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

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

July 18, 2012

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.

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 [17, 12, 35, 34] (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 [21, 1]. Using the principle of biconjugation to obtain tight local convex envelopes, [7, 24] 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 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 [35, 33], while at the same time introducing smoothness terms that are less affected by the underlying pixel grid orientation.
- In [7] and related work [20, 31] 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 [37] 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 4.2).
- The continuously derived labeling model [7] requires the smoothness cost to be a metric [5] (see also [20] for a discussion of the continuous setting). This is an unnecessary restriction as pointed out in Section 5.1.
- 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 4.3).

Thus, the results obtained in this work are of theoretical and practical interest. This manuscript is a substantially extended version of [38].

2 Background

In the following section we summarize the necessary background on discrete and continuous relaxations of multi-label problems. We refer to [3, 25] for a concise introduction to convex analysis, and to [35, 33] for an extensive review of Markov random field 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 $\mathbb{1}_C$ to denote the corresponding indicator function. i.e. $\mathbb{1}_C(x) = 0$ for $x \in C$ and $\infty$ 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 $\Delta$ def $= \{x : \sum x_i = 1, x_i \geq 0\}$. Finally, for an extended real-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ we denote its convex conjugate by $f^*(y) = \max_x x^Ty - f(x)$.

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

$$E_{\text{labeling}}(\Lambda) = \sum_{s \in \mathcal{V}} \theta^{(s)}_s + \sum_{(s,t) \in \mathcal{E}} \theta^{(s,t)}_{st},$$

(1)

where $\theta_s$ are the unary potentials and $\theta_{st}$ are the pairwise ones. Usually the label assignment $\Lambda$ 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

$$E_{\text{MRF}}(x) = \sum_{s} \theta^i_s x^i_s + \sum_{s,t,i,j} \theta^{ij}_{st} x^i_s x^j_t$$

(2)
subject to normalization constraints \( \sum_{i \in \{1, \ldots, L\}} x^i_s = 1 \) for each \( s \in V \) (one label needs to be assigned) and marginalization constraints \( \sum_t x^i_{st} = x^i_t \) and \( \sum_j x^j_{st} = x^j_t \). In general, enforcing \( x^i_s \in \{0, 1\} \) is NP-hard, hence the corresponding LP-relaxation is considered,

\[
E_{\text{LP-MRF}}(x) = \sum_{s,i} \theta^i_s x^i_s + \sum_{s,t} \sum_{i,j} \theta^j_{st} x^j_{st}
\]

\[
\text{s.t. } \sum_j x^j_{st} = x^i_t, \quad \sum_j x^j_{st} = x^i_t \quad x_s \in \Delta, \quad x^j_{st} \geq 0 \quad \forall s,t,i,j,
\]

There are several corresponding dual programs depending on the utilized (redundant) constraints. If we explicitly add the box constraints \( x^j_{st} \in [-1,1] \) the corresponding dual is

\[
E'_{\text{LP-MRF}}(p) = \sum_s \min \left \{ \theta^i_s + \sum_{t \in N(s)} p^i_{st} + \sum_{t \in N_s(t)} p^i_{st} \right \}
\]

\[
+ \sum_{t,i,j} \min \left \{ \theta^j_{st} - |p^j_{st} + p^j_{st-1}| \right \},
\]

where we defined \( N_t(s) := \{ t : (s, t) \in E \} \) and \( N_s(t) := \{ s : (s, t) \in E \} \). The particular choice of (redundant) box constraints \( x^j_{st} \in [-1,1] \) 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.

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 some classes of pairwise costs it is known that integral minimizers of \( E_{\text{LP-MRF}} \) can be expected [14, 27]. In other cases, the relaxations can be tightened by enriching the linear program [28, 29, 36].

### 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 [1], and is formulated as a primal-dual saddle-point energy

\[
E_{\text{superlevel}}(u, q) = \sum_{s,i} \theta^i_s (u^{i+1}_s - u^i_s) + \sum_{s,i} (q^i_s)^T \nabla u^i_s
\]

\[
\text{s.t. } u^i_s \leq u^{i+1}_s, \quad u^0_s = 0, \quad u^{L+1}_s = 1, \quad u^i_s \geq 0
\]

\[
\| \sum_{k=i}^{j-1} u^k_s \|_2 \leq \theta^{ij} \quad \forall s,i,j,
\]

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 \( u^{i+1}_s = 1 \) and \( u^i_s = 0 \). Consequently, \( u \in [0,1]^{|V|} \) in the discrete setting of a pixel grid.

\( q \in \mathbb{R}^{2|V|} \) are auxiliary variables. The stencil of \( \nabla \) depends on the utilized discretization, but usually forward differences are employed for \( \nabla \) (e.g. in [7, 31]). \( \theta^{ij} \) 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 \( \| \sum_{k=i}^{j-1} q^k_s \|_2 \leq \theta^{ij}_s \) 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 [6] requiring temporarily \( O(L^2) \) variables per pixel. In [31] 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 [37]. Again, by working with the primal formulation one can give a clearer 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 appendix or [31]):

\[
E_{\text{saddlepoint}}(x, p) = \sum_{s, i} \theta^i x^i_s + \sum_{s, i} (p_s^i)^T \nabla x^i_s
\]

s.t. \( \|p_s^i - p_s^j\|_2 \leq \theta^{ij}, x_s \in \Delta \quad \forall s, i, j, \)

\( x \) and \( p \) are of the same dimension as \( u \) and \( q \). By introducing “node marginals” \( x^i_s \) replacing the superlevel values \( u^i_s \), \( E_{\text{saddlepoint}} \) 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), \( \sum_{s, i} \theta^i x^i_s \), 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, \( \min_x f(x) + g(Ax) = \max_y -f^*(A^T y) - g^*(-z) \), 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 [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):

**Observation 1.** The primal of the saddlepoint energy \( E_{\text{saddlepoint}} \) (Eq. 5) is given by

\[
E_{\text{tight}}(x, y) = \sum_{s, i} \theta^i x^i_s + \sum_{s} \sum_{i, j < i} \theta^{ij} \|y^{ij}_s\|_2
\]

s.t. \( \nabla x^i_s = \sum_{j, j < i} y^{ij}_s - \sum_{j, j > i} y^{ji}_s, x_s \in \Delta \quad \forall s, i, \)

where \( y^{ij}_s \in \mathbb{R}^2 \) represents the transition gradient between a region with label \( i \) and the one with label \( j \). \( y^{ij}_s \) 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 \( y^{ij}_s \) and label gradients \( \nabla x^i_s \) and \( \nabla x^j_s \).

**Proof:** Since \( E_{\text{saddlepoint}} \) can be written as

\[
E_{\text{saddlepoint}}(x) = \sum_{s, i} \theta^i x^i_s + \sum_{s} \max_{p^i_s} (p^i_s)^T \nabla x^i_s
\]

s.t. \( \|p^i_s - p^{ij}_s\|_2 \leq \theta^{ij}, x_s \in \Delta, \)

we only need to consider the point-wise problem

\[
\max_{p^i_s} \sum_{i} (p^i_s)^T \nabla x^i_s \quad \text{subject to} \quad \|p^i_s - p^{ij}_s\|_2 \leq \theta^{ij}.
\]

We will omit the subscript \( s \) and derive the primal of

\[
\max_{p^i} \sum_{i} (p^i)^T \nabla x^i \quad \text{subject to} \quad \|p^i - p^j\|_2 \leq \theta^{ij} \quad \forall i < j.
\]
Fenchel duality \((-f^*(-A^T p) - g^*(p) \rightarrow f(y) + g(-Ay))\) leads to the primal
\[
\sum_{i,j<i} \theta_{ij} \left\| y_{ij} \right\|_2 \quad \text{subject to } Ay = \nabla x,
\] (9)
since the convex conjugate of \(f \equiv \| \cdot \|_2 \leq \theta \) is \(\| \cdot \|_2 \), and the conjugate of \(g \equiv a^T \cdot \), is \(\{ \cdot = a \} \). 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
\[
\sum_{j:j<i} y^{ji} - \sum_{j:j>i} y^{ij},
\] (10)
and the purely primal form of Eq. 8 is given by
\[
\min_{y_{ij}} \sum_{i,j<i} \theta_{ij} \left\| y_{ij} \right\|_2
\] s.t. \(\nabla x^s_i = \sum_{j:j<i} y^{ji} - \sum_{j:j>i} y^{ij} \). (11)

By replacing the inner maximization problem in Eq. 7 with this expression we obtain \(E_{\text{tight}}\).

The marginalization constraints in Eq. 6 consisting of 2\(L\) linear constraints per \(s \in V\) have only \(2(L - 1)\) degrees of freedom. Since \(\sum_i x^s_i = 1\) we have \(\sum_i \nabla x^s_i = 0\), and further the sum of all right hand sides, \(\sum_{i,j} y^{ij} - \sum_{j>i} y^{ij}\), is also 0, since the \(y^{ij}\) cancel out exactly.

Because \(x^s_i \in [0, 1]\) we have that \(\nabla x^s_i \in [-1, 1]^2\) and we can safely add the additional constraints \(y^{ij}_s \in [-1, 1]^2\) to obtain an equivalent convex program. We can interpret the variables \(y^{ij}_s\) such that \((y^{ij}_s)_1 = 1\) iff there is a horizontal transition from label \(i\) to label \(j\), and \((y^{ij}_s)_1 = -1\) if the reverse is the case (analogously for the vertical component \(y^{ij}_s)_2\)). Consequently, the \(y^{ij}_s\) variables correspond to signed pair-wise “pseudo-marginals”, and proper pseudo-marginals [33] can be obtained by setting (component-wise)
\[
x^{ij}_s := [y^{ij}_s]^+ \quad \text{and} \quad x^{ji}_s := -[y^{ij}_s]^-
\]
for \(i < j\). \(x^{ij}_s\) is e.g. given by \(x^{ij}_s = (x^i_s, x^j_s)^T - \sum_{j:j\neq i} x^{ij}_s\). Thus, the primal program equivalent to Eq. 6 (using the fact that \(\|y\|_2 = \|\|y\|_2 \) and \(\|y\| = \|y^+ - \|y\|^-\)), but purely stated in terms of non-negative pseudo-marginals, reads as
\[
E_{\text{tight-marginals}}(x) = \sum_{s,i} \theta^s_i x^s_i + \sum_{s,i,j \in V} \theta^s_{ij} \left\| x^s_i + x^s_j \right\|_2
\] s.t. \(\nabla x^s_i = \sum_{j:j\neq i} x^{ij}_s - \sum_{j:j\neq i} x^{ji}_s, x^s_i \in \Delta, x^{ij}_s \geq 0 \forall i \neq j\). (12)

This is very similar to the standard relaxation of MRFs (recall Eq. 3 after eliminating \(x^{ij}_s\) in the marginalization constraints\(^1\)), the only difference being the smoothness terms, which is
\[
\theta^s_{ij} \left\| x^s_i + x^s_j \right\|_2 \quad \text{instead of} \quad \theta^s_{ij} x^{ij}_s + \theta^s_{ji} x^{ji}_s.
\]

Note that \(\theta^s_{ij} x^{ij}_s + \theta^s_{ji} x^{ji}_s\) is equivalent to \(\theta^s_{ij} \left\| x^{ij}_s + x^{ji}_s \right\|_1\) (the anisotropic \(L_1\) norm), since \(x^{ij}_s, x^{ji}_s \geq 0\). Hence the primal model Eq. 12 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 \(x^{ij}_s\): \((x^{ij}_s)^T x^{ji}_s = 0\), i.e. \((x^{ij}_s) (x^{ji}_s)_1 = 0\) and \((x^{ij}_s)_2 (x^{ji}_s)_2 = 0\). 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 \(x^{ij}_s\) and \(x^{ji}_s\) (and therefore satisfying complementarity) without affecting the marginalization constraint. Hence, we can also replace \(\theta^s_{ij} \left\| x^{ij}_s + x^{ji}_s \right\|_2\) in the primal objective by
\[
\theta^s_{ij} \left\| x^{ij}_s \right\|_2^2.
\]
\(^1\)Note that the non-negativity constraint \(x^{ij}_s\) is also dropped, which will be further discussed in Section 5.1
since
\[
\|x_{ij}^s + x_{ji}^s\|_2 = \sqrt{(x_{ij}^s + x_{ji}^s)^2 + (x_{ij}^s + x_{ji}^s)^2} = \sqrt{\left<(x_{ij}^s)^2 + (x_{ji}^s)^2 + 2x_{ij}^s x_{ji}^s, x_{ij}^s + (x_{ji}^s)^2 + 2x_{ij}^s x_{ji}^s\right> = 0} = \|x_{ij}^s\|_2.
\]

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.

