# Advanced Structured Prediction

Editors:

**Sebastian Nowozin** Sebastian.Nowozin@microsoft.com
*Microsoft Research*
*Cambridge, CB1 2FB, United Kingdom*

**Peter V. Gehler** pgehler@tuebingen.mpg.de
*Max Planck Insitute for Intelligent Systems*
*72076 Tübingen, Germany*

**Jeremy Jancsary** jermyj@microsoft.com
*Microsoft Research*
*Cambridge, CB1 2FB, United Kingdom*

**Christoph H. Lampert** chl@ist.ac.at
*IST Austria*
*A-3400 Klosterneuburg, Austria*

This is a draft version of the author chapter.

The MIT Press
Cambridge, Massachusetts
London, England

# 1 Getting Feasible Variable Estimates From Infeasible Ones: MRF Local Polytope Study

**Bogdan Savchynskyy** bogdan.savchynskyy@iwr.uni-heidelberg.de
*University of Heidelberg*
*Heidelberg, Germany*

**Stefan Schmidt** schmidtsstefan@googlemail.com
*Heidelberg Engineering GmbH*
*Heidelberg, Germany*

*This chapter proposes a method for the construction of approximate* feasible *primal solutions from infeasible ones for large-scale optimization problems possessing certain separability properties. Whereas the infeasible primal estimates can typically be produced from (sub-)gradients of the dual function, it is often not easy to project them to the primal feasible set, since the projection itself has a complexity comparable to the complexity of the initial problem. We propose an alternative efficient method to obtain feasibility and show that its properties influencing the convergence to the optimum are similar to the properties of the Euclidean projection. We apply our method to the local polytope relaxation of inference problems for Markov Random Fields and discuss its advantages compared to existing methods.*

***Keywords:*** *Markov random fields, inference, primal bound, feasible estimate, optimizing projection, local polytope.*

## 1.1 Introduction

Convex relaxations of combinatorial problems, as appearing in computer vision, processing of medical data or analysis of transport networks, often contain millions of variables and hundreds of thousands of constraints. It is also quite common to employ their dual formulations to allow for more efficient optimization, which due to strong duality delivers also primal solutions. Indeed, approximate primal solutions can usually be reconstructed from (sub-)gradients of the dual objective. However, these are typically infeasible. Because of the problem size, only first order methods (based on the function and its (sub-)gradient evaluation only) can be applied. Since feasibility is not guaranteed up to the optimum, it is hardly attainable for such methods because of their slow convergence. The classical trick — (Euclidean) projection to the feasible set — can not be used efficiently because of the problem size.

A striking example of such a situation, which we explore in this chapter, is the reconstruction of feasible primal estimates for local polytope relaxations of Markov random field (MRF) inference problems (Schlesinger, 1976; Werner, 2007; Wainwright and Jordan, 2008), studied in Chapter **??**.

**Motivation: Why Feasible Relaxed Primal Estimates Are Needed.** It is often the case for convex relaxations of combinatorial problems that not a relaxed solution, but an integer approximation thereof is used in applications. Such integer primal estimates can be obtained from the dual ones due to the complementary slackness condition and using heuristic local search or rounding procedures (Werner, 2007; Kolmogorov, 2006; Ravikumar et al., 2010). However, such integer estimates do not converge to the optimum of the relaxed problem in general.

In contrast, a sequence of *feasible* solution estimates of the *relaxed problem* converging to the optimum guarantees vanishing of the corresponding duality gap, and hence (i) determines a theoretically sound stopping condition (Boyd and Vandenberghe, 2004); (ii) provides a basis for the comparison of different optimization schemes for a given problem; (iii) enables the construction of adaptive optimization schemes depending on the duality gap, for example adaptive step-size selection in subgradient-based schemes (Komodakis et al., 2011; Kappes et al., 2012) or adaptive smoothing selection procedures for non-smooth problems (Savchynskyy et al., 2012). Another example is the tightening of relaxations with cutting-plane based approaches (Sontag et al., 2008).

**Contribution.** We propose an efficient and well-scalable method for constructing feasible points from infeasible ones for a certain class of separable convex problems. The method guarantees convergence of the constructed feasible point sequence to the optimum of the problem if only this convergence holds for their infeasible counterparts. We theoretically and empirically show how this method works in a local polytope relaxation framework for MRF inference problems. We formulate and prove our results in a general way, which allows to apply them to arbitrary convex optimization problems having a similar separable structure.

### 1.1.1 Formulation of the Main Result

We start by stating the main result of the chapter for a separable linear programming problem. The result has a special form, which appears in the MRF energy minimization problem. This example illustrates the idea of the method and avoids shading it with numerous technical details. We refer to Section 1.2 and Section 1.3 for all proofs, special cases and generalizations.

Let $\langle \cdot, \cdot \rangle$ denote an inner product of two vectors in a Euclidean space. Let $\mathbb{R}^n_+$ denote the non-negative cone of the $n$-dimensional Euclidean space $\mathbb{R}^n$. Let $I = \{1, \dots, N\}$, $J = \{1, \dots, M\}$, be sets of integer indexes and $\mathcal{N}(j),\ j \in J$, be a collection of subsets of $I$. Let further $x \in \mathbb{R}^{nI}_+$ be a collection of $(x_i \in \mathbb{R}^n_+,\ i \in I)$ and $y \in \mathbb{R}^{mJ}_+$ denote $(y_j \in \mathbb{R}^m_+,\ j \in J)$. Let $A_{ij},\ i \in I,\ j \in J$, and $B_i,\ i \in I$, be full-rank matrices of dimensions $m \times n$ and $n \times k$ for some $k < n$ and let $c_i \in \mathbb{R}^k$. Consider the following separable linear programming problem in the standard form

$$\min_{\substack{x \in \mathbb{R}^{nI}_+ \\ y \in \mathbb{R}^{mJ}_+}} \sum_{i=1}^{N} \langle a_i, x_i \rangle + \sum_{j=1}^{M} \langle b_j, y_j \rangle \tag{1.1}$$

$$A_{ij} y_j = x_i,\ i \in \mathcal{N}(j),\ j \in J\,,$$

$$B_i x_i = c_i,\ i \in I\,.$$

Let $C$ be the feasible set of the problem (1.1) and the mapping $\mathcal{P}\colon \mathbb{R}^{nI}_+ \times \mathbb{R}^{mJ}_+ \to C$ be defined such that $\mathcal{P}(x, y) = (x', y')$, where

$$x'_i = \operatorname*{argmin}_{\tilde{x}_i \in \mathbb{R}^n_+} (x_i - \tilde{x}_i)^2 \text{ s.t. } B_i \tilde{x}_i = c_i,\ i \in I\,; \tag{1.2}$$

$$y'_j := \operatorname*{argmin}_{y_j \in \mathbb{R}^m_+} \langle b_j, y_j \rangle \text{ s.t. } A_{ij} y_j = x'_i,\ i \in \mathcal{N}(j)\,. \tag{1.3}$$

**The main result** of this chapter states that *from the convergence of* $(x^t, y^t) \in \mathbb{R}^{nI} \times \mathbb{R}^{mJ}$, $t = 1, 2, \dots \infty$, *to the set of optimal solutions of* (1.1) *it follows that* $\mathcal{P}(x^t, y^t)$ *converges to the set of optimal solutions as well.*

Please note that

- $\mathcal{P}(x^t, y^t)$ is always feasible due to its construction;
- in contrast to the Euclidean projection onto the set $C$, which constitutes a problem of size comparable to that of the initial one (1.1), to compute $\mathcal{P}(x^t, y^t)$ one has to solve many, but *small* quadratic and linear optimization problems (1.2)-(1.3), assuming that $n \ll I$, $m \ll J$ and $N(J) \ll I$. To this end such powerful, but not very well scalable tools as simplex or interior point methods can be used due to the small size of these problems.

In Section 1.2 we additionally show how the convergence speed of $\mathcal{P}(x^t, y^t)$ depends on coefficients $a_i$ and $b_i$.

Assuming that the set $C$ corresponds to the local polytope, variables $x_i$ and $y_i$ to unary and binary "max-marginals" and weights $a_i$ and $b_j$ to unary and pairwise potentials respectively, this result allows for an efficient estimation of feasible primal points from infeasible ones for MRF energy minimization algorithms, which has been considered as a non-trivial problem in the past (Werner, 2007).

### 1.1.2 Related Work on MRF Inference

The two most important inference problems for MRFs are maximum a posteriori (MAP) inference and marginalization (Wainwright and Jordan, 2008). Both are intractable in general and thus both require some relaxation. The simplest convex relaxation for both is based on exchanging the underlying convex hull of the feasible set, the marginal polytope, by an approximation called the local polytope, studied in Chapter **??**. However, even with this approximation the problems remain non-trivial, though solvable, at least theoretically. A series of algorithmic schemes were proposed to this end for the local polytope relaxations of both MAP (see Chapter **??** and works of Storvik and Dahl (2000), Komodakis et al. (2011), Schlesinger and Giginyak (2007), Ravikumar et al. (2010), Savchynskyy et al. (2011),Schmidt et al. (2011), Kappes et al. (2012), Savchynskyy et al. (2012), Martins et al. (2011)) and marginalization (Wainwright et al., 2005; Jancsary and Matz, 2011; Hazan and Shashua, 2010). It turns out that the corresponding dual problems have dramatically fewer variables and contain very simple constraints (Werner, 2007, 2009), hence they can even be formulated as unconstrained problems, as done by Schlesinger and Giginyak (2007) and Kappes et al. (2012). Therefore, most of the approaches address optimization of the dual objectives. A common difficulty for such approaches is the computation of a *feasible* relaxed primal estimate from the current dual one. *Infeasible* estimates can typically be obtained from the subgradients of the dual function, as shown by Komodakis et al. (2011), or from the gradients of the smoothed dual, as done by Johnson et al. (2007), Werner (2009), and Savchynskyy et al. (2011).

