Relative Interior Rule in Block-Coordinate Descent

Tomáš Werner, Daniel Průša, Tomáš Dlask
Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
{werner,prusapa1,dlaskto2}@fel.cvut.cz

Abstract

It is well-known that for general convex optimization problems, block-coordinate descent can get stuck in poor local optima. Despite that, versions of this method known as convergent message passing are very successful to approximately solve the dual LP relaxation of the MAP inference problem in graphical models. In attempt to identify the reason why these methods often achieve good local minima, we argue that if in block-coordinate descent the set of minimizers over a variable block has multiple elements, one should choose an element from the relative interior of this set. We show that this rule is not worse than any other rule for choosing block-minimizers. Based on this observation, we develop a theoretical framework for block-coordinate descent applied to general convex problems. We illustrate this theory on convergent message-passing methods.

1. Introduction

Block-coordinate descent (BCD) is an iterative optimization method which in every iteration finds a global optimum of the problem over a subset of variables, keeping the remaining variables fixed. For some problems, fixed points of BCD and cluster points of the sequence generated by it are global optima, see, e.g., [20] and the references therein. Focusing on convex problems, BCD can be made very efficient and scalable provided that optimality can be guaranteed, as in [13, 5, 1]. For general convex problems, BCD fixed/cluster points can be arbitrarily poor local minima (where 'local' is meant with respect to block-coordinate moves). Therefore, BCD is mostly regarded unsuitable for general convex problems.

An exception is the class of methods known as convergent message passing, used to approximately solve the linear-programming (LP) relaxation of the MAP inference problem in graphical models [19, 7] (frequently used to model low-level computer vision tasks such as denoising, segmentation or registration) and some other combinatorial problems [18]. These methods apply various forms of BCD to various forms of the dual LP relaxation, where the latter boils down to the unconstrained minimization of a piecewise-affine (hence non-differentiable) convex function. Examples are max-sum diffusion [11, 17, 23], TRW-S [8], MPLP [2], SRMP [9], and [3, 12]. For many problems from computer vision, TRW-S is typically faster than all competing methods and its fixed points are not far from global minima, especially for large sparse instances [19, 7]. This motivates us to study convergent message-passing methods independently of MAP inference, with the hope of extending them to a wider class of convex problems.

One might think that convergent message-passing methods are 'just' applications of BCD to suitable forms of the dual LP relaxation. However, this is not the whole explanation: we believe these methods have a single feature that allows them to achieve good local optima. In a BCD iteration, the minimizer over a variable block need not be unique and therefore a single minimizer must be chosen. We argue that this minimizer should be chosen from the relative interior of the set of all minimizers over the variable block. We call this the *relative interior rule*.

Based on this observation, we develop a theoretical framework for BCD applied to general convex problems. We distinguish three types of block-coordinate local minima: (ordinary) local minima, interior local minima, and pre-interior local minima. We show that the relative interior rule is not worse than any other rule to choose variable-block minimizers, in the following sense: starting from any non-pre-interior local minimum, BCD satisfying the relative interior rule inevitably improves the objective; starting from any pre-interior local minimum, BCD (not necessarily satisfying the relative interior rule) never improves the objective. Assuming a linear objective function, we show that local and interior local minima form sets of faces of the feasible set, which are closed under intersection. Inspired by the proof in [17] (revisited in [16, §8]), we prove convergence of BCD satisfying the relative interior rule to the set of pre-interior local minima. We show how well-known convergent message-passing methods fit in our theory. Here, local minimality conditions induced by the relative interior rule correspond to *local consistencies*, such as arc consistency [23] or weak-tree agreement [8]. We also sketch applications to some new problems.

2. Summary of Main Results

Suppose we want to minimize a convex function $f\colon V \to \mathbb{R}$ on a closed convex set $X \subseteq V$, where V is a finite-dimensional vector space over $\mathbb{R}$. For that, we consider the following coordinate-free generalization of block-coordinate descent. For brevity, for any $Y \subseteq V$ we will use $M_f(Y)$ to denote the set of all global minimizers of f on the set Y. Let $\mathcal{I}$ be a finite set of subspaces of V, representing permitted search directions. Having an estimate $x_n \in X$ of the minimum, the next estimate x_{n+1} is chosen such that

$$x_{n+1} \in M_f(X \cap (x_n + I_n)) \tag{1}$$

for[1] some $I_n \in \mathcal{I}$. Clearly, $f(x_{n+1}) \leq f(x_n)$. A point $x \in X$ satisfying

$$x \in M_f(X \cap (x + I)) \qquad \forall I \in \mathcal{I} \tag{2}$$

has the property that f cannot be improved by moving from x within X along any single subspace from $\mathcal{I}$. We call such a point a *local minimum* of f on X with respect to $\mathcal{I}$. When $\mathcal{I}$ and/or (X, f) is clear from context, we will speak only about a local minimum of f on X or just a local minimum. Note, we use the term 'local minimum' in a different meaning than is usual in optimization and calculus.

Coordinate descent and block-coordinate descent are special cases of this formulation. In the former, we have $V = \mathbb{R}^d$ and $\mathcal{I} = \{\operatorname{span}\{e_1\}, \dots, \operatorname{span}\{e_d\}\}$ where e_i denotes the ith standard basis vector of $\mathbb{R}^d$. In the latter, we have $V = \mathbb{R}^d$ and each element of $\mathcal{I}$ is the span of a subset of the standard basis of $\mathbb{R}^d$.

Recall [15, 4] that the *relative interior* of a convex set $X \subseteq V$ is the topological interior of X with respect to the affine hull of X. We will denote it by $\operatorname{ri} X$. We propose to modify condition (1) such that the minimum is always chosen from the relative interior of the current optimal set. Thus, condition (1) changes to

$$x_{n+1} \in \operatorname{ri} M_f(X \cap (x_n + I_n)). \tag{3}$$

A point x_{n+1} always exists because the relative interior of every non-empty convex set is non-empty. We call a point $x \in X$ that satisfies

$$x \in \operatorname{ri} M_f(X \cap (x + I)) \qquad \forall I \in \mathcal{I} \tag{4}$$

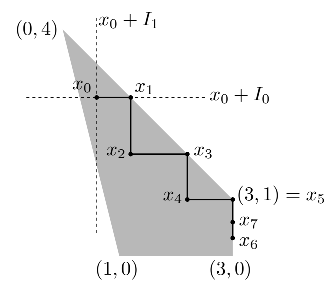
an *interior local minimum* of f on X with respect to $\mathcal{I}$. Clearly, every interior local minimum is a local minimum.

