What Is Optimized in Convex Relaxations for Multi-Label Problems: Connecting Discrete and Continuously-Inspired MAP Inference

Christopher Zach, Christian Häne, and Marc Pollefeys *Fellow*

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 (MRFs) on grids. It turns out that the continuous methods are non-linear extensions of the well-established 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. Finally, we generalize the underlying energy formulation from isotropic metric smoothness costs to arbitrary non-metric and orientation dependent smoothness terms.

Index Terms—Markov random fields, continuous labeling problems, convex relaxation, approximate inference

1 Introduction

Assigning labels to image pixels or 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) on an underlying pixel grid incorporating local label preference and smoothness in a local neighborhood. Since in general label assignment is NP-hard, finding the true solution is intractable and an approximate one is usually 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 [1], [2], [3], [4] (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 (continuous domain and label space), i.e. variations of the Mumford-Shah segmentation model [5], [6]. Using the principle of biconjugation to obtain tight local convex envelopes, [7], [8] obtains a convex relaxation of multi-label problems with isotropic and metric transition costs *in the continuous setting.* Subsequent discretization of this model to finite grids yields 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.

In summary the strong connection between LP relaxations for MRF inference and continuously inspired formulations has the following implications:

- It is possible to stay close to the well understood framework of LP relaxations for MRFs [3], [9], while at the same time introducing smoothness terms that are less affected by the underlying discrete pixel grid.
- In [7] and related work [10], [11] the objective to optimize is always a saddlepoint energy taking both primal and dual variables as arguments. Since the underlying optimization methods are iterative in their nature, a natural stopping criterion is the duality gap requiring the primal (Section 3.1) and the dual energy (Section 3.3).
- The GPU-accelerated method for real-time label assignment proposed in [12] is extended to truncated smoothness costs, and the connection to other convex relaxations is explored (Section 3.2), and also exploited to obtain a new optimization method (Section 4.1).
- The continuously derived labeling model [7] requires the smoothness cost to be a metric [13] (see also [10] for a discussion of the continuous setting). This is an unnecessary restriction as pointed out in Section 5.1.

- *C. Zach is with Microsoft Research Cambridge, UK. E-mail: chzach@microsoft.com*
- *C. Häne and M Pollefeys are with the Computer Vision and Geometry Group, ETH Zürich*

- Finally, a wider range of optimization methods becomes applicable for the continuously inspired formulations, since convex primal and dual programs can now be clearly stated. The ability to obtain different but equivalent dual programs by utilizing redundant primal constraints enables new options for minimization (Section 4.2).

Thus, the results obtained in this work are of theoretical and practical interest. In this exposition we restrict ourselves to the 2-dimensional setting with image domains being rectangular grids. It is straightforward to extend all energy models and results to higher dimensions. This manuscript is a substantially extended version of [14]. The main theoretical addition to [14] is Section 5 addressing multi-label problems with non-metric smoothness costs (Section 5.1) and general Finsler-type regularizers (Section 5.2). These smoothness regularizers appear in several applications, e.g. preference of piecewise smooth solutions naturally leads to a non-metric truncated quadratic penalizer. Favoring discontinuities in the solution which are aligned with e.g. strong edges in the image leads to smoothness terms discussed in Section 5.2 and visually demonstrated in Section 5.3. We further updated Section 4.2 to include an upper bound on the difference between the non-smooth original and smoothed energy suitable for accelerated optimization.

2 BACKGROUND

In the following section we summarize the necessary background on discrete and continuous relaxations of multi-label problems. We refer to [15], [16] for a concise introduction to convex analysis, and to [3], [9] for an extensive review of Markov random fields and maximum a posteriori (MAP) assignment.

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. The unit (probability) simplex (of appropriate dimension) is denoted by $\Delta \overset{\text{def}}{=} \{x : \sum_i x^i = 1, x^i \geq 0\}$. For an extended real-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ we denote its convex conjugate by $f^*(y) = \max_x x^T y - f(x)$. Finally, for a node (pixel) s on a 2-dimensional grid, we denote its left, right, up and down neighbor with $\mathsf{le}(s)$, $\mathsf{ri}(s)$, $\mathsf{up}(s)$, and $\mathsf{dn}(s)$, respectively.

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 $\Lambda : \mathcal{V} \to \{1, \dots, L\}$ 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 $\theta_{st}^{\cdot\cdot}$ 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

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

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,

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

The last two constraints will repeatedly throughout the manuscript, thus we introduce

$$\mathcal{C} \overset{\text{def}}{=} \left\{ \mathbf{x} : x_s \in \Delta,\, x_{st}^{ij} \geq 0 \;\; \forall (s,t) \in \mathcal{E},\, \forall i,j \right\}. \tag{4}$$

Since we focus on discrete image domains that are regular lattices, the set $\mathcal{E}$ consists of horizontal and vertical edges connecting neighboring pixels. In order to have a more intuitive correspondence between MRFs on discrete grids and continuously inspired formulations on the image plane as explained in Section 3.1, we introduce $x_s^{ij} \overset{\text{def}}{=} (x_{st}^{ij}, x_{sr}^{ij})^T$, where (s,t) is a horizontal edge originating at pixel s, and (s,r) is the respective vertical edge. Analogously, we also group θ_{st}^{ij} and θ_{sr}^{ij} to form θ_s^{ij}. Thus, the specialization of $E_{\text{LP-MRF}}$ for regular pixel grids can be stated as

$$\begin{aligned} E_{\text{Grid-LP-MRF}}(\mathbf{x}) &= \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j} (\theta_s^{ij})^T x_s^{ij} \\ \text{s.t. } x_s^i &= \sum_j x_{s,1}^{ij} \qquad x_s^i = \sum_j x_{\mathsf{le}(s),1}^{ji} \\ x_s^i &= \sum_j x_{s,2}^{ij} \qquad x_s^i = \sum_j x_{\mathsf{up}(s),2}^{ji} \qquad \mathbf{x} \in \mathcal{C} \end{aligned} \tag{5}$$

There are several corresponding dual programs of $E_{\text{LP-MRF}}$ depending on the utilized (redundant) constraints. If we explicitly add the box constraints $x_{st}^{ij} \in [0,1]$ the corresponding dual is

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

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 $x_{st}^{ij} \in [0,1]$ in the primal program leads to an exact penalizer for the usually obtained capacity constraints. If only non-negativity constraints on x_{st}^{ij} are enforced, one obtains more familiar dual programs incorporating capacity constraints (e.g. [3]). Different choices of primal constraints lead to different duals, we refer to Section 3.3 for further details.

Since $E_{\text{LP-MRF}}$ is a convex relaxation dropping integrality constraints, the solution of the relaxed problem may be fractional and therefore reveal little information, how labels should be assigned. Whether the relaxed solution is integral or not depends heavily on the shape of the pairwise potentials θ_s^{ij}. For some classes of pairwise costs it is known that integral minimizers of $E_{\text{LP-MRF}}$ can be expected [17], [18]. In other cases, the relaxations can be tightened by enriching the linear program [19], [20], [21].

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 [7]. 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 relaxation is inspired by the (continuous) analysis of Mumford-Shah like models [6], and is formulated as a primal-dual saddle-point energy

$$\begin{aligned} E_{\text{superlevel}}(\mathbf{u},\mathbf{q}) &= \sum_{s,i} \theta_s^i (u_s^{i+1} - u_s^i) + \sum_{s,i} (q_s^i)^T \nabla u_s^i \\ \text{s.t. } & u_s^i \leq u_s^{i+1},\, u_s^0 = 0,\, u_s^{L+1} = 1,\, u_s^i \geq 0 \\ & \Big\| \sum_{k=i}^{j-1} q_s^k \Big\|_2 \leq \theta^{ij} \qquad \forall s,i,j, \end{aligned} \tag{6}$$

which is minimized with respect to $\mathbf{u}$ and maximized with respect to $\mathbf{q}$. Here $\mathbf{u}$ 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. $\mathbf{q} \in \mathbb{R}^{2VL}$ are auxiliary variables. The stencil of ∇ depends on the utilized discretization, but usually forward differences are employed for ∇ (e.g. in [7], [11]). 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:

Remark 1. The saddle-point formulation in combination with the quadratic number of "capacity" constraints $\|\sum_{k=i}^{j-1} q_s^k\|_2 \leq \theta_{st}^{ij}$ makes it difficult to optimize efficiently. In [7] a nested, two-level iteration scheme is proposed, where the inner iterations are required to enforce the capacity constraints. The inner iterations correspond to Dykstra's projection algorithm [22] requiring temporarily $O(L^2)$ variables per pixel. In [11] 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.

Remark 2. The energy Eq. 6 handles triple junctions (i.e. nodes where at least 3 different phases meet) better than the (more efficient) approach proposed in [12]. Again, by working with the primal formulation one can give a clearer intuition why this is the case (see Section 3.2).

