What Is Optimized in Tight Convex Relaxations for Multi-Label Problems?

Christopher Zach
Microsoft Research Cambridge, UK
chzach@microsoft.com

Christian Häne
ETH Zürich, Switzerland
chaene@inf.ethz.ch

Marc Pollefeys
ETH Zürich, Switzerland
marc.pollefeys@inf.ethz.ch

Abstract

In this work we present a unified view on Markov random fields and recently proposed continuous tight convex relaxations for multi-label assignment in the image plane. These relaxations are far less biased towards the grid geometry than Markov random fields. It turns out that the continuous methods are non-linear extensions of the local polytope MRF relaxation. In view of this result a better understanding of these tight convex relaxations in the discrete setting is obtained. Further, a wider range of optimization methods is now applicable to find a minimizer of the tight formulation. We propose two methods to improve the efficiency of minimization. One uses a weaker, but more efficient continuously inspired approach as initialization and gradually refines the energy where it is necessary. The other one reformulates the dual energy enabling smooth approximations to be used for efficient optimization. We demonstrate the utility of our proposed minimization schemes in numerical experiments.

1. Introduction

Assigning labels to image regions e.g. in order to obtain a semantic segmentation, is one of the major tasks in computer vision. The most prominent approach to solve this problem is to formulate label assignment as Markov random field (MRF) incorporating local label preference and neighborhood smoothness. Since in general label assignment is NP-hard, finding the true solution is intractable and approximate ones are determined. One promising approach to solve MRF instances is to relax the intrinsically difficult constraints to convex outer bounds. There are currently two somewhat distinct lines of research utilizing such convex relaxations: the direction, that is mostly used in the machine learning community, is based on a graph representation of image grids and uses variations of dual block-coordinate methods [10, 9, 16, 15] (usually referred as message passing algorithms in the literature). The other set of methods is derived from the analysis of partitioning an image in the continuous setting, i.e. variations of the Mumford-Shah segmentation model [12, 1]. Using the principle of biconjugation to obtain tight convex envelopes, [5] obtains a convex relaxation of multi-label problems with generic (but metric) transition costs *in the continuous setting*. Subsequent discretization of this model to finite grids yields to strong results in practice, but it was not fully understood what is optimized in the discrete setting.

In this work we close the gap between convex formulations for MRFs and continuous approaches by identifying the latter methods as non-linear (but still convex) extensions of the standard LP relaxation of Markov random fields. This insight has several implications: (a) it becomes clearer why the model proposed in [5] is tighter than other relaxations proposed for similar labeling problems, and (b) a wider range of optimization methods becomes applicable, especially after obtaining equivalent convex programs utilizing redundant constraints. Thus, the results obtained in this work are of theoretical and practical interest.

2. Background

In the following section we summarize the necessary background on discrete and continuous relaxations of multi-label problems.

2.1. Notations

In this section we introduce some notation used in the following. For a convex set C we will use $\imath_C$ to denote the corresponding indicator function. i.e. $\imath_C(x) = 0$ for $x \in C$ and ∞ otherwise. We use short-hand notations $[x]_+$ and $[x]_-$ for $\max\{0, x\}$ and $\min\{0, x\}$, respectively. Finally, for an extended real-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ we denote its convex conjugate by $f^*(y) = \max_x x^T y - f(x)$.

2.2. Label Assignment, the Marginal Polytope and its LP Relaxation

In the following we will consider only labeling problems with unary and pairwise interactions between nodes. Let $\mathcal{V}$ be a set of $V = |\mathcal{V}|$ nodes and $\mathcal{E}$ be a set of edges connecting nodes from $\mathcal{V}$. The goal of inference is to assign labels Λ :

$\mathcal{V} \to \{1,\dots,L\}$ for all nodes $s \in \mathcal{V}$ minimizing the energy

$$E_{\text{labeling}}(\Lambda) = \sum_{s\in\mathcal{V}} \theta_s^{\Lambda(s)} + \sum_{(s,t)\in\mathcal{E}} \theta_{st}^{\Lambda(s),\Lambda(t)}, \tag{1}$$

where $\theta_s^{\cdot}$ are the unary potentials and $\theta_{st}^{\cdot}$ are the pairwise ones. Usually the label assignment Λ is represented via indicator vectors $x_s \in \{0,1\}^L$ for each $s \in \mathcal{V}$, and $x_{st} \in \{0,1\}^{L^2}$ for each $(s,t) \in \mathcal{E}$, leading to

$$E_{\text{MRF}}(x) = \sum_{s,i} \theta_s^i x_s^i + \sum_{s,t,i,j} \theta_{st}^{ij} x_{st}^{ij} \tag{2}$$

subject to normalization constraints $\sum_{i\in\{1,\dots,L\}} x_s^i = 1$ for each $s \in \mathcal{V}$ (one label needs to be assigned) and marginalization constraints $\sum_j x_{st}^{ij} = x_s^i$ and $\sum_i x_{st}^{ij} = x_t^j$. In general, enforcing $x_s^i \in \{0,1\}$ is NP-hard, hence the corresponding LP-relaxation is considered,

$$\begin{aligned} E_{\text{LP-MRF}}(x) &= \sum_{s,i} \theta_s^i x_s^i + \sum_{s,t}\sum_{i,j} \theta_{st}^{ij} x_{st}^{ij} \\ \text{s.t.} \ &\sum_j x_{st}^{ij} = x_s^i, \qquad \sum_j x_{st}^{ji} = x_t^i \\ &x_s \in \Delta, \qquad x_{st}^{ij} \ge 0 \qquad \forall s,t,i,j, \end{aligned} \tag{3}$$

