_{k-1} \setminus \{\, s \in S_{k-1} \mid s_{j_k} \neq 0 \,\} = S_k$.

We have shown that every consistent subset of S is contained in S_K, which is itself consistent. This means that $S_K = \operatorname{cons} S$. $\square$

To conclude, the sign relaxation of the condition $0 \in \operatorname{conv} X$ is that $\operatorname{cons}\{\, \operatorname{sgn} a \mid a \in X \,\} \neq \emptyset$.

4 Coordinate Minimization

Coordinate minimization of a function $f \colon \mathbb{R}^n \to \mathbb{R}$ is a method that, starting from an initial point $x = (x_1, \ldots, x_n)$, repeats the following iteration:

1. choose $j \in [n]$,
2. choose $x_j^* \in \operatorname*{argmin}_{x_j \in \mathbb{R}} f(x)$,
3. set $x_j \leftarrow x_j^*$.

Since the choices in the first two steps are not specified, this does not define a single algorithm but rather a class of algorithms (a 'method').

Further on, we focus on applying coordinate minimization to functions of the form (1).

For functions of the form (1), in Step 2 of coordinate minimization the univariate minimizer x_j^* is in general not unique. Therefore, some rule must be adopted to choose a unique minimizer. Let us restrict ourselves to deterministic rules. We show that some rules can behave poorly. Consider the following rule: choose the element of the set $\operatorname{argmin}_{x_j \in \mathbb{R}} f(x)$ that is nearest to x_j^{prev}, the j-th variable from the previous iteration. We refer to this as the *proximal rule*[3].

Example 1. Consider the function

$$f(x_1, x_2, x_3) = \max\{\, x_2 - x_3,\ x_3 - x_1,\ x_1 - x_2 \,\}, \tag{13}$$

which has minimum value 0, attained for any $x_1 = x_2 = x_3$. The point $x = (x_1, x_2, x_3) = (2, 1, 0)$ is fixed for coordinate minimization with the proximal rule, with value $f(x) = \max\{1, -2, 1\} = 1$. The subdifferential at this point is $\partial f(x) = \operatorname{conv} X$ where $X = \{(0, 1, -1), (1, -1, 0)\}$. Using the algorithm from Definition 3 we find that $\operatorname{cons}\{\, \operatorname{sgn} a \mid a \in X \,\} = \operatorname{cons}\{(0, 1, -1), (1, -1, 0)\} = \emptyset$. Thus, point x does not satisfy the sign relaxation of the condition $0 \in \partial f(x)$.

[3]This can be indeed seen as a proximal regularization of coordinate minimization: instead of function f we minimize the function $g(x, y) = f(x) + \mu\|x - y\|^2$ for a small $\mu > 0$. For this function, the univariate minimization would read $\operatorname{argmin}_{x_j \in \mathbb{R}}(f(x_1, \ldots, x_n) + \mu(x_j - x_j^{\text{prev}})^2)$, which has the unique minimizer given by the proximal rule. Cf. [1, Exercise 2.7.2].

Imitating max-sum diffusion, we propose a better rule, which we call the *unique rule*: in Step 2 of coordinate minimization, rather than minimizing the maximum of all affine functions, minimize the maximum of only those affine functions that depend on variable x_j. That is, minimize the function

$$\phi_j(x) = \max_{i \mid a_{ij} \neq 0} (a_i^T x + b_i) \tag{14}$$

where $\max_{i \mid a_{ij} \neq 0}$ denotes maximization over all $i \in [m]$ such that $a_{ij} \neq 0$. Further on, we will assume that the set $\{ \operatorname{sgn} a_i \mid i \in [m] \}$ is consistent. Under this assumption, for every $j \in [n]$ and every $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n \in \mathbb{R}$, the univariate function $x_j \mapsto \phi_j(x)$ has exactly one minimizer, x_j^*. This minimizer is the unique solution of the equation

$$\max_{i \mid a_{ij} < 0} (a_i^T x + b_i) = \max_{i \mid a_{ij} > 0} (a_i^T x + b_i). \tag{15}$$

To summarize, the iteration of the algorithm for coordinate $j \in [n]$ adjusts variable x_j to satisfy (15), keeping the other variables unchanged. A *fixed point* of the algorithm[4] is a point $x \in \mathbb{R}^n$ that satisfies (15) for all $j \in [n]$.