Remark 3. The energy in Eq. 6 can be rewritten in terms of (soft) indicator functions x_s per pixel, leading to the equivalent formulation (see the supplementary material or [11]):

$$\begin{aligned} E_{\text{saddlepoint}}(\mathbf{x},\mathbf{p}) &= \sum_{s,i} \theta_s^i x_s^i + \sum_{s,i} (p_s^i)^T \nabla x_s^i \\ \text{s.t. } & \big\| p_s^i - p_s^j \big\|_2 \leq \theta^{ij},\, x_s \in \Delta \qquad \forall s,i,j, \end{aligned} \tag{7}$$

$\mathbf{x}$ and $\mathbf{p}$ are of the same dimension as $\mathbf{u}$ and $\mathbf{q}$. By introducing "node marginals" x_s^i replacing the superlevel values u_s^i, $E_{\text{saddlepoint}}$ establishes already some connection to the local polytope relaxation for MRFs, $E_{\text{LP-MRF}}$ (Eq. 3), since the terms corresponding to the unary potentials (data costs), $\sum_{s,i} \theta_s^i x_s^i$, are the same in both models. Hence, $E_{\text{saddlepoint}}$ is the starting point for our further investigations in the next sections.

3 Convex Relaxations for Multi-Label MRFs Revisited

In this section we derive the connections between the standard LP relaxation for MRFs, $E_{\text{LP-MRF}}$, and the saddle-point energy $E_{\text{saddlepoint}}$, and further analyze the relation between $E_{\text{saddlepoint}}$, and a weaker, but more efficient relaxation. We will make heavy use of Fenchel duality, $\min_{\mathbf{x}} f(\mathbf{x}) + g(A\mathbf{x}) = \max_{\mathbf{z}} -f^*(A^T\mathbf{z}) - g^*(-\mathbf{z})$, where f and g are convex and l.s.c. functions, and A is a linear operator (matrix for finite dimensional problems). We refer e.g. to [15] 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. 6 and Eq. 7, respectively, were never analyzed from the purely primal viewpoint. Using Fenchel duality one can immediately state the primal form of Eq. 7, which has a more intuitive interpretation (detailed in Section 3.2):

Observation 1. *The primal of the saddlepoint energy $E_{\text{saddlepoint}}$ (Eq. 7) is given by*

$$\begin{aligned} E_{\textit{tight}}(\mathbf{x},\mathbf{y}) &= \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j} \theta^{ij} \|y_s^{ij}\|_2 \\ \textit{s.t. } \nabla x_s^i &= \sum_{j:j<i} y_s^{ji} - \sum_{j:j>i} y_s^{ij},\ x_s \in \Delta \quad \forall s,i, \end{aligned} \tag{8}$$

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} with label gradients ∇x_s^i.*

Proof: Since $E_{\text{saddlepoint}}$ can be written as

$$E_{\text{saddlepoint}}(\mathbf{x}) = \sum_{s,i} \theta_s^i x_s^i + \sum_s \max_{p_s^i} \sum_i (p_s^i)^T \nabla x_s^i \qquad (9)$$
$$\text{s.t. } \left\| p_s^i - p_s^j \right\|_2 \leq \theta^{ij},\ x_s \in \Delta,$$

we only need to consider the point-wise problem

$$\max_{p_s^i} \sum_i (p_s^i)^T \nabla x_s^i \qquad \text{subject to } \left\| p_s^i - p_s^j \right\|_2 \leq \theta^{ij}. \quad (10)$$

We will omit the subscript s and derive the primal of

$$\max_{p^i} \sum_i (p^i)^T \nabla x^i \qquad \text{s.t. } \left\| p^i - p^j \right\|_2 \leq \theta^{ij} \ \ \forall i < j.$$

Fenchel duality leads to the primal

$$\sum_{i,j:i<j} \theta^{ij} \left\| y^{ij} \right\|_2 \qquad \text{subject to } Ay = \nabla x, \qquad (11)$$

since the convex conjugate of $f \equiv \imath\{\|\cdot\|_2 \leq \theta\}$ is $\theta\|\cdot\|_2$, and the conjugate of $g \equiv a^T\cdot$ is $\imath\{\cdot = a\}$. The matrix $-A$ (which has rows corresponding to p^i and columns corresponding to y^{ij}) has a -1 entry at position (p^i, y^{ij}) (for $i < j$) and a +1 element at (p^j, y^{ij}) $(i > j)$. Thus, the i-th row of $-Ay$ reads as

$$\sum_{j:j<i} y^{ji} - \sum_{j:j>i} y^{ij}, \qquad (12)$$

and the purely primal form of Eq. 10 is given by

$$\min_{y_s^{ij}} \sum_{i,j:i<j} \theta^{ij} \left\| y_s^{ij} \right\|_2 \quad \text{s.t.} \quad \nabla x_s^i = \sum_{j:j<i} y_s^{ji} - \sum_{j:j>i} y_s^{ij}.$$

By replacing the inner maximization problem in Eq. 9 with this expression we obtain E_{tight}. □

Because $x_s^i \in [0,1]$ we have that $\nabla x_s^i \in [-1,1]^2$. Since among all solutions y_s^{ij} satisfying the marginalization constraints we search for the ones minimizing the smoothness cost, $\sum \theta^{ij}\|y^{ij}\|$, we can restrict y_s^{ij} to be in $[-1,1]^2$ without changing the set of minimizers. We can interpret the variables y_s^{ij} such 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 component $(y_s^{ij})_2$). Consequently, the y_s^{ij} variables correspond to *signed* pair-wise "pseudo-marginals", and proper pseudo-marginals [9] 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 $x_s^{ii} = (x_s^i, x_s^i)^T - \sum_{j:j\neq i} x_s^{ij}$. Thus, the primal program equivalent to Eq. 8 (using the fact that $\|y\|_2 = \||y|\|_2$ and $|y| = [y]_+ - [y]_-$), but purely stated in terms of non-negative pseudo-marginals, reads as

$$E_{\text{marginals}}(\mathbf{x}) = \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j} \theta^{ij} \left\| x_s^{ij} + x_s^{ji} \right\|_2 \quad (13)$$
$$\text{s.t.} \quad \nabla x_s^i = \sum_{j:j\neq i} x_s^{ji} - \sum_{j:j\neq i} x_s^{ij},$$

and $x_s \in \Delta$, $x_s^{ij} \geq 0$ for $i \neq j$. This is very similar to the standard relaxation of MRFs on regular lattices (recall $E_{\text{Grid-LP-MRF}}$ in Eq. 5, after eliminating x_{st}^{ii} in the marginalization constraints[1]), the only difference being the smoothness terms, which is

$$\theta^{ij} \left\| x_s^{ij} + x_s^{ji} \right\|_2 \ \text{ instead of } \ \theta^{ij} \left(x_{s,1}^{ij} + x_{s,2}^{ij} + x_{s,1}^{ji} + x_{s,2}^{ji} \right).$$

Note that the expression on the right is equivalent to $\theta^{ij}\|x_s^{ij} + x_s^{ji}\|_1$ (the anisotropic L_1 norm), since x_s^{ij} andx_s^{ji} are non-negative vectors. Hence the primal model Eq. 13 can be seen as isotropic extension of the standard model Eq. 5 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 $(x_s^{ij})_2 (x_s^{ji})_2 = 0$. If the complementarity conditions do not hold, the overall objective can be lowered by subtracting the component-wise 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 $\theta^{ij} \left\| \begin{matrix} x_s^{ij} \\ x_s^{ji} \end{matrix} \right\|_2$, since

$$\begin{aligned} \left\| x_s^{ij} + x_s^{ji} \right\|_2 &= \sqrt{(x_{s,1}^{ij} + x_{s,1}^{ji})^2 + (x_{s,2}^{ij} + x_{s,2}^{ji})^2} \\ &= \sqrt{\sum_{k=1,2} (x_{s,k}^{ij})^2 + (x_{s,k}^{ji})^2 + 2\, \underbrace{x_{s,k}^{ij} x_{s,k}^{ji}}_{=0}} \\ &= \left\| (x_{s,1}^{ij}, x_{s,2}^{ij}, x_{s,1}^{ji}, x_{s,2}^{ji})^T \right\| = \left\| \begin{matrix} x^{ij} \\ x^{ji} \end{matrix} \right\|_2 . \end{aligned}$$

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.

We conclude this section by discussing similarities and differences between $E_{\text{LP-MRF}}$ (Eq. 3) $E_{\text{tight}}/E_{\text{marginals}}$ (Eq. 8 and Eq. 13, respectively):

Remark 4. The smoothness terms in E_{tight} (and $E_{\text{marginals}}$) are non-linear, which is in contrast to the pairwise terms in $E_{\text{LP-MRF}}$. Further, depending on the employed discretization for ∇, the smoothness terms in E_{tight} depend on higher order cliques of x_s. If ∇ is discretized via one-sided finite differences, three neighboring nodes, s, $\mathsf{ri}(s)$, $\mathsf{dn}(s)$, contribute to the smoothness cost at node/pixel s. If ∇ is discretized using a staggered grid representation, a local 2×2 pixel grid constitutes the smoothness penalizer. Nevertheless, this is not equivalent to utilizing higher-order cliques in $E_{\text{LP-MRF}}$ to model the smoothness costs, since…

Remark 5. …in the continuously inspired energy E_{tight} one *is* interested in fractional values of x_s^i at label boundaries. This is the reason why continuously inspired approaches are claimed to be less affected by the underlying grid representation (so called "metrication artifacts"). In discrete MRF models one is interested

[1] Note that the non-negativity constraint x_s^{ii} is also dropped, which will be further discussed in Section 5.1

in an unambiguous label assignment at each node, i.e. in integral values for x_s^i. On the other hand, replacing the Euclidean norm in $E_{\text{marginals}}$ with the L^1-norm yields $E_{\text{LP-MRF}}$ (but with slightly different marginalization constraints). Overall, the LP relaxation of the discrete labeling problems and the continuously inspired one share the same underlying motivation. It is possible to strengthen the relaxation $E_{\text{LP-MRF}}$ to return integral solutions (e.g. by better outer bounds of the marginal polytope [23], [19], [20], [21]), but the impact of such tightening on $E_{\text{marginals}}$ (which has a non-linear objective function) is not immediately clear. What is clear is, that precisely relaxing the integrality assumption on the solutions makes continuously inspired formulations less prone to grid artifacts.

