Dual Decomposition for Joint Discrete-Continuous Optimization

Christopher Zach
Microsoft Research Cambridge

Abstract

We analyse convex formulations for combined discrete-continuous MAP inference using the dual decomposition method. As a consquence we can provide a more intuitive derivation for the resulting convex relaxation than presented in the literature. Further, we show how to strengthen the relaxation by reparametrizing the potentials, hence convex relaxations for discrete-continuous inference does not share an important feature of LP relaxations for discrete labeling problems: incorporating unary potentials into higher order ones affects the quality of the relaxation. We argue that the convex model for discrete-continuous inference is very general and can be used as alternative for alternation-based methods often employed for such joint inference tasks.

1 Introduction

Many problems in particular in low-level computer vision can be stated as

> *for each node in a random field (pixel, superpixel etc.) jointly determine a continuous unknown from $\mathbb{R}^d$ and an associated discrete label from $\{1, \ldots, L\}$ under respective smoothness assumptions.*

Problems falling in this category include the celebrated Mumford-Shah model for joint image segmentation and denoising [MS89, CV01], robust image processing methods in general (where the discrete state reflects whether data at a pixels is an inlier or an outlier, e.g. [BR96]), layered representations for optical flow [WA94, SSB10] and stereo [BSA98, KS04], dense depth computation with occlusion detection [GLY95], joint estimation of depth and semantic labels [LSR+10], and many others. These problems are usually formulated as Bayesian inference tasks and solved via energy optimization. *Joint* optimization over discrete and continuous unknowns is usually very difficult, but in many cases optimizing over either the discrete or the continuous unknowns can be efficiently done. Hence, authors usually propose an alternation-based (or block-coordinate) optimization method for such joint problems, which sometimes (but not always) corresponds to an expectation-maximization (EM) algorithm [Har58, DLR77]. A classic algorithm falling into this category is the K-means algorithm for clustering data [For65].The obvious drawback of such an approach is its susceptibility to bad initialization leading often to very poor or even degenerate solutions. Consequently, a method returning a good solution not requiring or unaffected by the choice of the initial values is highly desirable. In this work we build on the convex relaxation for discrete-continuous Markov random fields proposed in [ZK12] to tackle joint discrete-continuous problems particularly emerging in low-level computer vision. In [ZK12] the motivation for the proposed discrete-continuous convex model ("DC-MRF") is to solve MRFs with continuous label spaces and piecewise (but not globally) convex potentials. The utility of the DC-MRF model extends to far more general problems as later shown in this work. We summarize our contributions as follow:

1. We advocate the use of the convex DC-MRF model to a larger spectrum of joint discrete continuous problems usually solved by alternation. In summary, the DC-MRF model is applicable whenever the problem is convex after fixing the discrete labeling.
2. We derive the DC-MRF formulation from a dual decomposition/Langrangian relaxation perspective and demonstrate alternative decompositions (Section 3).
3. We strengthen the DC-MRF relaxation by realizing that moving lower-order potentials to higher

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

order cliques is beneficial for the relaxation. This is in stark contrast to linear programming relaxations of discrete labeling problems, where potentials can be *reparametrized* without affecting the relaxation (Sections 4.1 and 4.2).

4. We propose to optimize the dual energy in order to obtain a memory efficient minimization approach.

5. Finally, we provide a non-standard application that uses a *layered* denoising and inpainting model for depth data (Section 5).

We want to emphasize that we propose to solve a convex relaxation of a difficult optimization/inference problem. Thus, we can guarantee strong solutions whenever the convex relaxation is tight (or at least close to being tight). Other approaches often used to solve such difficult non-convex inference problems over continuous state spaces (besides alternating minimization methods) include continuation methods (e.g. graduated non-convexity [BZ87]), sampling-based belief propagation [IM09, PHMU11] and proposal-based algorithms (e.g. fusion moves [LRRB10]).