For functions in the form (1), coordinate minimization can be equivalently formulated in terms of the values $y = Ax + b$ of the affine functions, i.e., instead of updating the numbers $x_1, \ldots, x_n$ we update the numbers $y_1, \ldots, y_m$, while $x_1, \ldots, x_n$ are no longer explicitly kept. We first set $y = Ax + b$, where x is the initial point, and then repeat the following iteration:

1. choose $j \in [n]$,
2. choose $x_j^* \in \operatorname*{argmin}_{x_j \in \mathbb{R}} \max_{i \in [m]} (a_{ij} x_j + y_i)$,
3. set $y_i \leftarrow a_{ij} x_j^* + y_i$ for all $i \in [m]$.

Using the unique rule, in Step 2 we need to find the unique minimizer x_j^* of the univariate function $\max_{i \mid a_{ij} \neq 0} (a_{ij} x_j + y_i)$, which is the solution of the equation

$$\max_{i \mid a_{ij} < 0} (a_{ij} x_j + y_i) = \max_{i \mid a_{ij} > 0} (a_{ij} x_j + y_i). \tag{16}$$

After the iteration, we thus have

$$\max_{i \mid a_{ij} < 0} y_i = \max_{i \mid a_{ij} > 0} y_i. \tag{17}$$

A fixed point of the algorithm is a point $y \in \{ A(x + x') + b \mid x' \in \mathbb{R}^n \}$ that satisfies (17) for all $j \in [n]$. We introduce the following notations:

- Mapping $p_j \colon \mathbb{R}^m \to \mathbb{R}^m$ denotes the action of Steps 2 and 3 of coordinate minimization with the unique rule, formulated in terms of y. That is, for $y \in \mathbb{R}^m$ and $j \in [n]$, $p_j(y)$ is computed as follows: find the solution x_j^* of equation (16) and then set $y_i \leftarrow a_{ij} x_j^* + y_i$ for all $i \in [m]$.
- Further on, we assume that coordinates j in Step 1 of coordinate minimization are visited in the cyclic order. Let $p = p_n \circ p_{n-1} \circ \cdots \circ p_2 \circ p_1$ denote the action of the algorithm for one cycle.
- For $k \in \mathbb{N}$, let $p^k = p \circ \cdots \circ p$ (k-times) denote the action of the algorithm for k cycles.

In this notation, $y \in \mathbb{R}^m$ is a fixed point of the algorithm iff $p_j(y) = y$ for every $j \in [n]$, which holds iff $p(y) = y$.

Next we give several examples of the algorithm's behavior.

Example 2. Recall that coordinate minimization is not guaranteed to find a global minimum of a function of the form (1) because it is not differentiable. An example is the function

$$f(x_1, x_2) = \max\{ x_1 - 2x_2,\ x_2 - 2x_1 \},$$

which is unbounded but any point $x_1 = x_2$ is fixed for coordinate minimization. At any such point we have $0 \notin \partial f(x_1, x_2) = \operatorname{conv} X$ where $X = \{(1, -2), (-2, 1)\}$. The set $\{ \operatorname{sgn} a \mid a \in X \} = \{(1, -1), (-1, 1)\}$ is consistent. There is no difference between the proximal rule and unique rule, because univariate minimizers in Step 2 are unique. Coordinate minimization converges in one iteration.

[4] Note, such a point is a Nash equilibrium for penalty functions $\phi_1, \ldots, \phi_n \colon \mathbb{R}^n \to \mathbb{R}$.

The unique rule is not worse than the proximal rule, in the sense that every fixed point of coordinate minimization with the unique rule is also a fixed point of coordinate minimization with the proximal rule. The following example shows that the unique rule is in fact strictly better.

Example 3. Let us return to Example 1 and apply coordinate minimization with the unique rule to function (13). For $j = 1$, equation (16) reads $-x_1 + y_2 = x_1 + y_3$ with the solution $x_1^* = (y_2 - y_3)/2$, thus y is updated as $y_2 \leftarrow -x_1^* + y_2 = (y_2 + y_3)/2$ and $y_3 \leftarrow x_1^* + y_3 = (y_2 + y_3)/2$. Similarly we derive the iterations for $j = 2, 3$. In summary, we have

$$\begin{aligned} p_1(y_1, y_2, y_3) &= (\,y_1,\ (y_2 + y_3)/2,\ (y_2 + y_3)/2\,), \\ p_2(y_1, y_2, y_3) &= (\,(y_3 + y_1)/2,\ y_2,\ (y_3 + y_1)/2\,), \\ p_3(y_1, y_2, y_3) &= (\,(y_1 + y_2)/2,\ (y_1 + y_2)/2,\ y_3\,). \end{aligned}$$

We see that in every iteration, the algorithm takes a pair of the numbers y_1, y_2, y_3 and replaces both of them with their average. For any initial y_1, y_2, y_3, the sequence $(p^k(y_1, y_2, y_3))_{k\in\mathbb{N}}$ converges to the point with $y_1 = y_2 = y_3 = (y_1 + y_2 + y_3)/3$, i.e., to a minimizer of function (13).