In our analysis, another type of local minimum will appear: *pre-interior local minimum*. It will be defined later, now we just say that it is only a finite number of iterations (3) away from an interior local minimum.

Consider a sequence $(x_n)_{n=0}^\infty$ satisfying (1) resp. (3). To ensure that each search direction is always visited again after a finite number of iterations, we assume that the sequence $(I_n)_{n=0}^\infty$ contains each element of $\mathcal{I}$ an infinite number of times. For brevity, we will often write only (x_n) and (I_n) instead of $(x_n)_{n=0}^\infty$ and $(I_n)_{n=0}^\infty$. The following facts, proved in the sequel, show that methods satisfying (3) are not worse, in a precise sense, than methods satisfying (1):

- For every sequence (x_n) satisfying (3), if x_0 is an interior local minimum then x_n is an interior local minimum for all n (see Theorem 11).
- For every sequence (x_n) satisfying (3), if x_0 is a pre-interior local minimum then x_n is an interior local minimum for some n (see Corollary 14).
- For every sequence (x_n) satisfying (1), if x_0 is a pre-interior local minimum then $f(x_n) = f(x_0)$ for all n (see Theorem 13).
- For every sequence (x_n) satisfying (3), if x_0 is not a pre-interior local minimum then $f(x_n) < f(x_0)$ for some n (see Theorem 12).

As an illustrative example, consider coordinate descent applied to a simple linear program. Let $V = \mathbb{R}^2$, $X = \operatorname{conv}\{(1,0),(3,0),(3,1),(0,4)\}$, $f(x) = \langle -e_1, x\rangle$ (i.e., f is constant vertically and decreases to the right), and $\mathcal{I} = \{\operatorname{span}\{e_1\}, \operatorname{span}\{e_2\}\}$. See the picture:



The set of global minima is the line segment $[(3,0),(3,1)]$, the set of local minima is $[(3,0),(3,1)] \cup [(0,4),(3,1)]$, the set of interior local minima is $\{(0,4)\} \cup \operatorname{ri}[(3,0),(3,1)]$, and the set of pre-interior local minima is $\{(0,4)\} \cup [(3,0),(3,1)]$. The thick polyline depicts the first few points of a sequence (x_n) satisfying (3), assuming that the sequence (I_n) alternates between the two subspaces from $\mathcal{I}$. When starting from any point $x_0 \in X \setminus \{(0,4)\}$, every sequence (x_n) satisfying (3) leaves any non-interior local minimum after a finite number of iterations, while improving the objective function. Intuitively, this is because when the objective cannot be decreased by moving along any single subspace from $\mathcal{I}$, condition (3) at least enforces the point to move to a face of X of a higher dimension (if one exists), providing thus 'more room' to hopefully decrease the objective in future iterations. In contrast, condition (1) allows a sequence (x_n) to stay in any (possibly non-interior) local minimum forever. When starting from $x_0 = (0,4)$, every sequence satisfying (1) will stay in x_0 forever.

[1] For $x \in V$ and $I \subseteq V$, we denote $x + I = \{\, x + y \mid y \in I \,\}$.

We show (in Theorem 22) that after fixing the choices of minimizers in (3), under natural assumptions the sequence (x_n) satisfying (3) converges to the set of pre-interior local minima.

It is well-known that every convex optimization problem can be stated in the epigraph form, which has a linear objective: instead of minimizing $f(x)$ over $x \in X$, we minimize t over $(x,t) \in X \times \mathbb{R}$ subject to $f(x) \le t$. It is not hard to show (see the supplement) that the notions of (interior) local minima and the updates (1) and (3) remain 'the same' if we pass between the two formulations. To illustrate this, consider the case $X = V = \mathbb{R}^d$ and coordinate descent. In every iteration, we minimize $f(x_1,\dots,x_d)$ over a single variable x_i while in the epigraph form, we minimize t subject to $f(x_1,\dots,x_d) \le t$ over the pair (x_i,t). Clearly, both forms are equivalent. Therefore, in §3, §4 and §5 we will assume that f is linear.

3. Structure of the Set of Local Minima

It is well-known that the set $M_f(X)$ of global minima of a linear function f on a closed convex set X is an (exposed) face of X. We show that local resp. interior local minima also cluster to faces of X. Moreover, similarly as the set of all faces, we show that the set of faces containing local resp. interior local minima are closed under intersections.

For $x, y \in V$, we denote

$$[x,y] = \operatorname{conv}\{x,y\} = \{\, (1-\alpha)x + y \mid 0 \le \alpha \le 1 \,\}. \quad (5)$$

We have

$$\operatorname{ri}[x,y] = \{\, (1-\alpha)x + y \mid 0 < \alpha < 1 \,\}. \quad (6)$$

If $x \ne y$, then $[x,y]$ is a line segment and $\operatorname{ri}[x,y] = [x,y] \setminus \{x,y\}$. If $x = y$, then $[x,y] = \operatorname{ri}[x,y] = \{x\}$.

Let us recall basic facts about faces of a convex set [15, 4]. A *face* of a convex set $X \subseteq V$ is a convex set $F \subseteq X$ such that every line segment from X whose relative interior intersects F lies in F, i.e.,

$$x, y \in X, \quad F \cap \operatorname{ri}[x,y] \ne \emptyset \quad \Longrightarrow \quad x, y \in F. \quad (7)$$

The set of all faces of a closed convex set partially ordered by inclusion is a complete lattice, in particular it is closed under (possibly uncountable) intersections. For a point $x \in X$, let $F(X,x)$ denote the intersection of all faces (equivalently, the smallest face) of X containing x. For every $x, y \in X$,

$$\begin{aligned}
y &\in F(X,x) &&\iff && F(X,y) \subseteq F(X,x), && (8a)\\
y &\in \operatorname{ri} F(X,x) &&\iff && F(X,y) = F(X,x), && (8b)\\
y &\in \operatorname{rb} F(X,x) &&\iff && F(X,y) \subsetneq F(X,x), && (8c)
\end{aligned}$$

where $\operatorname{rb} X = X \setminus \operatorname{ri} X$ denotes the relative boundary of a convex set X. Equivalence (8b) shows that $F(X,x)$ is in fact the unique face of X having x in its relative interior. Note that (8c) follows from (8a) and (8b).