Even some approaches working in the primal domain (see Section **??** and works of Hazan and Shashua (2010), Martins et al. (2011) and Schmidt et al. (2011)) maintain infeasible primal estimates, whilst feasibility is guaranteed only in the limit.

Quite efficient primal schemes based on graph cuts, as proposed by Boykov et al. (2001), do not solve the problem in general and optimality guarantees provided by them are typically too weak. Hence we discuss neither these here, nor the widespread message passing and belief propagation (Kolmogorov, 2006; Weiss and Freeman, 2001) methods (discussed also in Chapter **??**), which also do not guarantee the attainment of the optimum of the relaxed problem.

**Feasible Primal Estimates.** The literature on obtaining feasible primal solutions for MRF inference problems from infeasible ones is not very vast. Apart from our conference papers (Savchynskyy et al. (2011); Schmidt et al. (2011); Savchynskyy et al. (2012)) describing special cases of our method in application to the MRF local polytope, we are aware of only three recent works contributing to this topic, by Schlesinger et al. (2011), Werner (2011). The most recent and practical method is described in Chapter **??**.

The method proposed by Schlesinger et al. (2011) is formulated in the form of an algorithm able to determine whether a given solution accuracy $\varepsilon$ is attained or not. To this end it restricts the set of possible primal candidate solutions and solves an auxiliary quadratic programming (QP) problem. However, this approach is unsuited to compute *the actually attained* $\varepsilon$ directly and the auxiliary QP in the worst case grows linearly with the size of the initial linear programming problem. Hence obtaining a feasible primal solution becomes prohibitively slow as the size of the problem gets larger.

Another closely related method was proposed by Werner (2011). It is, however, only suited to determine whether a given solution of the dual problem is an optimal one. This makes it non-practical, since the state-of-the-art methods achieve the exact solution of the considered problem only in the limit, after a potentially infinite number of iterations.

The very recent method described in Chapter **??** is simple, yet efficient. However, as we show in Section 1.2 (Theorem 1.4), our method applied on top of *any* other, including the one described in Chapter **??**, delivers better primal estimates, except for the cases when the estimates of the other method coincide with ours.

### 1.1.3 Content and Organization of the Chapter

In Section 1.2 we describe a general formulation and mathematical properties of the *optimizing projection* $\mathcal{P}(x, y)$, as already introduced for a special case in (1.2)-(1.3). We do this without relating it to inference in MRFs. This shows the generality of the method and keeps the exposition simple. Section 1.3 is devoted to local polytope relaxations of the MAP and marginalization inference problems for MRFs and specifies how the feasible estimates can be constructed for these. In Section 1.4 we discuss different optimization schemes for the local polytope relaxation for which the primal estimates can be reconstructed from the dual ones. Finally, Section 1.5 and Section 1.6 contain the experimental evaluation and conclusions, respectively.

## 1.2 Optimizing Projection

Let us denote by $\Pi_C \colon \mathbb{R}^n \to C$ an Euclidean projection to a set $C \subset \mathbb{R}^n$. Let $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ be two subsets of Euclidean spaces and $C \subset X \times Y$ be a closed convex set. We will denote as $C_X$ the set $\{x \in X \mid \exists y \in Y \colon (x, y) \in C\}$, that is the projection of $C$ to $X$.

The main definition of the chapter introduces the notion of *the optimizing projection* in its general form. A possible simplification and the corresponding discussion follow the definition.

**Definition 1.1.** *Let $f \colon X \times Y \to \mathbb{R}$ be a continuous convex function of two variables. The mapping $\mathcal{P}_{f,C} \colon X \times Y \to C$ such that $\mathcal{P}_{f,C}(x, y) = (x', y')$ defined as*

$$x' = \Pi_{C_X}(x)\,, \tag{1.4}$$

$$y' = \arg\min_{y \colon (x', y) \in C} f(x', y)\,, \tag{1.5}$$

*is called* optimizing projection *onto the set $C$ w.r.t. the function $f$.*

This definition provides the way to get *the feasible* point $(x', y') \in C$ from an arbitrary infeasible one $(x, y)$. Of course, getting just any feasible point is not a big issue in many cases. However, as we will see soon, the introduced optimizing projection possesses properties similar to the properties of a standard Euclidean projection, which makes it a useful tool in cases when its computation is easier than the one needed for the Euclidean projection. To this end both the partial projection (1.4) and the partial minimization (1.5) should be efficiently computable.

The role of projection (1.4) is to make $x$ “feasible”, i.e. to guarantee for $x'$

that there is at least one $y \in \mathcal{Y}$ such that $(x', y) \in C$, which guarantees the definition to be well-defined. If this condition holds already for $x$, it is easy to see that $x' = x$ and hence computing (1.4) is trivial. We will call such $x$ *feasible* w.r.t. $C$. Indeed, in (1.4) one can apply an arbitrary projection, since they all satisfy the mentioned property. However, we provide our analysis for Euclidean projections only.

**Example 1.1.** *Consider the linear programming problem* (1.1) *from the introduction. It is reasonable to construct an optimizing projection* $\mathcal{P}_{f,C}(x, y)$ *for it as in* (1.2)-(1.3)*, denoting with* $f$ *and* $C$ *the objective function and the feasible set of the problem* (1.1)*.*

We will deal with objective functions, which fulfill the following definition:

**Definition 1.2.** *A function* $f\colon X \times Y \to \mathbb{R}$ *is called Lipschitz-continuous w.r.t. its first argument* $x$*, if there exists a finite constant* $L_X(f) \geq 0$*, such that* $\forall y \in Y,\ x, x' \in X$*,*

$$|f(x, y) - f(x', y)| \leq L_X(f)\|x - x'\| \tag{1.6}$$

holds. Similarly $f$ is Lipschitz-continuous w.r.t.

- $y$ if $|f(x, y) - f(x, y')| \leq L_Y(f)\|y - y'\|$ for all $x \in X,\ y, y' \in Y$ and some constant $L_Y(f) \geq 0$;
- $z = (x, y)$ if $|f(x, y) - f(x', y')| \leq L_{XY}(f)\|z - z'\|$ for all $z, z' \in X \times Y$ and some constant $L_{XY}(f) \geq 0$ .

The following theorem specifies the main property of the optimizing projection, namely its continuity with respect to the optimal value of $f$.

**Theorem 1.1.** *Let* $f\colon X \times Y \to \mathbb{R}$ *be a continuous convex function and let* $f^*_C$ *be its minimum on the convex set* $C$*. Then*

- *for any* $z^t = (x^t, y^t) \in X \times Y$*,* $t = 0, \dots, \infty$*, from* $|f(x^t, y^t) - f^*_C| \xrightarrow{t\to\infty} 0$ *and* $\|z^t - \Pi_C(z^t)\| \xrightarrow{t\to\infty} 0$ *follows*

$$|f(\mathcal{P}_{f,C}(x^t, y^t)) - f^*_C| \xrightarrow{t\to\infty} 0\,. \tag{1.7}$$

- *for any* $z = (x, y) \in X \times Y$*, from Lipschitz-continuity of* $f$ *w.r.t. its second argument* $y$ *and feasibility of* $x$ *w.r.t.* $C$ *follows:*

$$|f(\mathcal{P}_{f,C}(x, y)) - f^*_C| \leq |f(x, y) - f^*_C| + L_Y(f)\|z - \Pi_C(z)\|\,. \tag{1.8}$$

- *for any* $z = (x, y) \in X \times Y$*, from Lipschitz-continuity of* $f$ *w.r.t. both its* *arguments $x$ and $y$ follows*

$$|f(\mathcal{P}_{f,C}(x,y)) - f^*_C| \leq |f(x,y) - f^*_C| + (L_X(f) + L_Y(f))\|z - \Pi_C(z)\| \,. \tag{1.9}$$

Theorem 1.1 basically states that if the sequence $z^t = (x^t, y^t) \in X \times Y,\ t = 1, \dots, \infty$, weakly converges to the optimum of $f$, then the same holds also for $\mathcal{P}_{f,C}(x^t, y^t)$. Moreover, for Lipschitz-continuous functions the rate of convergence is preserved up to a multiplicative constant. Please note that $\mathcal{P}_{f,C}(x,y)$ actually *does not depend on $y$*, the argument $y$ is needed only for the convergence estimates (1.9) and (1.8), but not for the optimizing projection itself.