The next example shows that there are functions for which coordinate minimization with the unique rule has no fixed point and even no point satisfying the sign-relaxed optimality condition.

Example 4. Let

$$f(x_1, x_2, x_3) = \max\{\, x_1 - x_2 - x_3,\ x_1 + 4,\ x_1 + x_2 + x_3,\ -x_1 + x_2 + 2 \,\}. \tag{18}$$

This function is not bounded from below. System (15), defining the fixed point condition, reads

$$\begin{aligned} \max\{\, x_1 - x_2 - x_3,\ x_1 + 4,\ x_1 + x_2 + x_3 \,\} &= -x_1 + x_2 + 2 \\ \max\{\, x_1 + x_2 + x_3,\ -x_1 + x_2 + 2 \,\} &= x_1 - x_2 - x_3 \\ x_1 + x_2 + x_3 &= x_1 - x_2 - x_3 \end{aligned}$$

The third equation implies $x_2 + x_3 = 0$, thus the system simplifies to

$$\begin{aligned} \max\{\, x_1,\ x_1 + 4,\ x_1 \,\} &= -x_1 + x_2 + 2 \\ \max\{\, x_1,\ -x_1 + x_2 + 2 \,\} &= x_1 \end{aligned}$$

The first equation simplifies to $x_1 + 4 = -x_1 + x_2 + 2$, hence $x_2 = 2x_1 + 2$. Plugging this to the second equation gives a contradiction. This shows that the algorithm has no fixed point.

We shall show that there is even no point $x \in \mathbb{R}^n$ such that $\operatorname{cons}\{\operatorname{sgn} a_i \mid i \in \operatorname{argmax}(Ax + b)\} \neq \emptyset$. For contradiction, suppose it is so. That is, there is $x \in \mathbb{R}^n$ and $I \subseteq [m]$ such that the set $\{\operatorname{sgn} a_i \mid i \in I\}$ is consistent and $\emptyset \neq I \subseteq \operatorname{argmax}(Ax + b)$. It can be checked that the only non-empty subset of $[m]$ for which the set $\{\operatorname{sgn} a_i \mid i \in I\}$ is consistent is $I = [m]$. But there is no x such that $\operatorname{argmax}(Ax + b) = [m]$, i.e., at which all four affine functions are active.

How will the algorithm behave in this case? For the initial point $(x_1, x_2, x_3) = (0, 0, 0)$, the first three iterations of the algorithm are

j	x_1	x_2	x_3	y_1	y_2	y_3	y_4
	0	0	0	0	4	0	2
1	−1	0	0	−1	3	−1	3
2	−1	−2	0	1	3	−3	1
3	−1	−2	2	−1	3	−1	1

The resulting values of y are the initial values minus one. Every later cycle will again decrease them by one, therefore algorithm therefore will decrease y unboundedly.

Example 5. Consider the function

$$f(x_1, x_2, x_3) = \max\{\, x_1 - x_2 - x_3,\ x_1 + 4,\ x_1 + x_2 + x_3,\ -x_1 + x_2 + 2,\ 0 \,\}, \tag{19}$$

which differs from (18) only by the extra zero function. As shown in the previous example, the algorithm has no fixed point and the first four components of the vector $y = (y_1, y_2, y_3, y_4, y_5)$ will diverge. Since after a few iterations y_1, y_2, y_3, y_4 become all negative but y_5 remains zero, the set $\operatorname{cons}\{\operatorname{sgn} a_i \mid i \in \operatorname{argmax} y\}$ becomes non-empty.

The following theorem shows that every fixed point of coordinate minimization with the unique rule satisfies the sign relaxation of the condition $0 \in \partial f(x)$.

Theorem 9. If (17) holds for every $j \in [n]$, then the set $\{ \operatorname{sgn} a_i \mid i \in \operatorname{argmax} y \}$ is consistent.