Lemma 1. *Let $X \subseteq V$ be a convex set. We have $x \in \operatorname{ri} X$ iff for every $y \in X$ there exists $u \in X$ such that $x \in \operatorname{ri}[y,u]$.*

Proof. The 'only-if' direction is immediate from the definition of relative interior. For the 'if' direction see, e.g., [15, Theorem 6.4]. □

Lemma 2. *Let $X, Y \subseteq V$ be closed convex sets such that $Y \subseteq X$. Let $x \in \operatorname{ri} Y$. Then*

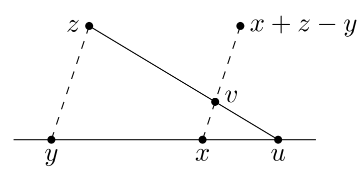
$$\begin{aligned}
y &\in Y &&\Longrightarrow && y \in F(X,x), && (9a)\\
y &\in \operatorname{ri} Y &&\Longrightarrow && y \in \operatorname{ri} F(X,x), && (9b)\\
y &\in \operatorname{rb} Y &&\Longrightarrow && y \in \operatorname{rb} F(X,x). && (9c)
\end{aligned}$$

Proof. For (9a), let $x \in \operatorname{ri} Y$ and $y \in Y$. Thus, by Lemma 1 there is $u \in Y$ such that $x \in \operatorname{ri}[u,y]$. Since $x \in F(X,x)$ and $y, u \in X$, the definition of face yields $y \in F(X,x)$. Implications (9b) and (9c) follow from (9a) and (8). □

Lemma 3. *Let $y, z, u \in V$ and $x \in \operatorname{ri}[u,y]$. Then we have $\operatorname{ri}[u,z] \cap \operatorname{ri}[x, x+z-y] \ne \emptyset$.*

Proof. Since $x \in \operatorname{ri}[u,y]$, there is $0 < \alpha < 1$ such that $x = (1-\alpha)u + \alpha y$ (note that if $y \ne u$ then α is unique). Let $v = (1-\alpha)u + \alpha z$, hence $v \in \operatorname{ri}[u,z]$. Subtracting the two equations yields $v = (1-\alpha)x + \alpha(x+z-y)$, hence $v \in \operatorname{ri}[x, x+z-y]$. □

The picture illustrates Lemma 3 for the points in a general position (i.e., y, z, u not collinear):



In the theorems in the rest of this section, the letter 'I' will always denote a subspace of V.

Theorem 4. *Let $x \in M_f(X \cap (x+I))$ and $y \in F(X,x)$. Then $y \in M_f(X \cap (y+I))$.*

Proof. Let $z \in X \cap (y+I)$. We need to prove that $f(y) \le f(z)$. Since $y \in F(X,x)$, by Lemma 1 there is $u \in X$ such that $x \in \operatorname{ri}[u,y]$. By Lemma 3, there is a point

$$v \in \operatorname{ri}[u,z] \cap \operatorname{ri}[x, x+z-y].$$

Since $z, u \in X$, by convexity of X we have $v \in X$. Since $z - y \in I$, we have $v \in x + I$. Since $x \in M_f(X \cap (x+I))$, we thus have $f(x) \le f(v)$, hence $f(x) \le f(x+z-y)$. Since $[x, x+z-y] = [y,z] + x - y$, by linearity of f we have $f(y) \le f(z)$. □

Corollary 5. *If x is a local minimum, then every point of $F(X, x)$ is a local minimum.*

Let us emphasize that if x and y are local minima and $y \in F(X, x)$, then we can have $f(y) \neq f(x)$.

Lemma 6. *Let $x \in \operatorname{ri} M_f(X \cap (x + I))$ and $y \in F(X, x)$. Then $M_f(X \cap (y + I)) \subseteq F(X, x)$.*

Proof. Let $z \in M_f(X \cap (y + I))$. By Theorem 4 we have $y \in M_f(X \cap (y + I))$, hence $f(z) = f(y)$. Since $y \in F(X, x)$, by Lemma 1 there is $u \in X$ such that $x \in \operatorname{ri}[u, y]$. By Lemma 3, there is a point

$$v \in \operatorname{ri}[u, z] \cap \operatorname{ri}[x, x + z - y].$$

Since $z, u \in X$ and $z - y \in I$, we have $v \in X \cap (x + I)$. Since $[x, x + z - y] = [y, z] + x - y$, by linearity of f we have $f(v) = f(x)$, hence $v \in M_f(X \cap (x + I))$. Lemma 2 yields $v \in F(X, x)$. Since $z, u \in X$, the definition of face yields $z \in F(X, x)$. $\square$

Lemma 7. *Let $x \in M_f(X \cap (x + I)) \subseteq F(X, x)$. Then $x \in \operatorname{ri} M_f(X \cap (x + I))$.*

Proof. Let $u \in M_f(X \cap (x + I))$. Hence $f(u) = f(x)$. Moreover, by Lemma 1 there is $v \in F(X, x)$ such that $x \in \operatorname{ri}[u, v]$. Since $u \in x + I$, we have $v \in x + I$. By linearity of f we have $f(v) = f(x)$, thus $v \in M_f(X \cap (x + I))$. By Lemma 1, $x \in \operatorname{ri} M_f(X \cap (x + I))$. $\square$

Theorem 8. *Let $Y \subseteq X$. Let $x \in \operatorname{ri} M_f(X \cap (x + I))$ for all $x \in Y$. Let $y \in \operatorname{ri} \bigcap_{x \in Y} F(X, x)$. Then $y \in \operatorname{ri} M_f(X \cap (y + I))$.*

Proof. Since $G = \bigcap_{x \in Y} F(X, x)$ is a face of X, we have $y \in \operatorname{ri} G$ iff $G = F(X, y)$. By Theorem 4, we have $y \in M_f(X \cap (y + I))$. By Lemma 6, $M_f(X \cap (y + I)) \subseteq G$. By Lemma 7, $y \in \operatorname{ri} M_f(X \cap (y + I))$. $\square$

Corollary 9. *Let $Y \subseteq X$ be a set of interior local minima. Then every relative interior point of the face $\bigcap_{x \in Y} F(X, x)$ is an interior local minimum.*

Corollary 10. *If x is an interior local minimum, then every point of $\operatorname{ri} F(X, x)$ is an interior local minimum.*

The results of this section can be summarized as follows:

- Let us call a face of X a *local minima face* if all its points are local minima. Since the set of faces of X is closed under intersection, it follows from Corollary 5 that the set of all local minima faces of X (assuming fixed f and $\mathcal{I}$) is closed under intersections.
- Let us call a face of X an *interior local minima face* if all its relative interior points are interior local minima. Corollary 9 shows that the set of all interior local minima faces of X is closed under intersections.