2 Background

2.1 Notations

We consider extended real-valued functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. The domain of f, $\mathrm{dom}(f)$, is $\{x \in \mathbb{R}^n : f(x) < \infty\}$, and we assume that $\mathrm{dom}(f) \neq \emptyset$. By allowing extended functions constraints on the feasible domain and infinite function values can be interchanged, and we will use the notation $\imath_C(x)$ to write a constraint $x \in C$ in functional form, i.e. $\imath_C(x) = 0$ iff $x \in C$ and ∞ otherwise. Equivalently, we also write $\imath\{x \in C\}$ or $\imath\{P(x)\}$, where P is some boolean predicate over x.

The convex conjugate of f, denoted by f^*, is defined as $f^*(y) = \sup_x x^T y - f(x)$. The biconjugate f^{**} is obtained by applying convex conjugation twice. It is known that for any function f the convex conjugate f^* is a lower-semicontinuous (l.s.c.) convex function, and $f^{**} = f$ iff f is convex and l.s.c. Otherwise f^{**} is the lower convex envelope of f, i.e. the supremum of all convex functions below the epigraph of f.

For a convex function f we denote the l.s.c. extension of its perspective $(x, y) \mapsto xf(y/x)$ to $x = 0$ by $f_\oslash$. $f_\oslash$ can be computed as the biconjugate of the standard perspective.

Finally, we focus on labeling problems with at most pairwise smoothness potentials. Thus, we assume a graph $G = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V}$ and edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is given. We will frequently use the shorthand notation $\sum_{s\sim t}$ for $\sum_{(s,t)\in\mathcal{E}}$, and denote the sets of (direct) ancestor and succesor nodes of s by $\mathrm{in}(s)$ and $\mathrm{out}(s)$, respectively. We use a compact notation to indicate nodes and edges (using subscripts, e.g. s or st) and states (using superscripts like i and ij). Hence, we write e.g. x^i_s for a pseudo-marginal instead of the more verbose variant $\mu_s(x_i)$.

2.2 The Convex Discrete-Continuous Model

In this section we briefly review the convex discrete-continuous formulation for inference proposed in [ZK12]. For families of convex functions $\{f^i_s\}_{s\in\mathcal{V}}$ and $\{f^{ij}_{st}\}_{(s,t)\in\mathcal{E}}$ (with $i, j \in \{1, \ldots, L\}$) the following objective is proposed,

$$E^{\text{orig}}_{\text{DC-MRF}}(\mathbf{x}, \mathbf{y}) = \sum_{s,i} (f^i_s)_\oslash(x^i_s, y^i_s) + \sum_{s\sim t}\sum_{i,j} (f^{ij}_{st})_\oslash(x^{ij}_{st}, y^{ij}_{st\to s}, y^{ij}_{st\to t}) \tag{1}$$

subject to the following marginalization constraints

$$\begin{aligned} x^i_s &= \sum_j x^{ij}_{st} & x^j_t &= \sum_i x^{ij}_{st} \\ y^i_s &= \sum_j y^{ij}_{st\to s} & y^j_t &= \sum_i y^{ij}_{st\to t} \end{aligned} \tag{2}$$

and simplex constraints $x_s \in \Delta^L$, $x_{st} \in \Delta^{L^2}$. The unknown vector $\mathbf{x}$ collects the pseudo-marginals (i.e. x_s indicates one-hot encoding of the assigned discrete state at node s), and the unknowns $\mathbf{y}$ represent the assigned continuous labels in the solution. $E^{\text{orig}}_{\text{DC-MRF}}$ above is stated for the very important case of at most pairwise interactions between labels, but can be extended to higher-order potentials in a straightforward manner. The DC-MRF model is an extension of the standard local-polytope relaxation for discrete labeling problems by allowing the unary and pairwise potentials now to be arbitrary piecewise convex functions.[1]

The formulation Eq. 1 is used in [ZK12] to model convex relaxations of non-convex continuous labeling tasks. In particular, the data term for a continuous labeling problem is allowed to be piecewise convex instead of globally convex. The discrete state obtained in the obtained discrete-continuous label assignment encodes which case in the piecewise convex definition of the unary costs is active at the minimizer.