where Δ denotes the unit (probability) simplex, $\Delta := \{x : \sum_i x^i = 1, x^i \ge 0\}$. There are several corresponding dual programs depending on the utilized (redundant) constraints. If we explicitly add the box constraints $x_{st}^{ij} \in [-1,1]$ the corresponding dual is

$$\begin{aligned} E^*_{\text{LP-MRF}}(p) = &\sum_s \min_i \Big\{\theta_s^i + \sum_{t\in N_t(s)} p^i_{st\to s} + \sum_{t\in N_s(t)} p^i_{ts\to s}\Big\} \\ &+ \sum_{s,t}\sum_{i,j} \min\big\{0, \theta_{st}^{ij} - |p^i_{st\to s} + p^j_{st\to t}|\big\}, \end{aligned}$$

where we defined $N_t(s) := \{t : (s,t) \in \mathcal{E}\}$ and $N_s(t) := \{s : (s,t) \in \mathcal{E}\}$. The particular choice of (redundant) box constraints $x_{st}^{ij} \in [-1,1]$ in the primal program leads to an exact penalizer for the usually obtained capacity constraints. Different choices of primal constraints lead to different duals, we refer to Section 3.3 for further details.

2.3. Continuously Inspired Convex Formulations for Multi-Label Problems

In this section we briefly review the convex relaxation approach for multi-label problems proposed in [5]. In contrast to the graph-based label assignment problem in Eq. 3, Chambolle et al. consider labeling tasks directly in the (2D) image plane. Their proposed convex relaxation is formulated as the primal-dual saddle-point energy

$$\begin{aligned} E_{\text{CCP-I}}(u,q) = &\sum_{s,i} \theta_s^i (u_s^{i+1} - u_s^i) + \sum_{s,i} (q_s^i)^T \nabla u_s^i \\ \text{s.t.} \ &u_s^i \le u_s^{i+1},\ u_s^0 = 0,\ u_s^{L+1} = 1,\ u_s^i \ge 0 \\ &\Big\|\sum_{k=i}^{j-1} q_s^k\Big\|_2 \le \theta^{ij} \qquad \forall s,i,j, \end{aligned} \tag{4}$$

which is minimized with respect to u and maximized with respect to q. Here u is a *super-level function* ideally transitioning from 0 to 1 for the assigned label, i.e. if label i should be assigned at node (pixel) s, we have $u_s^{i+1} = 1$ and $u_s^i = 0$. Consequently, $u \in [0,1]^{VL}$ in the discrete setting of a pixel grid. $q \in \mathbb{R}^{2VL}$ are auxiliary variables. The stencil of ∇ depends on the utilized discretization. We employ forward differences for ∇ as also used in [5] (unless noted otherwise). θ^{ij} are the transition costs between label i and j and can assumed to be symmetric w.l.o.g., $\theta^{ij} = \theta^{ji}$ and $\theta^{ii} = 0$. At this point we have a few remarks:

1. The saddle-point formulation in combination with the quadratic number of "capacity" constraints $\|\sum_{k=i}^{j-1} q_s^k\|_2 \le \theta_{st}^{ij}$ makes it difficult to optimize efficiently. In [5] a nested, two-level iteration scheme is proposed, where the inner iterations are required to enforce the capacity constraints. In [14] Lagrange multipliers for the dual constraints are introduced in order to avoid the nested iterations, leading to a primal-dual-primal scheme. In Section 3.1 we will derive the corresponding purely primal energy enabling a larger set of convex optimization methods to be applied to this problem.
2. The energy Eq. 4 handles triple junctions (i.e. nodes where at least 3 different phases meet) better than the (more efficient) approach proposed in [17]. Again, by working with the primal formulation one can give a clear intuition why this is the case (see Section 3.2).
3. The energy in Eq. 4 can be rewritten in terms of (soft) indicator functions x_s per pixel, leading to the equivalent formulation (see the supplementary material or [14]):

$$\begin{aligned} E_{\text{CCP-II}}(x,p) = &\sum_{s,i} \theta_s^i x_s^i + \sum_{s,i} (p_s^i)^T \nabla x_s^i \\ \text{s.t.} \ &\big\|p_s^i - p_s^j\big\|_2 \le \theta^{ij},\ x_s \in \Delta \qquad \forall s,i,j, \end{aligned} \tag{5}$$

x and p are of the same dimension as u and q.

3. Convex Relaxations for Multi-Label MRFs Revisited

In this section we derive the connections between the standard LP relaxation for MRFs, LP-MRF, and the saddle-point energy $E_{\text{CCP-II}}$, and further analyze the relation between $E_{\text{CCP-II}}$, and a weaker, but more efficient relaxation. We will make heavy use of Fenchel duality, $\min_x f(x) + g(Ax) = \max_z -f^*(A^T z) - g^*(-z)$, where f and g are conex and l.s.c. functions, and A is a linear operator (matrix for finite dimensional problems). We refer e.g. to [3] for a compact exposition of convex analysis.

3.1. A Primal View on the Tight Convex Relaxation

It seems that the saddle-point formulation in Eq. 4 and Eq. 5, respectively, were never analyzed from the purely primal viewpoint. Using Fenchel duality one can immediately state the primal form of Eq. 5, which has a more intuitive interpretation (detailed in Section 3.2):

$$E_{\text{tight}}(x,y) = \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j} \theta^{ij} \|y_s^{ij}\|_2 \qquad (6)$$
$$\text{s.t. } \nabla x_s^i = \sum_{j:j<i} y_s^{ji} - \sum_{j:j>i} y_s^{ij},\ x_s \in \Delta \quad \forall s, i,$$

where $y_s^{ij} \in \mathbb{R}^2$ represents the transition gradient between a region with label i and the one with label j. y_s^{ij} is 0 if there is no transition between i and j at node (pixel) s. The last set of constraints are the equivalent of marginalization constraints linking transition gradients y_s^{ij} and label gradients ∇x_s^i and ∇x_s^j. The derivation of Eq. 6 is given in the supplementary material.