We finally define one more type of local minimum: a point x is a *pre-interior local minimum* if $x \in F(X, y)$ for some interior local minimum y.

4. The Effect of Iterations

Here we prove properties of sequences (x_n) satisfying conditions (1) or (3) under various assumptions.

Theorem 11. *Let (x_n) be a sequence satisfying (3) such that x_0 is an interior local minimum. Then the following hold for all n: $f(x_n) = f(x_0)$, $x_n \in \operatorname{ri} F(X, x_0)$, and x_n is an interior local minimum.*

Proof. Suppose that for some n, x_n is an interior local minimum. Considering (3), by Lemma 2 we thus have $x_{n+1} \in \operatorname{ri} F(X, x_n)$. By Corollary 9, x_{n+1} is an interior local minimum. Since $x_n, x_{n+1} \in \operatorname{ri} M_f(X \cap (x_n + I_n))$, we have $f(x_{n+1}) = f(x_n)$. $\square$

Theorem 12. *Let (x_n) be a sequence satisfying (3) and $f(x_n) = f(x_0)$ for all n. Then the following hold: $x_n \in F(X, x_{n+1})$ for all n, x_n is an interior local minimum for some n, and x_0 is a pre-interior local minimum.*

Proof. Since $f(x_{n+1}) \leq f(x_n) = f(x_0)$ for all n, we have $f(x_{n+1}) = f(x_n)$ for all n. Combining this with (3) yields $x_n \in M_f(X \cap (x_n + I_n))$. Thus, for every n there are two possibilities:

- If $x_n \in \operatorname{ri} M_f(X \cap (x_n + I_n))$ then, by Lemma 2, we have $x_n \in \operatorname{ri} F(X, x_{n+1})$. By Theorem 8, we have $x_{n+1} \in \operatorname{ri} M_f(X \cap (x_{n+1} + I))$ for all $I \in \mathcal{I}$ such that $x_n \in \operatorname{ri} M_f(X \cap (x_n + I))$.
- If $x_n \in \operatorname{rb} M_f(X \cap (x_n + I_n))$ then, by Lemma 2, we have $x_n \in \operatorname{rb} F(X, x_{n+1})$.

In either case, $x_n \in F(X, x_{n+1})$. Moreover, if x_n is not an interior local minimum for some n, then after some finite number m of iterations the second case occurs (recall, we assume that (I_n) contains every element of $\mathcal{I}$ an infinite number of times), therefore $x_n \in \operatorname{rb} F(X, x_{n+m})$. But this implies $\dim F(X, x_{n+m}) > \dim F(X, x_n)$. If x_n were not an interior local minimum for any n, for some n we would have $\dim F(X, x_n) > \dim X$, which is impossible.

Since $x_n \in F(X, x_{n+1})$ for all n, the faces $F(X, x_0) \subseteq F(X, x_1) \subseteq \cdots$ form a non-decreasing chain. In particular, $x_0 \in F(X, x_n)$ for all n. Since x_n is an interior local minimum for some n, x_0 is a pre-interior local minimum. $\square$

Theorem 13. *Let (x_n) be a sequence satisfying (1) such that x_0 is a pre-interior local minimum, i.e., $x_0 \in F(X, x)$ for some interior local minimum x. Then for all n we have $x_n \in F(X, x)$ and $f(x_n) = f(x_0)$.*

Proof. We will use induction on n. The claim trivially holds for $n = 0$. We will show that for every n, $x_n \in F(X, x)$ implies $x_{n+1} \in F(X, x)$ and $f(x_{n+1}) = f(x_n)$.

Let $x_n \in F(X, x)$. By Lemma 1, there is $u \in X$ such that $x \in \operatorname{ri}[x_n, u]$. By Lemma 3, there is a point

$$v \in \operatorname{ri}[u, x_{n+1}] \cap \operatorname{ri}[x, x + x_{n+1} - x_n].$$

Since $u, x_{n+1} \in X$, we have $v \in X$. Since $x_{n+1} - x_n \in I_n$, we have $v \in x + I_n$. Since $x \in M_f(X \cap (x + I_n))$, this implies $f(x) \le f(v)$. Since $[x, x + x_{n+1} - x_n] = [x_n, x_{n+1}] + x - x_n$, by linearity of f we have $f(x_n) \le f(x_{n+1})$. But from (1) also $f(x_{n+1}) \le f(x_n)$, hence $f(x_{n+1}) = f(x_n)$. This implies $f(v) = f(x)$. Since $x \in \operatorname{ri} M_f(X \cap (x + I_n))$, we have $v \in M_f(X \cap (x + I_n))$. By Lemma 2, $v \in F(X, x)$. Since $u, x_{n+1} \in X$ and $v \in F(X, x)$, the definition of face gives $x_{n+1} \in F(X, x)$. $\square$

Corollary 14. *Let (x_n) be a sequence satisfying (3) such that x_0 is a pre-interior local minimum. Then there exists n such that x_n is an interior local minimum.*

Proof. Apply first Theorem 13 and then Theorem 12. $\square$

Corollary 15. *For every sequence (x_n) satisfying (3), x_0 is a pre-interior local minimum iff $f(x_n) = f(x_0)$ for all n.*

Proof. The 'if' direction follows from Theorem 12. The 'only-if' direction follows from Theorem 13. $\square$

5. Convergence

Here we examine convergence properties of sequences satisfying (3). We first give a general convergence result in §5.1 and then apply it to our situation in §5.2.