**Remark 1.1.** *Let us provide a bound similar to* (1.8) *for the Euclidean projection to get an idea how good the estimate* (1.8) *is:*

$$\begin{aligned} |f(\Pi_C(z)) - f^*_C| &\leq |f(\Pi_C(z)) - f(z)| + |f(z) - f^*_C| \\ &\leq |f(z) - f^*_C| + L_{XY}(f)\|z - \Pi_C(z)\| \,. \end{aligned} \tag{1.10}$$

*We see that bounds* (1.9) *and* (1.10) *for the optimizing mapping and Euclidean projection differ only by a constant factor: in the optimizing mapping, the Lipschitz continuity of the objective $f$ is considered w.r.t. to each variable $x$ and $y$ separately, whereas the Euclidean projection is based on the Lipschitz continuity w.r.t. the pair of variables* $(x, y)$*.*

The following technical lemma shows the difference between these two Lipschitz constants. Together with the next one it will be used in Section 1.3:

**Lemma 1.2.** *The linear function $f(x,y) = \langle a, x\rangle + \langle b, y\rangle$ is Lipschitz-continuous with Lipschitz constants $L_X(f) \leq \|a\|$, $L_Y(f) \leq \|b\|$ and $L_{XY}(f) \leq \sqrt{L_X(f)^2 + L_Y(f)^2}$.*

**Lemma 1.3.** *The function $f(z) = \langle a, z\rangle + \sum_{i=1}^N z_i \log z_i$, where* $\log$ *denotes the natural logarithm, is*

- *continuous on $[0,1]^N \ni z$ and*
- *Lipschitz-continuous on $[\varepsilon, 1]^N \ni z,\ \varepsilon > 0$, with Lipschitz-constant*

$$L_{XY}(f) \leq \|a\| + N|1 + \log\varepsilon| \,. \tag{1.11}$$

An important property of the optimizing projection is its *optimality.* Contrary to the Euclidean projection it can deliver better estimates even when applied to an already *feasible* point $(x, y) \in C$, which is stated by the following theorem.

**Theorem 1.4** (Optimality of optimizing projection)**.** *Let* $(x,y) \in C$*; then* $f(\mathcal{P}_{f,C}(x,y)) \leq f(x,y)$*, and the inequality holds strictly if* $y \notin \arg\min_{y'\colon (x,y')\in C} f(x,y')$*.*

The proof of the theorem is straightforward and follows from Definition 1.1 and the fact that $x' = x$.

## 1.3 MRF Inference and Optimizing Projections

In this section we consider optimization problems related to inference in MRFs and construct corresponding optimizing projections. We switch from the general mathematical notation used in the previous sections to the one specific for the considered field, in particular we mostly follow the book of Wainwright and Jordan (2008).

### 1.3.1 MAP-inference problem

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, where $\mathcal{V}$ is a finite set of nodes and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of edges. Let further $\mathcal{X}_v,\ v \in \mathcal{V}$, be finite *sets of labels*. The set $\mathcal{X} = \otimes_{v\in\mathcal{V}}\mathcal{X}_v$, where $\otimes$ denotes the Cartesian product, will be called *labeling set* and its elements $x \in \mathcal{X}$ are *labelings*. Thus each labeling is a collection $(x_v\colon v \in \mathcal{V})$ of labels. To shorten notation we will use $x_{uv}$ for a pair of labels $(x_u, x_v)$ and $\mathcal{X}_{uv}$ for $\mathcal{X}_u \times \mathcal{X}_v$. The collections of numbers $\theta_{v,x_v},\ v \in \mathcal{V},\ x_v \in \mathcal{X}_v$ and $\theta_{uv,x_{uv}},\ uv \in \mathcal{E},\ x_{uv} \in \mathcal{X}_{uv}$, will be called *unary* and *pairwise potentials*, respectively. The collection of all potentials will be denoted by $\theta$. The maximum a-posteriori (MAP) inference problem reads

$$\min_{x\in\mathcal{X}} E(x) := \sum_{v\in\mathcal{V}} \theta_v(x_v) + \sum_{uv\in\mathcal{E}} \theta_{uv}(x_u, x_v)\,, \tag{1.12}$$

and consists of finding a labeling with the smallest total potential (energy).

An alternative way of writing problem (1.12) is to express it in the form of a scalar product of the vector $\theta$ with a suitably constructed binary vector $\delta(x),\ x \in \mathcal{X}$: $\min_{x\in\mathcal{X}} \langle\theta, \delta(x)\rangle$.

The problem is NP-hard in general, hence it is commonly accepted to consider its convex relaxations. The one most widely used is its *local polytope* relaxation, defined in the following subsection.

### 1.3.2 Primal Relaxed MAP Problem

Denoting $\mathbb{R}^{\sum_{v\in\mathcal{V}}|\mathcal{X}_v|+\sum_{uv\in\mathcal{E}}|\mathcal{X}_{uv}|}$ as $\mathbb{R}(\mathbb{M})$ and the corresponding non-negative cone $\mathbb{R}_+^{\sum_{v\in\mathcal{V}}|\mathcal{X}_v|+\sum_{uv\in\mathcal{E}}|\mathcal{X}_{uv}|}$ as $\mathbb{R}_+(\mathbb{M})$, one writes (Schlesinger, 1976; Werner, 2007) the local polytope (linear programming) relaxation of a MAP inference problem as

$$\begin{aligned}
&\min_{\mu\in\mathbb{R}_+(\mathbb{M})} \sum_{v\in\mathcal{V}}\sum_{x_v\in\mathcal{X}_v}\theta_{v,x_v}\mu_{v,x_v} + \sum_{uv\in\mathcal{E}}\sum_{x_{uv}\in\mathcal{X}_{uv}}\theta_{uv,x_{uv}}\mu_{uv,x_{uv}}\\
&\text{s.t.}\ \begin{array}{l}\sum_{x_v\in\mathcal{X}_v}\mu_{v,x_v}=1,\ v\in\mathcal{V}\,,\\ \sum_{x_v\in\mathcal{X}_v}\mu_{uv,x_{uv}}=\mu_{u,x_u},\ x_u\in\mathcal{X}_u,\ uv\in\mathcal{E}\,,\\ \sum_{x_u\in\mathcal{X}_u}\mu_{uv,x_{uv}}=\mu_{v,x_v},\ x_v\in\mathcal{X}_v,\ uv\in\mathcal{E}\,.\end{array}
\end{aligned}\tag{1.13}$$

The constraints in (1.13) form the *local polytope*, later on denoted as $\mathcal{L}$. Slightly abusing notation, we will briefly write problem (1.13) as $\min_{\mu\in\mathcal{L}} E(\mu) := \min_{\mu\in\mathcal{L}}\langle\theta,\mu\rangle$ .

**Optimizing Projection.** We will denote as $\theta_w$ and $\mu_w,\ w\in\mathcal{V}\cup\mathcal{E}$, the collections of $\theta_{w,x_w}$ and $\mu_{w,x_w},\ x_w\in\mathcal{X}_w$, respectively. Hence the vectors $\theta$ and $\mu$ become collections of $\theta_w$ and $\mu_w,\ w\in\mathcal{V}\cup\mathcal{E}$. The $n$-dimensional simplex $\{x\in\mathbb{R}^n_+\colon\ \sum_{i=1}^n x_i=1\}$ will be denoted as $\Delta(n)$.

Problem (1.13) has a separable structure, that is, for suitably selected matrices $A_{uv}$ it can be written as

$$\begin{aligned}
&\min_{\mu\in\mathbb{R}(\mathbb{M})}\sum_{v\in\mathcal{V}}\langle\theta_v,\mu_v\rangle + \sum_{uv\in\mathcal{E}}\langle\theta_{uv},\mu_{uv}\rangle\\
&\text{s.t.}\ \begin{array}{ll}\mu_v\in\Delta(|\mathcal{X}_v|), & v\in\mathcal{V}\,,\\ A_{uv}\mu_{uv}=\mu_v,\ \mu_{uv}\ge 0, & uv\in\mathcal{E}\,.\end{array}
\end{aligned}\tag{1.14}$$

Note that under fixed $\mu_v$, the optimization of (1.14) splits into small independent subproblems, one for each $uv\in\mathcal{E}$. We will use this fact to compute the optimizing projection onto the local polytope $\mathcal{L}$ as follows.

Let $\mu_{\mathcal{V}}$ and $\mu_{\mathcal{E}}$ be collections of primal variables corresponding to graph nodes and edges respectively, i.e. $\mu_{\mathcal{V}}=(\mu_v,\ v\in\mathcal{V})$, $\mu_{\mathcal{E}}=(\mu_{uv},\ uv\in\mathcal{E})$ and $\mu=(\mu_{\mathcal{V}},\mu_{\mathcal{E}})$. The corresponding subspaces will be denoted by $\mathbb{R}(\mathbb{M}_{\mathcal{V}})$ and $\mathbb{R}(\mathbb{M}_{\mathcal{E}})$. Then according to (1.14) and Definition 1.1, the optimizing projection $\mathcal{P}_{E,\mathcal{L}}\colon\mathbb{R}(\mathbb{M}_{\mathcal{V}})\times\mathbb{R}(\mathbb{M}_{\mathcal{E}})\to\mathcal{L}$ maps $(\mu_{\mathcal{V}},\mu_{\mathcal{E}})$ to $(\mu'_{\mathcal{V}},\mu'_{\mathcal{E}})$ defined as

$$\mu'_v = \Pi_{\Delta(|\mathcal{X}_v|)}(\mu_v),\ v\in\mathcal{V}\,,\tag{1.15}$$

$$\mu'_{uv} = \arg\begin{array}{l}\min\limits_{\mu_{uv}\ge 0}\langle\theta_{uv},\mu_{uv}\rangle\\ \text{s.t. } A_{uv}\mu_{uv}=\mu'_v\end{array},\ uv\in\mathcal{E}\,.\tag{1.16}$$

Note that both (1.15) and (1.16) can be computed very efficiently. Projection to a simplex in (1.15) can be done e.g. by method proposed by Michelot (1986). The optimization problem in (1.16) constitutes a small-sized *trans-*

*portation problem* well-studied in linear programming, see e.g. the textbook of Bazaraa and Jarvis (1977).