We conclude this section by discussing similarities and differences between \(E_{\text{LP-MRF}}\) (Eq. 3) \(E_{\text{tight}}/E_{\text{tight-marginals}}\) (Eq. 6 and Eq. 12, respectively):

1. The smoothness terms in \(E_{\text{tight}}\) (and \(E_{\text{tight-marginals}}\)) are non-linear, which is in contrast to the pairwise terms in \(E_{\text{LP-MRF}}\). Further, depending on the employed discretization for \(\nabla\), the smoothness terms in \(E_{\text{tight}}\) depend on higher order cliques of \(x_s\). If \(\nabla\) is discretized via one-sided finite differences, three neighboring nodes, \(s, s + (1,0)^T, s + (0,1)^T\), contribute to the smoothness cost at node/pixel \(s\). If \(\nabla\) is discretized using a staggered grid representation, a local \(2 \times 2\) 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

2. in the continuously inspired energy \(E_{\text{tight}}\) one is interested in fractional values of \(x_i^s\) at label boundaries. This is the reason why continuously inspired approaches are claimed to be less afflicted by the underlying grid representation (so called “metrication artifacts”). In discrete MRF models one is interested in an unambiguous label assignment at each node, i.e. in integral values for \(x_i^s\). On the other hand, replacing the Euclidean norm in \(E_{\text{tight-marginals}}\) 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. If \(E_{\text{tight-marginals}}\) is tightened to return integral solutions (e.g. by better outer bounds of the marginal polytope \([32, 28, 29, 36]\)), one would obtain a higher-order discrete model. Hence, precisely relaxing the integrality assumption on the solutions makes continuously inspired formulations less prone to grid artifacts.

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

4. The fact, that the objective e.g. in \(E_{\text{tight-marginals}}\) 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. \([17, 18]\)) is not possible. Additionally, message passing methods based on dual coordinate descent \([17, 12, 35, 34, 13]\) are difficult to derive for non-linear smoothness terms. Hence, we use rather generic optimization methods for convex problems to optimize \(E_{\text{tight}}\) in Section 4.

3.2 Truncated Smoothness Costs

If the transition costs \(\theta_{ij}\) have no structure, then one has to employ the full representations Eq. 6 or 12. 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 \(y_{ij}^s\), where the star * indicates a wild-card symbol, i.e.
\[
y_{ij}^s = \sum_{j:|j-i| \geq T} y_{ij}^j - \sum_{j:|j-i| \geq T} y_{ji}^j.
\]
Thus, we can formulate a primal program using at most $O(TL)$ unknowns per pixel,

$$E_{\text{truncated}}(x, y) = \sum_{s,i} \theta^s_i x^i_s + \sum_s \sum_{i,j,i<j<T} \theta^{ij} ||y^{i,j}_s||_2 + \frac{\theta^*}{2} \sum_s \sum_i ||y^*_i||_2$$

$$\text{s.t. } \nabla x^i_s = \sum_{j:i<j<j+T} y^{i,i+1,j}_s - \sum_{j:i<j<j+T} y^{i,j}_s - y^*_i$$

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

$$E_{\text{truncated-linear}}(x, y) = \sum_{s,i} \theta^s_i x^i_s + \sum_s \sum_{i,j} ||y^{i,j}_s||_2 + \frac{\theta^*}{2} \sum_s \sum_i ||y^*_i||_2$$

$$\text{s.t. } \nabla x^i_s = y^{i-1,i}_s - y^{i,i+1}_s - y^*_i.$$

These models generalize the formulation proposed in [37] beyond the Potts smoothness cost. For the Potts model it is demonstrated in [7] that Eq. 13 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 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. 6 and 13 are equivalent. Further, for truncated smoothness costs $E_{\text{LP-MRF}}$ (Eq. 3) and the following reduced linear program,

$$E_{\text{reduced-LP-MRF}} = \sum_{s,i} \theta^s_i x^i_s + \sum_{(s,t) \in E} \left( \sum_{i,j:|i-j|<T} \theta^{ij} x^{i,j}_{st} + \frac{\theta^*}{2} \sum_i (x^{i,s}_st + x^{s,i}_st) \right)$$

$$\text{s.t. } x^i_s = \sum_{i,j:|i-j|<T} x^{i,j}_{st} + x^{i*}_st, x^i_s = \sum_{i,j:|i-j|<T} x^{i,j}_{st} + x^{s,i}_st,$$

are equivalent.

The proof shows the equivalence by setting up a transportation problem and is given in the appendix. 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. 13 and Eq. 5 for the discrete label configurations, we find out that for triple junctions the formulation in Eq. 13 underestimates the true cost: if label $i$ is assigned to a pixel $s$ and compare the smoothness costs assigned by Eq. 13 and Eq. 5 for the discrete label configurations, then we have $y^{k*}_s = (-1, -1)^T$, $y^{i*}_s = (1, 0)^T$ and $y^{k*}_s = (0, 1)^T$, and the smoothness contribution of $s$ according to Eq. 13 is

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

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

$$||-1||_2 + ||0||_2 = 2$$

(cf. Fig. 2(b)). It seems that Eq. 13 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. 13 \((E_{\text{truncated}})\) only interface normals \(y^i_s\) with respect to a particular label are maintained, whereas the tighter formulation Eq. 5 \((E_{\text{tight}})\) explicitly represents transition gradients \(y^i_j\) 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 \(E_{\text{tight}}\) 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 [4]) compare with the continuously inspired approaches Eq. 6 and Eq. 13 for the Potts smoothness model. Since in this case geo-cuts will approximate the interface boundary similar to Eq. 13, 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 \(\nabla\).

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. [12, 35, 17, 34]). Thus, we consider in this section dual formulations of the tight convex relaxation Eq. 6 and the more efficient, but weaker one Eq. 13.

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

\[
E^*_{\text{tight}}(p) = \sum_i \min_i \{\text{div} p^i_s + \theta^i_s\} \quad \text{s.t.} \quad \|p^i_s - p^j_s\|_2 \leq \theta^{ij},
\]

with the divergence \(\text{div} = -\nabla^T\) consistent with the discretization of the gradient. Note that we have redundant constraints on the primal variables \(y^i_j \in [-1, 1] \times [-1, 1]\) (since \(x^i_j \in [0, 1]\)). One could compute the dual of \(\theta^{ij}\|y^i_j\|_2 + \tau\{\|y^i_j\|_\infty \leq 1\}\), but because of its radial symmetry the constraint
\[ \|g^i_s\|_2 \leq \sqrt{2} \] seems to be more appropriate. Via \( (x \mapsto \theta |x| + v_{[0,1]}(x))^* (y) = \max_{x \in [0,1]} \{ xy - \theta |x| \} = B \max \{ 0, |y| - \theta \} \) and the radial symmetry of terms in \( g^i_s \) we obtain for the dual energy in this setting
\[
E^*_{\text{tight-I}}(p) = \sum_s \min_i \{ \text{div} p^i_s + \theta^i_s \} + \sum_s \sum_{i,j<i} \sqrt{2} \min \{ 0, \theta^{ij} - \| p^i_s - p^j_s \|_2 \},
\] (17)
which has the same overall shape as \( E^*_\text{LP-MRF} \) in Section 2.2. In contrast to Eq. 16 the dual energy Eq. 17 uses an exact penallizer 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, which is obtained by explicitly introducing a Lagrange multiplier \( q_s \) for the normalization constraints \( \sum_i x^i_s = 1, \)
\[
E^*_{\text{tight-II}}(p, q) = \sum_s q_s + \sum_s \sum_i \text{div} p^i_s + \theta^i_s - q_s \] + \sum_s \sum_{i,j<i} \sqrt{2} \min \{ 0, \theta^{ij} - \| p^i_s - p^j_s \|_2 \},
\] (18)
Eq. 18 is much easier to smooth than Eq. 16 (which can be smoothed via a numerically delicate log-barrier) or Eq. 17 (where the exact minimum can be replaced by a soft-minimum, e.g. using log-sum-exp). We discuss appropriate smoothing of Eq. 18 and corresponding optimization in Section 4. Further, since for every optimal \( (p^*, q^*) \) the objective remains the same for \( (p^* + \delta \lambda, q^*) \) for a \( \delta \in \mathbb{R} \), \( E^*_{\text{tight-II}} \) 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 \( p^i_s \), e.g. \( \sum_s \sum_i p^i_s \) = 0.

For completeness we also state the dual of the weaker relaxation Eq. 13 in the constrained form:
\[
E^*_{\text{truncated}}(p) = \sum_s \min_i \{ \text{div} p^i_s + \theta^i_s \} \tag{19}
\]
\[
\text{s.t.} \quad \| p^i_s - p^j_s \|_2 \leq \theta^{ij} \quad \forall s, i, j \quad \| p^i_s \| \leq \theta^i / 2 \quad \forall s, i.
\]
In the dual the constraints set in Eq. 19 is a superset of the constraints in the tight relaxation Eq. 16 (since \( \| p^i_s \| \leq \theta^i / 2 \) implies \( \| p^i_s - p^j_s \|_2 \leq \theta^{ij} \)), hence we have \( E^*_{\text{truncated}} \leq E^*_{\text{tight-I}} \) 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. 16) via block coordinate methods is more difficult, and deriving message passing algorithms appears not promising. In the appendix 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. 16 these conditions translate to
\[
(x^*)_s \in \partial \max_i \{ -\text{div}(p^i_s)^* - \theta^i_s \} \quad \text{and} \quad (y^*)_s \in \partial \{ \| (p^i_s)^* - (p^j_s)^* \|_2 \leq \theta^{ij} \}.
\]
The first condition means, that \( -\text{div}(p^i_s)^* - \theta^i_s < \max \{ -\text{div}(p^j_s)^* - \theta^j_s \} \) 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 \( \| (p^i_s)^* - (p^j_s)^* \|_2 \leq \theta^{ij} \) implies \( (y^*)_s^j = 0 \) (there is no transition between label \( i \) and \( j \) at pixel \( s \)). If \( \| (p^i_s)^* - (p^j_s)^* \|_2 = \theta^{ij} \) we have \((y^*)_s^j \propto (p^i_s)^* - (p^j_s)^*\). These generalized complementarity 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 Optimization Methods

The primal (Eqs. 6 and 12) and dual (Eqs. 16 and 17) 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 [19, 31] have closed form iterations, but require $O(L^2)$ variables. This is also the case if e.g. Douglas-Rachford splitting [11] (see also the recent survey in [9]) is applied either on the primal problem Eq. 6 or on the always finite dual Eq. 17. 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. 13. 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 smoothed version of the dual energy Eq. 17, and gradually reduces the smoothness parameter (and the allowed time step).