Proof. For every $j \in [n]$, (17) implies that

$$\exists i \in \operatorname{argmax} y\colon a_{ij} < 0 \quad \Longleftrightarrow \quad \exists i' \in \operatorname{argmax} y\colon a_{i'j} > 0. \tag{20}$$

Indeed, if the common value of both sides in (17) is equal to [less than] $\max y$, both sides of (20) are true [false]. Condition (20) is equivalent to condition (11) for the set $\{ \operatorname{sgn} a_i \mid i \in \operatorname{argmax} y \}$. □

For an initial point $y \in \mathbb{R}^m$, consider the sequence of vectors $(p^k(y))_{k\in\mathbb{N}}$. Although we believe that, under some reasonably weak assumptions, this sequence converges to a fixed point, we are not able to prove this. Following [20], we formulate and prove a weaker result.

For $\epsilon \geq 0$, let

$$\operatorname{argmax}^\epsilon y = \{ i \in [m] \mid y_i + \epsilon \geq \max y \} \tag{21}$$

denote the set of ϵ-maximal components of a vector $y \in \mathbb{R}^m$. We now define function $e\colon \mathbb{R}^m \to \mathbb{R}_+$ by

$$e(y) = \inf\{ \epsilon \geq 0 \mid \operatorname{cons}\{ \operatorname{sgn} a_i \mid i \in \operatorname{argmax}^\epsilon y \} \neq \emptyset \}. \tag{22}$$

This function measures how much point y violates the condition[5] $\operatorname{cons}\{ \operatorname{sgn} a_i \mid i \in \operatorname{argmax} y \} \neq \emptyset$, which is the sign relaxation of the condition $0 \in \operatorname{conv}\{ a_i \mid i \in \operatorname{argmax} y \} = \partial f(x)$.

Theorem 10. Let $0 \in \operatorname{rint} \operatorname{conv}\{a_1, \dots, a_m\}$. Let $y \in \mathbb{R}^m$. Then $\lim_{k\to\infty} e(p^k(y)) = 0$.

Proof. In the appendix. □

4.1 Sum of Maxima of Affine Functions

We have applied coordinate minimization to pointwise maximum of affine functions (1). But what if we want to apply it to a convex piecewise-affine function in the form of a sum of pointwise maxima of affine function, i.e.,

$$f(x) = \sum_i \max_j (a_{ij}^T x + b_{ij}), \tag{23}$$

where $a_{ij} \in \mathbb{R}^n$ and $b_{ij} \in \mathbb{R}$. It turns out that minimizing function (23) can be easily transformed to minimizing a function (1). One way to do that is as follows.

Theorem 11. For every $\alpha_1, \dots, \alpha_m \in \mathbb{R}$,

$$\frac{1}{m} \sum_{i\in[m]} \alpha_i = \min_{u_1+\cdots+u_m=0} \; \max_{i\in[m]} (\alpha_i + u_i). \tag{24}$$

Proof. Clearly, at optimum all expressions under the maximum will have the same value. Let us denote this common value by $b = \alpha_i + u_i$, so $\alpha_i = b - u_i$. Now $\sum_i \alpha_i = mb - \sum_i u_i = mb$. □

Using Theorem 11, minimizing function (23) can be transformed to minimizing the function

$$g(x, u) = \max_{i,j} (a_{ij}^T x + b_{ij} + u_i)$$

subject to $\sum_i u_i = 0$. Minimization over u can be done in closed form, interlacing iterations of coordinate minimization over x.

[5] In [20], 'consistency' means $\operatorname{cons}\{ \operatorname{sgn} a_i \mid i \in \operatorname{argmax} y \} \neq \emptyset$, rather than consistency in the sense of our Definition 1.

A Proof of Theorem 10

A.1 Properties of the Algorithm

Proposition 12. For every $y \in \mathbb{R}^m$ and $j \in [n]$ we have $\max p_j(y) \le \max y$.

Proof. This just says that coordinate minimization never increase the objective function. $\square$

For $y \in \mathbb{R}^m$ and $z \in \mathbb{R}$, we denote

$$I(y, z) = \{\, i \in [m] \mid y_i \ge z \,\}. \tag{25}$$

In particular, note that $\operatorname{argmax} y = I(y, \max y)$ and $\operatorname{argmax}^\epsilon y = I(y, \max y - \epsilon)$.

Proposition 13. Let $y \in \mathbb{R}^m$ and $j \in [n]$.