[1] Note that in Eq. 1 we dropped the additional constraints $0 \leq y^i_s \leq x^i_s$ etc. explicitly stated in [ZK12], since they are not necessary unless boundedness of the continuous unknowns is requested (which then can be incorporated into f^i_s and f^{ij}_{st}, respectively).

3 Deriving $E^{\text{orig}}_{\text{DC-MRF}}$ via Dual Decomposition

The derivation given in [ZK12] of $E^{\text{orig}}_{\text{DC-MRF}}$ and the respective marginalization constraints is rather "constructive" by explicitly stating allowed configurations and considering their respective convex hull. In this section we aim for a more "analytic" derivation based on the principle of dual decomposition, which was successfully applied for discrete inference (e.g. [KPT11, SGJ11]). In order to simplify the notation we consider discrete-continuous labeling problems with only pairwise potentials, i.e.

$$E_{\text{DC-Labeling}}(x,z) = \sum_{s\sim t} \sum_{i,j} x^{ij}_{st} f^{ij}_{st}(z_s, z_t)$$

subject to $x^{ij}_{st} \in \{0,1\}$ and marginalization constraints. $z_s \in \mathbb{R}$ is the continuous unknown at each node s. Note that $E_{\text{DC-Labeling}}$ is difficult to solve due to the integrality constraints on x and the generally non-convex products appearing in the objective. Consequently, even relaxing the integrality constraints to simplex constraints on x would in general lead to a non-convex problem. The standard approach for dual decomposition is to introduce smaller, easy-to-solve subproblems and enforce consistency between the respective solutions. One way to apply the dual decomposition principle for $E_{\text{DC-Labeling}}$ is to treat each problem on an edge $(s,t) \in \mathcal{E}$ as subproblem, and therefore introduce local copies $z_{st\to s}$ and $z_{st\to t}$ of z_s and z_t for each edge (s,t). By introducing multipliers $\mu_{st\to s}$ and $\mu_{st\to t}$ for the consistency constraints $z_{st\to s} = z_s$ and $z_{st\to t} = z_t$ we obtain the following Lagrangian,

$$\begin{aligned}
L_{\text{DD-I}}(x,z;\mu) &= \sum_{s\sim t} \sum_{i,j} x^{ij}_{st} f^{ij}_{st}(z_{st\to s}, z_{st\to t}) \\
&\quad + \sum_{s\sim t} \big(\mu_{st\to s}\big(z_s - z_{st\to s}\big) + \mu_{st\to t}\big(z_t - z_{st\to t}\big) \\
&= \sum_{s\sim t} \left(\sum_{i,j} x^{ij}_{st} f^{ij}_{st}(z_{st\to s}, z_{st\to t}) - \mu^T_{st} z_{st} \right) \\
&\quad + \sum_s z_s \left(\sum_{t\in \text{out}(s)} \mu_{st\to s} + \sum_{t\in \text{in}(s)} \mu_{ts\to s} \right).
\end{aligned}$$

We also use the short-hand notation $z_{st} = (z_{st\to s}, z_{st\to t})^T$ (and similar for μ_{st}). It can be shown (after introducing Lagrange multipliers for the consistency between x_s and x_{st}, see the supplementary material) that the induced convex primal problem reads as

$$E_{\text{DC-DD-I}}(x,y) = \sum_{s,t} \sum_{i,j} (f^{ij}_{st})_\oslash (x^{ij}_{st}, y^{ij}_{st\to s}, y^{ij}_{st\to t}) \quad (3)$$

subject to

$$\begin{aligned}
x^i_s &= \sum_j x^{ij}_{st} & x^j_t &= \sum_i x^{ij}_{st} \\
y_s &= \sum_{ij} y^{ij}_{st\to s} & y_t &= \sum_{ij} y^{ij}_{st\to t}
\end{aligned} \quad (4)$$