5.1. General Convergence Result

Let $p\colon X \to X$ and $f\colon X \to \mathbb{R}$ be continuous functions. Let (x_n) be a sequence satisfying

$$x_{n+1} = p(x_n) \qquad \forall n = 0, 1, \ldots. \tag{10}$$

Let

$$X^* = \{\, x \in X \mid f(p(x)) = f(x) \,\}. \tag{11}$$

Theorem 16. *If the sequence $(f(x_n))_{n=0}^{\infty}$ is convergent, then every cluster point[2] of the sequence (x_n) is in X^*.*

Proof. Let x be a cluster point of the sequence (x_n), i.e., for some strictly increasing sequence (k_n) we have

$$\lim_{n\to\infty} x_{k_n} = x. \tag{12}$$

Applying the continuous map p to (12) yields

$$p\big(\lim_{n\to\infty} x_{k_n}\big) = \lim_{n\to\infty} p(x_{k_n}) = \lim_{n\to\infty} x_{k_n+1} = p(x). \tag{13}$$

We show that

$$\begin{aligned} f(x) &= \lim_{n\to\infty} f(x_{k_n}) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} f(x_{k_n+1}) \\ &= f(p(x)). \end{aligned}$$

The first and last equality hold by applying the continuous function f to equality (12) and (13). The second and third equality hold because the sequence $(f(x_n))$ converges, thus every its subsequence converges to the same number. $\square$

[2]A *cluster point* (also known as *limit point* or *accumulation point*) of a sequence is the point of convergence of its converging subsequence.

Let $d\colon X^2 \to \mathbb{R}_+$ be a metric on X. Let

$$d(Y, x) = \inf_{y \in Y} d(x, y) \tag{14}$$

denote the distance of a point $x \in X$ from a set $Y \subseteq X$.

Lemma 17. *For every $Y \subseteq X$, the function $d(Y, \cdot)$ is Lipschitz.*

Proof. For every $x, y \in X$ and $z \in Y$ we have $d(Y, x) \le d(x, z) \le d(x, y) + d(y, z)$. Taking inf over $z \in Y$ on the right gives $d(Y, x) \le d(x, y) + d(Y, y)$. Swapping x and y gives $|d(Y, x) - d(Y, y)| \le d(x, y)$. $\square$

Lemma 18. *A sequence of real numbers is convergent if it is bounded and has a unique cluster point.*

Proof. Let x be a cluster point of a bounded sequence (x_n). Suppose (x_n) does not converge to x. Then for some $\epsilon > 0$, for every m there is $n > m$ such that $|x_n - x| > \epsilon$. So (x_n) has a subsequence (y_n) such that $|y_n - x| > \epsilon$ for all n. As (y_n) is bounded, it has a convergent subsequence, (z_n). But (z_n) clearly cannot converge to x, a contradiction. $\square$

Theorem 19. *If the sequence $(f(x_n))$ is convergent and the sequence (x_n) is bounded, then $\lim_{n\to\infty} d(X^*, x_n) = 0$.*

Proof. Since the function $d(X^*, \cdot)$ is Lipschitz and the sequence (x_n) is bounded, the sequence $(d(X^*, x_n))$ is bounded. Thus, it has a convergent subsequence, $(d(X^*, y_n))$ where (y_n) is a subsequence of (x_n). By Lemma 18, it suffices to show that $\lim_{n\to\infty} d(X^*, y_n) = 0$.

As a subsequence of (x_n), the sequence (y_n) is bounded. Thus it has a convergent subsequence[3], (z_n). Thus,

$$x = \lim_{n\to\infty} z_n \tag{15}$$

is a cluster point of (x_n). We claim that

$$0 = d(X^*, x) = \lim_{n\to\infty} d(X^*, z_n) = \lim_{n\to\infty} d(X^*, y_n).$$

The first equality holds by Theorem 16. The second equality is obtained by applying the continuous function $d(X^*, \cdot)$ to (15). The last equality holds because the sequence $(d(X^*, y_n))$ is convergent, hence its subsequence $(d(X^*, z_n))$ converges to the same number. $\square$

Note, Theorem 19 does not imply that (x_n) converges to any point, it only says that (x_n) converges to the set X^*. Neither it implies that the map p has a fixed point. We remark that Theorem 19 remains true if the function $d(X^*, \cdot)$ is replaced by any Lipschitz function $e\colon X \to \mathbb{R}$ such that $e(x) = 0$ iff $x \in X^*$. One such function was proposed for max-sum diffusion in [17], see also [16, §8].

[3]Because (x_n) is contained in a closed convex subset X of a finite-dimensional real vector space V.

5.2. Convergence for the Relative Interior Rule

To apply this result to sequences satisfying the relative interior rule, we fix the choice of minimizers in (3) by assuming that for each $I \in \mathcal{I}$, a continuous map $p_I\colon X \to X$ is given that satisfies

$$p_I(x) \in \operatorname{ri} M_f(X \cap (x+I)) \tag{16}$$

for every $x \in X$. We further assume that the elements of $\mathcal{I}$ in (3) are visited in a (quasi-)cyclic order. In one such iteration cycle, all elements of $\mathcal{I}$ are visited (some possibly more than once), in a fixed order defined by a surjective map $\sigma\colon \{1,\dots,m\} \to \mathcal{I}$, where $m \geq |\mathcal{I}|$. The action of the iteration cycle is thus described by the map

$$p_\sigma = p_{\sigma(1)} \circ \cdots \circ p_{\sigma(m)}. \tag{17}$$

We finally define the map p from §5.1 to be

$$p = (p_\sigma)^{k+1} \quad \text{where} \;\; k = \dim X \tag{18}$$

(i.e., p is the composition of p_σ with itself $(k+1)$-times).

In Theorem 12, the sequence (I_n) was assumed to contain every element of $\mathcal{I}$ an infinite number of times. But our (quasi-)cyclic order has a stronger property: each element of $\mathcal{I}$ is always visited again after at most m iterations. Thus, Theorem 12 can be strengthened as follows:

Theorem 20. *Let $x \in X$ and $f(p(x)) = f(x)$. Then $p(x)$ is an interior local minimum and x is a pre-interior local minimum.*

Proof. Similarly to the proof of Theorem 12, it holds that:

- If x is an interior local minimum, then $x \in \operatorname{ri} F(X, p_\sigma(x))$.
- If x is not an interior local minimum, then $x \in \operatorname{rb} F(X, p_\sigma(x))$, so $\dim F(X, p_\sigma(x)) > \dim F(X, x)$.