Remark 6. The difference between the marginalization constraints of $E_{\text{LP-MRF}}$ and E_{tight} and the implications are discussed in detail in Section 5.1.

Remark 7. The fact, that the objective e.g. in $E_{\text{marginals}}$ is non-linear also implies, that many efficient optimization strategies developed for $E_{\text{LP-MRF}}$ are not applicable. In particular, decomposing the image grid graph in a (small) set of trees and exactly solving MAP inference on trees as a subroutine (e.g. [1], [24]) is not possible. Additionally, message passing methods based on dual coordinate descent [1], [2], [3], [4], [25] are difficult to derive for non-linear smoothness terms. Hence, we use rather generic optimization methods for convex problems to optimize E_{tight} in Section 4.

3.2 Truncated Smoothness Costs

If the transition costs θ^{ij} have no structure, then one has to employ the full representations Eq. 8 or 13. 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,

$$\begin{aligned} E_{\text{truncated}}(\mathbf{x},\mathbf{y}) &= \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j<i+T} \theta^{ij} \|y_s^{ij}\|_2 \\ &\quad + \frac{\theta^*}{2} \sum_s \sum_i \|y_s^{i*}\|_2 \\ \text{s.t. } \nabla x_s^i &= \sum_{j:i-T<j<i} y_s^{ji} - \sum_{j:i<j<i+T} y_s^{ij} - y_s^{i*} \quad (14) \end{aligned}$$

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 $O(L)$:

$$\begin{aligned} E_{\text{trunc-linear}}(\mathbf{x},\mathbf{y}) &= \sum_{s,i} \theta_s^i x_s^i + \sum_{s,i} \|y_s^{i,i+1}\|_2 + \frac{\theta^*}{2} \sum_{s,i} \|y_s^{i*}\|_2 \\ \text{s.t. } \nabla x_s^i &= y_s^{i-1,i} - y_s^{i,i+1} - y_s^{i*}. \quad (15) \end{aligned}$$

These models generalize the formulation proposed in [12] beyond the Potts smoothness cost. For the Potts model it is demonstrated in [7] that Eq. 14 is a weaker relaxation than Eq. 7 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 2. *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. 8 and 14 are equivalent. Further, we have equivalence between $E_{\text{LP-MRF}}$ (Eq. 3) and the following reduced linear program:*

$$\begin{aligned} E_{\textit{red-LP-MRF}} &= \sum_{s,i} \theta_s^i x_s^i + \sum_{(s,t)\in\mathcal{E}} \sum_{i,j:|i-j|<T} \theta^{ij} x_{st}^{ij} \\ &\quad + \sum_{(s,t)\in\mathcal{E}} \frac{\theta^*}{2} \sum_i (x_{st}^{i*} + x_{st}^{*i}) \quad (16) \\ \textit{s.t. } \; x_s^i &= \sum_{j:|i-j|<T} x_{st}^{ij} + x_{st}^{i*}, \;\; x_t^j = \sum_{i:|i-j|<T} x_{st}^{ij} + x_{st}^{*j}. \end{aligned}$$

The proof shows the equivalence by setting up a transportation problem and is given in the supplementary material. 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 we are unaware of previous explicit use of the described reduced construction.

The situation is different in the Euclidean norm setting, such that equivalence does not hold anymore. 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. 14 and Eq. 7 for the discrete label configurations, we find out that for triple junctions the formulation in Eq. 14 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 $y_s^{i*} = (-1,-1)^T$, $y_s^{j*} = (1,0)^T$ and $y_s^{k*} = (0,1)^T$, and the smoothness contribution of s according to Eq. 14 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. 7 are $y_s^{ij} = (-1,0)^T$ and

$y_s^{ik} = (0,-1)^T$, 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. 14 is a weaker model than Eq. 7 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. 14 ($E_{\text{truncated}}$) only interface normals y_s^{i*} with respect to a particular label are maintained, whereas the tighter formulation Eq. 7 (E_{tight}) explicitly represents transition gradients y_s^{ij} for all label combinations (i,j). Another way to express the difference between the formulations is, that $E_{\text{truncated}}$ 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 (recall Obs. 2).

One might ask how graph cuts with larger neighborhoods (geo-cuts [26]) compare with the continuously inspired approaches Eq. 8 and Eq. 14 for the Potts smoothness model. Since in this case geo-cuts will approximate the interface boundary similar to Eq. 14, 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. [2], [3], [1], [4]). Thus, we consider in this section dual formulations of the tight convex relaxation Eq. 8 and the more efficient, but weaker one Eq. 14.

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

$$E^*_{\text{tight-I}}(\mathbf{p}) = \sum_s \min_i \{\operatorname{div} p_s^i + \theta_s^i\} \text{ s.t. } \|p_s^i - p_s^j\|_2 \le \theta^{ij}, \tag{17}$$

with the divergence $\operatorname{div} = -\nabla^T$ consistent with the discretization of the gradient. If ∇ is e.g. computed via forward differences, div is based on backward ones. Note that we have redundant constraints on the primal variables $y_s^{ij} \in [-1,1] \times [-1,1]$ (since $x_s^i \in [0,1]$). One could compute the dual of $\theta^{ij}\|y_s^{ij}\|_2 + \imath\{\|y_s^{ij}\|_\infty \le 1\}$, but because of its radial symmetry the constraint $\|y_s^{ij}\|_2 \le \sqrt{2}$ seems to 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

$$\begin{aligned} E^*_{\text{tight-II}}(\mathbf{p}) &= \sum_s \min_i \{\operatorname{div} p_s^i + \theta_s^i\} \\ &\quad + \sum_s \sum_{i,j:i<j} \sqrt{2}\min\left\{0, \theta^{ij} - \|p_s^i - p_s^j\|_2\right\}, \end{aligned} \tag{18}$$

which has the same overall shape as $E^*_{\text{LP-MRF}}$ in Section 2.2. In contrast to Eq. 17 the dual energy Eq. 18 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 first-order optimization methods). We finally state a variant of the dual energy, which is obtained by explicitly introducing a Lagrange multiplier q_s for the normalization constraints $\sum_i x_s^i = 1$,

$$\begin{aligned} E^*_{\text{tight-III}}(\mathbf{p},\mathbf{q}) &= \sum_s q_s + \sum_{s,i}\left[\operatorname{div} p_s^i + \theta_s^i - q_s\right]_- \\ &\quad + \sum_s \sum_{i,j:i<j} \sqrt{2}\min\left\{0, \theta^{ij} - \|p_s^i - p_s^j\|_2\right\}. \end{aligned} \tag{19}$$

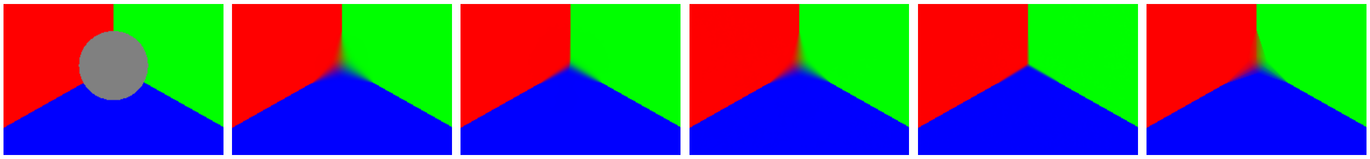
Eq. 19 is much easier to smooth than Eq. 17 (which can be smoothed via a numerically delicate log-barrier) or Eq. 18 (where the exact minimum can be replaced by a soft-minimum, e.g. using log-sum-exp). We discuss appropriate smoothing of Eq. 19 and corresponding optimization in Section 4. Further, since for every optimal $(\mathbf{p}^*, \mathbf{q}^*)$ the objective remains the same for $(\mathbf{p}^* + \mathbf{1}\delta, \mathbf{q}^*)$ for a $\delta \in \mathbb{R}$, $E^*_{\text{tight-III}}$ has at least a one-dimensional space of solutions. In order to remove this degree of freedom in the solution, one can add a constraint on the average value of p_s^i, e.g. $\sum_s \sum_i p_s^i = 0$.

For completeness we also state the dual of the weaker relaxation Eq. 14 in the constrained form:

$$\begin{aligned} E^*_{\text{truncated}}(\mathbf{p}) &= \sum_s \min_i \{\operatorname{div} p_s^i + \theta_s^i\} \\ \text{s.t. } & \|p_s^i - p_s^j\|_2 \le \theta^{ij} \qquad \forall s, \forall i,j : |i-j| < T \\ & \|p_s^i\| \le \theta^*/2 \qquad \forall s, i. \end{aligned} \tag{20}$$

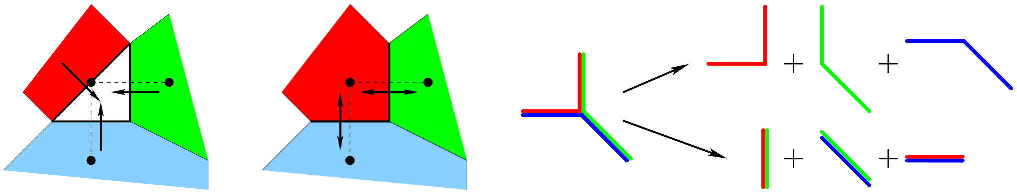
In the dual the constraints set in Eq. 20 is a superset of the constraints in the tight relaxation Eq. 17 (since $\|p_s^i\| \le \theta^*/2$ implies $\|p_s^i - p_s^j\|_2 \le \theta^*$), hence we have $E^*_{\text{truncated}} \le E^*_{\text{tight-I}}$ for their respective optimal solutions (recall that the dual energies are maximized w.r.t. $\mathbf{p}$).