4.1 A Baseline Method: SDMM in the Dual

We use the following variation of $E^*_\text{tight-I}$ (Eq. 18) as the objective to be optimized via the simultaneous-direction method of multipliers (SDMM, e.g. [9]), where the penalizers are replaced by constraints (i.e. the redundant bound constraints $\|x_{ij}\| \leq \sqrt{2}$ are neglected):

$$-E^*_\text{tight-IV}(p, q) = -\sum_s q_s + \sum_{s,i} \left[ q_s - \text{div} p_s^i - \theta_s^i \right]_+ + \sum_s \sum_{i,j<i} t \left[ \|p_s^i - p_s^j\|_2 \leq \theta_s^{ij} \right].$$

(20)

We state the convex negated energy, which is minimized with respect to $p$ and $q$, since it directly fits into the SDMM framework for convex minimization. The SDMM algorithm to minimize $-E^*_\text{tight-IV}$ can now be stated as outlined in Algorithm 1.

Fix $\gamma > 0$, initialize $Q_s^i$, $P_s^i$, $\bar{P}_s^i$, $\hat{P}_s^i$, $\hat{y}_s^i$, $\hat{Z}_s^i$,...

for $n = 1, \ldots$
do

for $s \in V, i = 1, \ldots, L$ do

$q_s \leftarrow \frac{1}{\gamma} \left( \gamma + \sum_i (Q_s^i - \hat{y}_s^i) \right)$

$v_1 \leftarrow \left( P_{s+(1,0)}^i - \bar{Z}_{s+(1,0)}^i \right)_1, v_2 \leftarrow \left( P_{s+(0,1)}^i - \bar{Z}_{s+(0,1)}^i \right)_2$

$p_s^i \leftarrow \frac{1}{L+1} \left( P_s^i - Z_s^i + v + \sum_{j:j \neq i} (P_{s}^{ij} - \hat{y}_s^{ij}) \right)$

for $s \in V, i = 1, \ldots, L$ do

$v_1 \leftarrow \left( P_{s+(1,0)}^i \right)_1, v_2 \leftarrow \left( P_{s-(0,1)}^i \right)_2$

$(Q_s^i, P_s^i, \bar{P}_s^i) \leftarrow \text{prox}_{\gamma f_i} (q_s + \hat{x}_s^i, p_s^i + \bar{Z}_s^i)$

$\hat{x}_s^i \leftarrow \hat{x}_s^i + q_s - Q_s^i$

$\bar{Z}_s^i \leftarrow \bar{Z}_s^i + p_s^i - \bar{P}_s^i$

for $s \in V, i = 1, \ldots, L$ do

for $i = 1, \ldots, L$ do

$(P_s^i, \hat{y}_s^i) \leftarrow \text{prox}_{\gamma \theta_i} (p_s^i + \hat{y}_s^i, P_s^i + \hat{y}_s^i)$

$\hat{y}_s^{ij} \leftarrow \hat{y}_s^{ij} + P_s^{ij}$

$\hat{y}_s^{ji} \leftarrow \hat{y}_s^{ji} + P_s^{ji}$

end

end

Algorithm 1: Dual SDMM
Since SDMM is a particular instance of Douglas-Rachford splitting, which is itself a proximal method, the algorithm uses proximity operations. These proximity steps generalize the projection onto convex sets and are generally defined for a proper convex l.s.c. function $f$,

$$\text{prox}_f(x) \overset{\text{def}}{=} \arg \min_x \frac{1}{2} \|x - z\|^2 + f(x).$$

It is known that the minimizer exists and is unique, hence $\text{prox}_f$ is a proper function. The two utilized proximity steps in Algorithm 1, $\text{prox}_{s_i}$ and $\text{prox}_{s_i^j}$, can be derived by case analysis as

$$\text{prox}_{s_i}(q, p, \tilde{p}) = \begin{cases} (q - \gamma, p + \theta \mathbf{1}, \tilde{p} - \gamma \mathbf{1}) & \text{if } r \overset{\text{def}}{=} q - 1T p + 1T \tilde{p} - \theta \mathbf{1} \leq 5\gamma \\ (q - \frac{\gamma}{\theta}, p + \frac{\gamma}{\theta} \mathbf{1}, \tilde{p} - \frac{\gamma}{\theta} \mathbf{1}) & \text{otherwise}, \end{cases}$$

$$\text{prox}_{s_i^j}(p^i, p^j) = \begin{cases} (p^i + \lambda(p^i - p^j), p^j - \lambda(p^i - p^j)) & \text{if } \|p^i - p^j\|_2 > \theta \overset{\text{def}}{=} \lambda \overset{\text{def}}{=} \frac{\gamma}{\gamma - \gamma} \\ (p^i, p^j) & \text{otherwise}. \end{cases}$$

The auxiliary variables $\hat{x}_a$ and $\hat{y}_a^j$ are not named arbitrarily, since it turns out, that $x_a \overset{\text{def}}{=} \hat{x}_a / \gamma$ and $y_a^j \overset{\text{def}}{=} \hat{y}_a^j / \gamma$ are the primal variables appearing in $E_{\text{tight}}$. Further, if $\hat{y}_a^j$ and $\hat{y}_a^j$ are initialized such that $\hat{y}_a^j = -\hat{y}_a^j$, then this holds also during the iterations, hence only half of the $\hat{y}_a^j$ needs to be stored. This is easy to see, since $\text{prox}_{s_i^j}$ moves $p^i$ and $p^j$ exactly in opposite direction, i.e. $\text{prox}_{s_i^j}(p^i, p^j) = (p^i + v, p^i - v)$ for some vector $v$. Hence, we have updates $\hat{y}_a^j \leftarrow \hat{y}_a^j - v$ and $\hat{y}_a^j \leftarrow \hat{y}_a^j + v$ in the algorithm, leaving the invariant $\hat{y}_a^j = -\hat{y}_a^j$. Finally, by projecting $p^i$ and $p^j$ directly to satisfy the constraint $\|p^i - p^j\|_2 \leq \theta^j$, we avoid introducing additional variables to represent $p^i - p^j$ as e.g. done in [31]. In order to simplify the presentation we omitted handling of boundary pixels in Algorithm 1, and we refer to the publicly available code instead.

SDMM in the dual works best for problems with a relatively small number of labels, since the $O(L^2)$ variables must be still maintained per pixel. Nevertheless, it appears to work superior e.g. to SDMM for the primal energy.

### 4.2 Subsequent Refinement of the Efficient 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 $y^*_s \neq 0$ 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 $A_s$ denote the set of labels with $y^*_s \neq 0$, and at potentially problematic triple junctions we have $|A_s| \geq 3$. The underestimation of the primal smoothness translates to unnecessarily strong restrictions on $p^i_s$ for $i \in A_s$, i.e. all constraints $\|p^i_s\|_2 \leq \theta^i / 2$ are strongly active for $i \in A_s$ (recall that $y^*_s \neq 0$ is a generalized Lagrange multiplier for $\|p^i_s\|_2 \leq \theta^i / 2$). Consequently, replacing the constraints $\|p^i_s\|_2 \leq \theta^i / 2$ by the weaker ones of the corresponding tight relaxation $\|p^i_s - p^\gamma_s\| \leq \theta^i$ for all $i \in A_s$ allows the dual energy to increase. In the primal this means, that for active labels $i$ the indiscriminative transition gradient $y^*_s$ is substituted by explicit transition variables $y_s^j$ (for $j > i$) and $y_s^i$ (for $j < i$). The marginalization constraint of $E_{\text{truncated}}$ (Eq. 13)

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

is replaced by one in Eq. 6,

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

for active labels $i \in A_s$. 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 3. If for a primal solution \((x^*, y^*)\) of the augmenting procedure the set of active labels \(A_s = \{i : (y^*)_{i}^s \neq 0\}\) 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 \(E_{\text{tight}}\).

Proof: We show that if we are given an optimal primal/dual solution pair generated by the refinement procedure satisfying the assumption stated in the observation, a primal-dual pair of optimality certificates can be constructed for the tight model, \(E_{\text{tight}}\).

Note that the only difference between the dual of the tight model (recall Eq. 16),

\[
E_{\text{tight},1}(p) = \sum_s \min_i \{\text{div} \, p_s^i + \theta_s^i\} \quad \text{s.t.} \quad \|p_s^i - p_s^j\|_2^2 \leq \theta_{ij}^s,
\]

and the weaker model for truncated costs (Eq. 19),

\[
E_{\text{truncated}}^*(p) = \sum_s \min_i \{\text{div} \, p_s^i + \theta_s^i\}
\]

\[
\text{s.t.} \quad \|p_s^i - p_s^j\|_2^2 \leq \theta_{ij}^s
\]

\[
\|p_s^i\| \leq \theta_s^* \quad \forall s, \forall i, j : |i - j| < T
\]

\[
\text{s.t.} \quad \|p_s^i\| \leq \theta_s^* \quad \forall s, i,
\]

is the set of constraints. We assume that \(\theta_{ij}^s = \theta_s^*\) for \(|i - j| > T\) in Eq. 16 and that \(\theta_s^* \geq \theta_{ij}^s\), since we consider truncated smoothness cost. Consequently we have that the constraints in Eq. 19 are a superset of those in Eq. 16, due to \(\|p_s^i\| \leq \theta_s^* / 2\) implies \(\|p_s^i - p_s^j\| \leq \theta_s^*\). The essential fact to prove observation 2 is, that if only two phase transitions are active, i.e. \(y_{i1}^s \neq 0\) and \(y_{i2}^s \neq 0\) for some \(i_1\) and \(i_2\), it must hold that \(y_{i1}^s = -y_{i2}^s\) (the boundary normal of the entering phase must be opposite to the one of the leaving phase). This can be easily seen and is intuitive for the Potts smoothness cost. Extending that fact to general truncated smoothness priors can be seen as follows:

\[
0 = \nabla \sum_i x_s^i = \sum_i \nabla x_s^i = \sum_i \left( \sum_{j:T<j<i} y_s^{ij} - \sum_{j:i<j<T} y_s^{ji} - y_s^{i1} + y_s^{i2} \right)
\]

\[
= \sum_{i,j:i<j<T} y_s^{ij} - \sum_{i,j:T<j<i} y_s^{ij} - y_s^{i1} + y_s^{i2}
\]

\[
= \sum_{i,j:i<j<T} y_s^{ij} - y_s^{i1} + y_s^{i2}
\]

\[
= -y_s^{i1} - y_s^{i2}.
\]

Note that from the normalization constraint, \(\sum_i x_s^i = 1\), it follows that \(\nabla \sum_i x_s^i = 0\). Further, by assumption we have \(y_{i1}^s = 0\) for \(i \neq i_1, i_2\). First order optimality conditions \(y_s^{ij} \in \partial \theta_s\|p_s^i\|_2 \leq \theta_s^* / 2\) (i.e. \(y_{i1}^s \propto -p_s^i\) and \(y_{i2}^s \propto -p_s^i\)) imply that \(p_s^i = -p_s^i\). Together with \(\|p_s^i\| = \|p_s^j\| = \theta_s^* / 2\) we obtain \(\|p_s^i - p_s^j\| = \theta_s^*\).