Therefore, if $f(p(x)) = f(x)$ and $p(x)$ were not an interior local minimum, we would have $\dim F(X, p(x)) > \dim X$, a contradiction. Since $x \in F(X, p(x))$, x is a pre-interior local minimum. □

Combining Theorems 13 and 20, we see that the set (11) contains all pre-interior local minima and only them. The objective function f is convex on V, hence continuous. For a sequence (x_n) defined by (10) and (18), Theorems 16 and 19 thus imply the following:

Corollary 21. *If the sequence $(f(x_n))$ is convergent, then every cluster point of (x_n) is a pre-interior local minimum.*

Corollary 22. *If the sequence (x_n) is bounded and the sequence $(f(x_n))$ is convergent, then (x_n) converges to the set of pre-interior local minima.*

As the sequence $(f(x_n))$ is non-increasing, it is convergent if f is bounded below on X. Trivially, the sequence (x_n) is bounded if the set X is bounded. But, since $(f(x_n))$ is non-increasing, there is a weaker (and hence more useful) sufficient condition: (x_n) is bounded if the level set

$$X_0 = \{\, x \in X \mid f(x) \leq f(x_0) \,\} \tag{19}$$

is bounded.

6. Application to MAP Inference

Here we show how the above general results manifest themselves in convergent message-passing methods for MAP inference. MAP inference in a graphical model (with pairwise factors) leads to the problem

$$F(\theta) = \max_{x\colon V \to L} \Big[\sum_{i \in V} \theta_i(x_i) + \sum_{\{i,j\} \in E} \theta_{ij}(x_i, x_j) \Big] \tag{20}$$

where (V, E) with $E \subseteq \binom{V}{2}$ is an undirected graph, L is a label set, and $\theta_i\colon L \to \mathbb{R}$ and $\theta_{ij}\colon L^2 \to \mathbb{R}$ are weight functions (adopting that $\theta_{ij}(x, y) = \theta_{ji}(y, x)$).

The objective of (20) is preserved by replacing weights θ with reparameterized weights θ^δ given by

$$\theta_i^\delta(x) = \theta_i(x) - \sum_{j \in N_i} \delta_{ij}(x), \tag{21a}$$

$$\theta_{ij}^\delta(x, y) = \theta_{ij}(x, y) + \delta_{ij}(x) + \delta_{ji}(y), \tag{21b}$$

where δ is the vector of 'messages' $\delta_{ij}\colon L \to \mathbb{R}$ ($i \in V$, $j \in N_i$), and $N_i = \{\, j \in V \mid \{i, j\} \in E \,\}$ is the set of neighbors of vertex i. In particular, $F(\theta) = F(\theta^\delta)$ for all δ.

Many LP-based MAP inference algorithms minimize a convex piecewise-affine upper bound on (20) over reparameterizations. Two such bounds are

$$U_1(\theta) = \sum_{i \in V} \max_{x \in L} \theta_i(x) + \sum_{\{i,j\} \in E} \max_{x,y \in L} \theta_{ij}(x, y),$$

$$U_2(\theta) = \max\Big\{ \max_{i \in V} \max_{x \in L} \theta_i(x) \;,\; \max_{\{i,j\} \in E} \max_{x,y \in L} \theta_{ij}(x, y) \Big\}.$$

Clearly,

$$F(\theta) \leq U_1(\theta) \leq n U_2(\theta)$$

where $n = |V| + |E|$. Minimizing $U_1(\theta^\delta)$ or $U_2(\theta^\delta)$ over δ can be seen as a dual LP relaxation of (20). If the graph (V, E) is connected, at optimum we have $U_1(\theta^\delta) = n U_2(\theta^\delta)$, so these two relaxations are equivalent. For details see, e.g., [23, 22].

It is known [23] that coordinate-wise local minima of these problems can be characterized by *arc consistency*. For a weight vector θ, define the boolean vector $\bar\theta$ by

$$\bar\theta_i(x) = [\![\theta_i(x) = \max_{x'} \theta_i(x')]\!],$$

$$\bar\theta_{ij}(x, y) = [\![\theta_{ij}(x, y) = \max_{x', y'} \theta_{ij}(x', y')]\!].$$

A boolean vector $\bar{\theta}$ is arc consistent if

$$\bar{\theta}_i(x) = \bigvee_y \bar{\theta}_{ij}(x,y) \tag{22}$$

for all $i \in V$, $j \in N_i$, $x \in L$ (where $\vee$ denotes disjunction). We say that $\bar{\theta}$ has a *non-empty arc consistency closure* iff it has a non-empty arc consistent subset (i.e., there is a non-zero arc consistent boolean vector $\bar{\theta}' \leq \bar{\theta}$, where $\leq$ is meant component-wise). It can be checked that δ is an interior local minimum of the above problems iff $\bar{\theta}^\delta$ is arc consistent, and δ is a pre-interior local minimum iff $\bar{\theta}^\delta$ has a non-empty arc consistency closure. Therefore, (pre-)interior local minimality generalizes arc consistency.

6.1. Max-Sum Diffusion

The max-sum diffusion update [17, 23] chooses a variable $\delta_{ij}(x)$ and changes its value so that the equality

$$\theta_i^\delta(x) = \max_y \theta_{ij}^\delta(x,y) \tag{23}$$

becomes satisfied. We show that this minimizes $U_2(\theta^\delta)$ over $\delta_{ij}(x)$, satisfying the relative interior rule. Indeed,

$$U_2(\theta^\delta) = \max\{\underbrace{a - \delta_{ij}(x)}_{\theta_i^\delta(x)},\ \underbrace{b + \delta_{ij}(x)}_{\max_y \theta_{ij}^\delta(x,y)},\ c\}, \tag{24}$$

where a, b, c are constants w.r.t. $\delta_{ij}(x)$. The max-sum diffusion update minimizes (24) over $\delta_{ij}(x)$ by choosing $\delta_{ij}(x)$ such that $a - \delta_{ij}(x) = b + \delta_{ij}(x)$, which has the unique solution $\delta_{ij}(x) = \frac{1}{2}(a-b)$. When the set of minimizers of (24) over $\delta_{ij}(x)$ is an interval (rather than a singleton), this point is in the middle of this interval, hence in its relative interior.

We observed that modifying the update such that $\delta_{ij}(x)$ is chosen elsewhere (not in the middle) inside the optimal interval did not affect the algorithm much. However, updates choosing $\delta_{ij}(x)$ to be one of the endpoints of the optimal interval typically got stuck in a very poor (non-pre-interior) local minimum, even for very small instances.