- If the set $\{\, \operatorname{sgn} a_i \mid i \in I(y, \max y) \,\}$ is consistent in coordinate j, then $I(p_j(y), \max y) = I(y, \max y)$.
- Otherwise, $I(p_j(y), \max y) = I(y, \max y) \setminus \{\, i \mid a_{ij} \ne 0 \,\}$.

Proof. Denote $y' = p_j(y)$ and

$$\alpha^- = \max_{i \mid a_{ij} < 0} y_i, \qquad \alpha^+ = \max_{i \mid a_{ij} > 0} y_i, \qquad \alpha' = \max_{i \mid a_{ij} < 0} y'_i = \max_{i \mid a_{ij} \ne 0} y'_i = \max_{i \mid a_{ij} > 0} y'_i.$$

It follows from the definition of p_j that $y'_i = y_i$ for all i with $a_{ij} = 0$ and that α' lies between α^- and α^+.

By Definition 1, the set $\{\, \operatorname{sgn} a_i \mid i \in I(y, \max y) \,\}$ is consistent in coordinate j iff

$$\exists i \in I(y, \max y)\colon a_{ij} < 0 \iff \exists i' \in I(y, \max y)\colon a_{i'j} > 0. \tag{26}$$

Using the above observations, we see that:

- If both sides of (26) are true, then $\alpha^- = \alpha^+ = \alpha'$, hence $y' = y$, hence $I(y', \max y) = I(y, \max y)$.
- If both sides of (26) are false, then $\alpha^-, \alpha^+, \alpha' < \max y$. Hence $y_i, y'_i < \max y$ for all i with $a_{ij} \ne 0$. Hence $I(y', \max y) = I(y, \max y)$.
- If, say, the LHE of (26) is true and the RHS is false, then $\alpha^- = \max y$ and $\alpha^+, \alpha' < \max y$. Hence $y'_i < \max y$ for all i with $a_{ij} \ne 0$. Hence $I(y', \max y) = I(y, \max y) \setminus \{\, i \mid a_{ij} \ne 0 \,\}$. $\square$

Proposition 14. For every $y \in \mathbb{R}^m$ and every $k \ge m$ we have

$$\{\, \operatorname{sgn} a_i \mid i \in I(p^k(y), \max y) \,\} = \operatorname{cons}\{\, \operatorname{sgn} a_i \mid i \in I(y, \max y) \,\}.$$

Proof. Proposition 13 shows that the algorithm in fact enforces consistency (see Definition 3) of the set $\{\, \operatorname{sgn} a_i \mid i \in I(p^k(y), \max y) \,\}$. Moreover, in every m iterations each coordinate is visited and hence the set shrinks unless it is already consistent. As the set initially has no more than m elements, after m applications of p the set stops shrinking and becomes consistent. $\square$

Proposition 15. Let $y \in \mathbb{R}^m$.

(a) If $\operatorname{cons}\{\, \operatorname{sgn} a_i \mid i \in \operatorname{argmax} y \,\} \ne \emptyset$, then for every $k \in \mathbb{N}$ we have $\max p^k(y) = \max y$.

(b) If $\operatorname{cons}\{\, \operatorname{sgn} a_i \mid i \in \operatorname{argmax} y \,\} \ne \emptyset$, then the set $\{\, \operatorname{sgn} a_i \mid i \in \operatorname{argmax} p^m(y) \,\}$ is consistent.

(c) If $\operatorname{cons}\{\, \operatorname{sgn} a_i \mid i \in \operatorname{argmax} y \,\} = \emptyset$, then $\max p^m(y) < \max y$.

Proof. By Proposition 12, for every $k \in \mathbb{N}$ we have $\max p^k(y) \le \max y$. By (25) we have:

- If $I(p^k(y), \max y) \ne \emptyset$, then $\max p^k(y) = \max y$.
- If $I(p^k(y), \max y) = \emptyset$, then $\max p^k(y) < \max y$.

Noting that $\operatorname{argmax} y = I(y, \max y)$, the claims now follow from Proposition 14. $\square$

A.2 Continuity and Boundedness

Proposition 16. Let $a_1, \dots, a_m \in \mathbb{R}$ be such that $\{\operatorname{sgn} a_i \mid i \in [m]\} = \{-1, 1\}$. Then the function $\xi\colon \mathbb{R}^m \to \mathbb{R}$ given by

$$\xi(y) = \xi(y_1, \dots, y_m) = \operatorname*{argmin}_{x \in \mathbb{R}} \max_{i \in [m]} (a_i x + y_i) \tag{27}$$

is continuous.