Since $x_s^i \in [0,1]$ we have that $\nabla x_s^i \in [-1,1]^2$ and we can safely add the additional constraints $y_s^{ij} \in [-1,1]^2$ to obtain an equivalent convex program. We obtain the interpretation that e.g. $(y_s^{ij})_1 = 1$ iff there is a horizontal transition from label i to label j, and $(y_s^{ij})_1 = -1$ if the reverse is the case (analogously for the vertical direction). Consequently, the y_s^{ij} variables correspond to *signed* binary pseudo-marginals, and proper pseudo-marginals can be obtained by setting (component-wise)

$$x_s^{ij} := [y_s^{ij}]_+ \qquad \text{and} \qquad x_s^{ji} := -[y_s^{ij}]_-$$

for $i < j$. x_s^{ii} is e.g. given by $x_s^{ii} = x_s^i - \sum_{j:j\neq i} x_s^{ij}$. Thus, the primal program equivalent to Eq. 6, but purely stated in terms of non-negative pseudo-marginals, reads as

$$E(x) = \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j} \theta^{ij} \left\|x_s^{ij} + x_s^{ji}\right\|_2 \qquad (7)$$
$$\text{s.t. } \nabla x_s^i = \sum_{j:j\neq i} x_s^{ji} - \sum_{j:j\neq i} x_s^{ij},\ x_s \in \Delta,\ x_s^{ij} \geq 0.$$

This is very similar to the standard relaxation of MRFs (recall Eq. 3 after eliminating x_{st}^{ii} in the marginalization constraints), the only difference being the smoothness terms, which is

$$\theta^{ij} \left\|x_s^{ij} + x_s^{ji}\right\|_2 \qquad \text{instead of} \qquad \theta^{ij} x_s^{ij} + \theta^{ij} x_s^{ji}.$$

Note that $\theta^{ij} x_s^{ij} + \theta^{ij} x_s^{ji}$ is equivalent to $\theta^{ij}\|x_s^{ij} + x_s^{ji}\|_1$ (the anisotropic L_1 norm), since $x_s^{ij}, x_s^{ji} \geq 0$. Hence the primal model Eq. 7 can be seen as isotropic extension of the standard model Eq. 3 for regular image grids. Further, we have a complementarity condition for every optimal solution x_s^{ij}: $(x_s^{ij})^T x_s^{ji} = 0$, i.e. $(x_s^{ij})_1 (x_s^{ji})_1 = 0$ and $(x_s^{ij})_2 (x_s^{ji})_2 = 0$. It is easy to see that if the complementarity conditions do not hold, the overall objective can be lowered by subtracting the componentwise minimum from x_s^{ij} and x_s^{ji} (and therefore satisfying complementarity) without affecting the marginalization constraint. Hence, we can also replace $\theta^{ij}\|x_s^{ij} + x_s^{ji}\|_2$ in the primal objective by

$$\theta^{ij} \left\| \begin{matrix} x_s^{ij} \\ x_s^{ji} \end{matrix} \right\|_2 .$$

Finally, observe that all primal formulations have a number of unknowns that is quadratic in the number of labels L. This is not surprising since the number of constraints on the dual variables is $O(L^2)$ per node.

3.2. Truncated Transition Costs

If the transition costs θ^{ij} have no structure, then one has to employ the full representations Eq. 6 or 7. In this section we consider the important case of truncated smoothness costs, i.e. $\theta^{ij} = \theta^*$ if $|i-j| \geq T$ for some T, and $\theta^{ij} < \theta^*$ if $|i-j| < T$. The two most important examples in this category are the Potts smoothness model ($T = 1$), and truncated linear costs with $\theta^{ij} = \min\{|i-j|, \theta^*\}$.

It is tempting to combine the transition gradients corresponding to "large" jumps from label i to label j with $|i-j| \geq T$ into one vector y_s^{i*}, where the star $*$ indicates a wild-card symbol, i.e.

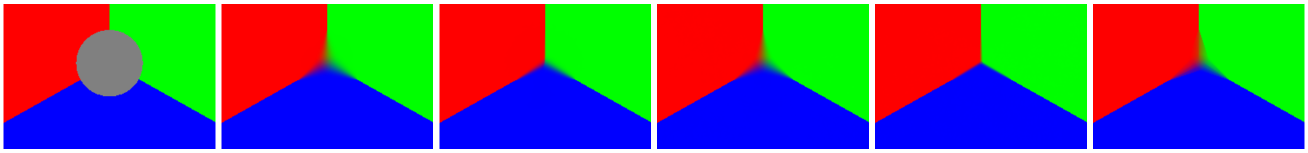
$$y_s^{i*} = \sum_{j:j-i\geq T} y_s^{ij} - \sum_{j:i-j\geq T} y_s^{ji}.$$