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. 17) 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.



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

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



(a) (b) (c)

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

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 one, we state the generalized KKT conditions (see e.g. [15], Ch. 3): if we have the primal energy $E(\mathbf{x}) = f(\mathbf{x}) + g(A\mathbf{x})$ for convex f and g, and a linear map A, the dual energy is (subject to a qualification constraint) $E^*(\mathbf{z}) = -f^*(A^T\mathbf{z}) - g^*(-\mathbf{z})$. Further, a primal dual pair $(\mathbf{x}^*, \mathbf{z}^*)$ is optimal iff $\mathbf{x}^* \in \partial f^*(A^T\mathbf{z}^*)$ and $A\mathbf{x}^* \in \partial g^*(-\mathbf{z}^*)$. For the tight relaxation Eq. 17 these conditions translate to

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

The first condition means, that $-\operatorname{div}(p^*)^j_s - \theta^j_s < \max_i\{-\operatorname{div}(p^*)^i_s - \theta^i_s\}$ for a label j implies $(x^*)^j_s = 0$ (label j is strictly not assigned in the optimal solution at s). If $-\operatorname{div}(p^*)^j_s - \theta^j_s$ is strictly smaller than the maximum, an infinitesimal change of $-\operatorname{div}(p^*)^j_s$ does not affect the maximal value, hence the j-component of the subdifferential $\partial \max_i\{-\operatorname{div}(p^*)^i_s - \theta^i_s\}$ is 0. The second condition states, that $\|(p^*)^i_s - (p^*)^j_s\|_2 < \theta^{ij}$ implies $(y^*)^{ij}_s = 0$ (there is no transition between label i and j at pixel s). If $\|(p^*)^i_s - (p^*)^j_s\|_2 = \theta^{ij}$ we have $(y^*)^{ij}_s \propto (p^*)^i_s - (p^*)^j_s$. 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, $A\mathbf{x}^* \in \partial g^*(-\mathbf{z}^*)$, just implies that the primal solution has to satisfy the normalization and marginalization constraints.

4 Optimization Methods

The primal (Eqs. 8 and 13) and dual (Eqs. 17 and 18) 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 [7] 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 [27], [11] have closed form iterations, but require $O(L^2)$ variables. This is also the case if e.g. Douglas-Rachford splitting [28] (see also the recent survey in [29]) is applied either on the primal problem Eq. 8 or on the always finite dual Eq. 18. 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. 14. 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. 18, and gradually reduces the smoothness parameter (and the allowed time step).

4.1 Iterated Refinement of the Truncated 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 $E_{\text{truncated}}$ can only be potentially strengthened where three or more phases meet, i.e. at pixels s such that $y^{i*}_s \neq 0$ 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 $j < i$).[2] The marginalization constraint of $E_{\text{truncated}}$ (Eq. 14)

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

is replaced by one in Eq. 8, $\nabla x_s^i = \sum_{j<i} y_s^{ji} - \sum_{j>i} y_s^{ij}$, 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. 8 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 3. *If for a primal solution* $(\mathbf{x}^*, \mathbf{y}^*)$ *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* $\mathbf{x}^*$ *is also optimal for* $E_{\textit{tight}}$.

The full proof is in the supplementary material. In the proof the optimal primal variables $(\mathbf{x}^*, \mathbf{y}^*)$ are extented to a feasible primal solution of E_{tight}, and it is shown that the dual unknowns are still a certificate for optimality.

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 $\mathcal{A}_s = \{1, \ldots, L\}$. In practice only a few augmentation steps are necessary leading to a $\approx 10\%$ increase of memory requirements over the efficient model Eq. 14. We use the primal-dual method [30] 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. 17 or 18) 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. [7] utilize a 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. 19), 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 [31]: 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 $f_\varepsilon = (f^* + \varepsilon\|\cdot\|^2/2)^*$. We employ a quadratic prox-function for smoothing rather e.g. the entropic one utilized in [32], [33] for similar inference problems. It turns out that our smooth energy yields better approximation bounds to the unsmoothed energy than the one used in the aforementioned work (see below).

In order to have convex instead of concave terms, we minimize $-E^*_{\text{tight-III}}$ with respect to $\mathbf{p}$ and $\mathbf{q}$,

$$\begin{aligned} -E^*_{\text{tight-III}}(\mathbf{p}, \mathbf{q}) = \sum_s -q_s + \sum_{s,i} \big[q_s - \operatorname{div} p_s^i - \theta_s^i\big]_+ \\ + \sum_s \sum_{i,j:i<j} \sqrt{2}\big[\|p_s^i - p_s^j\|_2 - \theta^{ij}\big]_+. \end{aligned} \quad (21)$$

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

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

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

$$h_\varepsilon^\theta(z) = \begin{cases} 0 & \text{if } \|z\| \leq \theta \\ \frac{(\|z\| - \theta)^2}{2\varepsilon} & \text{if } \theta \leq \|z\| \leq \theta + \sqrt{2}\varepsilon \\ \sqrt{2}(\|z\| - \theta) - \varepsilon & \text{if } \|z\| \geq \theta + \sqrt{2}\varepsilon. \end{cases} \quad (22)$$

[2] This techniques resembles *column generation* methods to solve large-scale linear programs.

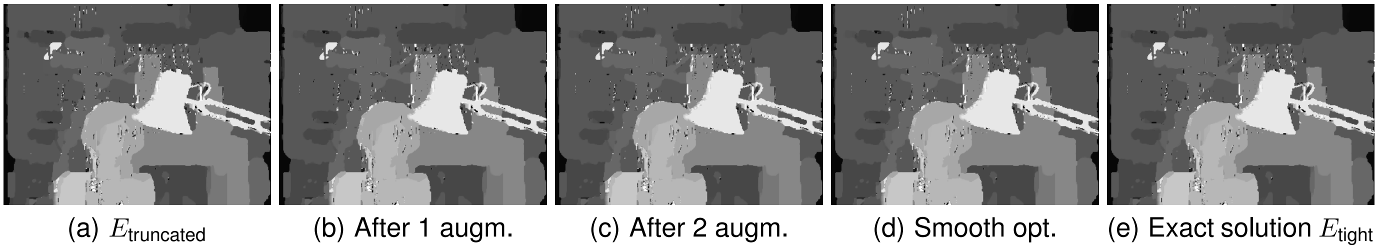

(a) $E_{\text{truncated}}$ (b) After 1 augm. (c) After 2 augm. (d) Smooth opt. (e) Exact solution E_{tight}

Fig. 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.

We refer to the supplementary material for the derivation. Overall, the smooth energy corresponding to Eq. 21 reads as

$$-E^*_{\text{tight-III},\varepsilon}(\mathbf{p},\mathbf{q}) = \sum_s -q_s + \sum_{s,i} \left[q_s - \operatorname{div} p^i_s - \theta^i_s\right]_{+,\varepsilon} + \sum_s \sum_{i,j:i<j} h^{\theta^{ij}}_\varepsilon(p^i_s - p^j_s). \tag{23}$$

By construction (adding a quadratic penalizer in the primal) we always have $E^*_{\text{tight-III},\varepsilon}(\mathbf{p},\mathbf{q}) \geq E^*_{\text{tight-III}}(\mathbf{p},\mathbf{q})$ (or $-E^*_{\text{tight-III},\varepsilon}(\mathbf{p},\mathbf{q}) \leq -E^*_{\text{tight-III}}(\mathbf{p},\mathbf{q})$). We can provide an upper bound on the approximation error:

Observation 4. *For an optimal solution* $(\mathbf{p}^*,\mathbf{q}^*)$ *of* $E^*_{\textit{tight-III},\varepsilon}$ *we have*

$$E^*_{\textit{tight-III},\varepsilon}(\mathbf{p}^*,\mathbf{q}^*) - E^*_{\textit{tight-III}}(\mathbf{p}^*,\mathbf{q}^*) \leq \frac{3\varepsilon|\mathcal{V}|}{2}, \tag{24}$$

where $|\mathcal{V}|$ *is the number of nodes in the underlying graph.*

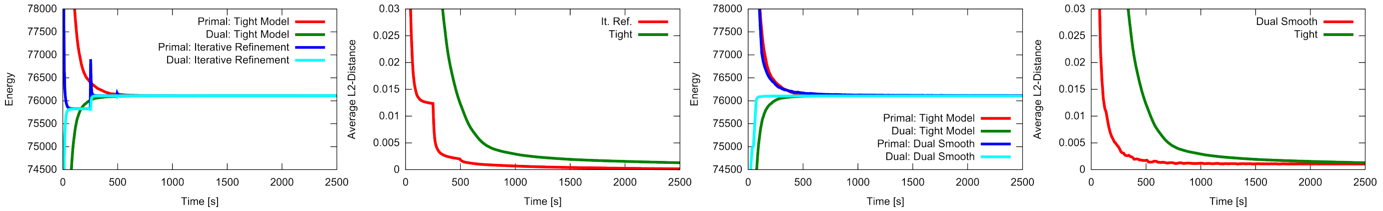
The proof is given in the supplementary material. Note that, in contrast to [32], [33], the upper bound is independent of the number of labels. Given a desired accuracy δ to the optimal non-smooth energy, a necessary smoothing parameter ε is given by $\varepsilon \leq \frac{2\delta}{3|\mathcal{V}|}$. This bound, $3|\mathcal{V}|\varepsilon/2$, is a worst-case bound and often not tight in practice. The proof reveals that the smoother the resulting labeling the closer $E^*_{\text{tight-III},\varepsilon}$ is to $E^*_{\text{tight-III}}$.