In the following we assume \(i_1 < i_2\) w.l.o.g. Given now the primal solution obtained from the refinement approach, we construct a feasible primal solution for the tight energy, i.e. we have to determine \(y_{ij}^s\) for \(i, j : |i - j| \geq T\). We set in this case \(y_{i1}^{ij} = y_{i1}^{ij*}\), and \(y_{i2}^{ij} = 0\) for \(i, j : |i - j| \geq T\) otherwise. It can be easily checked that this choice for \(y_{ij}^{ij}\) satisfies the marginalization constraints, i.e. one half of the optimality conditions. The dual variables \(p\) are a certificate for optimality, since \(y_{i1}^{ij*} \neq 0\) implies \(\|p_s^i - p_s^j\| = \theta_s^*\) (i.e. the inequality constraint is tight), and for \(i, j : |i - j| \geq T\) we have \(y_{ij}^s = 0\) and \(\|p_s^i - p_s^j\| \leq \theta_s^*\). Overall, the other half of optimality conditions, \(y_{ij}^s \neq 0 \implies \|p_s^i - p_s^j\| = \theta_{ij}^s\), and we have shown optimality of the constructed solution with respect to the tight energy \(E_{\text{tight}}\). □

On planar grids at most four regions can meet in a single node (only 3 if \(\nabla\) 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 \(A_s = \{1, \ldots, L\}\). In practice only a few augmentation steps are necessary leading to a \(\approx 10\%\) increase of memory requirements over the efficient model Eq. 13. We use the primal-dual method [8] 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.
After 1 augmentation
approximation bounds to the unsmoothed energy than the one used in the aforementioned work (see one utilized in [15, 26] for similar inference problems. It turns out that our smooth energy yields better by

\[ f = \frac{\text{grad}}{\epsilon} \]

of its gradient is presented in [22]: for a non-smooth (convex) function \( E(x) \), the gradient 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. 18), 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 [22]; for a non-smooth (convex) function \( f \) and a smoothing parameter \( \epsilon > 0 \), a smooth version \( f_\epsilon \) of \( f \) with Lipschitz-continuous gradient (and Lipschitz constant \( 1/\epsilon \)) is given by \( f_\epsilon = (f^* + \epsilon \| \cdot \|_2^2/2)^* \). We employ a quadratic prox-function for smoothing rather e.g. the entropic one utilized in [15, 26] 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 \( p \) and \( q \),

\[
-E^*_\text{tight-III}(p, q) = \sum_s -q_s + \sum_s \sum_i [q_s - \text{div} p_s - \theta^i_s]^+] + \sum_s \sum_{i,j} \sqrt{2} [\| p_s - p_t \|_2^2 - \theta^{ij}] + .
\]

(21)

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

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

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

\[
h^{\theta}_\epsilon(z) = \begin{cases} 
0 & \text{if } \| z \| \leq \theta \\
 (\| z \| - \theta)^2 & \text{if } \theta \leq \| z \| \leq \theta + \sqrt{2}\epsilon \\
\sqrt{2}(\| z \| - \theta) - \epsilon & \text{if } \| z \| \geq \theta + \sqrt{2}\epsilon.
\end{cases}
\]

(22)

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

\[
-E^*_\text{tight-III,\epsilon}(p, q) = \sum_s -q_s + \sum_s \sum_i [q_s - \text{div} p_s - \theta^i_s]^+]_{+\epsilon} + \sum_s \sum_{i,j} h^{\theta}_{\epsilon}(p_s - p_t).
\]

(23)

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

\[
\epsilon \log \left(1 + e^{x/\epsilon} \right) \rightarrow_\epsilon [x]_+.
\]
Observation 4. For an optimal solution \((p^*, q^*)\) of \(E_{\text{tight-III}, \varepsilon}^*\) we have

\[
E_{\text{tight-III}, \varepsilon}^*(p^*, q^*) - E_{\text{tight-III}, \varepsilon}^*(p^*, q^*) \leq \frac{3\varepsilon|V|}{2},
\]

where \(|V|\) is the number of nodes in the underlying graph.

The proof is given in the appendix. Note that, in contrast to [15, 26], the upper bound is independent of the number of labels. In practice, the bound \(3|V|\varepsilon/2\) seems to be relatively tight. Given a desired accuracy \(\delta\) to the optimal non-smooth energy, a necessary smoothing parameter \(\varepsilon\) is given by \(\varepsilon \leq \frac{2\delta}{\sqrt{|V|}}\).

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 \(L \leq \|A\|_F^2 L_f\), where \(L_f\) is the Lipschitz constant of \(\nabla f\) 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 \((p, q)\) to their appearances in the respective summands (see the appendix 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 [2], 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 \(\delta\) 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 \(\varepsilon\) 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.

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 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 \(E_{\text{tight-marginals}}\). In Section 5.1 the “metrification” of the smoothness costs and 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. 12, the slightly different marginalization constraints appearing in Eq. 3 and Eq. 12, respectively, yield to different behaviors. In the standard local polytope relaxation Eq. 3 the marginalization constraints reads as

\[ \sum_j x_{st}^j = x_s^i, \quad \sum_j x_{st}^j = x_t^i, \quad x_{st}^{ij} \geq 0. \]  

(25)

One can eliminate \( x_s^{ii} \) to arrive at “differential” marginalization constraints,

\[ x_t^i - x_s^i = \sum_{j,j \neq i} x_{st}^{ij} - \sum_{j,j \neq i} x_{st}^{ji}, \]  

(26)

but in Eq. 12 corresponding to the primal of \( E_{\text{saddlepoint}} \) also the non-negativity constraint \( x_s^{ii} = x_s^i - \sum_{j,j \neq i} x_{st}^{ij} \geq 0 \) is dropped. The lack of the non-negativity constraint on \( x_s^{ii} \) implies that any non-metric smoothness costs \( \theta_{ij}^s \) is implicitly converted into a metric via the following construction: assume that \( \theta_{ij}^s + \theta_{ii}^s < \theta_{ii}^s + 2 \). If \( x_s^i = 1 \) and \( x_s^{i+2} = 1 \) (i.e., we have a jump from label \( i \) to \( i + 2 \) along edge \( (s,t) \)), then the desired smoothness cost is \( \theta_{ij}^s + 2 \). By setting \( x_{st}^{i+1} = x_{st}^{i+1, i+2} = 1 \) and \( x_{st}^{i+1, i+2} = -1 \) the differential marginalization constraints Eq. 26 are still satisfied, but the contribution of edge \((s,t)\) to the smoothness cost is now \( \theta_{ij}^s + \theta_{ii}^s < \theta_{ii}^s + 2 \). 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. 4) for non-metric pairwise costs. Consequently, \( E_{\text{saddlepoint}} \) is not suitable to solve labeling problems with non-metric smoothness priors such as (i) truncated quadratic means, that permuting label values potentially leads to different values of \( E \) on the labels such that e.g. a jump from label \( i \) to \( i + 1 \) is “larger” than one from \( i \) to \( i + 2 \). 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., \( \theta_{ii}^s = \theta_{ii}^s = 0 \)), 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 overall 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 \( x_{s}^{i+1} = x_{s}^{i+1,i+2} = \cdots = x_{s}^{i-1,j} = 1 \) and \( x_{s}^{i+1,i+2} = x_{s}^{i+2} = \cdots = x_{s}^{i-1,j-1} = -1 \) 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. 12),

\[
E_{\text{tight-marginals-II}}(x) = \sum_{s,i} \theta_{ii}^s x_s^i + \sum_{s,j,i<j} \theta_{ij}^s \|x_s^{ij} + x_s^{ji}\|_2 \quad \text{s.t.} \quad \begin{align*}
x_s^i &= \sum_j x_{s}^{ij}, \quad x_t^i = \sum_j x_{t}^{ij}, \quad x_{s}^{ij} \geq 0, \quad x_{t}^{ij} \geq 0, \quad \sum_j x_{s}^{ij} = 1. \\
x_s^i &= \sum_{j,j=1} x_{s}^{ij}, \quad x_t^i = \sum_{j,j=1} x_{t}^{ij}, \quad x_{s}^{ij} \geq 0, \quad x_{t}^{ij} \geq 0, \quad \sum_j x_{s}^{ij} = 1. \\
\end{align*}
\]  

(27)

In contrast to \( E_{\text{tight}} \) (Eq. 6) and \( E_{\text{tight-marginals}} \) (Eq. 12) the objective value is invariant under label permutation: if \( \sigma \) is a permutation in \( \{1, \ldots, L\} \), then for any feasible \( x \) we have that \( x^\sigma \overset{\text{def}}{=} (x_s^{\sigma(i)}, x_s^{\sigma(i)}, x_t^{\sigma(i)}, x_t^{\sigma(i)}, \sigma(i)) \) is also feasible and has the same energy value for permuted costs \( \theta^\sigma \overset{\text{def}}{=} (\theta_s^{\sigma(i)}, \theta_s^{\sigma(i)}, \sigma(i), \sigma(i)) \). 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 \( E_{\text{tight}}/E_{\text{tight-marginals}} \) and \( E_{\text{tight-marginals-II}} \) for a small stereo instance with the (non-metric) smoothness costs \( \theta_{ii}^s = 0, \theta_{ii}^s+1 = 1 \), and \( \theta_{ij} = 10 \) for \( j > i + 1 \),
respectively, in Fig. 5. Observe that \( E_{\text{tight}} \) is essentially “blind” to the true cost of larger discontinuities, and the result in Fig. 5(a) shows many more abrupt label changes than Fig. 5(b). The result in Fig. 5(a) corresponds to a solution with truncated linear smoothness (with truncation point \( \theta^* = 10 \)).

![Fig. 5: Stereo result using absolute color differences and a non-metric discontinuity model. The visual results and the energy values obtained by minimizing \( E_{\text{tight}} \) and \( E_{\text{tight-marginals-II}} \) are quite significant. The final energy values of \( E_{\text{tight}} \) is much smaller than the one for \( E_{\text{tight-marginals-II}} \) due to dropping the \( x_{i}^{s,j} \geq 0 \) constraints in the former.](image)

We derive the dual of \( E_{\text{tight-marginals-II}} \) in the following. Due to the complementary constraints \( x_{i}^{s,j} \perp x_{s}^{j} \) holding for every optimal solution \( x \) of \( E_{\text{tight-marginals-II}} \) we can rewrite the smoothness term as

\[
\sum_{s} \sum_{i,j<i} \theta_{i,j} \left\| x_{i}^{s,j} \right\|_{2}^{2},
\]

(recall Section 3.1). We introduce Lagrange multipliers (dual variables) \( p_{s}^{i} \), \( p_{s}^{j} \), and \( p_{s}^{i} \) for the respective marginalization constraints. In order to derive the dual of \( E_{\text{tight-marginals-II}} \), we need the following two facts:

**Observation 5.** Let \( E(x) = f(x) + \iota \{ Ax = 0 \} \). Then the dual energy is \( E^*(p) = -f^*(A^T p) \).

We have for \( g(x) := \iota \{ Ax = 0 \} \) that \( g^*(y) = \iota \{ y \in \text{im}(A^T) \} \). Fenchel duality yields

\[ E^*(y) = -f^*(y) - g^*(-y) = -f^*(y) - \iota \{ -y \in \text{im}(A^T) \}, \]

i.e. \( y = A^T p \) for some \( p \). Overall we get \( E^*(p) = -f^*(A^T p) \).