Thus, we can formulate a primal program using at most $O(TL)$ unknowns per pixel,

$$E_{\text{fast}}(x,y) = \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j<i+T} \theta^{ij}\|y_s^{ij}\|_2$$
$$+ \frac{\theta^*}{2} \sum_s \sum_i \|y_s^{i*}\|_2 \qquad (8)$$
$$\text{s.t. } \nabla x_s^i = \sum_{j:i-T<j<i} y_s^{ji} - \sum_{j:i<j<i+T} y_s^{ij} - y_s^{i*}$$

and $x_s \in \Delta$. Since a large jump is represented twice via y^{i*} and y^{j*}, the truncation value appears as $\theta^*/2$ above. For the truncated linear smoothness cost the number of required unknowns reduces further to $O(L)$:

$$E_{\text{fast}}(x,y) = \sum_{s,i} \theta_s^i x_s^i + \sum_{s,i} \|y_s^{i,i+1}\|_2 + \frac{\theta^*}{2} \sum_{s,i} \|y_s^{i*}\|_2$$
$$\text{s.t. } \nabla x_s^i = y_s^{i-1,i} - y_s^{i,i+1} - y_s^{i*}. \qquad (9)$$



(a) Input image (b) Forward differences (c) Forward differences (d) Staggered grid (e) Staggered grid (f) Geo-cut

Figure 1. The triple junction inpainting example. (b) and (d) use the weaker relaxation E_{fast}, and (c) and (e) are the results of E_{tight}. The geo-cut solution with a 32-neighborhood is shown in (f).

These models generalize the formulation proposed in [17] beyond the Potts smoothness cost. It is demonstrated in [5] that Eq. 8 is a weaker relaxation than Eq. 5 if three regions with different labels meet (see also Fig. 1). Before we analyze the difference between those models, we state an equivalence result:

Observation 1 *If we use the 1-norm $\|\cdot\|_1$ in the smoothness term instead of the Euclidean one (i.e. we consider the standard LP relaxation of MRFs using horizontal and vertical edges), the formulations in Eqs. 6 and 8 are equivalent.*

More generally, one can collapse the pairwise pseudo-marginals for standard MRFs on graphs in the case of truncated pairwise potentials, leading to substantial reductions in memory requirements. We presume this fact has probably been used in the MRF community, but for completeness we provide a proof in the supplementary material.

The situation is different in the Euclidean norm setting. In the following we consider the Potts smoothness cost. If we use forward differences for the gradient and compare the smoothness costs assigned by Eq. 8 and Eq. 5 for the discrete label configurations, we find out that for triple junctions the formulation in Eq. 8 underestimates the true cost: if label i is assigned to a pixel s, and labels j and k are assigned to the forward neighbors (see Fig. 2), then we have $y^{i*}_s = (-1,-1)^T$, $y^{j*}_s = (1,0)^T$ and $y^{k*}_s = (0,1)^T$, and the smoothness contribution of s according to Eq. 8 is

$$\frac{1}{2}\left(\left\|\begin{matrix}-1\\-1\end{matrix}\right\|_2 + \left\|\begin{matrix}1\\0\end{matrix}\right\|_2 + \left\|\begin{matrix}0\\-1\end{matrix}\right\|_2\right) = 1 + \frac{\sqrt{2}}{2}$$

(see also Fig. 2(a)). On the other hand, the transition gradients according to Eq. 5 are $y^{ij}_s = (-1,0)^T$ and $y^{ik}_s = (0,-1)^T$, and its smoothness contribution is

$$\left\|\begin{matrix}-1\\0\end{matrix}\right\|_2 + \left\|\begin{matrix}0\\-1\end{matrix}\right\|_2 = 2$$

(cf. Fig. 2(b)). It seems that Eq. 8 is a weaker model than Eq. 5 due to the different cost contributions, but the deeper reason is, that the former formulation cannot enforce that all adjacent regions have opposing boundary normals. In the model Eq. 8 (E_{fast}) only interface normals y^{i*}_s with respect to a particular label are maintained, whereas the tighter formulation Eq. 5 (E_{tight}) explicitly represents transition gradients y^{ij}_s for all label combinations (i,j). Another way to express the difference between the formulations is, that E_{fast} penalizes the length of segmentation boundaries (thereby being agnostic to neighboring labels), and E_{tight} accumulates the length of interfaces between each pair of regions separately (i.e. label transitions have the knowledge of both involved labels, see also Fig. 2(c)). The two models are different (after convexification) when using a Euclidean length measure, but not when using an anisotropic L^1 length measure.

One might ask how graph cuts with larger neighborhoods (geo-cuts [4]) compare with the continuously inspired approaches Eq. 6 and Eq. 8 for the Potts smoothness model. Since in this case geo-cuts will approximate the interface boundary similar to Eq. 8, similar results are expected (which is experimentally confirmed in Fig. 1(f)). In Fig. 1(d) and (e) we illustrate the (beneficial) impact of using a staggered grid discretization (instead of forward differences) for the gradient ∇.

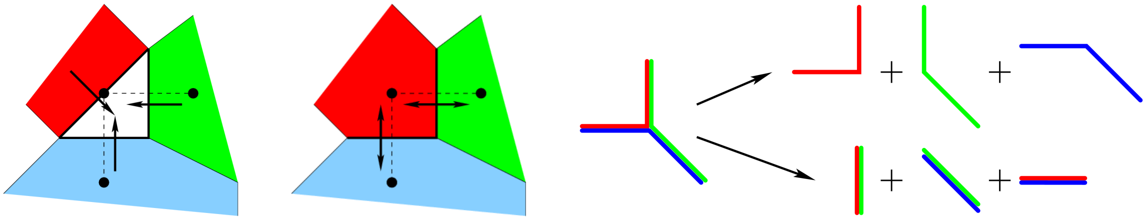
3.3. The Dual View

A standard approach for efficient minimization of MRF energies is to optimize the dual formulation instead of the primal one. Recalling Section 2.2 we observe that the dual energies have a number of unknowns that scales linearly with the number of labels (and nodes), but a quadratic number of terms (recall $E^*_{\text{LP-MRF}}$). Consequently, block coordinate methods for optimizing the dual are very practical, and those methods are often referred as message passing approaches (e.g. [9, 16, 10, 15]). Thus, we consider in this section dual formulations of the tight convex relaxation Eq. 6 and the more efficient, but weaker one Eq. 8.