In order to apply FISTA, we need an estimate of the respective Lipschitz value: 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 $L \leq \|A\|_2^2 L_f$, 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 $\|A\|_2 \leq 5(L+1)$ for the matrix A mapping $(\mathbf{p},\mathbf{q})$ 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. 23 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. 23. Due to its guaranteed fast convergence of the objective we employ the accelerated proximal gradient method proposed in [34], known as "fast iterated shrinkage thresholding algorithm" or FISTA. In Fig. 4(c) and (d) we report the energy evolution of Eq. 23 and the Euclidean distance to a converged, ground-truth solution, respectively. For a given accuracy δ in the obtained energy, FISTA achieves this accuracy in $O(1/\delta)$ iterations. Unfortunately, the obtained upper bound on the required number of iterations is very loose, due to the large hidden constant (which is also instance-dependent). Hence, we apply a two-stage "annealing" approach, where an approximate dual solution is initially found by setting ε to a relatively large value aiming for a 10% accuracy in the final energy. Since the true optimal energy is not known, we use the best-cost energy ignoring smoothness terms as lower bound for the true optimal energy. After obtaining an initial approximate solution, we soft-restart FISTA with the desired accuracy of the energies. We aim for 0.5% accuracy in the final values between the optimal non-smooth and smooth energies, but the obtained energies are much closer in practice. A clear advantage of using a smoothed energy and a first order optimal method like FISTA is the trivial implementation on GPUs, where we can expect speedups of two orders of magnitude.

Optimizing the isotropic smoothness cost (Euclidean norm) appears to converge much slower than the anisotropic (L^1) term. Fig. 5 illustrates the evolution of the primal-dual gap for the (nonlinear) isotropic formulation of $E_{\text{tight-III},\varepsilon}$ and the otherwise equivalent energy with $\|\cdot\|_2$ replaced by $\|\cdot\|_1$. After a comparable performance to a about 1% relative duality gap, closing the gap is much slower for the isotropic formulation than for the anisotropic one. We hypothesize that in this stage mostly the level curves of the solution are adjusted, but the label assignment itself is unaffected. Nevertheless, reaching a conservative duality gap is much harder for isotropic smoothness terms.

5 Extensions

In this section we describe two extensions for the smoothness terms of the labeling energy Eq. 27. Both are based on the established connection discussed in Section 3.1 between the $E_{\text{LP-MRF}}$ and $E_{\text{marginals}}$. In Section 5.1 the "metrification" of the smoothness costs and



(a) E_{tight} vs. $E_{\text{trunc.+ref.}}$ (b) E_{tight} vs. $E_{\text{trunc.+ref.}}$ (c) E_{tight} vs. $E_{\text{tight-III},\varepsilon}$ (d) E_{tight} vs. $E_{\text{tight-III},\varepsilon}$

Fig. 4. Evolution of the energies and respective Euclidean distances to a converged ground truth solution for the tight model Eq. 7, the refinement strategy (a,b), and FISTA applied on $E_{\text{tight-III},\varepsilon}$ (c,d).

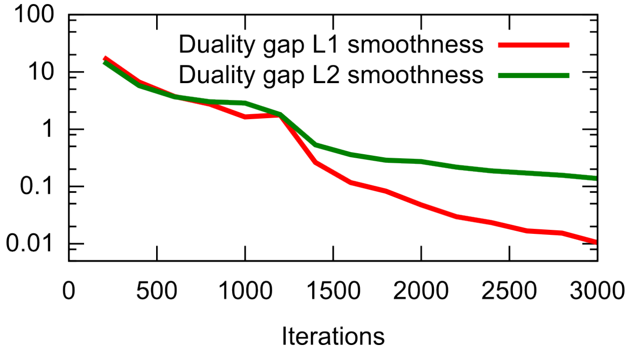


Fig. 5. Evolution of the relative duality gap (in percent) for isotropic and anisotropic regularizers.

its cause is discussed, and in Section 5.2 extensions to more general direction-dependent smoothness terms is provided.

5.1 Non-Metric Smoothness Costs

Besides the non-linearity of the smoothness terms in Eq. 13, the slightly different marginalization constraints appearing in Eq. 3 and Eq. 13, respectively, yield to different behaviors. In the standard local polytope relaxation Eq. 3 the marginalization constraints read as

$$\sum_j x_{st}^{ij} = x_s^i, \qquad \sum_j x_{st}^{ji} = x_t^i, \qquad x_{st}^{ij} \geq 0. \tag{25}$$

One can eliminate x_s^{ii} to arrive at "differential" marginalization constraints,

$$x_t^i - x_s^i = \sum_{j:j\neq i} x_{st}^{ji} - \sum_{j:j\neq i} x_{st}^{ij}, \tag{26}$$

but in Eq. 13 corresponding to the primal of $E_{\text{saddlepoint}}$ *also the non-negativity constraint* $x_{st}^{ii} = (x_s^i, x_s^i)^T - \sum_{j:j\neq i} x_{st}^{ij} \geq 0$ *is dropped.* The lack of the non-negativity constraint on x_{st}^{ii} implies that any non-metric smoothness costs θ_{st}^{ij} is implicitly converted into a metric via the following construction: assume that $\theta_{st}^{ii} + \theta_{st}^{i,i+1} < \theta_{st}^{i,i+2}$. If $x_s^i = 1$ and $x_t^{i+2} = 1$ (i.e. we have a jump from label i to $i+2$ along edge (s,t)), then the desired smoothness cost is $\theta_{st}^{i,i+2}$. By setting $x_{st}^{i,i+1} = x_{st}^{i+1,i+2} = 1$ and $x_{st}^{i+1,i+1} = -1$ the differential marginalization constraints Eq. 26 are still satisfied, but the contribution of edge (s,t) to the smoothness cost is now $\theta_{st}^{ii} + \theta_{st}^{i,i+1} < \theta_{st}^{i,i+2}$. The argument can be generalized to any transition from label i to label k, thus the true smoothness cost is potentially underestimated in all models derived from $E_{\text{saddlepoint}}$ (or $E_{\text{superlevel}}$, recall Eq. 6) for non-metric pairwise costs. Consequently, $E_{\text{saddlepoint}}$ is not suitable to solve labeling problems with non-metric smoothness priors auch as (i) truncated quadratic costs or (ii) inclusion of a "null" or "background" label with constant transition costs to all other "object" or "foreground" labels. Dropping the non-negativity constraints $x_{st}^{ii} \geq 0$ implicitly introduces an order on the labels such that e.g. a jump from label i to $i+2$ is "larger" than one from i to $i+1$. This also means, that permuting label values potentially leads to different values of $E_{\text{saddlepoint}}$, which is not the case for $E_{\text{LP-MRF}}$.

An instructive example is also the following: do not penalize label jumps of height at most one (i.e. $\theta^{ii} = \theta^{i,i+1} = 0$), and use arbitrary but strictly positive smoothness costs otherwise ($\theta^{ij} > 0$ for $|i-j| > 1$). Then the contribution of the smoothness term to the overal objective is always 0 for every solution, since any jump from label i to label $j > i$ can avoid the positive discontinuity cost by setting $x_s^{i,i+1} = x_s^{i+1,i+2} = \cdots = x_s^{j-1,j} = 1$ and $x_s^{i+1,i+1} = x_s^{i+2,i+2} = \cdots = x_s^{j-1,j-1} = -1$ in order to satisfy the (differential) marginalization constraints Eq. 26.

Using the standard marginalization constraints Eq. 25 or, equivalently, adding the constraint $x_s^i - \sum_{j:j\neq i} x_{st}^{ij} \geq 0$ (element-wise) to Eq. 26 resolves the issue. We restate the stronger primal energy on the 2D image grid (corresponding to Eq. 13),

$$\begin{aligned} E_{\text{marginals-II}}(\mathbf{x}) &= \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j} \theta^{ij} \left\| x_s^{ij} + x_s^{ji} \right\|_2 \\ \text{s.t. } \; x_s^i &= \sum_j x_{s,1}^{ij} \qquad x_s^i = \sum_j x_{\text{le}(s),1}^{ji} \\ x_s^i &= \sum_j x_{s,2}^{ij} \qquad x_s^i = \sum_j x_{\text{up}(s),2}^{ji} \qquad \mathbf{x} \in \mathcal{C}. \end{aligned} \tag{27}$$

In contrast to E_{tight} (Eq. 8) and $E_{\text{marginals}}$ (Eq. 13) the objective value is invariant under label permutation: if σ is a permutation in $\{1, \ldots, L\}$, then for any feasible x we have that $x^\sigma \overset{\text{def}}{=} (x_s^{\sigma(i)}, x_s^{\sigma(i),\sigma(j)})$ is also feasible and has the same energy value for permuted costs $\theta^\sigma \overset{\text{def}}{=} (\theta_s^{\sigma(i)}, \theta_s^{\sigma(i),\sigma(j)})$. Hence, optimal solutions are unaffected by the exact mapping between label semantics (defining the unary and pairwise costs) and label indices.