Let us apply Theorem 1.1 and Lemma 1.2 to the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ introduced by (1.15)-(1.16). According to these, the convergence rate of a given sequence $\mu^t \in \mathbb{R}(\mathbb{M})$ in the worst case slows down by a factor $L_{\mathbb{M}_{\mathcal{V}}}(E) + L_{\mathbb{M}_{\mathcal{E}}}(E) \leq \|\theta_{\mathcal{V}}\| + \|\theta_{\mathcal{E}}\|$. This factor can be quite large, but since the optimum $E^*$ grows together with the value $\|\theta_{\mathcal{V}}\| + \|\theta_{\mathcal{E}}\|$, its influence on the obtained *relative* accuracy is typically much lower than the value itself.

**Remark 1.2.** *However, if $\theta$ contains "infinite" numbers, typically assigned to pairwise factors $\theta_{\mathcal{E}}$ to model "hard" constraints, both optimizing* and *Euclidean projections can be quite bad, which is demonstrated by the following simple example: $\mathcal{V} = \{v, u\}$, $\mathcal{E} = \{uv\}$, $\mathcal{X}_v = \mathcal{X}_u = \{0, 1\}$, $\theta_{00} = \theta_{11} = \theta_{01} = 0$, $\theta_{10} = \infty$. If now $\mu_{v,1} > \mu_{u,1}$, optimizing w.r.t. $\mu_{uv}$ leads to $\theta_{10} \cdot \mu_{vu,10} = \infty \cdot (\mu_{v,1} - \mu_{u,1})$, whose value can be arbitrary large, depending on the actual numerical value approximating $\infty$. And since neither the optimizing projection nor the Euclidean one take into account the actual values of pairwise factors when assigning values to $\mu_{\mathcal{V}}$, the relation $\mu_{v,1} > \mu_{u,1}$ is not controlled.*

We provide a numerical simulation related to infinite values of pairwise potentials in Section 1.5.

**Remark 1.3** (Higher order models and relaxations)**.** *The generalization of the optimizing projection* (1.15)-(1.16) *for both higher order models, and higher order local polytopes (Wainwright and Jordan, 2008, Sec. 8.5) is quite straightforward. The underlying idea remains the same: one has to fix a subset of variables such that the resulting optimization problem splits into a number of small ones.*

**Remark 1.4** (Efficient representation of the relaxed primal solution)**.** *Note that since the pairwise primal variables $\mu_{\mathcal{E}}$ can be easily recomputed from the unary ones $\mu_{\mathcal{V}}$, it is sufficient to store only the latter if one is not interested in specific values of pairwise variables $\mu_{\mathcal{E}}$. Because of possible degeneracy, there may exist multiple vectors $\mu_{\mathcal{E}}$ optimizing the energy $E$ for a given $\mu_{\mathcal{V}}$.*

### 1.3.3 Relaxed Dual MAP Problem

In this section we consider the Lagrange dual to the problem (1.13). Let us denote as $\mathcal{N}(v) = \{u \in \mathcal{V}\colon uv \in \mathcal{E}\}$ the set of neighboring nodes of a node $v \in \mathcal{V}$. We consider the dual variable $\nu \in \mathbb{R}(\mathbb{D})$ to consist of the following groups of coordinates: $\nu_v$, $v \in \mathcal{V}$; $\nu_{uv}$, $uv \in \mathcal{E}$; and $\nu_{v \to u, x_v}$,

$v \in \mathcal{V},\ u \in \mathcal{N}(v),\ x_v \in \mathcal{X}_v$. In this notation the dual to (1.13) reads

$$\max_{\nu \in \mathbb{R}(\mathbb{D})} \sum_{v\in\mathcal{V}} \nu_v + \sum_{uv\in\mathcal{E}} \nu_{uv} \tag{1.17}$$
$$\text{s.t.}\ \begin{array}{l} \theta_{v,x_v} - \sum_{u\in\mathcal{N}(v)} \nu_{v\to u,x_v} \geq \nu_v\ ,\ v\in\mathcal{V},\, x_v \in \mathcal{X}_v\ ,\\ \theta_{uv,x_{uv}} + \nu_{u\to v,x_u} + \nu_{v\to u,x_v} \geq \nu_{uv},\ uv \in \mathcal{E}\,, x_{uv}\in\mathcal{X}_{uv}\,. \end{array}$$

We will use the notation $\mathcal{U}(\nu) := \sum_{v\in\mathcal{V}}\nu_v + \sum_{uv\in\mathcal{E}}\nu_{uv}$ for the objective function of (1.17).

**Optimizing Projection.** The dual (1.17) possesses clear separability as well: after fixing all variables except $\nu_v,\ v\in\mathcal{V}$, and $\nu_{uv},\ uv\in\mathcal{E}$, the optimization splits into a series of small and straightforward minimizations over a small set of values

$$\nu_v = \min_{x_v\in\mathcal{X}_v} \theta_{v,x_v} - \sum\nolimits_{u\in\mathcal{N}(v)} \nu_{v\to u,x_v},\ v\in\mathcal{V}\,, \tag{1.18}$$
$$\nu_{uv} = \min_{x_{uv}\in\mathcal{X}_{uv}} \theta_{uv,x_{uv}} + \nu_{u\to v,x_u} + \nu_{v\to u,x_v},\ uv\in\mathcal{E}\,. \tag{1.19}$$

The formula (1.18) can be applied directly for each $v\in\mathcal{V}$ and (1.19) accordingly for each $uv\in\mathcal{E}$.

We denote by $\mathbb{D}$ the dual feasible set defined by the constraints of (1.17). We split all dual variables into two groups. The first one will contain ”messages” $\nu_\to = (\nu_{v\to u},\ v\in\mathcal{V},\ u\in\mathcal{N}(v))$, that are variables, which reweight unary and pairwise potentials leading to an improvement in the objective. The vector space containing all possible values of these variables will be denoted as $\mathbb{R}(\mathbb{D}_\to)$. The second group will contain lower bounds on optimal reweighted unary and pairwise potentials $\nu_0 = (\nu_w,\ w\in\mathcal{V}\cup\mathcal{E})$. The total sum of their values constitutes the dual objective. All possible values of these variables will form the vector space $\mathbb{R}(\mathbb{D}_0)$. Hence the optimizing projection $\mathcal{P}_{\mathcal{U},\mathbb{D}}\colon \mathbb{R}(\mathbb{D}_\to)\times\mathbb{R}(\mathbb{D}_0)\to\mathbb{R}(\mathbb{D})$ maps $(\nu_\to,\nu_0)$ to $(\nu'_\to,\nu'_0)$ as

$$\nu'_{v\to u} = \nu_{v\to u},\ v\in\mathcal{V},\ u\in\mathcal{N}(v)\,, \tag{1.20}$$
$$\nu'_v = \min_{x_v\in\mathcal{X}_v} \theta_{v,x_v} - \sum\nolimits_{u\in\mathcal{N}(v)} \nu'_{v\to u,x_v},\ v\in\mathcal{V}\,, \tag{1.21}$$
$$\nu'_{uv} = \min_{x_{uv}\in\mathcal{X}_{uv}} \theta_{uv,x_{uv}} + \nu_{u\to v,x_u} + \nu'_{v\to u,x_v},\ uv\in\mathcal{E}\,. \tag{1.22}$$

Equation (1.20) corresponds to the projection (1.4), which has the form $\Pi_{\mathbb{R}(\mathbb{D}_\to)}(\nu_\to) = \nu_{\to 0}$ and is thus trivial.

Applying Theorem 1.1 and Lemma 1.2 to the optimizing projection $\mathcal{P}_{\mathcal{U},\mathbb{D}}$ yields that the convergence of the projected $\nu^t$ slows down no more than by a factor $L_{\mathbb{D}_0} \leq |\sqrt{\mathcal{V}}| + |\sqrt{\mathcal{E}}|$ and does not depend on the potentials $\theta$. However, since the optimal energy value often grows proportionally to

$|\mathcal{V}| + |\mathcal{E}|$, the influence of the factor on the estimated related precision is typically insignificant.