The dual energy of E_{tight} can be derived (via Fenchel duality) as

$$E^*_{\text{tight-I}}(p) = \sum_s \min_i\{\operatorname{div} p^i_s + \theta^i_s\} \text{ s.t. } \|p^i_s - p^j_s\|_2 \le \theta^{ij}, \tag{10}$$

with the divergence $\operatorname{div} = -\nabla^T$ consistent with the discretization of the gradient. Note that we have redundant constraints on the primal variables $y^{ij}_s \in [-1,1] \times [-1,1]$ (since $x^i_s \in [0,1]$). One could compute the dual of $\theta^{ij}\|y^{ij}_s\|_2 + \imath\{\|y^{ij}_s\|_\infty \le 1\}$, but because of its radial symmetry the constraint $\|y^{ij}_s\|_2 \le \sqrt{2}$ seems to


(a)

(b)

(c)

Figure 2. Three regions meet in one grid point. (a) The situation as handled in E_{fast}. (b) How E_{tight} sees this situation. (c) The different counting of region boundaries. Top row: E_{fast} simply sums the lengths of region boundaries. Bottom row: E_{tight} considers interfaces between each pair of regions separately.

be more appropriate. Via $\left(x \mapsto \theta|x| + \imath_{[0,B]}(x)\right)^* (y) = \max_{x\in[0,B]} \{xy - \theta|x|\} = B\max\{0, |y| - \theta\}$ and the radial symmetry of terms in y_s^{ij} we obtain for the dual energy in this setting

$$E^*_{\text{tight-II}}(p) = \sum_s \min_i \{\operatorname{div} p_s^i + \theta_s^i\} + \sum_s \sum_{i,j:i<j} \sqrt{2}\min\left\{0, \theta^{ij} - \|p_s^i - p_s^j\|_2\right\}, \tag{11}$$

which has the same overall shape as $E^*_{\text{LP-MRF}}$ in Section 2.2. In contrast to Eq. 10 the dual energy Eq. 11 uses an exact penalizer on the constraints and always provides a finite value, which can be useful in some cases (e.g. to compute the primal-dual gap in order to have a well-established stopping criterion when using iterative optimization first-order methods). We finally state a variant of the dual energy,

$$E^*_{\text{tight-III}}(p, q) = \sum_s q_s + \sum_{s,i} \left[\operatorname{div} p_s^i + \theta_s^i - q_s\right]_- + \sum_s \sum_{i,j:i<j} \sqrt{2}\min\left\{0, \theta^{ij} - \|p_s^i - p_s^j\|_2\right\}, \tag{12}$$

Eq. 12 is much easier to smooth than Eq. 10 (which can be smoothed via a numerically more delicate log-barrier) or Eq. 11 (where the exact minimum can be replaced by a soft-minimum, e.g. using log-sum-exp). We discuss appropriate smoothing of Eq. 12 and corresponding optimization in Section 4.

For completeness we also state the dual of the weaker relaxation Eq. 8 in the constrained form:

$$E^*_{\text{fast}}(p) = \sum_s \min_i \{\operatorname{div} p_s^i + \theta_s^i\} \tag{13}$$
$$\text{s.t. } \|p_s^i - p_s^j\|_2 \le \theta^{ij} \qquad \forall s, \forall i,j : |i-j| < T$$
$$\|p_s^i\| \le \theta^*/2 \qquad \forall s, i.$$

In the dual the constraints set in Eq. 13 is a superset of the constraints in the tight relaxation Eq. 10, hence we have $E^*_{\text{fast}} \le E^*_{\text{tight-I}}$ for their respective optimal solutions (recall that the dual energies are maximized with respect to p).

In contrast to LP-MRF formulations we have non-linear capacity constraints in the duals presented above. Thus, optimizing these dual energies (in particular Eq. 10) via block coordinate methods is more difficult, and deriving message passing algorithms appears not promising. In the supplementary material we present the detailed derivations of the dual energies stated above and report additional forms of the dual energy.

3.4. First-Order Optimality Conditions

In order to ensure optimality of a primal-dual pair and to construct e.g. the primal solution from the dual ones, we state the generalized KKT conditions (see e.g. [3], Ch. 3): if we have the primal energy $E(x) = f(x) + g(Ax)$ for convex f and g, and a linear map A, the dual energy is (subject to a qualification constraint) $E^*(z) = -f^*(A^T z) - g^*(-z)$. Further, a primal dual pair (x^*, y^*) is optimal iff $x^* \in \partial f^*(A^T z^*)$ and $Ax^* \in \partial g^*(-z^*)$. For the tight relaxation Eq. 10 these conditions translate to

$$(x^*)_s \in \partial \max_i \{-\operatorname{div}(p^*)_s^i - \theta_s^i\} \qquad \text{and}$$
$$(y^*)_s^{ij} \in \partial\imath\left\{\left\|(p^*)_s^i - (p^*)_s^j\right\|_2 \le \theta^{ij}\right\}.$$

The first condition means, that $-\operatorname{div}(p^*)_s^j - \theta_s^j < \max_i\{-\operatorname{div}(p^*)_s^i - \theta_s^i\}$ for a label j implies $(x^*)_s^j = 0$ (label j is strictly not assigned in the optimal solution at s). The second condition states, that $\|(p^*)_s^i - (p^*)_s^j\|_2 < \theta^{ij}$ implies $(y^*)_s^{ij} = 0$ (there is no transition between label i and j at pixel s). If $\|(p^*)_s^i - (p^*)_s^j\|_2 = \theta^{ij}$ we have $(y^*)_s^{ij} \propto (p^*)_s^i - (p^*)_s^j$. These generalized complementary slackness constraints can be used to set many values in the primal solution to 0. The second part of the KKT conditions, $Ax^* \in \partial g^*(-z^*)$, just implies that the primal solution has to satisfy the normalization and marginalization constraints.