Proof. On a neighborhood of any point $y \in \mathbb{R}^m$, function ξ depends only on the coordinates for which the affine functions $a_i x + y_i$ are active at y. Moreover, we can move the minimum to the origin without loss of generality. Therefore, to show that function (27) is continuous on $\mathbb{R}^m$, it suffices to show that the function of the form (27) is continuous at the point $y = 0$.

For any $y \in \mathbb{R}^m$, $\xi(y)$ is the x-coordinate of the intersection of the graphs of two affine functions $a_i x + y_i$ and $a_j x + y_j$, one with negative and one with positive slope. Thus, $\xi(y) = (y_i - y_j)/(a_j - a_i)$ for some (i, j) such that $a_i < 0$ and $a_j > 0$. Therefore,

$$\|y\|_\infty \le \delta \quad \Longrightarrow \quad |\xi(y)| \le \max_{i \mid a_i < 0} \max_{j \mid a_j > 0} \frac{|y_i - y_j|}{a_j - a_i} \le \delta \max_{i \mid a_i < 0} \max_{j \mid a_j > 0} \frac{2}{a_j - a_i}.$$

This shows that function ξ is continuous at $y = 0$. ☐

Proposition 17. For every $j \in [n]$, the map p_j is continuous.

Proof. Map p_j is continuous because it is a composition of function ξ from Proposition 16 and the affine map $y_i \mapsto a_{ij} x_j + y_i$ (as given by Step 3 of coordinate minimization). ☐

Proposition 18. Let $0 \in \operatorname{rint} \operatorname{conv}\{a_1, \dots, a_m\}$. Let $y \in \mathbb{R}^n$. Then the sequence $(p^k(y))_{k \in \mathbb{N}}$ is bounded.

Proof. By Proposition 12, for every k we have $p^k(y) \le \max y$. The claim now follows from Proposition 2. ☐

Proposition 19. Let $y, y' \in \mathbb{R}^m$, $\epsilon \ge 0$, $\delta \ge 0$, and $\|y - y'\|_\infty \le \delta$. Then $\operatorname{argmax}^\epsilon y \subseteq \operatorname{argmax}^{\epsilon + 2\delta} y'$.

Proof. Let $\|y - y'\|_\infty \le \delta$, i.e., $-\delta \le y_i - y'_i \le \delta$ for every i. This implies $-\delta \le \max y - \max y' \le \delta$. By these inequalities, for every i we have the implication

$$\max y - y_i \le \epsilon \quad \Longrightarrow \quad \max y' - y'_i \le \epsilon + 2\delta. \tag{28}$$

By (21), this means that $\operatorname{argmax}^\epsilon y \subseteq \operatorname{argmax}^{\epsilon + 2\delta} y'$. ☐

Proposition 20. The function e is continuous.

Proof. Let $\|y - y'\|_\infty \le \delta$. By (22), the set $\{\operatorname{sgn} a_i \mid i \in \operatorname{argmax}^{e(y)} y\}$ has a non-empty consistent subset. By Proposition 19, the set $\{\operatorname{sgn} a_i \mid i \in \operatorname{argmax}^{e(y) + 2\delta} y'\}$ has the same consistent subset, therefore $e(y') \le e(y) + 2\delta$. Similarly we prove that $e(y) \le e(y') + 2\delta$. Thus $|e(y) - e(y')| \le 2\delta$. ☐

Proposition 21. Let $\operatorname{cons}\{\operatorname{sgn} a_i \mid i \in [m]\} \neq \emptyset$. Then for every $y \in \mathbb{R}^n$ we have $e(y) \le \max y - \min y$.

Proof. Let $\epsilon = \max y - \min y$. Then clearly $\operatorname{argmax}^\epsilon y = [m]$, hence $\operatorname{cons}\{\operatorname{sgn} a_i \mid i \in \operatorname{argmax}^\epsilon y\} \neq \emptyset$. ☐

A.3 Convergence

Using the preparations from §A.1 and §A.2, we now prove the main convergence result. We will do it by reformulating [20, Theorem 1].

Let $q = p^m$ denote m cycles of coordinate minimization. In this section, y_k will denote a vector from $\mathbb{R}^m$, rather than the k-th component of a vector $y \in \mathbb{R}^m$. Sequences such as $(y_k)_{k \in \mathbb{N}}$ will be denoted in short as (y_k).

Recall that an *accumulation point* of a sequence is the limit point of its convergent subsequence.