### 1.3.4 Entropy-Smoothed Primal Problem

Let $H\colon \mathbb{R}^n_+ \to \mathbb{R}$ be *an entropy* function defined as $H(z) = -\sum_{i=1}^n z_i \log z_i$ and the dimensionality $n$ defined by the dimensionality of the input. The problem

$$\min_{\mu \in \mathbb{R}_+(\mathbb{M})} \hat{E} := \min_{\mu \in \mathbb{R}_+(\mathbb{M})} \langle \theta, \mu \rangle - \sum_{w \in \mathcal{V} \cup \mathcal{E}} c_w H(\mu_w)$$
$$\text{s.t.} \begin{array}{l} \sum_{x_v \in \mathcal{X}_v} \mu_{v,x_v} = 1,\ v \in \mathcal{V}\,, \\ \sum_{x_v \in \mathcal{X}_v} \mu_{uv,x_{uv}} = \mu_{u,x_u},\ x_u \in \mathcal{X}_u,\ uv \in \mathcal{E}\,, \\ \sum_{x_u \in \mathcal{X}_u} \mu_{uv,x_{uv}} = \mu_{v,x_v},\ x_v \in \mathcal{X}_v,\ uv \in \mathcal{E}\,, \end{array} \tag{1.23}$$

is closely related to the primal relaxed one (1.13), and arises e.g. when one applies the smoothing technique (Nesterov, 2004; Jojic et al., 2010; Savchynskyy et al., 2011, 2012; Hazan and Shashua, 2010) or considers approximations for marginalization inference (Wainwright and Jordan, 2008; Wainwright et al., 2005; Jancsary and Matz, 2011). We refer to the works of Heskes (2004), Weiss et al. (2007) and Hazan and Shashua (2010) for a description of the sufficient conditions for convexity of (1.23). Assuming a precision $\varepsilon = 10^{-16}$ to be sufficient for practical needs, we equip (1.23) with an additional set of box constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$, where $|\mathbb{M}|$ is the dimensionality of the vector $\mu$. This is done to obtain a finitely large Lipschitz constant according to Lemma 1.3.

**Optimizing projection.** Denoting the local polytope $\mathcal{L}$ augmented with the additional box-constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$ as $\hat{\mathcal{L}}$, we define the corresponding optimizing projection $\mathcal{P}_{\hat{E},\hat{L}}(\mu)$ as

$$\mu'_v \quad = \Pi_{\Delta(|\mathcal{X}_v|) \cap [\varepsilon,1]^{|\mathcal{X}_v|}}(\mu_v),\ v \in \mathcal{V}\,, \tag{1.24}$$

for $uv \in \mathcal{E}$:

$$\mu'_{uv} = \arg \min_{\mu_{uv} \in [\varepsilon,1]^{|\mathcal{X}_{uv}|}} \langle \theta_{uv} - c_{uv} \log(\mu_{uv}), \mu_{uv} \rangle \quad \text{s.t. } A_{uv}\mu_{uv} = \mu'_v\,, \tag{1.25}$$

where $\log z$, $z \in \mathbb{R}^n$, is defined coordinate-wise. By applying Theorem 1.1 and Lemma 1.3 one obtains that the convergence rate of a given sequence $\mu^t \in \mathbb{R}(\mathbb{M})$ in the worst case slows down by a factor $\|\theta_{\mathcal{V}}\| + \|\theta_{\mathcal{E}}\| + \sum_{w \in \mathcal{V} \cup \mathcal{E}} |\mathcal{X}_w||1 + \log \varepsilon|$, where the last term constitutes a difference to the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ for the primal MAP-inference problem (1.13).

**Remark 1.5.** *Indeed, the additional constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$ are needed only for the theoretical analysis of the projected estimate $\mathcal{P}_{\hat{E},\hat{\mathcal{L}}}(\mu)$, to show that when the true marginals $\mu$ become close to* 0*, the optimizing projection (and in fact the Euclidean one also) behaves worse.*

*However, there is no reason to use these constraints in practice: according to Theorem 1.1, the projected feasible estimates will converge to the optimum of the problem together with the non-projected infeasible ones even without the box constraints, due to continuity of the entropy $H$. It is only the speed of convergence of the projected estimates, which will decrease logarithmically. Moreover, omitting the box constraints $\mu \in [\varepsilon, 1]^{|\mathbb{M}|}$ simplifies the computations* (1.24) *and* (1.25)*. The first one then corresponds to the projection onto the simplex, and the second one to a small-sized* entropy minimization*, efficiently solvable by the Newton method after resorting to its corresponding* smooth *and* unconstrained *dual problem.*

*Moreover, we suggest to threshold $\mu_v$ by setting $\mu_{v,x_v}$ to zero if it is less than the precision $\varepsilon$. That decreases the size of the subproblem* (1.25) *and allows to avoid numerical problems.*

## 1.4 Optimizing Projection in Algorithmic Schemes

In the previous sections, we concentrated on the way to compute the optimizing projection, assuming that a weakly converging (but infeasible) sequence is given. In this section, we briefly discuss how these infeasible sequences can be generated.

### 1.4.1 Prox-Point Primal-Dual Algorithms

In the simplest case, the (infeasible) estimates $\mu^t$ for the primal (1.13) and $\nu^t$ for the dual (1.17) problems are generated by an algorithm itself on each iteration $t$, as is typical for primal-dual algorithms. These algorithms address the relaxed problem (1.13) in its saddle-point formulation

$$\max_{\mu \geq 0} \min_{\nu} \quad \left\{ \langle -b, \nu \rangle + \left\langle \mu, A^\top \nu \right\rangle - \langle \theta, \mu \rangle \right\} . \tag{1.26}$$

The matrix $A$ corresponds to equality constraints in (1.13). Some of the methods (described in Section **??** and works of Martins et al. (2011), Fu and Banerjee (2013)) additionally approach (1.26) with prox-terms of the form $\|A\mu - b\|^2$ or $\|A^\top \nu - \theta\|^2$. Some of these algorithms maintain feasible dual estimates $\nu^t$ as in Section **??** and in works of Martins et al. (2011) and Fu and Banerjee (2013), whereas others do not, as done by Schmidt et al. (2011).

However, to the best of our knowledge, none of these algorithms maintains feasibility of the primal estimates $\mu^t$ with respect to the problem (1.13). One can obtain the feasible estimates, as well as the duality gap estimation, by applying the optimizing projection $\mathcal{P}_{E,\mathcal{L}}(\mu^t)$ defined by (1.15)-(1.16) and – if needed – $\mathcal{P}_{\mathcal{U},\mathbb{D}}(\nu^t)$ defined by (1.20)-(1.22), respectively.

### 1.4.2 Dual Decomposition Based Algorithms

There is an alternative way to formulate a dual problem to (1.13), based on the Lagrangian or dual decomposition. This technique allows to construct particularly efficient inference algorithms. We will review the reconstruction of primal estimates for these algorithms in this section.

For the sake of brevity we consider the case, where the master-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be covered by two *acyclic* subgraphs $\mathcal{G}^i = (\mathcal{V}^i, \mathcal{E}^i)$, $i = 1, 2$, such that each edge of $\mathcal{G}$ is covered only once, and each vertex twice, i.e. by either subgraph: $\mathcal{V}^1 = \mathcal{V}^2 = \mathcal{V}$, $\mathcal{E}^i \cup \mathcal{E}^2 = \mathcal{E}$ and $\mathcal{E}^1 \cap \mathcal{E}^2 = \emptyset$. An example is a grid graph, which allows such a decomposition into two subgraphs corresponding to its rows and columns.

Introducing

$$\theta^i_{uv} = \begin{cases} \theta_{uv}, & uv \in \mathcal{E}^i \\ 0, & uv \notin \mathcal{E}^i \end{cases}, \ i = 1, 2\,, \tag{1.27}$$

and assuming $\theta^1_{v,x_v} + \theta^2_{v,x_v} = \theta_{v,x_v}$, $\forall v \in \mathcal{V}, x_v \in \mathcal{X}_v$, which can be rewritten in a parametric way as $\theta^1_{v,x_v} = \frac{\theta_{v,x_v}}{2} + \lambda_{v,x_v}$ and $\theta^2_{v,x_v} = \frac{\theta_{v,x_v}}{2} - \lambda_{v,x_v}$, $\lambda_{v,x_v} \in \mathbb{R}$, one obtains a lower bound

$$\min_{x\in\mathcal{X}} E(\theta, x) = \min_{x\in\mathcal{X}} \langle \theta, \delta(x) \rangle \geq \max_{\lambda\in\mathbb{R}(\Lambda)} \sum_{i=1}^{2} \min_{x\in\mathcal{X}} \left\langle \theta^i, \delta(x) \right\rangle = \min_{\mu\in\mathcal{L}} \langle \theta, \mu \rangle \ . \tag{1.28}$$

Here $\mathbb{R}(\Lambda) := \mathbb{R}^{\sum_{v\in\mathcal{V}} |\mathcal{X}_v|}$. The last equality is not straightforward and holds for a decomposition of $\mathcal{G}$ to *arbitrary* acyclic subgraphs. We refer to the work of Komodakis et al. (2011) for its proof.

The unconstrained concave, but non-smooth problem

$$\max_{\lambda\in\mathbb{R}(\Lambda)} U(\lambda) := \max_{\lambda\in\mathbb{R}(\Lambda)} \sum_{i=1}^{2} \min_{x\in\mathcal{X}} \left\langle \theta^i, \delta(x) \right\rangle \tag{1.29}$$

is dual to the relaxed problem (1.13).