4. Scalable Optimization Methods

The primal (Eqs. 6 and 7) and dual (Eqs. 10 and 11) programs of the tight relaxation are non-smooth convex and concave energies, and therefore any convex optimization method able to handle non-smooth programs is in theory suitable for minimizing these energies. The major complication with the tight convex relaxation is, that it requires either a quadratic number of unknowns per pixel in the primal (in terms of the number of labels) or has a quadratic number of coupled constraints (respectively penalizing terms) in the dual. The nested optimization procedure proposed in [5] is appealing in terms of memory requirements (since only a linear number of unknowns is maintained per pixel, although the inner reprojection step consumes temporarily $O(L^2)$ variables), but as any other nested iterative approach it comes with difficulties determining when to stop the inner iterations. On the other hand, the methods described in [11, 14] have closed form iterations, but require $O(L^2)$ variables. This is also the case if e.g. Douglas-Rachford splitting [8] (see also the recent survey in [7]) is applied either on the primal problem Eq. 6 or on the always finite dual Eq. 11. We propose two methods for efficiently solving the tight relaxation: the first one addresses truncated smoothness costs (Section 3.2) and starts with solving the efficient (but slightly weaker) model Eq. 8. It subsequently identifies potential triple junctions and switches locally to the tight relaxation until convergence. The second proposed method applies a forward-backward splitting-like method on a smoothened version of the dual energy Eq. 11, and gradually reduces the smoothness parameter (and the allowed time step).

4.1. Subsequent Refinement of the Efficient Model

Our first proposed method to solve the tight convex relaxation in an efficient way is based on the intuition given in Section 3.2: the weaker relaxation E_{weak} can only be potentially strengthened where three or more phases meet, i.e. at pixels s such that $y_s^{i*} \neq 0$ for at least three labels i. For these pixels the weaker model underestimates the true smoothness costs and does not guarantee consistency of boundary normals (recall Fig. 2). For a pixel s let $\mathcal{A}_s$ denote the set of labels with $y_s^{i*} \neq 0$, and at potentially problematic triple junctions we have $|\mathcal{A}_s| \geq 3$. The underestimation of the primal smoothness translates to unnecessarily strong restrictions on p_s^i for $i \in \mathcal{A}_s$, i.e. all constraints $\|p_s^i\| \leq \theta^*/2$ are strongly active for $i \in \mathcal{A}_s$ (recall that $y_s^{i*} \neq 0$ is a generalized Lagrange multiplier for $\|p_s^i\| \leq \theta^*/2$). Consequently, replacing the constraints $\|p_s^i\| \leq \theta^*/2$ by the weaker ones of the corresponding tight relaxation $\|p_s^i - p_s^j\| \leq \theta^*$ for all $i \in \mathcal{A}_s$ allows the dual energy to increase. In the primal this means, that for active labels i the indiscriminative transition gradient y_s^{i*} is substituted by explicit transition variables y_s^{ij} (for $j > i$) and y_s^{ji} (for $j < i$). The marginalization constraint of E_{fast} (Eq. 8)

$$\nabla x_s^i = \sum_{j:i-T<j<i} y_s^{ji} - \sum_{j:i<j<i+T} y_s^{ij} - y_s^{i*}$$

is replaced by one in Eq. 6,

$$\nabla x_s^i = \sum_{j<i} y_s^{ji} - \sum_{j>i} y_s^{ij}$$

for active labels $i \in \mathcal{A}_s$. After augmenting the energy for the problematic pixels, a new minimizer is determined. In practice most problematic pixels are fixed after the first augmentation step, but not all, and there is no guarantee (verified by experiments) that a global solution of the tight model Eq. 6 is already reached after just one augmentation. Hence, the augmentation procedure is repeated until no further refinement is necessary. This approach is guaranteed to find a global minimum of the tight relaxation:

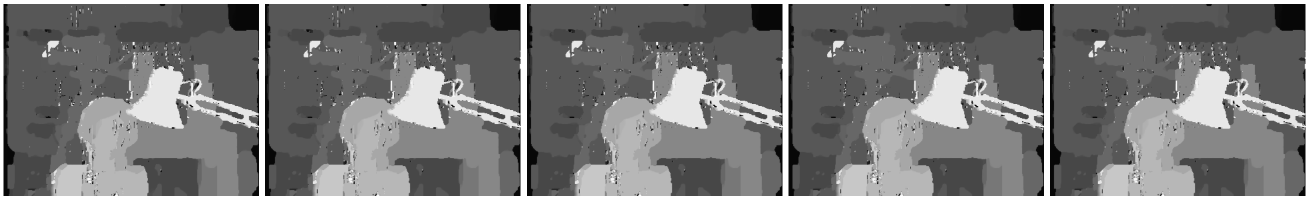
Observation 2 *If for a primal solution (x^*, y^*) of the augmenting procedure the set of active labels $\mathcal{A}_s = \{i : (y^*)_s^{i*} \neq 0\}$ has at most two elements for all pixels $s \in \Omega$ (i.e. at most two different labels meet at "non-augmented" pixels), then x^* is also optimal for E_{tight}.*

This means that all potential triple or higher-order junctions have been addressed by the augmentation steps. The correctness of this observation can be shown by verifying the first-order optimality conditions, i.e. that (x^*, y^*) together with the dual variables p^* forms an optimal primal-dual pair (recall Section 3.4, see the supplementary material for details).