Corollary 22 assumes that the sequence of message vectors δ during diffusion is bounded. Though this has been always observed, the proof is unknown. This technical issue can be fixed as follows: rather than minimizing $U_2(\theta^\delta)$ over δ, we minimize $U_2(\theta')$ over $\theta' \in X$ where X consists of vectors θ^δ for all possible δ. It can be shown that this reformulation preserves the relative interior rule. Then (19) is the level set $X_0 = \{\, \theta' \in X \mid U_2(\theta') \leq u_0 \,\}$ where u_0 is the initial value of the upper bound. This set is bounded by a simple argument: if some component of θ^δ decreases by changing δ, then, by (21), some other component inevitably increases. Thus Corollary 22 applies, showing that vectors θ^δ converge to a pre-interior local minimum of U_2 on X.

At any fixed point θ^δ of max-sum diffusion (where (23) holds globally), $\bar{\theta}^\delta$ is arc consistent, hence θ^δ is an interior local minimum. Though empirically the vectors θ^δ converge to a fixed point, a proof of this is unknown [23].

6.2. MPLP

The MPLP update [2] chooses an edge $\{i,j\} \in E$ and changes the variables δ_{ij}, δ_{ji} so that the equalities[4]

$$\theta_i^\delta(x) = \max_y \bigl[\theta_{ij}^\delta(x,y) + \theta_j^\delta(y)\bigr], \tag{25a}$$

$$\theta_j^\delta(y) = \max_x \bigl[\theta_{ij}^\delta(x,y) + \theta_i^\delta(x)\bigr] \tag{25b}$$

become satisfied for all $x, y \in L$. This update minimizes $U_1(\theta^\delta)$ over the variables δ_{ij}, δ_{ji} subject to $\theta_{ij}^\delta(x,y) \leq 0$ (in fact, (25) implies $\max_{x,y} \theta_{ij}^\delta(x,y) = 0$).

In contrast to max-sum diffusion, the MPLP update can choose a point on the relative boundary of the minimal set of this problem, hence it does not satisfy the relative interior rule. But notice that at MPLP fixed points (when (25) holds for all $\{i,j\} \in E$ and $x, y \in L$) we have

$$\bar{\theta}_i^\delta(x) = \bigvee_y \bigl[\bar{\theta}_{ij}^\delta(x,y) \wedge \bar{\theta}_j^\delta(y)\bigr] \tag{26}$$

for all $i \in V$, $j \in N_i$, $x \in L$. This implies that the arc consistency closure of $\bar{\theta}^\delta$ is non-empty, hence θ^δ is a pre-interior local minimum. It can be shown (if p is the composition of MPLP updates, as in §5.2) that (11) is still the set of pre-interior local minima and hence Theorem 19 applies.

6.3. Potts Problem

If $\theta_{ij}(x,y) = -[\![x = y]\!]$ in (20), we speak about the Potts problem. In that case, the dual LP relaxation can be simplified [14]: minimize $U_1(\theta^\delta)$ over δ subject to

$$\delta_{ij}(x) + \delta_{ji}(x) = 0, \tag{27a}$$

$$-\tfrac{1}{2} \leq \delta_{ij}(x) \leq \tfrac{1}{2}. \tag{27b}$$

Though ignoring these constraints would not change the optimal value of $U_1(\theta^\delta)$, it is interesting to try and design a coordinate descent method which includes them. This is a challenge because, to our knowledge, no convergent message-passing methods for problems with inequality constraints (here, (27b)) have been proposed so far.

Constraints (27) imply that $\max_{x,y} \theta_{ij}^\delta(x,y) = 0$ for all $\{i,j\} \in E$, thus the pairwise terms in $U_1(\theta^\delta)$ can be ignored. After orienting the graph (V,E) arbitrarily (so that $E \subseteq V^2$), we can eliminate constraint (27a) by keeping the variables $\delta_{ij}(x)$ only for $(i,j) \in E$, and write (21a) as

$$\theta_i^\delta(x) = \theta_i(x) + \sum_{(i,j)\in E} \delta_{ij}(x) - \sum_{(j,i)\in E} \delta_{ji}(x). \tag{28}$$

We propose the update

$$\begin{aligned} \delta_{ij}(x) := \tfrac{1}{2} h(\max_{y\neq x} \theta_i^\delta(y) - \theta_i^\delta(x) + \delta_{ij}(x)) - {}\\ \tfrac{1}{2} h(\max_{y\neq x} \theta_j^\delta(y) - \theta_j^\delta(x) - \delta_{ij}(x)) \end{aligned} \tag{29}$$

[4] We write the MPLP update in a different form than in [2]. Simple algebra shows (see the supplementary material) that (25) implies the update as stated in [2, Proposition 1].

where $h(t) = \min\{\frac{1}{2}, \max\{t, -\frac{1}{2}\}\}$ is the projection of t onto the interval $[-\frac{1}{2}, \frac{1}{2}]$. It can be shown that this update minimizes $U_1(\theta^\delta)$ subject to (27b) and satisfies the relative interior rule. We observed that on toy image segmentation instances (4 labels, 20×20 pixels), updates (29) converged to global optima and runtimes were similar to max-sum diffusion. See the supplement for proofs and details.

6.4. Max-marginal Averaging

Here we consider the Lagrangean decomposition framework [6, 10] for problem (20), understanding that it also includes TRW-S [8]. We will write (20) as

$$F(\theta) = \max_{x \in L^V} \langle \theta, \phi(x) \rangle \tag{30}$$

where $\phi\colon L^V \to \{0,1\}^I$ is a suitable feature map and I is the set of features (labels and label pairs) [21]. An upper bound on (30) is constructed by decomposition to subproblems. A subproblem $s \in S$ has weights $\theta^s \in \mathbb{R}^I$. Assuming

$$\theta = \sum_{s \in S} \theta^s \tag{31}$$

and swapping max and sum in (30), we obtain two bounds

$$F(\theta) = F\Big(\sum_{s \in S} \theta^s\Big) \le \sum_{s \in S} F(\theta^s) \le |S| \max_{s \in S} F(\theta^s). \tag{32}$$

The subproblem weights are constrained by

$$\theta_i^s = 0 \qquad \forall s \in S,\ i \in I \setminus I^s \tag{33}$$

where each set $I^s \subseteq I$ is such that the function $F(\theta^s)$ is tractable to evaluate (e.g., I^s can define a subtree of (V, E)). We want to minimize one of the upper bounds (32) over the variables θ_i^s subject to (31) and (33).

For I and ϕ induced by (20) and natural choices of sets I^s (such as the rows and columns of an image), the numbers $F(\theta^s)$ can always be made the same for all $s \in S$ while keeping (31) and (33). Therefore, the two upper bounds in (32) coincide at optimum.