We illustrate the difference between E_{tight} (resp. $E_{\text{marginals}}$) and $E_{\text{marginals-II}}$ for a small stereo instance with the (non-metric) smoothness costs $\theta^{ii} = 0$, $\theta^{i,i+1} = 1$, and $\theta^{ij} = 10$ for $j > i+1$, respectively, in Fig. 6. Observe that E_{tight} is essentially "blind" to the true cost of larger discontinuities, and the result in Fig. 6(a) shows many more abrupt label changes than Fig. 6(b). The particular choice of for the smoothness term strongly penalizes larger discontinuities leading to an overly smooth result, but at the same time makes the difference between Figs. 6(a) and (b) clearly visible. The result in Fig. 6(a) essentially corresponds to a solution with truncated linear smoothness (see Fig. 6(c)) with truncation point $\theta^* = 10$.

In the following we state the dual of $E_{\text{marginals-II}}$ and refer to the supplementary material for its derivation:

$$E^*_{\text{marginals-II}}(\mathbf{p}) = \sum_s \min_i \left\{\theta^i_s + \text{flow}\, p^i_s\right\} \qquad (28)$$

$$\text{s.t.} \quad \left\| \begin{matrix} [p^i_{s\leftarrow} + p^j_{s\rightarrow}]_+ \\ [p^i_{s\uparrow} + p^j_{s\downarrow}]_+ \\ [p^j_{s\leftarrow} + p^i_{s\rightarrow}]_+ \\ [p^j_{s\uparrow} + p^i_{s\downarrow}]_+ \end{matrix} \right\|_2 \le \theta^{ij}, \qquad \begin{aligned} p^i_{s\leftarrow} + p^i_{s\rightarrow} &\le 0 \\ p^i_{s\uparrow} + p^i_{s\downarrow} &\le 0, \end{aligned}$$

where $\text{flow}\, p^i_s \overset{\text{def}}{=} p^i_{s\leftarrow} + p^i_{\text{le}(s)\rightarrow} + p^i_{s\uparrow} + p^i_{\text{up}(s)\downarrow}$.

The difference between $E^*_{\text{marginals-II}}$ and $E^*_{\text{tight-I}}$ (Eq. 17), is that the latter enforces $p^i_{s\leftarrow} + p^i_{s\rightarrow} = 0$ and $p^i_{s\uparrow} + p^i_{s\downarrow} = 0$: in this case one has

$$\left\| \begin{matrix} [p^i_{s\leftarrow} + p^j_{s\rightarrow}]_+ \\ [p^i_{s\uparrow} + p^j_{s\downarrow}]_+ \\ [p^j_{s\leftarrow} + p^i_{s\rightarrow}]_+ \\ [p^j_{s\uparrow} + p^i_{s\downarrow}]_+ \end{matrix} \right\|_2 \le \theta^{ij} \iff \left\| \begin{matrix} [p^i_{s\leftarrow} - p^j_{s\leftarrow}]_+ \\ [p^i_{s\uparrow} - p^j_{s\uparrow}]_+ \\ [p^j_{s\leftarrow} - p^i_{s\leftarrow}]_+ \\ [p^j_{s\uparrow} - p^i_{s\uparrow}]_+ \end{matrix} \right\|_2 \le \theta^{ij}$$

$$\iff \left\| \begin{matrix} p^i_{s\leftarrow} - p^j_{s\leftarrow} \\ p^i_{s\uparrow} - p^j_{s\uparrow} \end{matrix} \right\|_2 \le \theta^{ij},$$

since $\left\|([x]_+, [-x]_+)^T\right\| = \|x\|$. Enforcing equality constraints instead of inequality ones as in $E^*_{\text{marginals-II}}$ implies that $\max_p E^*_{\text{marginals-II}}(\mathbf{p}) \ge \max_p E^*_{\text{tight-I}}(\mathbf{p})$, but equality needs not (and will not) hold in general. We finish this section with a remark:

Remark 8. Instead of using the Euclidean norm, $\|\cdot\|_2$, in $E_{\text{marginals-II}}$ (Eq. 27), one can employ any p-norm ($p \ge 1$) in the smoothness term. If we define $\delta \overset{\text{def}}{=} \min\{x^{ij}_s, x^{ji}_s\} > 0$ (element-wise), we have

$$\sum_k \left(\left(x^{ij}_s + x^{ji}_s\right)_k - 2\delta_k\right)^p \le \sum_k \left(\left(x^{ij}_s + x^{ji}_s\right)_k\right)^p$$

with strict inequality when some $\delta_k > 0$ (due to the strict monotonicity of $(\cdot)^p$). Consequently, if $\delta \ne 0$, we have

$$\left\|x^{ij}_s + x^{ji}_s - 2\delta\right\|_p < \left\|x^{ij}_s + x^{ji}_s\right\|_p.$$

If we assume that $\theta^{ii} = 0$, then every optimal solution of $E_{\text{marginals-II}}$ using the p-norm naturally satisfies the complementarity conditions. Otherwise the overall objective can be reduced by increasing x^{ii}_s and x^{jj}_s by δ and decreasing x^{ij}_s and x^{ji}_s, respectively, in order to satisfy the marginalization constraints. Due to the complementarity of x^{ij}_s and x^{ji}_s, $\theta^{ij}\|x^{ij}_s + x^{ji}_s\|_p$ can be rewritten as $\theta^{ij} \left\| \begin{matrix} x^{ij}_s \\ x^{ji}_s \end{matrix} \right\|_p$, leading to dual constraints of the form

$$\left\| \begin{matrix} [p^i_{s\leftarrow} + p^j_{s\rightarrow}]_+ \\ [p^i_{s\uparrow} + p^j_{s\downarrow}]_+ \\ [p^j_{s\leftarrow} + p^i_{s\rightarrow}]_+ \\ [p^j_{s\uparrow} + p^i_{s\downarrow}]_+ \end{matrix} \right\|_q \le \theta^{ij}$$

with $1/p + 1/q = 1$. This reduces e.g. to the standard LP relaxation on a grid with 4-neighborhoods for $p = 1$.

5.2 Direction-Dependent Smoothness

In some applications it is desirable to penalize region boundaries depending on the location and on the orientation of the discontinuity. In [35] a saddle-point formulation was proposed in order to generalize Eq. 7 beyond isotropic smoothness terms. We start by replacing the isotropic smoothness costs, $\sum_s \sum_{i<j} \theta^{ij}_s \|y^{ij}_s\|_2$, in Eq. 8 with the following term,

$$\sum_s \sum_{i,j:i<j} \phi^{ij}_s(y^{ij}_s),$$

where $\phi^{ij}_s(\cdot)$ is a convex, and positively 1-homogeneous function. Since ϕ^{ij}_s can vary with the pixel and the involved labels, the cost of a label transition can now be modeled depending on the location (pixel), the source and the destination label, and on the attained transition direction. In the dual programs the capacity constraints $\|p^i_s - p^j_s\|_2 \le \theta^{ij}$ are replaced by constraints of the form

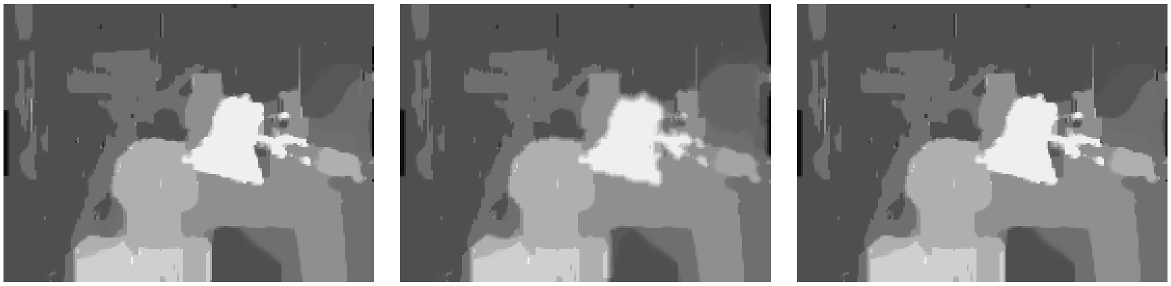
$$p^i_s - p^j_s \in W_{\phi^{ij}_s},$$

where $W_{\phi^{ij}_s}$ is sometimes called the Wulff shape of ϕ^{ij}_s (see e.g. [36], [37], [38]). This follows from the fact that the convex conjugate of a positively 1-homogeneous function is the indicator function of a suitable convex set. The full convex problem in the generalized setting reads as

$$E_{\text{Finsler}}(\mathbf{x}, \mathbf{y}) = \sum_{s,i} \theta^i_s x^i_s + \sum_s \sum_{i,j:i<j} \theta^{ij} \phi^{ij}_s\left(y^{ij}_s\right)$$

$$\text{s.t.}\ \nabla x^i_s = \sum_{j:j<i} y^{ji}_s - \sum_{j:j>i} y^{ij}_s,\ x_s \in \Delta \qquad (29)$$