and simplex constraints $x_s \in \Delta^L$, $x_{st} \in \Delta^{L^2}$. Note that Eq. 4 above is almost identical to Eq. 2 with the only difference the lack of state-specific unknowns y^i_s and their respective constraints. Aside from this, $E_{\text{DC-DD-I}}$ is identical to $E^{\text{orig}}_{\text{DC-MRF}}$. Since the constraints are weaker in $E_{\text{DC-DD-I}}$ than in $E^{\text{orig}}_{\text{DC-MRF}}$, we have

$$\min_{x,y} E_{\text{DC-DD-I}} \le \min_{x,y} E^{\text{orig}}_{\text{DC-MRF}}.$$

Essentially, $E_{\text{DC-DD-I}}$ is weaker than $E^{\text{orig}}_{\text{DC-MRF}}$ since in the starting point for dual decomposition we had one global continuous unknown z_{st} per edge (s,t) *agnostic to the picked state*. Thus, straightforward application of dual decomposition does *not* provide the desired result, Eq 1 together with the respective constraints.

Another approach for dual decomposition is to utilize very "fine-grained" constraints, e.g. $z_s = z^{ij}_{st\to s}$, which are only active if the respective clique state ij is picked for the edge (s,t), i.e. $x^{ij}_{st} = 1$. The fact that consistency constraints should be only conditionally active can be formulated as bilinear constraints $x^{ij}_{st}(z_s - z^{ij}_{st\to s}) = 0$ and $x^{ij}_{st}(z_t - z^{ij}_{st\to t}) = 0$, respectively. We do not pursue this particular decomposition for the following two reasons: (i) the number of dual variables would be quadratic in the number of states, therefore losing the benefits of the compact representation using the dual program, and (ii) the induced convex primal program has a large number of unknowns and constraints (depending exponentially on the degree of nodes). Consequently, we rule out this relaxation in the current work.

It turns out that the right dual decomposition formulation to arrive at $E^{\text{orig}}_{\text{DC-MRF}}$ is the following: we introduce local, state-specific copies $z^i_{st\to s}$ and $z^i_{st\to t}$ together with constraints $z_s = z^i_{st\to s}$ and $z_t = z^i_{st\to t}$. Note that e.g. the constraint $z^i_s = z^i_{st\to s}$ is only active if label i is picked at node s (i.e. $x^i_s = 1$), hence the constraint can be written again as bilinear one, $x^i_s(z_s - z^i_{st\to s}) = 0$, We have a similar constraint connecting z_t and $z^i_{st\to t}$, $x^i_t(z_t - z^i_{st\to t}) = 0$. Consequenly the Lagrangian now reads as

$$\begin{aligned}
L_{\text{DD}}(x,z;\mu) &= \sum_{s\sim t} \sum_{i,j} x^{ij}_{st} f^{ij}_{st}(z_{st\to s}, z_{st\to t}) \\
&\quad + \sum_{s\sim t} \sum_i \mu^i_{st\to s} x^i_s \big(z_s - z^i_{st\to s}\big)
\end{aligned}$$

$$+\sum_{s\sim t}\sum_{i}\mu^i_{st\to t}x^i_t\big(z_t - z^i_{st\to t}\big).$$

It is shown in the supplementary material that this particular choice of dual decomposition yields the convex relaxation $E^{\text{orig}}_{\text{DC-MRF}}$ stated in Eq. 1 (with all $f^i_s \equiv 0$). If f^i_s are not identical to zero, introducing local copies z^i_s, $z^i_{st\to s}$, and $z^i_{st\to t}$ (all representing z_s), together with respective Lagrange multipliers yields exactly to the relaxation $E^{\text{orig}}_{\text{DC-MRF}}$. In view of the remarks in the subsequent Section 4.1 it is always favorable to incorporate the unary potentials into the higher-order ones, hence we do not consider the case $f^i_s \not\equiv 0$ explicitly. Consequently, we are interested in minimizing problems of the shape

$$E_{\text{DC-MRF}}(\mathbf{x},\mathbf{y}) = \sum_{s\sim t}\sum_{i,j}(f^{ij}_{st})_{\oslash}(x^{ij}_{st}, y^{ij}_{st\to s}, y^{ij}_{st\to t}) \tag{5}$$

subject to the marginalization constraints as in $E^{\text{orig}}_{\text{DC-MRF}}$ (Eq. 1).