On planar grids at most four regions can meet in a single node (only 3 if ∇ is discretized via one-sided finite differences), one expects the augmentation procedure to terminate with only few pixels being enhanced. In theory, more phases could meet in a single pixel, since we have to allow fractional values for x_s^i. In a few cases (pixels) we observed $\mathcal{A}_s = \{1, \dots, L\}$. In practice only a few augmentation steps are necessary leading to a $\approx 10\%$ increase of memory requirements over the efficient model Eq. 8. We use the primal-dual method [6] for minimization. See Figs. 3(a-c) and 4(a,b) for the intermediate results and energy evolution, respectively. All methods reach relatively fast a solution that is visually similar to the fully converged one, but achieving a significantly small relative duality gap (e.g. $< 0.01\%$) is computationally much more expensive for all methods.

4.2. Smoothing-Based Optimization

Recall that the dual energies of the tight relaxation (Eq. 10 or 11) have only $O(L)$ unknowns per pixel, but a quadratic number of constraints/terms in the objective. In terms of efficient memory use, a purely dual or primal-dual method is desirable. Chambolle et al. [5] utilize a



(a) E_{fast} (b) After 1 augmentation (c) After 2 augmentations (d) Smooth optimization (e) Exact solution E_{tight}

Figure 3. Stereo result using absolute color differences and the Potts discontinuity model. We want to emphasize, that not the quality of the obtained disparity map, but the equivalence between (c), (d) and (e) is of importance.

primal-dual method requiring the projection into the non-trivial feasible set. This projection has no closed form solution and needs to be solved via inner iterations (requiring temporarily $O(L^2)$ variables per pixel). The dual energies, e.g. $E^*_{\text{tight-III}}$ with only penalizer terms (recall Eq. 12), allows to smoothen the dual energy in a numerically robust way. A principled way to smooth non-smooth functions with bounds on the Lipschitz constant of its gradient is presented in [13]: for a non-smooth (convex) function f and a smoothing parameter $\varepsilon > 0$, a smooth version f_ε of f with Lipschitz-continuous gradient (and Lipschitz constant $1/\varepsilon$) is given by $f_\varepsilon = (f^* + \varepsilon\|\cdot\|^2/2)^*$. In order to have convex instead of concave terms, we minimize $-E^*_{\text{tight-III}}$ with respect to p and q,

$$-E^*_{\text{tight-III}}(p,q) = \sum_s -q_s + \sum_{s,i}\left[q_s - \operatorname{div} p^i_s - \theta^i_s\right]_+ + \sum_s \sum_{i,j:i<j} \sqrt{2}\left[\|p^i_s - p^j_s\|_2 - \theta^{ij}\right]_+. \tag{14}$$

The second and third sums are non-smooth. First, the $[\cdot]_+ = \max(0,\cdot)$ expressions in the second sum can be replaced by a soft-maximum function. Especially in the machine learning literature the logistic soft-hinge, $\varepsilon \log\left(1 + e^{x/\varepsilon}\right) \overset{\varepsilon\to 0}{\to} [x]_+$, is often employed, but the exponential and logarithm functions are slow to compute and require special handling for very small ε. Similar to the Huber cost, which is a smooth version of the magnitude function, the smooth version of $[\cdot]_+$ can be easily derived as

$$[x]_{+,\varepsilon} := \begin{cases} 0 & x \le 0 \\ x - \varepsilon/2 & x \ge \varepsilon \\ x^2/2\varepsilon & 0 \le x \le \varepsilon. \end{cases}$$

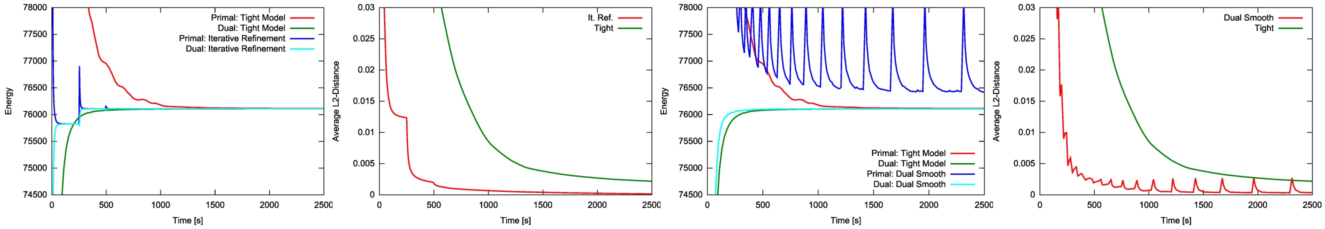
Obtaining a smooth variant of expressions of the shape $h^\theta(z) := \sqrt{2}[\|z\|_2 - \theta]_+$ appearing in the last summation is more involved, but can be shown to be

$$h^\theta_\varepsilon(z) = \begin{cases} 0 & \text{if } \|z\| \le \theta \\ \frac{(\|z\|-\theta)^2}{2\varepsilon} & \text{if } \theta \le \|z\| \le \theta + \sqrt{2}\varepsilon \\ \sqrt{2}(\|z\| - \theta) - \varepsilon & \text{if } \|z\| \ge \theta + \sqrt{2}\varepsilon. \end{cases}$$