In view of [39] we call the location and direction dependent regularizer a Finsler metric. As pointed out also in [35] this energy shares the problem of converting non-metric smoothness costs into metric ones with Eq. 8 (which is due to the lack of non-negativity constraints $x^{ii}_s \ge 0$ as explained in the previous section). Unfortunately, in contrast to Section 5.1 we cannot simply introduce non-negative pseudo-marginals x^{ij}_s and replace $\phi^{ij}_s(y^{ij}_s)$ by $\phi^{ij}_s(x^{ij}_s + x^{ji}_s)$, since (among other problems) $x^{ij}_s + x^{ji}_s$ is symmetric. A transition between label i and j in a particular direction will be penalized



(a) E_{tight}, energy ≈ 1843.5 (b) $E_{\text{marginals-II}}$, energy ≈ 1913 (c) Trunc. linear, energy ≈ 1848

Fig. 6. Stereo result using absolute color differences and a non-metric discontinuity model. The visual results and the energy values obtained by minimizing E_{tight} and $E_{\text{marginals-II}}$ are quite significant. The final energy values of E_{tight} is much smaller than the one for $E_{\text{marginals-II}}$ due to dropping the $x_s^{ii} \geq 0$ constraints in the former. (c) depicts the solution obtained by optimizing the corresponding truncated linear smoothness using $E_{\text{marginals-II}}$.

exactly like the opposite jump. Further, the argument to ϕ_s^{ij} is always in the non-negative quadrant, thus the shape of ϕ_s^{ij} outside the positive quadrant is ignored. Surprisingly, substituting $y_s^{ij} = x_s^{ij} - x_s^{ji}$ in Eq. 8 does not weaken the relaxation, and we arrive at the following convex program (after adding standard marginalization constraints as in Section 5.1, or equivalently $x_s^{ii} \geq 0$):

$$E_{\text{Finsler-II}}(\mathbf{x}) = \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j} \phi_s^{ij}\left(x_s^{ij} - x_s^{ji}\right) \quad (30)$$

subject to the same constraints as in $E_{\text{marginals-II}}$ (Eq. 27). In contrast to e.g. Eq. 13 we lose complementarity between x_s^{ij} and x_s^{ji} in optimal solutions. Nevertheless, any minimizer x^* of $E_{\text{Finsler-II}}$ can be converted into a solution $\tilde{x}$ satisfying the complementarity conditions $\tilde{x}_s^{ij} \perp \tilde{x}_s^{ji}$. We set $\delta_s^{ij} \overset{\text{def}}{=} \delta_s^{ji} \overset{\text{def}}{=} \min\{(x^*)_s^{ij}, (x^*)_s^{ij}\}$ (elementwise) for $i < j$, $\tilde{x}_s^{ij} \overset{\text{def}}{=} (x^*)_s^{ij} - \delta_s^{ij}$ for $i \neq j$, and $(x^*)_s^{ii} \overset{\text{def}}{=} (x^*)_s^{ii} + \sum_{j:j\neq i} \delta_s^{ij}$. Node marginals stay the same, $\tilde{x}_s^i \overset{\text{def}}{=} (x^*)_s^i$. Clearly, we have $\tilde{x}_s^{ij} \perp \tilde{x}_s^{ji}$ by construction and the marginalization constraints are still satisfied. Obviosuly, the unary terms are unaffected since $\tilde{x}_s^i = (x^*)_s^i$. Further, the smoothness costs also remain the same, since

$$\tilde{x}_s^{ij} - \tilde{x}_s^{ji} = (x^*)_s^{ij} - \delta_s^{ij} - (x^*)_s^{ji} + \delta_s^{ij} = (x^*)_s^{ij} - (x^*)_s^{ji}.$$

Overall, we constructed a solution $\tilde{x}$ with the same objective value and satisfying the complementarity constraints. The downside of the formulation in Eq. 30 is, that the set of minimizers is enlarged leading to slightly inferior convergence speed of iterative convex optimization methods. At least for Riemann-type smoothness costs $\phi_s^{ij} : \mathbb{R}^2 \to \mathbb{R}_0^+$ induced by quadratic forms one can find a higher-dimensional extension $\Phi : \mathbb{R}^4 \to \mathbb{R}_0^+$ similar to the conversion from $\|x_s^{ij} + x_s^{ji}\|$ to $\|(x_s^{ij}, x_s^{ji})^T\|$ from Section 3.1. We refer to the supplementary material for the construction.

5.3 Numerical Results

General non-metric and direction-dependent smoothness terms are useful to encode preferred boundaries between semantic categories. For instance, the interface between ground and empty space is typically horizontal, and such priors can be encoded using appropriate positively 1-homogeneous penalizers ϕ_s^{ij}. We developed a joint 3D reconstuction and semantic segmentation approach [40] building on the formulation $E_{\text{Finsler-II}}$ from the previous section. However, in this section we continue to use the dense stereo problem and describe a more classic application for location and orientation-dependent smoothness. It is well known that incorporating strong edges in the input image is usually improving the computational stereo result at true depth boundaries. We modulate an underlying smoothness term (which we choose to be a truncated linear or quadratic pairwise costs in our experiments) with a Riemann metric induced by the local edge structure, leading to the following instance of $E_{\text{Finsler-II}}$ (Eq. 30),

$$E_{\text{stereo}}(\mathbf{x}) = \sum_{s,i} \theta_s^i x_s^i + \sum_s \sum_{i,j:i<j} \theta^{ij} \psi_s\left(x_s^{ij} - x_s^{ji}\right) \quad (31)$$

subject to the same constraints as in Eq. 30. Here, $\theta^{ij} = \min\{\theta^*, |i-j|\}$ (truncated linear) or $\theta^{ij} = \min\{\theta^*, (i-j)^2\}$ (truncated quadratic), respectively. Since the number of labels in stereo applications tends to be large, we introduce "wildcard" variables x_s^{i*} and x_s^{*i} (in analogy to Section 3.2), which leads to a slightly weaker but tractable convex problem,

$$\begin{aligned} E_{\text{red-stereo}}(\mathbf{x}) = \sum_{s,i} \theta_s^i x_s^i + \frac{\theta^*}{2} \sum_{s,i} \left(\psi_s\left(x_s^{i*}\right) + \psi_s\left(x_s^{*i}\right)\right) \\ + \sum_s \sum_{i,j:i<j,|i-j|<T} \theta^{ij} \psi_s\left(x_s^{ij} - x_s^{ji}\right) \quad (32) \end{aligned}$$

subject to

$$\begin{aligned} x_s^i &= \sum_{j:|i-j|<T} x_{s,1}^{ij} + x_{s,1}^{i*} & x_s^i &= \sum_{j:|i-j|<T} x_{\mathsf{le}(s),1}^{ji} + x_{\mathsf{le}(s),1}^{*i} \\ x_s^i &= \sum_{j:|i-j|<T} x_{s,2}^{ij} + x_{s,2}^{i*} & x_s^i &= \sum_{j:|i-j|<T} x_{\mathsf{up}(s),2}^{ji} + x_{\mathsf{up}(s),2}^{*i} \end{aligned}$$

and $\mathbf{x} \geq 0$, $\sum_i x_s^i = 1$, for an appropriate value of T. Let $\tilde{I}$ be the smoothed version of the left image I_L (we use a piecewise smooth approximation $\tilde{I}$), and let $\nabla\tilde{I}_s^\perp$ be the orthogonal vector to the image gradient $\nabla\tilde{I}$ at pixel s, then we define

$$\psi_s(y) \overset{\text{def}}{=} \sqrt{y^T D_s y}$$

with

$$D_s \overset{\text{def}}{=} \frac{1}{\|\nabla\tilde{I}_s^\perp\|^2+\mu}\left(\nabla\tilde{I}_s^\perp(\nabla\tilde{I}_s^\perp)^T+\mu I\right). \qquad (33)$$

This particular normalization of the diffusion tensor leads to $\psi_s(y) \leq \|y\|_2$ with its maximum attained for $y \perp \nabla\tilde{I}_s$. $\mu > 0$ is a parameter to guarantee that D_s is strictly positive definite and is set to $1/100$. This choice of D_s makes jumps at strong image edges regardless of the orientation never more expensive than in textureless regions, i.e. strong edges only optionally reduce the smoothness cost. As data term we utilize $\lambda\,\mathrm{BT}(I_L(x), I_R(x-(i,0)^T)$, where BT is the sampling insensitive matching cost from [41]. After rewriting $\sqrt{(y^T D_s y)}$ as $\left\|L_s^T y\right\|_2$, where L_s is the Cholesky factor of D_s, we introduce respective dual variables and use the primal-dual algorithm [30] to determine the minimizer.

Fig. 7 illustrates the influence of a Riemann-type regularizer over a purely isotropic and uniform smoothness term agnostic to image edges. We show results for unmodified ($D_s = I$, Fig. 7(a)) and modulated (D_s as in Eq. 33, Fig. 7(b)) regularizers. The truncated linear smoothness term favors (as expected) piecewise constant solutions, whose discontinuities are better aligned with strong image edges in Fig. 7(b). Similar observations hold for a truncated quadratic pairwise term (which favors piecewise smooth solutions) as displayed in Figs. 7(c) and (d). We set $\lambda = 6$ for the truncated linear cost and $\lambda = 3$ for the truncated quadratic one to obtain visually similar results.

6 Conclusion

In [7] the question is raised, whether there is a simple primal representation of the convex relaxation Eq. 6 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 convex relaxations derived from a continuous setting. Both models can benefit from the established connection: discrete approaches can largely avoid the grid bias intrinsic in grid-based graphs by using isotropic regularizers, and some shortcoming of continuously inspired formulations can be fixed by a better understanding of the relation to discrete approaches for MAP inference.