**Observation 6.** Let \( f(x) = \theta \| x \|_2 + \iota \geq_0(x) \). Then \( f^*(y) = \iota \{ \| y \|_2 \leq \theta \} \), where \( [y]_+ = ([y]_+)_i = (\max(0, y_i))_i \).

We show that \( f^{**}(x) = f(x) \). We have

\[ f^{**}(x) = \max_{y} \left\{ \sum_{i} x_i y_i - \iota \{ \| y \|_2 \leq \theta \} \right\}. \]

If for an \( i \) \( x_i < 0 \), we can set \( y_i \to -\infty \) without violating the constraint. Hence \( f^{**}(x) = \infty \) if \( x \notin \mathbb{R}^N \). \( x \geq 0 \) implies \( y \geq 0 \) (otherwise the value of \( x^T y \) is suboptimal), and we have in this case (since \( [y]_+ = y \))

\[ f^{**}(x) = \max_{y} \left\{ x^T y - \iota \{ \| y \|_2 \leq \theta \} \right\} = \max_{y} \left\{ x^T y - \iota \{ y \geq 0 \} \right\} = \theta \| x \|_2. \]

Overall we have \( f^{**}(x) = \theta \| x \|_2 + \iota \geq_0(x) = f(x) \).

Since \( x_{s,j}^{i} \) appears in constraints corresponding to \( p_{s}^{i} \), \( p_{s}^{j} \), and \( p_{s}^{j} \), and \( x_{i}^{s} \) is contained (with -1 coefficient) in constraints corresponding to \( p_{s}^{i} \), \( p_{s}^{j} \), \( p_{s}^{j} \), and \( p_{s}^{j} \), we obtain the following dual program:

\[
E_{\text{tight-marginals-II}}^*(p) = \sum_{s} \min_{i} \left\{ \theta_{i}^{s} + p_{s}^{i} + p_{s}^{j}(0) + p_{s}^{j} + p_{s}^{j} \right\}
\]

subject to

\[
\left\| [p_{s}^{i} + p_{s}^{j}] \right\|_{2}^{2} \leq \theta_{i,j}, \quad p_{s}^{i} + p_{s}^{j} \leq 0, \quad p_{s}^{j} + p_{s}^{j} \leq 0.
\]
The last constraints, \( p_{s \rightarrow}^{i} + p_{s \leftarrow}^{i} \leq 0 \) and \( p_{s \uparrow}^{i} + p_{s \downarrow}^{i} \leq 0 \), come from the fact, that \( x_{s}^{ii} \) does only appear in the primal objective via \( x_{s}^{ii} \geq 0 \). If \( \theta^{ii} > 0 \), these constraints are replaced by
\[
\left\| \frac{p_{s \rightarrow}^{i} + p_{s \leftarrow}^{i}}{p_{s \uparrow}^{i} + p_{s \downarrow}^{i}} \right\|_{2} \leq \frac{\sqrt{2}}{\theta^{ii}} \iff \left\| \frac{p_{s \rightarrow}^{i} + p_{s \leftarrow}^{i}}{p_{s \uparrow}^{i} + p_{s \downarrow}^{i}} \right\|_{2} \leq \theta^{ii}.
\]

The difference between \( E_{\text{tight-marginals-II}}^{*} \) and \( E_{\text{tight-I}}^{*} \) (Eq. 16), is that the latter enforces \( p_{s \rightarrow}^{i} + p_{s \leftarrow}^{i} = 0 \) and \( p_{s \uparrow}^{i} + p_{s \downarrow}^{i} = 0 \): in this case one has
\[
\left\| \frac{p_{s \rightarrow}^{i} + p_{s \leftarrow}^{i}}{p_{s \uparrow}^{i} + p_{s \downarrow}^{i}} \right\|_{2} \leq \frac{\sqrt{2}}{\theta^{ij}} \iff \left\| \frac{p_{s \rightarrow}^{i} + p_{s \leftarrow}^{i}}{p_{s \uparrow}^{i} + p_{s \downarrow}^{i}} \right\|_{2} \leq \theta^{ij},
\]

since \( \|(x_{s}^{i} + [-x_{s}^{i}])^{T}\| = \|x\| \). Enforcing equality constraints instead of inequality ones as in \( E_{\text{tight-marginals-II}}^{*} \) implies that \( \max_{p} E_{\text{tight-marginals-II}}^{*}(p) \geq \max_{p} E_{\text{tight-I}}^{*}(p) \), but equality needs not (and will not) hold in general. We finish this section with a few remarks:

1. Instead of using the Euclidean norm, \( \| \cdot \|_{2} \), in \( E_{\text{tight-marginals-II}}^{*} \) (Eq. 27), one can employ any \( p \)-norm \((p \geq 1)\) in the smoothness term. If we define \( \delta \overset{\text{def}}{=} \min\{x_{s}^{ij}, x_{s}^{ji}\} > 0 \) (element-wise), we have
\[
\sum_{k} \left( (x_{s}^{ij} + x_{s}^{ji})_{k} - 2\delta_{k} \right)^{p} \leq \sum_{k} \left( (x_{s}^{ij} + x_{s}^{ji})_{k} \right)^{p}
\]

and with strict inequality when some \( \delta_{k} > 0 \) (due to the strict monotonicity of \((\cdot)^{p}\)). Consequently, if \( \delta \neq 0 \), we have
\[
\left\| x_{s}^{ij} + x_{s}^{ji} - 2\delta \right\|_{p} < \left\| x_{s}^{ij} + x_{s}^{ji} \right\|_{p}.
\]

If we assume that \( \theta^{ii} = 0 \), then every optimal solution of \( E_{\text{tight-marginals-II}}^{*} \) using the \( p \)-norm naturally satisfies the complementarity conditions. Otherwise the overall objective can be reduced by increasing \( x_{s}^{ij} \) and \( x_{s}^{ji} \) by \( \delta \) and decreasing \( x_{s}^{ij} \) and \( x_{s}^{ji} \), respectively, in order to satisfy the marginalization constraints. Due to the complementarity of \( x_{s}^{ij} \) and \( x_{s}^{ji} \), \( \theta^{ij}\left\| x_{s}^{ij} + x_{s}^{ji} \right\|_{p} \) can be rewritten as
\[
\theta^{ij}\left\| \frac{x_{s}^{ij}}{x_{s}^{ji}} \right\|_{p},
\]
leading to dual constraints of the form
\[
\left\| \frac{p_{s \rightarrow}^{i} + p_{s \leftarrow}^{i}}{p_{s \uparrow}^{i} + p_{s \downarrow}^{i}} \right\|_{q} \leq \theta^{ij},
\]

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

2. As in Section 3.3 several different instances of the dual energy can be formulated, depending on the utilized primal constraints. A variant of a dual energy to \( E_{\text{tight-marginals-II}}^{*} \) using penalizers rather
than constraints is the following:

\[ E_{\text{penalizing}}(p) = \sum_{s} \min_{i,j} \left\{ \theta_{s}^{i} + p_{s}^{i+} + p_{s}^{i-1,0+} + p_{s}^{i-1,-1} \right\} \]

\[- \sqrt{2} \sum_{s} \sum_{i,j:i < j} \left[ \left[ p_{s}^{i+} + p_{s}^{j-} \right] + \right] - \theta^{ij} \]

\[- \sum_{s} \sum_{i,j} \left( \left[ p_{s}^{i+} + p_{s}^{j-} \right] + \right), \]

(29)

where \( \|x_{s}^{ij} + x_{s}^{ji}\|_2 \leq \sqrt{2} \) and \( x_{s}^{ij} \in [0, 1] \) are explicitly incorporated. We also used the following fact (proven in the appendix):

**Observation 7.** Let \( f(x) = \theta \|x\|_2 + v\{\|x\|_2 \leq B\} + \iota_{\geq 0}(x) \). Then \( f^{*}(y) = B [\|y\|_2 - \theta]_+ \), where \( [y]_+ = (|y_i|)^T = (\max\{0, y_i\})^T \).

### 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 [30] a saddle-point formulation was proposed in order to generalize Eq. 5 beyond isotropic smoothness terms. We start by replacing the isotropic smoothness costs, \( \sum_{s} \sum_{i<j} \theta_{s}^{ij} \|y_{s}^{ij}\|_2 \), in Eq. 6 with the following term,

\[ \sum_{s} \sum_{i,j:i < j} \phi_{s}^{ij}(y_{s}^{ij}), \]

where \( \phi_{s}^{ij}(\cdot) \) is a convex, and positively 1-homogeneous function. Since \( \phi_{s}^{ij} \) 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 \( \|p_{s}^{i} - p_{s}^{j}\|_2 \leq \theta^{ij} \) are replaced by constraints of the form

\[ p_{s}^{i} - p_{s}^{j} \in W_{\phi_{s}^{ij}}, \]

where \( W_{\phi_{s}^{ij}} \) is sometimes called the Wulff shape of \( \phi_{s}^{ij} \) (see e.g. [23, 39, 40]). 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

\[ E_{\text{tight generalized}}(x, y) = \sum_{s,i} \theta_{s}^{i} x_{s}^{i} + \sum_{s} \sum_{i,j:i < j} \theta_{s}^{ij} \phi_{s}^{ij}(y_{s}^{ij}) \]

\[ \text{s.t. } \nabla x_{s}^{i} = \sum_{j:j<i} g_{s}^{ij} - \sum_{j:j>i} g_{s}^{ji}, \ x_{s} \in \Delta \]  

\[ \forall s, i, \]

(30)

As pointed out also in [30] this energy shares the problem of converting non-metric smoothness costs into metric ones with Eq. 6 (which is due to the lack of non-negativity constraints \( x_{s}^{ij} \geq 0 \) as pointed out in the previous section). Unfortunately, in contrast to Section 5.1 we cannot simply introduce non-negative pseudo-marginals \( x_{s}^{ij} \) and replace \( \phi_{s}^{ij}(y_{s}^{ij}) \) by \( \phi_{s}^{ij}(x_{s}^{ij} + x_{s}^{ji}) \), since (among other problems) \( x_{s}^{ij} + x_{s}^{ji} \) is symmetric. A transition between label \( i \) and \( j \) in a particular direction will be penalized exactly like the opposite jump. Further, the argument to \( \phi_{s}^{ij} \) is always in the non-negative quadrant, thus the shape of \( \phi_{s}^{ij} \) outside the positive quadrant is ignored. Surprisingly, substituting \( y_{s}^{ij} = x_{s}^{ij} - x_{s}^{ji} \) in Eq. 6 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 \( x_{s}^{ii} \geq 0 \):

\[
E_{\text{tight-generalized-II}}(x) = \sum_{s,i} \theta_{s} x_{s}^{i} + \sum_{s} \sum_{i,j, \neq j} \phi_{s}^{ij} (x_{s}^{ij} - x_{s}^{ji}) \quad \text{s.t.} \quad \tag{31}
\]

In contrast to e.g. Eq. 12 we lose complementarity between \( x_{s}^{ij} \) and \( x_{s}^{ji} \) in optimal solutions. Nevertheless, any minimizer \( x^{*} \) of \( E_{\text{tight-generalized-II}} \) can be converted into a solution \( \tilde{x} \) satisfying the complementarity conditions \( \tilde{x}_{s}^{ij} \perp \tilde{x}_{s}^{ji} \). We set \( \delta_{s}^{ij} \overset{\text{def}}{=} \delta_{s}^{ji} \overset{\text{def}}{=} \min \{ (x_{s}^{*})_{s}^{ij}, (x_{s}^{*})_{s}^{ji} \} \) (element-wise) for \( i \neq j \) and \( (x_{s}^{*})_{s}^{ij} \overset{\text{def}}{=} (x_{s}^{*})_{s}^{ji} + \sum_{j, \neq i} \delta_{s}^{ij} \). Node marginals stay the same, \( \tilde{x}_{s}^{i} \overset{\text{def}}{=} (x_{s}^{*})_{s}^{i} \). Clearly, we have \( \tilde{x}_{s}^{ij} \perp \tilde{x}_{s}^{ji} \) by construction and the marginalization constraints are still satisfied. Obviously, the unary terms are unaffected since \( \tilde{x}_{s}^{i} = (x_{s}^{*})_{s}^{i} \). Further, the smoothness costs also remain the same, since

\[
\tilde{x}_{s}^{ij} - \tilde{x}_{s}^{ji} = (x_{s}^{*})_{s}^{ij} - \delta_{s}^{ij} - (x_{s}^{*})_{s}^{ji} + \delta_{s}^{ij} = (x_{s}^{*})_{s}^{ij} - (x_{s}^{*})_{s}^{ji}.
\]

Overall, we constructed a solution \( \tilde{x} \) with the same objective value and satisfying the complementarity constraints.