In the following two paragraphs we will provide several different expressions for computing $\mu^{it}_w$, $w \in \mathcal{V}^i \cup \mathcal{E}^i$, $i = 1, 2$, which will serve as coordinates of the *infeasible* primal sequences converging to the optimum of the problems (1.13) or (1.23), respectively. Although multiple ways of constructing such a sequence out of these coordinates are possible, we will use the following scheme in our experiments:

$$\overline{\mu}^t_{\mathcal{V}} = \frac{1}{2}(\mu^{1t}_{\mathcal{V}} + \mu^{2t}_{\mathcal{V}}), \quad \overline{\mu}^t_{uv} = \begin{cases} \mu^{1t}_{uv}, & uv \in \mathcal{E}^1 \\ \mu^{2t}_{uv}, & uv \in \mathcal{E}^2 \end{cases}, \ uv \in \mathcal{E}\,. \tag{1.30}$$

To transform the sequences into feasible ones we will apply corresponding ($\mathcal{P}_{E,\mathcal{L}}$ or $\mathcal{P}_{\hat{E},\hat{L}}$) optimizing projections to $\overline{\mu}^t$.

**Subgradient and Bundle methods.** Sub-gradient method of Shor et al. (1985)

$$\lambda^{t+1} = \lambda^t + \tau^t \frac{\partial U}{\partial \lambda}(\lambda^t), \text{ where } \tau^t \to 0 \text{ and } \sum_{t=1}^{\infty} \tau^t = \infty \tag{1.31}$$

was one of the first optimization algorithms with convergence guarantees, independently applied by Storvik and Dahl (2000) and later by Schlesinger and Giginyak (2007) and Komodakis et al. (2007) to tackle (1.29). It is based on the fact that the subgradient $\frac{\partial U}{\partial \lambda} = \delta_{\mathcal{V}}(x^{*1}) - \delta_{\mathcal{V}}(x^{*2})$, where $x^{*i} = \arg\min_{x \in \mathcal{X}} \left\langle \theta^i(\lambda), \delta(x) \right\rangle$, is efficiently computable by dynamic programming when graphs $\mathcal{G}^i$ are acyclic.

It is shown by Larsson et al. (1999) and later applied by Komodakis et al. (2011) (see also (Sontag et al., 2011, Sec.1.7.1)) that both *time-averaged*

$$\mu^{it}_w := \frac{\sum_{k=1}^t \delta_w(x^{*i,k})}{t}, \ w \in \mathcal{V}^i \cup \mathcal{E}^i \tag{1.32}$$

and *step-size averaged* labelings

$$\mu^{it}_w := \frac{\sum_{k=1}^t \tau^k \delta_w(x^{*i,k})}{\sum_{k=1}^t \tau^k}, \ w \in \mathcal{V}^i \cup \mathcal{E}^i\,, \tag{1.33}$$

where $x^{*i} = \arg\min_{x \in \mathcal{X}} \left\langle \theta^i(\lambda^t), \delta(x) \right\rangle$ and $t$ denotes the iteration counter of the algorithm (1.31), can be used to construct a primal sequence converging to the optimum of the primal relaxed problem (1.13). Sequences $\overline{\mu}^t$, constructed as in (1.30) out of $\mu^{it}_w$ defined either by (1.32) or by (1.33), are infeasible however, i.e. they do not fulfill constraints of (1.13) up to the optimum. They can be turned into feasible ones with the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ defined by (1.15)-(1.16).

The coordinates of a converging infeasible primal sequence for the bundle method can be constructed as $\mu^{it}_w := \frac{\sum_{k=1}^t \xi^k \delta_w(x^{*i,k})}{\sum_{k=1}^t \xi^k}$, $w \in \mathcal{V}^i \cup \mathcal{E}^i$, where coefficients $\xi^k$ are the weights of the $k$-th subgradient in the bundle (Kappes et al., 2012, eq. 23).

**Smoothing/Marginalization Inference.** Another group of optimization algorithms (Savchynskyy et al., 2011, 2012; Hazan and Shashua, 2010; Ravikumar et al., 2010; Johnson et al., 2007) overcomes the non-smoothness of the dual problem (1.29) by smoothing it prior to optimization. To this end the 'min' operation in (1.29) is replaced by the well-known 'log-sum-exp' (or negative soft-max) function (Rockafellar and Wets, 2004; Nesterov, 2004) yielding

$$\hat{U}_\rho(\lambda) := -\sum_{i=1}^{2} \rho \log \sum_{x \in \mathcal{X}} \exp\left\langle -\theta^i(\lambda)/\rho, \delta(x)\right\rangle, \quad \rho > 0\,. \tag{1.34}$$

This approximation becomes tighter as $\rho$ decreases, as stated by the well-known inequality $\hat{U}_\rho(\lambda) + 2\rho \log|\mathcal{X}| \geq U(\lambda) \geq \hat{U}_\rho(\lambda)$. Maximization of $\hat{U}_\rho$ over $\mathbb{R}(\Lambda)$ is dual to minimization of the entropy-smoothed energy $\hat{E}$ over $\mathcal{L}$ (for certain coefficients $c_w$) defined in (1.23) and hence is used also for an approximate marginalization inference (Wainwright et al., 2005; Jancsary and Matz, 2011).

Let us define coordinates of the primal sequence as

$$\mu^{it}_{w,x_w} := \frac{\sum\limits_{x' \in \mathcal{X}, x'_w = x_w} \exp\left\langle -\theta^i(\lambda^t)/\rho, \delta(x')\right\rangle}{\exp(-\hat{U}^i_\rho(\lambda^t)/\rho)}, \; w \in \mathcal{V}^i \cup \mathcal{E}^i\,, \tag{1.35}$$

where $\lambda^t$ converges to the optimum of $\hat{U}_\rho$ as $t \to \infty$. Note that the $\mu^{it}$ correspond to sum-prod marginals of the subgraphs $\mathcal{G}^i$, and are efficiently computable by dynamic programming when $\mathcal{G}^i$ are acyclic. It is known (Savchynskyy et al., 2011) that the sequence $\overline{\mu}^t$ constructed from $\mu^{it}_{w,x_w}$ as in (1.30) converges to the optimum of (1.23) as $t \to \infty$. Application of the optimizing projection $\mathcal{P}_{\hat{E},\hat{L}}(\mu)$ defined in Section 1.3.4 turns the *infeasible* sequence $\overline{\mu}^t$ into a feasible one.

**Remark 1.6.** *If the final objective of the optimization is not the entropy-smoothed primal problem* (1.23)*, but the primal MAP-inference* (1.13)*, and the smoothing is used as an optimization tool to speed up or guarantee convergence (Savchynskyy et al., 2011, 2012; Hazan and Shashua, 2010; Johnson et al., 2007), one can obtain even better primal bounds at a lower computational cost. Namely, the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ can be applied to approximate the optimal solution of the primal MAP-inference problem* (1.13)*. Denote $\hat{\mu}' = (\hat{\mu}'_\mathcal{V}, \hat{\mu}'_\mathcal{E}) = \mathcal{P}_{\hat{E},\hat{\mathcal{L}}}(\mu_\mathcal{V}, \mu_\mathcal{E})$ and $\mu' = (\mu'_\mathcal{V}, \mu'_\mathcal{E}) = \mathcal{P}_{E,\mathcal{L}}(\mu_\mathcal{V}, \mu_\mathcal{E})$.*

*Ignoring the box-constraints according to the recommendations of Remark 1.5, from the definitions* (1.15) *and* (1.24) *it follows that $\hat{\mu}'_\mathcal{V} = \mu'_\mathcal{V}$, and thus due to* (1.16) *and* (1.25)*, $E(\mu') \leq E(\hat{\mu}')$. This means that the projection $\mathcal{P}_{E,\mathcal{L}}$ is preferable for approximating the minimum of $E$ over $\mathcal{L}$ even* *in the case when the smoothed problem* (1.23) *was optimized, rather than the original non-smooth* (1.13). *As an additional benefit, one obtains faster convergence of the projection even from the worst-case analysis, due to a better estimate of the Lipschitz constant for the function $E$ compared to the function $\hat{E}$, as provided by Lemmas 1.2 and 1.3.*

### 1.4.3 Non-smooth Coordinate Descent: TRWS, MPLP and others

We are not aware of methods for reconstructing primal solutions of the relaxed MAP-inference problem (1.13) from dual estimates for non-smooth coordinate descent based schemes like TRW-S of Kolmogorov (2006) and MPLP of Globerson and Jaakkola (2007). Indeed, these schemes do not solve the relaxed MAP problem in general, hence even if one would have such a method at hand, it would not guarantee convergence of the primal estimates to the optimum.

## 1.5 Experimental Analysis and Evaluation

The main goal of this section is to show how Theorem 1.1 works in practice. To this end we provide three different experiments. All they address the relaxed MAP inference problem (1.13) and include reconstruction of *feasible* primal estimates for it. Additionally we refer to the works of Savchynskyy et al. (2011), Schmidt et al. (2011), Savchynskyy et al. (2012) for experiments with an extended set of benchmark data.

In the first experiment we show convergence of the feasible primal estimates for three different algorithms. In the second one we show advantages of the feasible relaxed primal estimates over integral primal estimates for efficient adaptive algorithms. Finally, the third experiment shows that the bounds (1.8)- (1.9) allow a qualitative prediction of the objective value in the (feasible) projected point.