The starting point for continuous convex relaxations is [6], where a saddle-point energy with a continuous image domain and a continuous label space is discussed. We do not know whether it is easy to state the corresponding primal program in such a continuous setting. Eq. 8 provides the answer in the discretized setting. There seem to be several sources of difficulties, e.g. 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_s^i is not smooth. Analyzing the continuous setting and further extensions of Eq. 8 are subject to future work. [3]

References

[1] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, 2006.

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

[3] T. Werner, "A linear programming approach to max-sum problem: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 7, 2007.

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

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

[6] 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*, vol. 16, no. 3, pp. 299–333, 2003.

[7] A. Chambolle, D. Cremers, and T. Pock, "A convex approach for computing minimal partitions," Ecole Polytechnique, Tech. Rep., 2008.

[8] T. Pock, A. Chambolle, D. Cremers, and H. Bischof, "A convex relaxation approach for computing minimal partitions," in *Proc. CVPR*, 2009.

[9] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, pp. 1–305, 2008.

[10] J. Lellmann and C. Schnörr, "Continuous multiclass labeling approaches and algorithms," Heidelberg University, Tech. Rep., 2010.

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

[12] C. Zach, D. Gallup, J.-M. Frahm, and M. Niethammer, "Fast global labeling for real-time stereo using multiple plane sweeps," in *Proc. VMV*, 2008.

[13] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[14] C. Zach, C. Häne, and M. Pollefeys, "What is optimized in tight convex relaxations for multi-label problems?" in *Proc. CVPR*, 2012.

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

[16] R. T. Rockafellar, *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*. Princeton University Press, Dec. 1996.

[17] H. Ishikawa, "Exact optimization for Markov random fields with convex priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1333–1336, 2003.

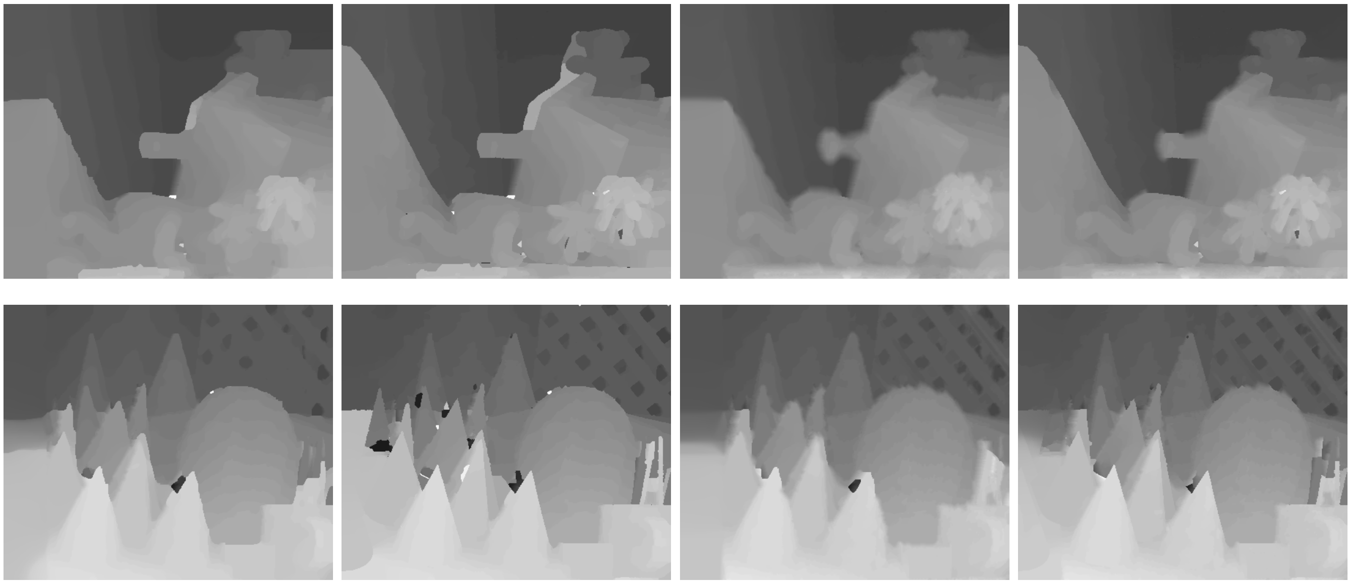
[18] D. Schlesinger, "Exact solution of permuted submodular MinSum problems," in *Proc. Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2007, pp. 28–38.

[19] D. Sontag and T. Jaakkola, "New outer bounds on the marginal polytope," in *NIPS*, 2007.

[20] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss, "Tightening LP relaxations for MAP using message passing," in *Proc. Uncertainty in Artificial Intelligence (UAI)*, 2008.

[21] T. Werner, "Revisiting the linear programming relaxation approach to Gibbs energy minimization and weighted constraint satisfaction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1474–1488, 2010.

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



(a) Plain trunc. linear (b) Riemann trunc. linear (c) Plain trunc. quadr. (d) Riemann trunc. quadr.

Fig. 7. Stereo results for the “teddy” (top row) and “cones” (bottom row) dat sets introduced in [42]. 1st column: uniform truncated linear smoothness. 2nd column: Riemann-metric modulated truncated linear smoothness. 3rd column: uniform truncated quadratic regularizer. 4th column: Riemann-metric modulated truncated quadratic smoothness.

[22] J. P. Boyle and R. L. Dykstra, “A method for finding projections onto the intersection of convex sets in Hilbert spaces,” *Lecture Notes in Statistics*, vol. 37, pp. 28–47, 1986.

[23] M. Wainwright and M. I. Jordan, “Log-determinant relaxation for approximate inference in discrete Markov random fields,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6, p. 20992109, 2006.

[24] N. Komodakis, N. Paragios, and G. Tziritas, “MRF optimization via dual decomposition: Message-passing revisited,” in *Proc. ICCV*, 2007.

[25] T. Hazan and A. Shashua, “Norm-prodcut belief propagtion: Primal-dual message-passing for lp-relaxation and approximate-inference,” *IEEE Trans. on Information Theory*, vol. 56, no. 12, pp. 6294–6316, 2010.

[26] Y. Boykov and V. Kolmogorov, “Computing geodesics and minimal surfaces via graph cuts,” in *Proc. ICCV*, 2003, pp. 26–33.

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

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

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

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

[31] Y. Nesterov, “Smooth minimization of non-smooth functions,” *Math. Programming*, vol. 103, pp. 127–152, 2005.

[32] V. Jojic, S. Gould, and D. Koller, “Accelerated dual decomposition for MAP inference,” in *ICML*, 2010, pp. 503–510.

[33] B. Savchynskyy, J. H. Kappes, S. Schmidt, and C. Schnörr, “A study of Nesterov’s scheme for Lagrangian decomposition and MAP labeling,” in *Proc. CVPR*, 2011, pp. 1817–1823.

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

[35] E. Strekalovskiy and D. Cremers, “Generalized ordering constraints for multilabel optimization,” in *Proc. ICCV*, 2011.

[36] S. Osher and S. Esedoglu, “Decomposition of images by the anisotropic Rudin-Osher-Fatemi model,” *Comm. Pure Appl. Math.*, vol. 57, pp. 1609–1626, 2004.

[37] C. Zach, M. Niethammer, and J.-M. Frahm, “Continuous maximal flows and Wulff shapes: Application to MRFs,” in *Proc. CVPR*, 2009, pp. 1911–1918.

[38] C. Zach, L. Shan, and M. Niethammer, “Globally optimal Finsler active contours,” in *Pattern Recognition*, ser. Lecture Notes in Computer Science, vol. 5748, 2009, pp. 552–561.

[39] J. Melonakos, E. Pichon, S. Angenent, and A. Tannenbaum, “Finsler active contours,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 412–423, 2008.

[40] C. Häne, C. Zach, A. Cohen, R. Angst, and M. Pollefeys, “Joint 3D scene reconstruction and class segmentation,” in *Proc. CVPR*, 2013.

[41] S. Birchfield and C. Tomasi, “A pixel dissimilarity measure that is insensitive to image sampling,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 401–406, 1998.

[42] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” in *Proc. CVPR*, 2003, pp. 195–202.

Christopher Zach (PhD 2007 TU Graz) is currently a post-doctoral researcher at Microsoft Research Cambridge. He was previously a post-doctoral researcher at UNC-Chapel Hill (2008–2009) and a senior researcher at ETH Zürich (2009–2011). His research interests are convex methods in computer vision, structure from motion, dense reconstruction from images, and computer vision on graphics processing units.

Christian Häne received his BSc and MSc in computer science from ETH Zürich in 2010 and 2011, respectively. He is currently a graduate student at ETH Zürich in the Computer Vision and Geometry Group. His research interests include convex methods for dense 3D reconstruction and the application of these methods to challenging scenarios.

Marc Pollefeys (PhD 1999 KU Leuven) is a full professor in the Dept. of Computer Science of ETH Zurich since 2007. His main research interest is the modeling of static and dynamic environments from images. Prof. Pollefeys received a Marr prize, NSF CAREER award, Packard Fellowship and ERC grant. He is the General Chair for ECCV 2014 and was a Program Co-Chair for CVPR 2009. He is on the Editorial Board of IJCV, an associate editor for the IEEE PAMI and is a Fellow of the IEEE