Theorem 22. Let $q\colon \mathbb{R}^m \to \mathbb{R}^m$ be continuous. Let $y \in \mathbb{R}^m$. Let the sequence $(\max q^k(y))$ be convergent. Then every accumulation point y^* of the sequence $(q^k(y))$ satisfies $\max q(y^*) = \max y^*$.

Proof. For brevity, denote $y_k = q^k(y)$. Let y^* be an accumulation point of (y_k), thus

$$\lim_{l\to\infty} y_{k(l)} = y^* \tag{29}$$

for some strictly increasing function $k\colon \mathbb{N} \to \mathbb{N}$. Applying the continuous map q to equality (29) yields

$$\lim_{l\to\infty} q(y_{k(l)}) = \lim_{l\to\infty} y_{k(l)+1} = q(y^*), \tag{30}$$

where we used that $q(y_{k(l)}) = y_{k(l)+1}$. Now

$$\max y^* = \lim_{l\to\infty} \max y_{k(l)} = \lim_{k\to\infty} \max y_k = \lim_{l\to\infty} \max y_{k(l)+1} = \max q(y^*). \tag{31}$$

The first and last equality follow from applying the continuous function $\max\colon \mathbb{R}^m \to \mathbb{R}$ (defined by (2a)) to equalities (29) and (30). The second and third equality hold because the sequence $(\max y_k)$ is convergent and thus every its subsequence converges to the same number. $\square$

The following fact is well-known from analysis:

Proposition 23. Let (a_k) be a bounded sequence. If every convergent subsequence of (a_k) converges to a point a, then the sequence (a_k) converges to a.

Proof. Suppose (a_k) does not converge to a. Then for some $\epsilon > 0$, for every k_0 there is $k > k_0$ such that $\|a_k - a\| > \epsilon$. So there is a subsequence (b_k) such that $\|b_k - a\| > \epsilon$ for all k. As (b_k) is bounded, by Bolzano-Weierstrass it has a convergent subsequence, (c_k). But (c_k) clearly cannot converge to a, a contradiction. $\square$

Theorem 24. Let $q\colon \mathbb{R}^m \to \mathbb{R}^m$ and $e\colon \mathbb{R}^m \to \mathbb{R}$ be continuous such that, for every $y \in \mathbb{R}^m$:

1. $\max q(y) \leq \max y$,
2. $\max q(y) = \max y$ implies $e(y) = 0$,
3. the sequences $(q^k(y))$, $(\max q^k(y))$ and $(e(q^k(y)))$ are bounded.

Then for every $y \in \mathbb{R}^m$ we have $\lim\limits_{k\to\infty} e(q^k(y)) = 0$.

Proof. Denote $y_k = q^k(y)$. The sequence $(\max y_k)$ is bounded and non-increasing, therefore convergent. By Theorem 22, every accumulation point y^* of (y_k) satisfies $\max q(y^*) = \max y^*$. This implies $e(y^*) = 0$.

By Proposition 23, now it suffices to show that every convergent subsequence of the sequence $(e(y_k))$ converges to 0. So let (z_k) be a subsequence of (y_k) such that $\lim_{k\to\infty} e(z_k) = e^*$.

- If (z_k) is convergent, then $y^* = \lim_{k\to\infty} z_k$ is an accumulation point of (y_k), therefore $e(y^*) = 0$. Applying the continuous function e to this limit yields $e(y^*) = \lim_{k\to\infty} e(z_k) = e^* = 0$.
- If (z_k) is not convergent, by Bolzano-Weierstrass it has a convergent subsequence, (w_k). As (w_k) is also a subsequence of (y_k), by the above reasoning we have $\lim_{k\to\infty} e(w_k) = 0$. But because $\lim_{k\to\infty} e(z_k) = e^*$, every subsequence of $(e(z_k))$ converges to e^*. As (w_k) is a subsequence of (z_k), this implies $e^* = 0$. $\square$

Theorem 24 implies Theorem 10. Indeed, map q is continuous because it is a composition of maps p_j which are continuous by Proposition 17. Function e is continuous by Proposition 20. Condition 1 holds by Proposition 12 and Condition 2 by Proposition 15(c). The sequences $(q^k(y))$ and $(\max q^k(y))$ are bounded by Proposition 18. The sequence $(e(q^k(y)))$ is bounded by Proposition 21.

We remark that Theorem 24 has a wider applicability, to prove convergence to local consistency for several other message-passing algorithms. For that, the functions max, q and e need to be replaced by appropriate functions in these algorithms and they must satisfy the assumptions of he theorem.