The downside of the formulation in Eq. 31 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 \( \phi_{s}^{ij} : \mathbb{R}^{2} \rightarrow \mathbb{R}_{+}^{2} \) induced by quadratic forms one can find a higher-dimensional extension \( \Phi : \mathbb{R}^{4} \rightarrow \mathbb{R}_{+}^{2} \) similar to the conversion from \( \|x_{s}^{ij}, x_{s}^{ji}\| \) to \( \|(x_{s}^{ij}, x_{s}^{ji})\|^{T} \) from Section 3.1. For the choice of \( \phi_{s}^{ij}(y) = \sqrt{y^{T}C_{s}^{ij}y} \) we can state the following analogon to \( E_{\text{tight-marginals-II}} \) (Eq. 27):

\[
E_{\text{tight-marginals-riemann}}(x) = \sum_{s,i} \theta_{s} x_{s}^{i} + \sum_{s} \sum_{i,j, \neq j} \sqrt{\left( \frac{x_{s}^{ij}}{x_{s}^{ji}} \right)^{T} A_{s}^{ij} \left( \frac{x_{s}^{ij}}{x_{s}^{ji}} \right)} \quad \text{s.t.} \quad \tag{32}
\]

where (with \( C_{s}^{ij} = (a \ b \ c \ d) \))

\[
A_{s}^{ij} \overset{\text{def}}{=} \begin{pmatrix} a & c & 0 & -c \\ c & b & -c & 0 \\ 0 & -c & a & c \\ -c & 0 & c & b \end{pmatrix},
\]

See Appendix E why this choice of a 4-dimensional Mahalanobis norm leads to equivalent solutions satisfying the complementarity conditions.

### 6 Conclusion

In [7] 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 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 [1], 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. 6 provides the answer in the discretized setting. There seem to be several sources of difficulties, e.g. the marginalization constraint in its difference form, \( \nabla x^i = \sum_{j<i} y^{ij} - \sum_{j>i} y^{ij} \), would read just as a linear PDE, but there is the complication that \( x^i_1 \) is not smooth. Analyzing the continuous setting and further extensions of Eq. 6 are subject to future work.²

A Converting Between Superlevel and Indicator Representations

In this section we show the equivalence between Eq. 4 and Eq. 5. In the main paper we subsequently focus on Eq. 5.

We use \( u^i_s \) to denote the superlevel representation and \( x^i_s \) for the indicator representation of a label assignment, i.e. \( x_s = \partial_i u_s \), where we use backward differences for \( \partial_i \) and \( u^i_0 = 0 \) as boundary condition. With these definitions we obtain \( x^i_1 = u^i_s \) and \( x^i_i = u^i_s - u^{i-1}_s \), which is desired. We have (in 2 dimensions, but this generalizes to any dimension)

\[
\nabla_x \partial_i u_s = \begin{pmatrix}
(u^i_{s+1,0}) & (u^i_{s+1,0}) - (u^i_s - u^{i-1}_s) \\
(u^i_{s+1,0}) - (u^i_s - u^{i-1}_s) & (u^i_{s+1,0}) - (u^i_s - u^{i-1}_s)
\end{pmatrix} = \partial_i \nabla_x u_s.
\]

Since we have \( x_s = \partial_i u_s \),

\[
\max_{p_s \in C} \langle p_s, \nabla_x x_s \rangle = \max_{p_s \in C} \langle p_s, \partial_i \nabla_x u_s \rangle = \max_{p_s \in C} \langle \partial_i^T p_s, \nabla_x u_s \rangle,
\]

where \( C \) is the constraint set \( C = \{ p \, : \|p^i - p^j\| \leq \theta^j \} \). Explicitly we have

\[
\partial_i u^k_s = \begin{cases} 
1 & \text{if } k = 1 \\
-1 & \text{if } 1 < k < L.
\end{cases}
\]

For a 6-label problem the matrix corresponding to \( \partial_i \) is

\[
\begin{pmatrix}
1 & -1 & 1 \\
-1 & -1 & 1 \\
-1 & 1 & -1 \\
1 & -1 & 1
\end{pmatrix}
\]

For the adjoint operator \( \partial_i^T \) we have

\[
\partial_i^T p^k_s = \begin{cases} 
p^k_s - p^{k+1}_s & \text{if } 1 \leq k < L \\
p^L_s & \text{if } k = L.
\end{cases}
\]

The solution of \( \partial_i^T p_s = q_s \) is of the form \( p^k_s = \sum_{j=k}^L q^j_s \) (like an antiderivative), and the constraints expressed in terms of \( q_s \) are

\[
\theta^j \geq \|p^i_s - p^j_s\| = \| \sum_{l=i}^L q^j_s - \sum_{l=i}^L q^j_s \| = \left\{ \begin{array}{ll}
\| \sum_{i=1}^{j-1} q^i_s \| & \text{if } i < j, \\
\| \sum_{i=1}^{j-1} q^i_s \| & \text{if } i > j,
\end{array} \right.
\]

which are exactly the constraints used in the super-level representation Eq. 4.

²Code is available at http://www.inf.ethz.ch/personal/chaene/. Supported by the 4DVideo ERC Starting Grant Nr. 210806.
B Proof of observation 2

For convenience we restate the claim: 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 13 are equivalent. Further, for truncated smoothness costs \( E_{\text{LP-MRF}} \) (Eq. 3) and the following reduced linear program,

\[
E_{\text{reduced-LP-MRF}} = \sum_{s,t} \theta_s^i x_s^i + \sum_{(s,t) \in E} \left( \sum_{i,j:|i-j|<T} \theta^i_{ij} x_{st}^{ij} + \frac{\theta^*}{2} \sum_{i} (x_{st}^{i+} + x_{st}^{i-}) \right) \\
\text{s.t. } x_s^i = \sum_{i,j:|i-j|<T} x_{st}^{ij} + x_{st}^{j*} \quad x_s^j = \sum_{i,j:|i-j|<T} x_{st}^{ji} + x_{st}^{i*}
\]

are equivalent.

**Proof:** We are going to show the equivalence of \( E_{\text{LP-MRF}} \) and \( E_{\text{reduced-LP-MRF}} \) for smoothness costs with \( \theta^i_{ij} = 0 \) and \( \theta^* = \theta^* \) if \( |i-j| \geq T \) for some \( T \). We will demonstrate that an optimal solution for \( E_{\text{LP-MRF}} \) can be converted to a feasible solution of \( E_{\text{reduced-LP-MRF}} \) with the same objective value and vice versa. We have the full model,

\[
E_{\text{truncated-LP-MRF}} = \sum_{s,t} \theta_s^i x_s^i + \sum_{(s,t) \in E} \left( \sum_{i,j:|i-j|<T} \theta^i_{ij} x_{st}^{ij} + \theta^* \sum_{i,j:|i-j|\geq T} x_{st}^{ij} \right) \\
\text{s.t. } x_s^i = \sum_{i,j:|i-j|<T} x_{st}^{ij} + x_{st}^{j*} \quad x_s^j = \sum_{i,j:|i-j|\geq T} x_{st}^{ji} + x_{st}^{i*}
\]

subject to the marginalization constraints \( \sum_j x_{st}^{ij} = x_s^i \) and \( \sum_i x_{st}^{ij} = x_s^j \). We used that \( \theta^i_{ij} = \theta^* \) for \( |i-j| \geq T \) (where \( T \) is the truncation point) and \( \theta^* < \theta^* \). If we have a minimizer of \( E_{\text{truncated-LP-MRF}} \), we can easily construct a solution of \( E_{\text{reduced-LP-MRF}} \) with the same overall objective by setting

\[
x_{st}^{ij*} = \sum_{j:|i-j|\geq T} x_{st}^{ij} \quad \text{and} \quad x_{st}^{j*} = \sum_{i:|i-j|\geq T} x_{st}^{ij}.
\]

since the pairwise truncated smoothness costs are the same

\[
\frac{\theta^*}{2} \sum_i x_{st}^{i+} + \frac{\theta^*}{2} \sum_j x_{st}^{j+} = \theta^* \sum_i \sum_{j:|i-j|\geq T} x_{st}^{ij} + \frac{\theta^*}{2} \sum_j \sum_{i:|i-j|\geq T} x_{st}^{ji} = \theta^* \sum_{i,j:|i-j|\geq T} x_{st}^{ij}
\]

Note that the modified marginalization constraints of \( E_{\text{reduced-LP-MRF}} \) are also satisfied by construction. If we have a minimizer \( x \) of \( E_{\text{reduced-LP-MRF}} \), we have to construct a solution \( \tilde{x} \) of \( E_{\text{reduced-LP-MRF}} \) with the same objective satisfying the standard marginalization constraint. We set

\[
\tilde{x}_s^i = x_s^i \quad \text{and} \quad \tilde{x}_s^j = x_s^j \quad \forall i,j : |i-j| < T.
\]

Determining \( x_{st}^{ij} \) for \( i,j : |i-j| \geq T \) is more difficult. In the following we consider a particular edge \( s,t \) and omit the subscript. We use the north-west corner rule (e.g. [16]) to assign \( \tilde{x}_{ij} \) for \( i,j : |i-j| \geq T \):

\[
\tilde{x}_{i+} \leftarrow x_{i+}^* \\
\tilde{x}_{j+} \leftarrow x_{j+}^*
\]

**while some \( \tilde{x}_{ij} \) is not assigned do**

**Choose** \( i \) and \( j \) (with \( |i-j| \geq T \)) such that \( \tilde{x}_{ij} \) is not assigned

\[
\tilde{x}_{ij} \leftarrow \min \{ \tilde{x}_{i+}, \tilde{x}_{j+} \} \\
\tilde{x}_{i+} \leftarrow \tilde{x}_{i+} - \tilde{x}_{ij} \\
\tilde{x}_{j+} \leftarrow \tilde{x}_{j+} - \tilde{x}_{ij}
\]

\[
\{ \tilde{x}_{ij} \geq 0 \} \\
\{ \tilde{x}_{i+} \geq 0 \} \\
\{ \tilde{x}_{j+} \geq 0 \}
\]

\[
\{ x_i = \sum_{j: (i,j) \text{ assigned}} \tilde{x}_{ij} + \tilde{x}_{i+} \} \\
\{ x_j = \sum_{i: (i,j) \text{ assigned}} \tilde{x}_{ij} + \tilde{x}_{j+} \} 
\]
The updates ensure that $\hat{x}^i$, $\bar{x}^i$ and $\bar{x}^j$ stay non-negative and that the following modified marginalization constraints are still satisfied after each iteration:

$$
\dot{x}_i = \sum_{i,j:|j-i|<T} \dot{x}^i + \sum_{i,j:|i-j|\geq T} \dot{x}^j = \sum_{j} \hat{x}^j + \bar{x}^i + \bar{x}^j
$$