In [8, 6], the upper bound is minimized by 'max-marginal averaging'. The max-marginal of the function $\langle \theta, \phi(x) \rangle$ associated with a feature $i \in I$ is the number

$$F_i(\theta) = \max_{x\colon \phi_i(x)=1} \langle \theta, \phi(x) \rangle. \tag{34}$$

The update chooses $i \in I$ and changes the variables $(\theta_i^s)_{s \in S_i}$ so that the max-marginals $F_i(\theta^s)$ become the same for all $s \in S_i$, where $S_i = \{\, s \in S \mid i \in I^s \,\}$. We show that this update minimizes $\max_s F(\theta^s)$ over $(\theta_i^s)_{s \in S_i}$, satisfying the relative interior rule.

It follows from (34) that $F_i(\theta)$ depends on θ_i linearly: $F_i(\theta) = a + \theta_i$ where a does not depend on θ_i. By (30) and (34), $F(\theta) = \max\{b, F_i(\theta)\}$ where b does not depend on θ_i. Hence,

$$\max_s F(\theta^s) = \max\{\, \max_{s \in S} \underbrace{(a^s + \theta_i^s)}_{F_i(\theta^s)},\ c\,\}$$

where a^s and c do not depend on θ_i^s. This is to be minimized over variables $(\theta_i^s)_{s \in S}$ subject to (31)+(33). It is easy to check that the condition that the numbers $a^s + \theta_i^s$ be the same for all $s \in S_i$ determines the variables $(\theta_i^s)_{s \in S}$ uniquely, and that these variables are a solution from the relative interior of the optimal set of this problem. For thee subproblems, pre-interior local minima correspond to weak-tree agreement from [8].

For our feasible set given by (31)+(33), the level set (19) is bounded by a similar argument as in §6.1, so Corollary 22 shows convergence to the set of pre-interior local minima.

7. Application to Weighted Vertex Cover

An important question is whether our theory can lead to practical algorithms for large-scale optimization of some new convex problems, unrelated to MAP inference. As a preliminary step in this direction, we propose a coordinate descent update for the LP relaxation of the minimum vertex cover problem. This LP relaxation reads

$$\min_{x\colon V \to [0,1]} \sum_{i \in V} \theta_i x_i \quad \text{s.t.}\ \ x_i + x_j \ge 1\ \forall \{i,j\} \in E \tag{35}$$

where (V, E) is an undirected graph with node weights $\theta\colon V \to \mathbb{R}_+$. The dual problem reads

$$\max_{y\colon E \to \mathbb{R}_+} \Big(\sum_{\{i,j\} \in E} y_{ij} + \sum_{i \in V} \min\Big\{\theta_i - \sum_{j \in N_i} y_{ij}, 0\Big\}\Big). \tag{36}$$

To optimize the dual problem over a single variable $y_{ij} \ge 0$, we propose the update

$$y_{ij} = \tfrac{1}{2}(\max\{\theta_i - a_i^{-j}, 0\} + \max\{\theta_j - a_j^{-i}, 0\}) \tag{37}$$

where $a_i^{-j} = \sum_{k \in N_i \setminus \{j\}} y_{ik}$ and symmetrically for a_j^{-i}. The dual restricted to a single variable is a piecewise-affine function with breakpoints $\theta_i - a_i^{-j}$ and $\theta_j - a_j^{-i}$. To show that update (37) satisfies the relative interior rule, it suffices to analyze possible cases for the signs of these breakpoints.

This method reached global optimality of the dual LP relaxation for all 41 minimum vertex cover instances from [24], for which the vertex weights were sampled i.i.d. as the absolute values of a Gaussian. On all of the instances, the method was faster than the simplex algorithm. Details can be found in the supplement.

Acknowledgment. This work has been supported by the Czech Science Foundation project 19-09967S, the OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765, and the CTU student grant SGS19/170/OHK3/3T/13.

References

[1] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007.

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

[3] Tamir Hazan and Amnon 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.

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

[5] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *25th Intl. Conference on Machine Learning (ICML)*, pages 408–415, 2008.

[6] Jason K. Johnson, Dmitry M. Malioutov, and Alan S. Willsky. Lagrangian relaxation for MAP estimation in graphical models. In *45th Allerton Conference on Communication, Control and Computing*, 2007.

[7] Jörg H. Kappes, Bjoern Andres, Fred A. Hamprecht, Christoph Schnörr, Sebastian Nowozin, Dhruv Batra, Sungwoong Kim, Bernhard X. Kausler, Thorben Kröger, Jan Lellmann, Nikos Komodakis, Bogdan Savchynskyy, and Carsten Rother. A comparative study of modern inference techniques for structured discrete energy minimization problems. *Intl. J. of Computer Vision*, 115(2):155–184, 2015.

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

[9] Vladimir Kolmogorov. A new look at reweighted message passing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(5):919–930, May 2015.

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

[11] 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.

[12] Talya Meltzer, Amir Globerson, and Yair Weiss. Convergent message passing algorithms: a unifying view. In *Conf. on Uncertainty in Artificial Intelligence*, pages 393–401, 2009.

[13] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, April 1998.

[14] Daniel Průša and Tomáš Werner. LP relaxation of the Potts labeling problem is as hard as any linear program. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1469–1475, 2017.

[15] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, 1970.

[16] Bogdan Savchynskyy. Discrete graphical models – an optimization perspective. *Foundations and Trends in Computer Graphics and Vision*, 11(3-4):160–429, 2019.

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

[18] Paul Swoboda, Jan Kuske, and Bogdan Savchynskyy. A dual ascent framework for Lagrangean decomposition of combinatorial problems. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4950–4960, 2017.

[19] Richard Szeliski, Ramin Zabih, Daniel Scharstein, Olga Veksler, Vladimir Kolmogorov, Aseem Agarwala, Marshall Tappen, and Carsten Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008.

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

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

[22] Tomáš Werner. A linear programming approach to max-sum problem: A review. Technical Report CTU-CMP-2005-25, Czech Technical University in Prague, December 2005.

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

[24] Ke Xu, Frédéric Boussemart, Fred Hemery, and Christophe Lecoutre. A simple model to generate hard satisfiable instances. In *19th Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 337–342, 2005. Benchmark: `http://sites.nlsde.buaa.edu.cn/~kexu/benchmarks/graph-benchmarks.htm`.