Acknowledgement

This work has been supported by the Czech Science Foundation grant 16-05872S.

References

[1] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 2nd edition, 1999.

[2] C. Bessiere. Constraint propagation. In *Handbook of Constraint Programming*, chapter 3. Elsevier, 2006.

[3] M. C. Cooper, S. de Givry, M. Sanchez, T. Schiex, M. Zytnicki, and T. Werner. Soft arc consistency revisited. *Artificial Intelligence*, 174(7-8):449–478, 2010.

[4] E. Freuder and A. K. Mackworth. Constraint satisfaction: An emerging paradigm. In *Handbook of Constraint Programming*, chapter 2. Elsevier, 2006.

[5] A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Neural Information Processing Systems*, pages 553–560, 2008.

[6] T. Hazan and A. Shashua. Convergent message-passing algorithms for inference over general graphs with convex free energies. In *Conf. on Uncertainty in Artificial Intelligence*, pages 264–273, 2008.

[7] J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis.* Grundlehren Text Editions. Springer, 2004.

[8] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, T. Kröger, J. Lellmann, N. Komodakis, B. Savchynskyy, and C. Rother. A comparative study of modern inference techniques for structured discrete energy minimization problems. *Intl. J. of Computer Vision*, 115(2):155–184, 2015.

[9] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(10):1568–1583, 2006.

[10] V. Kolmogorov. A new look at reweighted message passing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(5), May 2015.

[11] N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):531–552, 2011.

[12] V. K. Koval and M. I. Schlesinger. Dvumernoe programmirovanie v zadachakh analiza izobrazheniy (Two-dimensional programming in image analysis problems). *Automatics and Telemechanics*, 8:149–168, 1976. In Russian.

[13] V. A. Kovalevsky and V. K. Koval. A diffusion algorithm for decreasing the energy of the max-sum labeling problem. Glushkov Institute of Cybernetics, Kiev, USSR. Unpublished, approx. 1975.

[14] A. Mackworth. Constraint satisfaction. In *Encyclopaedia of Artificial Intelligence*, pages 285–292. Wiley, 1991.

[15] A. L. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing. An augmented Lagrangian approach to constrained MAP inference. In *Intl. Conf. on Machine Learning*, pages 169–176, 2011.

[16] T. Meltzer, A. Globerson, and Y. Weiss. Convergent message passing algorithms: a unifying view. In *Conf. on Uncertainty in Artificial Intelligence*, pages 393–401, 2009.

[17] P. Meseguer, F. Rossi, and T. Schiex. Soft constraints. In *Handbook of Constraint Programming*, chapter 9. Elsevier, 2006.

[18] B. Savchynskyy. A bundle approach to efficient MAP-inference by Lagrangian relaxation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1688–1695, 2012.

[19] B. Savchynskyy, S. Schmidt, J. H. Kappes, and C. Schn orr. Efficient MRF energy minimization via adaptive diminishing smoothing. In *Conf. on Uncertainty in Artificial Intelligence*, pages 746–755, 2012.

[20] M. I. Schlesinger and K. Antoniuk. Diffusion algorithms and structural recognition optimization problems. *Cybernetics and Systems Analysis*, 47:175–192, 2011.

[21] M. I. Schlesinger and V. V. Giginjak. Solving (max,+) problems of structural pattern recognition using equivalent transformations. *Upravlyayushchie Sistemy i Mashiny (Control Systems and Machines), Kiev, Naukova Dumka*, 1 and 2, 2007. In Russian, English translation available on www.

[22] D. Sontag, D. K. Choe, and Y. Li. Efficiently searching for frustrated cycles in MAP inference. In *Conf. on Uncertainty in Artificial Intelligence*, pages 795–804, Corvallis, Oregon, 2012.

[23] P. Swoboda, J. Kuske, and B. Savchynskyy. A dual ascent framework for lagrangean decomposition of combinatorial problems. In *The IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[24] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwal, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields. In *Eur. Conf. on Computer Vision*, pages II: 16–29, 2006.

[25] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, June 2001.

[26] S. Živný. *The Complexity of Valued Constraint Satisfaction Problems.* Cognitive Technologies. Springer, 2012.

[27] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[28] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(7):1165–1179, July 2007.

[29] T. Werner. Revisiting the linear programming relaxation approach to Gibbs energy minimization and weighted constraint satisfaction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(8):1474–1488, August 2010.