We show that all $\bar{x}^i$ and $\bar{x}^j$ are 0 after termination of this algorithm. First, it cannot be that $\bar{x}^i > 0$ and $\bar{x}^j > 0$ for some $i$ and $j$: if this is the case for $i, j : |i-j| < T$, we can increase $\hat{x}^i$ and simultaneously strictly lowering the overall smoothness cost, thus contradicting that our initial solution was optimal. If $\bar{x}^i > 0$ and $\bar{x}^j > 0$ for some $i, j : |i-j| \geq T$, this contradicts the instructions $(\hat{x}^j \leftarrow \min\{\bar{x}^i, \bar{x}^j\})$, $\bar{x}^i \leftarrow \bar{x}^i - \bar{x}^j$, $\bar{x}^j \leftarrow \bar{x}^j - \bar{x}^j$ in the algorithm above, which sets one of $\bar{x}^i$ or $\bar{x}^j$ to zero. W.l.o.g. some of the $\bar{x}^i$ are strictly greater than 0, but all $\bar{x}^j$ are 0. We have

$$
1 = \sum_i \dot{x}_i = \sum_i \sum_j (\dot{x}^i + \bar{x}^i) = \sum_j \hat{x}^j + \bar{x}^i + \bar{x}^j = 1 + \bar{x}^i,
$$

which is a contradiction. Hence all $\bar{x}^i$ and $\bar{x}^j$ have to be 0 at the end of the algorithm. We further have

$$
\sum_{j:|j-i|<T} \dot{x}^j = x^i \quad \text{and} \quad \sum_{i,j:|i-j|\geq T} \dot{x}^j = x^j
$$

and the pairwise smoothness costs are the same for $x$ and $\hat{x}$ (similar to Eq. 36) and both overall objectives for $E_{\text{truncated-LP-MRF}}(\hat{x})$ and $E_{\text{reduced-LP-MRF}}(x)$ coincide. Thus, we have proved the observation.

C Notes on smoothing-based optimization

C.1 A smooth version of $h^\theta(z) = B[\|z\|_2 - \theta]_+$

By construction we know that the convex conjugate of $h^\theta$ is given by

$$
(h^\theta)^*(x) = \theta\|x\|_2 + t\{\|x\| \leq B\}.
$$

Thus, a smooth version of $h^\theta$ is the convex conjugate of

$$
(h^\theta_t)^*(x) = \theta\|x\|_2 + t\{\|x\| \leq B\} + \frac{\varepsilon}{2}\|x\|_2^2.
$$

Consequently,

$$
h^\theta_t(z) = \max_{x:\|x\|_2 \leq B} x^T z - \theta\|x\|_2 - \frac{\varepsilon}{2}\|x\|_2^2.
$$

If we fix $\|x\|$, then an $x$ colinear with $z$ is maximizing the expression, hence we can reduce the problem by restricting $x$ to be $x = cz$ for some $c \geq 0$. Hence, the above maximization problem is equivalent to

$$
h^\theta_t(z) = \max_{c \geq 0: c\|z\|_2 \leq B} c\|z\|_2^2 - \varepsilon\theta\|z\|_2 - \frac{\varepsilon c^2}{2}\|z\|_2^2.
$$

We have $h^\theta_t(0) = 0$, and in the following we assume $z \neq 0$, i.e. $\|z\|_2 > 0$. We have to analyze three cases:

- $c \in (0, B/\|z\|_2)$: First order conditions on $c$ yield

$$
\|z\|_2^2 - \theta\|z\|_2 - \varepsilon c\|z\|_2^2 = 0
$$

i.e.

$$
c = \frac{\|z\|_2 - \theta}{\varepsilon\|z\|_2} \quad \text{and} \quad h^\theta_t(z) = \frac{1}{2\varepsilon}(\|z\|_2 - \theta)^2
$$

in this case. Note that $c > 0$ if $\|z\|_2 > \theta$. 

22
c = 0: This case is effective if \(\|z\|_2 \leq \theta\), and in this case we have
\[
h^\theta_\varepsilon(z) = 0.
\]

\(c = B/\|z\|_2\): In this case we obtain
\[
h^\theta_\varepsilon(z) = B(\|z\|_2 - \theta) - \frac{\varepsilon}{2} B^2.
\]

This case is in effect if \(c = \frac{\|z\|_2 - \theta}{\|z\|_2} \geq B\), i.e. \(\|z\| \geq \theta + B\varepsilon\). Setting \(B = \sqrt{2}\) yields the expression for \(h^\theta_\varepsilon\) in Eq. 22.

Overall we obtain the smooth version of \(h^\theta\) as stated in the main text.

### C.2 Proof of observation 4

We repeat the claim for convenience: for an optimal solution \((p, q)\) of \(E^\ast_{\text{tight-III}, \varepsilon}\) we have

\[
E^\ast_{\text{tight-III}, \varepsilon}(p, q) - E^\ast_{\text{tight-III}}(p, q) \leq \frac{3\varepsilon|\mathcal{V}|}{2},
\]

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

**Proof:** We have that \(q\) are the Lagrange multipliers for \(\sum x^s_i = 1\) and \(p\) are the multipliers for the (differential) marginalization constraints in the primal, respectively. Hence, there are primal variables \((x, y)\) satisfying the normalization and marginalization constraints, and

\[
E^\ast_{\text{tight-III}, \varepsilon}(p, q) = E^\ast_{\text{tight-III}, \varepsilon}(x, y)
\]

\[
= \sum_{s, i} \theta^s_i x^s_i + \sum_s \sum_{i,j<i} \theta^{ij} \|y^{ij}_s\|_2 + \frac{\varepsilon}{2} \sum_s \sum_{i,j<i} (x^s_i)^2 + \frac{\varepsilon}{2} \sum_s \sum_i (|x^s_i|)^2
\]

subject to \(x_s \in \Delta\) and \(\nabla x^s_i = \sum_{j:j<i} y^{ij}_s - \sum_{j:j>i} y^{ji}_s\). Since \(x_s\) is in the unit simplex, \(\sum_i (x^s_i)^2 \leq 1\) (every element of \(\Delta\) has at most unit Euclidean length) and \(\frac{\varepsilon}{2} \sum_i (x^s_i)^2\) is therefore bounded by \(\varepsilon|\mathcal{V}|/2\). In order to bound the second quadratic term, in view of \(E^\ast_{\text{tight-marginals}}\) (Eq.12) we rewrite the smoothness terms in the primal energy as

\[
\sum_{s} \sum_{i,j<i} \theta^{ij} \frac{\|x^s_{ij}\|_2}{2} + \frac{\varepsilon}{2} \sum_{s} \sum_{i,j<i} (\|x^s_{ij}\|_2^2 + \|x^s_{ij}\|_2^2)
\]

subject to \(\nabla x^s_i = \sum_{j:j \neq i} x^s_{ij} - \sum_{j:j \neq i} x^s_{ji}\). For \(i, j\) with \(i < j\) we introduced \(x^s_{ij} = [y^{ij}_s]_+\) and \(x^s_{ji} = [y^{ji}_s]_-\). By introducing \(x^s_{ij} = x^s_{ij} - \sum_{j:j \neq i} x^s_{ij} \in \mathbb{R}^2\) the differential marginalization constraints are equivalent to

\[
x^s_i = \sum_j x^s_{ij}, \quad x^s_i = \sum_{j} x^s_{ij},
\]

with \(\sum_{i,j} x^s_{ij} = 1_2\) and \(x^s_{ij} \geq 0\) for \(i \neq j\). Using the complementarity of \(x^s_{ij}\) and \(x^s_{ji}\) (and \(x^s_{ij}\) is 2-dimensional) we obtain that

\[
\sum_{i,j<i} (\|x^s_{ij}\|_2^2 + \|x^s_{ji}\|_2^2) \leq 2,
\]

hence the second quadratic term in \(E^\ast_{\text{tight-III}, \varepsilon}\) is bounded by \(\varepsilon|\mathcal{V}|\). Note that adding the constraints \(x^s_i \geq 0\) does not change the bound, hence it also applies to the stronger relaxation for non-metric smoothness costs discussed in Section 5.1.
C.3 Bound on the operator norm of $A$

To get the Lipschitz constant we again look at the $A$ matrix and get an upper bound for $\| A \|_2$ via $\| A \|_2^2 \leq \| A \|_1 \| A \|_\infty$. Note that $\| A \|_1$ is the maximum absolute column sum, and $\| A \|_\infty$ is the maximum absolute row sum. The columns of $A$ are indexed by the unknowns $(p^s_i)_i$, $(p^q_i)_q$, and $q_s$, and the rows of $A$ correspond to the terms in $E^*_{\text{tight-HIII}}$ (or its smooth version),

$$E^*_{\text{tight-HIII}}(p, q) = \sum_s q_s + \sum_{s,i} \left[ \text{div } p^s_i + \theta^i_s - q_s \right]_+ + \sum_s \sum_{i,j} \sqrt{2} \min \left\{ 0, \theta^{ij} - \| p^s_i - p^s_j \|_2 \right\}.$$ 

Since all occurrences of $p^s_i$ and $q_s$ have a $+1$ or $-1$ coefficient, it is sufficient to just count the occurrences of each variable. Since at most 5 variables appear in one term (rows corresponding to $[\text{div } p^s_i + \theta^i_s - q_s]_+$), we have $\| A \|_\infty = 5$. $q_s$ appears in $L + 1$ terms (in $q_s$ and in $\sum_i [\text{div } p^s_i + \theta^i_s - q_s]_+$), and e.g. $(p^s_i)_i$ occurs also at most in $L + 1$ terms (in the divergence terms with respect to $s$ and $s - (1,0)$ and in $L - 1$ expressions $\sum_i, i < j \sqrt{2} \min \left\{ 0, \theta^{ij} - \| p^s_i - p^s_j \|_2 \right\}$), hence $\| A \|_1 = L + 1$. Overall we have the bound $\| A \|_2 \leq 5(L + 1)$.