For the experiments we used our own implementations of the First Order Primal-Dual Algorithm (acronym `FPD`) of Chambolle and Pock (2010) (originally proposed by Pock et al. (2009)) as described by Schmidt et al. (2011), the adaptive diminishing smoothing algorithm `ADSAL` proposed by Savchynskyy et al. (2012), the dual decomposition based subgradient ascent `SG` with an adaptive step-size rule (Kappes et al., 2012, eq.17) and primal estimates based on time-averaged subgradients (see Section 1.4.2), and finally Nesterov's accelerated gradient ascent method `NEST` applied to the smoothed dual decomposition-based objective studied by Savchynskyy et al. (2011). All implementations are based on data structures of the OpenGM library by Andres et al. (2012).

The optimizing projection to the local polytope w.r.t. to the MAP-energy (1.15)-(1.16) is computed using our implementation of a specialization of the simplex algorithm for transportation problems (Bazaraa and Jarvis, 1977). We adopted an elegant method by Bland (1977), also discussed in the textbook of Papadimitriou and Steiglitz (1998), to avoid cycling. The source code of the solver is available as a part of the OpenGM library.

**Feasible Primal Bound Estimation.** In the first experiment, we demonstrate that for all three groups of methods discussed in Section 1.4 our method efficiently provides feasible primal estimates for the MAP inference problem (1.13). To this end we generated a $256 \times 256$ grid model with 4 variable states ($|\mathcal{X}_v| = 4$) and potentials $\theta$ randomly distributed in the interval $[0, 1]$. We solved the LP relaxation of the MAP inference problem (1.13) with `FPD` as a representative of methods dealing with infeasible primal estimates, `ADSAL` as the fastest representative of smoothing-based algorithms and the subgradient method `SG`. The corresponding plot is presented in Figure 1.1 (left). We note that for *all* algorithms the time needed to compute the optimizing projection $\mathcal{P}_{E,\mathcal{L}}$ did not exceed the time needed to compute the subgradient/gradient of the respective dual function and typically required 0.01-0.02 s on a 3GHz machine. The generated dataset is not LP tight, hence the obtained relaxed primal solution has a significantly lower energy than the integer one. In contrast to the situation when only non-relaxed integer primal estimates would be computed, the primal and dual bounds of the relaxed problem converge to the same limit value. Due to the feasibility of both primal and dual estimates, the primal and dual objective functions' values bound the optimal value of the relaxed problem from above and below, respectively.

**Relaxed Primal Estimates for Adaptive Algorithms** We demonstrate the practical usefulness of feasible relaxed primal estimates with the diminishing smoothing `ADSAL` algorithm of Savchynskyy et al. (2012). It optimizes the smooth dual (1.34) with a degree of smoothing $\rho$, which decreases with the estimated duality gap. In the original work of Savchynskyy et al. (2012) the primal bound is computed with the optimizing projection as described in Remark 1.6. To demonstrate its importance we substituted this computation with an estimation of an integer solution by rounding (as done by Kolmogorov (2006) in the TRW-S algorithm). Figure 1.1 (right) shows the difference between convergence of the original `ADSAL` algorithm and its modification on the randomly generated $25 \times 25$ grid model with 4 labels. Since the gap between the integer and the dual bounds does not vanish,

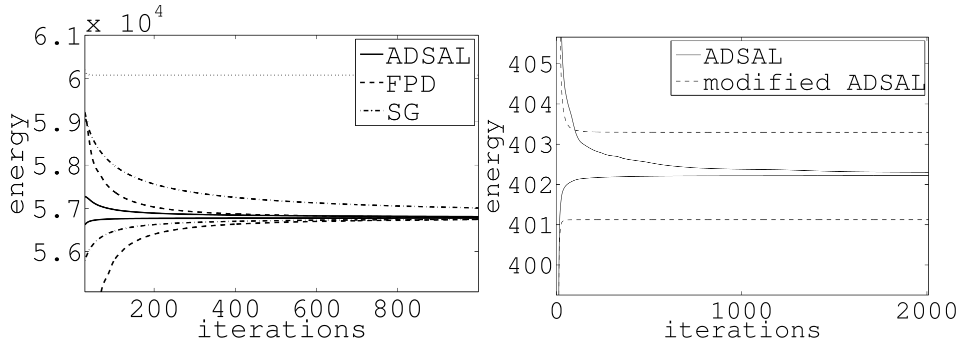

**Figure 1.1**: Left: Convergence of the primal (upper curves) and dual (lower curves) bounds to the same optimal limit value for `ADSAL`, `FPD` and `SG` algorithms. The obtained integer bound is plotted as a dotted line.
Right: Convergence of the original `ADSAL` algorithm and its modification. In the modified algorithm an integral labeling provides a primal bound for the smoothing update, whereas in the original algorithm a feasible *relaxed* primal estimate is used.

the smoothing does not vanish either and the overall algorithm gets stuck in a suboptimal point, whereas the original algorithm based on the relaxed primal estimate converges to the optimum.

**Evaluation of Convergence Estimates.** The third experiment is devoted to the evaluation of the convergence estimates (1.8)-(1.9) provided by Theorem 1.1. To this end, we generated four LP-tight grid-structured datasets with known optimal labeling. We refer to the work of Schmidt et al. (2011) for a description of the generation process. The resulting unary and pairwise potentials were distributed in the interval $[-10, 10]$. We picked a random subset of edges not belonging to the optimal labeling and assigned them "infinite" values. We created four datasets with "infinities" equal to 10 000, 100 000, 1 000 000 and 10 000 000 and ran `NEST` for inference. According to Theorem 1.1 the energy $E$ evaluated on projected feasible estimates $\mathcal{P}_{E,\mathcal{L}}(\overline{\mu}^t_{\mathcal{V}}, \overline{\mu}^t_{\mathcal{E}})$, $t = 1, \ldots, \infty$, where the *infeasible* estimates $\overline{\mu}^t$ were constructed as in Section 1.4.2, can be represented as

$$E(\mathcal{P}_{E,\mathcal{L}}(\overline{\mu}^t_{\mathcal{V}}, \overline{\mu}^t_{\mathcal{E}})) = F(\overline{\mu}^t) + L_Y(E)\|\overline{\mu}^t - \Pi_{\mathcal{L}}\overline{\mu}^t\| \tag{1.36}$$

for a suitably selected function $F$. Since `NEST` is a purely dual method and "infinite" pairwise potentials did not contribute significantly to the values and gradients of the (smoothed) dual objective, the infeasible primal estimates $\overline{\mu}^t$ (with $t$ denoting an iteration counter) were the same for all four different approximations of the infinity value. Since according to Lemma 1.2 the Lipschitz constant $L_Y(E)$ is asymptotically proportional to the norm of the pairwise potentials $\|\theta_{\mathcal{E}}\|$ we plotted the values $\log E(\mathcal{P}_{E,\mathcal{L}}(\overline{\mu}^t_{\mathcal{V}}, \overline{\mu}^t_{\mathcal{E}}))$ as a

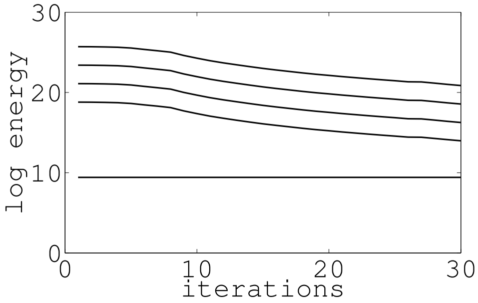

**Figure 1.2**: Convergence of the obtained primal feasible solution for four datasets which differ only by the values used as "infinity". The energy values are plotted in logarithmic scale. From bottom to top: optimal log-energy, primal bounds corresponding to infinity values equal to 10 000, 100 000, 1 000 000 and 10 000 000.

function of $t$ for all four datasets in Figure 1.2. As predicted by Theorem 1.1 the corresponding energy values differ by approximately a factor of 10, as the "infinite" values do. Due to the logarithmic energy scale this difference corresponds to equal log-energy distances between the curves in Figure 1.2.

## 1.6 Conclusions

We presented an efficient and quite general optimizing projection method for computing feasible primal estimates for dual and primal-dual optimization schemes. The method provides convergence guarantees similar to those of the Euclidean projection, but contrary to the latter allows for efficient computations if the feasible set and the objective function possess certain separability properties. As any optimization tool it has also certain limitations related to the Lipschitz continuity of the primal objective, however, exactly the same limitations are characteristic also for the Euclidean projection. Hence they can not be considered as particular disadvantages of this method, but must rather be considered as disadvantages of projection methods in general. These limitations can only be overcome by constructing algorithms that intrinsically maintain feasible primal estimates during iterations. The construction of such algorithms has to be addressed in future work.

## Acknowledgements

This work has been supported by the German Research Foundation (DFG) within the program “Spatio-/Temporal Graphical Models and Applications in Image Analysis”, grant GRK 1653.