We refer to the supplementary material for the derivation. Overall, the smooth energy corresponding to Eq. 14 reads as

$$-E^*_{\text{tight-III},\varepsilon}(p,q) = \sum_s -q_s + \sum_{s,i}\left[q_s - \operatorname{div} p^i_s - \theta^i_s\right]_{+,\varepsilon} + \sum_s \sum_{i,j:i<j} h^{\theta^{ij}}_\varepsilon(p^i_s - p^j_s). \tag{15}$$

By using the chain rule, $\nabla_x f(Ax) = A^T \nabla_y f(y)|_{y=Ax}$, for a differentiable function f and a matrix A, the upper bound of the Lipschitz constant of $\nabla_x f(Ax)$ is given by $L \le \|A\|_2^2 L_f$, where L_f is the Lipschitz constant of ∇f and $\|A\|_2$ is the respective operator norm of A. Consequently, the Lipschitz constant of $\nabla E^*_{\text{tight-III}}$ can be bounded by $5(L+1)/\varepsilon$, since $\|A\|_2 \le 5(L+1)$ for the matrix A mapping (p,q) to their appearances in the respective summands (see the supplementary material for details). Thus, the largest allowed timestep in forward-backward splitting and related accelerated gradient methods is required to be less or equal than $\varepsilon/(5(L+1))$ in order to have convergence guarantees. Note that Eq. 15 is completely smooth and the backward step e.g. in forward-backward splitting is a no-op. We considered and implemented different dual energies leading to a smooth and a non-smooth term in the objective, but none of these appears to be superior to Eq. 15. Hence, in Fig. 4(c) and (d) we report the energy evolution of Eq. 15 using the proximal gradient algorithm proposed in [2], and the Euclidean distance to a converged, ground-truth solution, respectively. Surprisingly, while the dual energy and the distance to the true solution seems to favor the smoothing-based approach over a primal-dual implementation for E_{tight}, the primal energy evolution is clearly inferior. Our conjecture is, that the marginalization constraints in the primal are only slowly satisfied in the smooth formulation. The recurring peaks in the energy and distance graphs Fig. 4(c,d) are due to adjustment of ε in an annealing scheme. A clear advantage of using FISTA for the smooth energy is the trivial implementation on GPUs, where we expect speedups of two orders of magnitude.



(a) E_{tight} vs. E_{fast}+refinement

(b) E_{tight} vs. E_{fast}+refinement

(c) E_{tight} vs. $E_{\text{tight-III},\varepsilon}$

(d) E_{tight} vs. $E_{\text{tight-III},\varepsilon}$

Figure 4. Evolution of the energies and respective Euclidean distances to a converged ground truth solution for the tight model Eq. 5, the refinement strategy (a,b), and FISTA applied on $E_{\text{tight-III},\varepsilon}$ (c,d).

5. Conclusion

In [5] the question is raised, whether there is a simple primal representation of the convex relaxation Eq. 4 for multi-label problems. In this work we are able to give an intuitive answer to that question *at least in the discrete, finite-dimensional setting*. Thus, there is now a clearer understanding what the tight convex formulation optimizes on a discrete image grid, and how to improve the computational efficiency. There are strong links between the local polytope relaxation for MRFs and the ones derived in a continuous and infinite-dimensional setting. We do not know whether it is easy to state the primal program in the continuous domain in a similar way to Eq. 6. For instance, the marginalization constraint in its difference form, $\nabla x^i = \sum_{j<i} y^{ji} - \sum_{j>i} y^{ij}$, would read just as a linear PDE, but there is the complication that x^i_s is not smooth. Analyzing the continuous setting and further extensions of Eq. 6 are subject to future work.[1]

References

[1] G. Alberti, G. Bouchitté, and G. D. Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Calc. Var. Partial Differential Equations*, 16(3):299–333, 2003.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2:183–202, 2009.

[3] J. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2000.

[4] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *Proc. ICCV*, pages 26–33, 2003.

[5] A. Chambolle, D. Cremers, and T. Pock. A convex approach for computing minimal partitions. Technical report, Ecole Polytechnique, 2008.

[6] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems withApplications to Imaging. *Journal of Mathematical Imaging and Vision*, pages 1–26, 2010.

[7] P. L. Combettes and J.-C. Pesquet. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing, pages 185–212. Springer, 2011.

[8] J. Eckstein and D. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming*, 55:293–318, 1992.

[9] A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *NIPS*, 2007.

[10] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1568–1583, 2006.

[11] J. Lellmann, D. Breitenreicher, and C. Schnörr. Fast and exact primal-dual iterations for variational problems in computer vision. In *Proc. ECCV*, 2010.

[12] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.

[13] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Programming*, 103:127–152, 2005.

[14] E. Strekalovskiy, B. Goldluecke, and D. Cremers. Tight convex relaxations for vector-valued labeling problems. In *Proc. ICCV*, 2011.

[15] Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Uncertainty in Artificial Intelligence*, 2007.

[16] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(7), 2007.

[17] C. Zach, D. Gallup, J.-M. Frahm, and M. Niethammer. Fast global labeling for real-time stereo using multiple plane sweeps. In *Vision, Modeling and Visualization Workshop (VMV)*, 2008.

[1] Code is available at `http://www.inf.ethz.ch/personal/chaene/`.

Supported by the 4DVideo ERC Starting Grant Nr. 210806.