C.4 Extracting the primal solution from the smooth dual

We recall the smooth dual energy from Eq. 23 and indicate the correspondence between the terms in the dual energy and the respective primal variable,

$$-E^*_{\text{tight-HIII}, \varepsilon}(p, q) = \sum_s -q_s + \sum_{s,i} \left[ q_s - \text{div } p^s_i + \theta^i_s \right]_+ + \sum_s \sum_{i,j} h^{ij}_\varepsilon (p^s_i - p^s_j).$$ (38)

First order optimality conditions require that the corresponding primal unknowns are given by

$$x^s_i = \frac{d}{dz} [z - \theta^s_i]_+ |_{z = q_s - \text{div } p^s_i}$$

and

$$y^{ij}_\varepsilon = \nabla_x h^{ij}_\varepsilon(z).$$

This allows to obtain primal estimates for iterative dual optimization methods, but the marginalization constraints between $x_s$ and $y_s$ will be only fulfilled after convergence. In order to obtain useful primal solutions prior to convergence e.g. to have tight duality gaps, we extract the primal variables as follows: since extracting

$$x^s_i = \Pi_{[0,1]} \left( \frac{q_s - \text{div } p^s_i - \theta^i_s}{\varepsilon} \right)$$

will not satisfy $\sum_i x^s_i = 1$ prior to convergence, we temporarily adjust $q_s$ such that the extracted node marginals satisfy the normalization constraint. We will denote this Lagrange multiplier $\hat{q}_s$, and it should be chosen such that

$$\hat{x}^s_i(\hat{q}_s) \overset{\text{def}}{=} \Pi_{[0,1]} \left( \frac{\hat{q}_s - \text{div } p^s_i - \theta^i_s}{\varepsilon} \right)$$

satisfies $\sum_i \hat{x}^s_i(\hat{q}_s) = 1$. Note that $\sum_i \hat{x}^s_i(\hat{q}_s)$ is monotonically increasing, $\sum_i \hat{x}^s_i$ will be 0 for too small values of $\hat{q}_s$, and $L$ for large values of $\hat{q}_s$. Further, it is a piece-wise linear function with a finite number of breakpoints. The procedure to determine $\hat{q}_s$ is somewhat similar to the projection into the unit simplex (e.g. [10]): let $\tilde{\theta}^s_k$ denote the ascending sequence of reparametrized potentials $\theta^s_i + \text{div } p^s_i$. Under the hypothesis that exactly $m$ (out of $L$) $x^s_i$ are non-zero, $\hat{q}_s$ is determined from the condition

$$\sum_{i=1}^m \frac{\hat{q}_s - \tilde{\theta}^s_k}{\varepsilon} = 1,$$
i.e. \( \hat{q}_s = (\sum_{i=1}^{m} \hat{\theta}_i^{(k)} + \varepsilon)/m \). Further, \( \hat{q}_s \leq \bar{\theta}_s^{(k)} \) for \( k > m \) is required in order to have \( \hat{x}_s^{(k)}(\hat{q}_s) = 0 \) (the node marginal for the label with the \( k \)-th smallest reparametrized cost) for \( k > m \). \( \hat{x}_s^{(k)} = 0 \) implies \( \hat{x}_s^{(k+1)} = 0 \). Thus, the desired \( m \) is the largest index, such that

\[
\frac{1}{m} \left( \sum_{i=1}^{m} \hat{\theta}_i^{(k)} + \varepsilon \right) < \bar{\theta}_s^{(m+1)}.
\]

Since \( \hat{x}_s^{(1)} \) will be 1 for \( \hat{q}_s \geq \bar{\theta}_s^{(1)} + \varepsilon \), it is sufficient to restrict the search only to reparametrized costs \( \hat{\theta}_i^{(k)} \leq \bar{\theta}_s^{(1)} + \varepsilon \). Once the node marginals \( x^*_s \) are extracted, the pairwise marginals can be obtained by solving a modified transport problem independently for each \( s \). Each problem reads as

\[
\min_{y^{ij}_{s}} \sum_{i,j,i<j} \hat{\theta}_{ij}^{s} \|y^{ij}_{s}\| \quad \text{s.t.} \quad \nabla x^i_{s} = \sum_{j:j<i} y^{ij}_{s} - \sum_{j:j>i} y^{ij}_{s} - \nabla x^i_{s}.
\]

Note that only \( L-1 \) of the marginalization constraints are independent (since \( \sum_{i} \nabla x^i_{s} = 0 \)). We use the first \( L-1 \) constraints and express \( y_s^{L} \) in terms of the others and \( \nabla x^i_{s} \),

\[
y_s^{i,L} = \sum_{j:j<i} y^{ij}_{s} - \sum_{j:i<j} y^{ij}_{s} - \nabla x^i_{s}.
\]

Therefore the above problem reduces to an unconstrained non-smooth convex problem

\[
\min_{y^{ij}_{s}} \sum_{i,j,i<j} \hat{\theta}_{ij}^{s} \|y^{ij}_{s}\| + \sum_{i:i \in [1:L]} \sum_{j:j<i} \hat{\theta}_{ij}^{L} \|y^{ij}_{s}\| - \sum_{j:j>i} \hat{\theta}_{ij}^{L} \|y^{ij}_{s}\| - \nabla x^i_{s},
\]

which we solve by an iterative proximal method. Eliminating \( y_s^{i,L} \) has the benefit, that \( y^{ij}_{s} \) satisfies the marginalization constraints even prior to convergence.

### D A proof of observation 7

We recall the claim: Let \( f(x) = \theta \|x\|_2 + \iota\{\|x\|_2 \leq B\} + \iota_{\geq 0}(x) \). Then \( f^*(y) = B \|y\|_2 - \theta \|y\|_2 \| \), where \( \|y\|_2 = (\|y\|_2^T \|y\|_2)^{1/2} = (\max\{0, y_i\})^T \).

**Proof:** We compute the biconjugate \( f^{**}(x) \). If some \( x_i < 0 \), then we can let \( y_i \rightarrow -\infty \), which will not increase \( f^*(y) \), hence \( x^T y - f^*(y) \rightarrow \infty \) in this case. Thus, \( f^{**}(x) = \infty \) if \( x \notin \mathbb{R}^N_+ \). If \( x \geq 0 \) an optimal \( y \) is also \( \geq 0 \). We have in this case \( x \geq 0 \):

\[
f^{**}(x) = \max_{y \geq 0} x^T y - B \|y\|_2 - \theta \|y\|_2 \|_+
\]

\[
= \max_{y \geq 0} x^T y - B \|y\|_2 - \theta \|y\|_2 \|
\]

\[
= \max \left\{ \max_{y \geq 0: \|y\|_2 \leq \theta} x^T y, \max_{y \geq 0: \|y\|_2 > \theta} x^T y - B(\|y\|_2 - \theta) \right\}
\]

\[
= \max \left\{ \theta \|x\|_2, \max_{y \geq 0: \|y\|_2 > \theta} x^T y - B\|y\|_2 + B\theta \right\}
\]

\[
= \max \left\{ \theta \|x\|_2, \max_{k:k \|x\|_2 \leq \theta} k\|x\|_2^2 - Bk\|x\|_2 + B\theta \right\}
\]

\[
= \max \left\{ \theta \|x\|_2, \max_{k:k \|x\|_2 \leq \theta} k\|x\|_2(\|x\| - B) + B\theta \right\}
\]

\[
= \max \left\{ \theta \|x\|_2, \theta \|x\|_2 + \iota\{\|x\|_2 \leq B\} \right\}
\]

\[
= \theta \|x\|_2 + \iota\{\|x\|_2 \leq B\}.
\]

Overall we have \( f^{**}(x) = \theta \|x\|_2 + \iota\{\|x\|_2 \leq B\} + \iota_{\geq 0}(x) = f(x) \). \( \square \)
E  Higher-dimensional embedding of $\phi(y) = \sqrt{y^TCy}$

Let $\phi : \mathbb{R}^2 \to \mathbb{R}^n_+$, $\phi(y) = \sqrt{y^TCy}$ with $C = \left( \begin{smallmatrix} c & b \\ -c & 0 \end{smallmatrix} \right)$ positive definite. One is interested in a function $\Phi : \mathbb{R}^4 \to \mathbb{R}^n_+$ with the following properties (for $x_1^+, x_2^+, x_1^-, x_2^- \geq 0$)

$$
\begin{align*}
\Phi \left( \begin{array}{c}
x_1^+ \\
x_2^+ \\
0 \\
0
\end{array} \right) &= \phi \left( \begin{array}{c}
x_1^+ \\
x_2^+
\end{array} \right) \\
\Phi \left( \begin{array}{c}
x_1^- \\
x_2^- \\
x_1^0 \\
x_2^0
\end{array} \right) &= \phi \left( \begin{array}{c}
x_1^- \\
x_2^-
\end{array} \right)
\end{align*}
$$

and

$$
\begin{align*}
\Phi \left( \begin{array}{c}
x_1^+ + \delta \\
x_2^+ + \delta \\
x_1^- + \delta \\
x_2^- + \delta
\end{array} \right) > \Phi \left( \begin{array}{c}
x_1^+ \\
x_2^+ \\
x_1^- \\
x_2^-
\end{array} \right)
\end{align*}
$$

for $\delta > 0$. The first property ensures that $\Phi$ reduces to $\phi$ in the appropriate cases, and the second property enforces that complementary solutions $x^+ \perp x^-$ (i.e. $x_1^+ x_1^- = 0$ and $x_2^+ x_2^- = 0$) are strictly cheaper than non-complementary assignments. We set

$$
A \equiv \begin{pmatrix} a & c & 0 & -c \\ c & b & -c & 0 \\ 0 & -c & a & c \\ -c & 0 & c & b \end{pmatrix},
$$

then one choice of $\Phi$ is $\Phi(x) = \sqrt{x^T Ax}$. Note that $A$ is in general not positive definite, but for for $x_1^+, x_2^+, x_1^-, x_2^- \geq 0$ we have

$$
\begin{align*}
x^T Ax &= a(x_1^+)^2 + a(x_1^-)^2 + b(x_2^+)^2 + b(x_2^-)^2 + 2c(x_1^+ x_2^++ x_1^- x_2^- - x_1^+ x_2^- - x_1^- x_2^+) \\
&= a(x_1^+)^2 + a(x_1^-)^2 + b(x_2^+)^2 + b(x_2^-)^2 + 2c(x_1^+ x_2^- - x_2^- x_2^-) - 2c(x_1^+ x_2^- - x_2^- x_2^-) \\
&\geq a(x_1^+ - x_1^-)^2 + b(x_2^+ - x_2^-)^2 + 2c(x_1^+ - x_1^-)(x_2^+ - x_2^-) \\
&= (x_1^+ - x_1^-)^T C (x_1^+ - x_1^-) \geq 0.
\end{align*}
$$

since e.g. $(x_1^+)^2 + (x_1^-)^2 \geq (x_1^+)^2 + (x_1^-)^2 - 2x_1^+ x_1^- = (x_1^+ - x_1^-)^2$ for non-negative $x_1^+$ and $x_1^-$, and $C$ was assumed to be positive definite. Hence, $\Phi$ is well-defined. Further, by construction the first property is satisfied, e.g.

$$
\Phi \left( \begin{array}{c}
x_1^+ \\
0 \\
x_2^-
\end{array} \right) = \sqrt{a(x_1^+)^2 + b(x_2^-)^2 - 2c x_1^+ x_2^-} = \phi \left( \begin{array}{c}
x_1^+ \\
x_2^- 
\end{array} \right).
$$

The second property follows from (with $\delta > 0$)

$$
\begin{align*}
\Phi \left( \begin{array}{c}
x_1^+ + \delta \\
x_2^+ + \delta \\
x_1^- + \delta \\
x_2^- + \delta
\end{array} \right) &= \sqrt{a(x_1^+ + \delta)^2 + a(x_1^- + \delta)^2 + b(x_2^+ + \delta)^2 + b(x_2^- + \delta)^2 + 2c(x_1^+ - x_1^-)(x_2^+ - x_2^-)} \\
&> \sqrt{a(x_1^+)^2 + a(x_1^-)^2 + b(x_2^+)^2 + b(x_2^-)^2 + 2c(x_1^+ - x_1^-)(x_2^+ - x_2^-)}.
\end{align*}
$$

26
References


28