### Proofs

**Theorem 1.1**.

*Proof.* We will denote $(x^p, y^p) = z^p = \Pi_C(z)$ and $(x', y') = \mathcal{P}_{f,C}(x, y)$. Note that

- from $f^*_C \leq f(x', y') \leq f(x', y'')$ for any $y'' \in Y$ such that $(x', y'') \in C$ it follows that

$$f^*_C \leq f(x', y') \leq f(x', y^p)\,, \tag{1.37}$$

- from $\|z - z^p\| = \sqrt{\|x - x^p\|^2 + \|y - y^p\|^2}$ it follows that

$$\|y - y^p\| \leq \|z - z^p\| \text{ and } \|x - x^p\| \leq \|z - z^p\|\,. \tag{1.38}$$

- according to (1.4) $x' = \Pi_{C_X}(x) = \arg\min_{\tilde{x}\in C_X} \|x - \tilde{x}\|$ and hence

$$\|x - x'\| \leq \|x - x^p\| \tag{1.39}$$

since $x^p \in C_X$. Combining this with (1.38) we obtain

$$\|x - x'\| \leq \|z - z^p\|\,. \tag{1.40}$$

- The triangle inequality $\|x' - x^p\| \leq \|x' - x\| + \|x - x^p\|$ and (1.39) applied to $x^{t'} := \Pi_{C_X}(x^t)$ and $(x^{tp}, y^{tp}) := \Pi_C(x^t, y^t)$ in place of $x'$ and $(x^p, y^p)$ respectively suggest that from $\|(x^t, y^t) - \Pi_C(x^t, y^t)\| \xrightarrow{t\to\infty} 0$ follows that

$$\|x^{t'} - x^{tp}\| \xrightarrow{t\to\infty} 0 \text{ and } \|x - x^{t'}\| \xrightarrow{t\to\infty} 0\,. \tag{1.41}$$

Implication (1.7) follows from (1.41), continuity of $f$ and inequality

$$\begin{aligned}|f(\mathcal{P}_{f,C}(x^t, y^t)) - f^*_C| = |f(x^{t'}, y^{t'}) - f^*_C| &\overset{(1.37)}{\leq} |f(x^{t'}, y^{tp}) - f^*_C| \\ &\leq |f(x^{t'}, y^{tp}) - f(x^{t'}, y^t)| + |f(x^{t'}, y^t) - f^*_C|\,. \end{aligned} \tag{1.42}$$

Implication (1.8) follows from

$$\begin{aligned}|f(\mathcal{P}_{f,C}(x,y)) - f_C^*| &= |f(x',y') - f_C^*| \overset{(1.37)}{\leq} |f(x',y^p) - f_C^*| \\ &\leq |f(x',y^p) - f(x',y)| + |f(x',y) - f_C^*| \leq L_Y(f)\|y - y^p\| + |f(x',y) - f_C^*| \\ &\overset{(1.38)}{\leq} L_Y(f)\|z - z^p\| + |f(x',y) - f_C^*|\,. \qquad (1.43)\end{aligned}$$

assuming that $x' = x$.

Implication (1.9) follows from (1.43) and Lipschitz-continuity of $f$ w.r.t. $x$:

$$\begin{aligned}|f(\mathcal{P}_{f,C}(x,y)) - f_C^*| &\overset{(1.43)}{\leq} L_Y(f)\|z - z^p\| + |f(x',y) - f_C^*| \\ &\leq L_Y(f)\|z - z^p\| + |f(x',y) - f(x,y)| + |f(x,y) - f_C^*| \\ &\leq L_Y(f)\|z - z^p\| + L_X(f)\|x' - x\| + |f(x,y) - f_C^*| \\ &\overset{(1.40)}{\leq} L_Y(f)\|z - z^p\| + L_X(f)\|z - z^p\| + |f(x,y) - f_C^*| \\ &= (L_Y(f) + L_X(f))\|z - z^p\| + |f(x,y) - f_C^*|\,. \qquad (1.44)\end{aligned}$$

□

**Lemma 1.2**.

*Proof.* All three Lipschitz-constants are derived from the Cauchy-Bunyakovsky-Schwarz inequality $\langle c, \nu\rangle \leq \|c\| \cdot \|\nu\|,\ c, \nu \in \mathbb{R}^N$, applied respectively to $x$, $y$ and $z = (x,y)$ in place of $\nu$. □

**Lemma 1.3**.

*Proof.* The function $f_i(z_i) = z_i \log z_i$ of a single variable is differentiable on $[\varepsilon, M]$ and its derivative $f_i'(z_i) = 1 + \log z_i$ is monotone increasing, hence $f_i(z_i)$ is convex. This implies $f_i(z_i) - f_i(z_i') \leq f_i'(z_i)(z_i - z_i')$ and $|f_i(z_i) - f_i(z_i')| \leq |f_i'(z_i)||(z_i - z_i')|$. Taking into account that due to monotonicity $|f_i'(z_i)| \leq \max\{|1+\log\varepsilon|, |1+\log M|\}$ for $z_i \in [\varepsilon, M]$, and using the fact that $L(f_1+f_2) \leq L(f_1)+L_f(f_2)$ together with Lemma 1.2, one obtains (1.11). □

## 1.7 References

B. Andres, T. Beier, and J. H. Kappes. OpenGM: A C++ library for discrete graphical models. Technical report, arXiv:1206.0111, 2012.

M. S. Bazaraa and J. J. Jarvis. *Linear Programming and Network Flows.* Wiley, 1977.

R. G. Bland. New finite pivoting rules for the simplex method. *Mathematics of Operations Research*, pages 103–107, 1977.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, USA, 2004.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1222–1239, 2001.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, pages 1–26, 2010.

Q. Fu and H. W. A. Banerjee. Bethe-ADMM for tree decomposition based parallel MAP inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2013.

A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Proceedings of the Conference on Neural Information Processing Systems*, 2007.

T. Hazan and A. Shashua. Norm-product belief propagation: Primal-dual message-passing for approximate inference. *IEEE Transactions on Information Theory*, 56(12):6294 –6316, Dec. 2010.

T. Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16(11):2379–2413, 2004.

J. Jancsary and G. Matz. Convergent decomposition solvers for tree-reweighted free energies. In *Proceedings of the Conference on Artificial Intelligence and Statistics*, 2011.

J. K. Johnson, D. Malioutov, and A. S. Willsky. Lagrangian relaxation for MAP estimation in graphical models. In *Proceedings of the Allerton Conference on Communication, Control and Computation*, 2007.

V. Jojic, S. Gould, and D. Koller. Accelerated dual decomposition for MAP inference. In *Proceedings of the International Conference on Machine Learning*, pages 503–510, 2010.

J. H. Kappes, B. Savchynskyy, and C. Schnörr. A bundle approach to efficient MAP-inference by Lagrangian relaxation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2012.

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

N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *Proceedings of the International Conference on Computer Vision*, 2007.

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

T. Larsson, M. Patriksson, and A.-B. Strömberg. Ergodic, primal convergence in dual subgradient schemes for convex programming. *Mathematical Programming*, 86:283–312, 1999.

A. F. T. 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 *Proceedings of the International Conference on Machine Learning*, 2011.

C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of Rn. *Journal of Optimization Theory and Applications*, 50

(1):195–200, 1986.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, Ser. A(103):127–152, 2004.

C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity.* Mineola, N.Y. : Dover Publications, 2nd edition, 1998.

T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the piecewise smooth Mumford-Shah functional. In *Proceedings of the International Conference on Computer Vision*, 2009.

P. Ravikumar, A. Agarwal, and M. Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *Journal of Machine Learning Research*, 11:1043–1080, 2010.

R. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer, 2nd edition, 2004.

B. Savchynskyy, J. Kappes, S. Schmidt, and C. Schnörr. A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

B. Savchynskyy, S. Schmidt, J. Kappes, and C. Schnörr. Efficient MRF energy minimization via adaptive diminishing smoothing. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 746–755, 2012.

M. Schlesinger and V. Giginyak. Solution to structural recognition (max,+)-problems by their equivalent transformations. in 2 parts. *Control Systems and Computers*, (1-2), 2007.

M. Schlesinger, E. Vodolazskiy, and N. Lopatka. Stop condition for subgradient minimization in dual relaxed (max,+) problem. In *Proceedings of the Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2011.

M. I. Schlesinger. Syntactic analysis of two-dimensional visual signals in the presence of noise. *Kibernetika*, (4):113–130, July-August 1976.

S. Schmidt, B. Savchynskyy, J. Kappes, and C. Schnörr. Evaluation of a first-order primal-dual algorithm for MRF energy minimization. In *Proceedings of the Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2011.

N. Z. Shor, K. C. Kiwiel, and A. Ruszcayski. *Minimization methods for non-differentiable functions.* Springer-Verlag New York, Inc., 1985.

D. Sontag, T. Meltzer, A. Globerson, Y. Weiss, and T. Jaakkola. Tightening LP relaxations for MAP using message-passing. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 503–510, 2008.

D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. *Optimization for Machine Learning*, 1, 2011.

G. Storvik and G. Dahl. Lagrangian-based methods for finding MAP solutions for MRF models. *IEEE Transactions on Image Processing*, 9(3):469–479, 2000.

M. Wainwright, T. Jaakkola, and A. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51:2313–2335, 2005.

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.

Y. Weiss and W. T. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, 47(2):736–744, 2001.

Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2007.

T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7), July 2007.

T. Werner. Revisiting the decomposition approach to inference in exponential families and graphical models. Technical report, CMP, Czech TU, 2009.

T. Werner. How to compute primal solution from dual one in MAP inference in MRF? *Control Systems and Computers*, (2), March-April 2011.