3.1 The Dual of $E_{\text{DC-MRF}}$ and its Interpretation

By introducing Lagrange multipliers ("messages") $p^i_{st\to s}$, $p^i_{st\to t}$, $q^i_{st\to s}$, and $q^i_{st\to t}$, for the constraints on x and y one can derive a particular dual $E^*_{\text{DC-MRF}}(p,q)$ as

$$\sum_{s\sim t}\min_{i,j}\Big\{p^i_{st\to s} + p^j_{st\to t} - (f^{ij}_{st})^*(q^i_{st\to s}, q^j_{st\to t})\Big\} \tag{6}$$

subject to the following "flow conservation" constraints,

$$\begin{aligned}\sum_{t\in\text{out}(s)} p^i_{st\to s} + \sum_{t\in\text{in}(s)} p^i_{ts\to s} &= 0\\ \sum_{t\in\text{out}(s)} q^i_{st\to s} + \sum_{t\in\text{in}(s)} q^i_{ts\to s} &= 0\end{aligned} \tag{7}$$

for all s and i. The derivation is given in the supplementary material. The interpretation of this dual energy is an extension of the one for discrete inference: the expressions $p^i_{st\to s}+p^j_{st\to t}$ adjust (or *reparametrize*) the potentials *in order to favor agreement of discrete states* between the subproblems (on edges) in terms of x. Thus, the correcting terms $p^i_{st\to s} + p^j_{st\to t}$ have the same meaning as in the dual decomposition approach for discrete MRFs.

The additional dual variables $q^i_{st\to s}$ and $q^j_{st\to t}$ modify the slope of the potential function (i.e. add a linear term to $(f^{ij}_{st})_{\oslash}$), effectively adjusting the location of the minimizer. Recall that (with $q^{ij}_{st} = (q^i_{st\to s}, q^j_{st\to t})$)

$$\begin{aligned}-(f^{ij}_{st})^*(q^{ij}_{st}) &= -\max_z\Big\{(q^{ij}_{st})^T z - f^{ij}_{st}(z)\Big\}\\ &= \min_z\Big\{f^{ij}_{st}(z) - (q^{ij}_{st})^T z\Big\},\end{aligned}$$

hence by modifying q^{ij}_{st} the location the minimizer z (and the respective objective value) is adjusted. Overall, optimizing the dual variables q *leads to agreement of the continuous unknowns* y for the subproblems in the primal energy.

As usual in relaxations for labeling problems, the dual energy given in Eq. 6 is not unique, and different dual programs can be obtained by enforcing additional (redundant) constraints in the primal formulation. The advantage of optimizing the dual $E^*_{\text{DC-MRF}}$ is the same as for inference with discrete states: the number of unknowns grows only linearly with the number of states. This comes with some cost: the number of (non-smooth) terms in the objective (or the number of constraints, depending on the exact shape of the dual) grows quadratically with the number of states.

3.2 Optimization

Despite the compactness of the dual energy Eq. 6 it turns out that it is a rather difficult energy to optimize. We used the following "generic" approach to optimize the dual: by introducing an explicit Lagrange multipliers ν_{st} for the normalization constraints, $\sum_{ij} x^{ij}_{st} = 1$, and incorporating the bounds $x^{ij}_{st} \in [0,1]$ a different dual only using exact penalization terms can be obtained,

$$\begin{aligned}E^*_{\text{DC-MRF}}(\mathbf{p},\mathbf{q},\mathbf{r}) &= -\sum_{s\sim t}\nu_{st}\\ &+\sum_{s\sim t}\sum_{i,j}\Big[\nu_{st} + p^i_{st\to s} + p^j_{st\to t} - (f^{ij}_{st})^*(q^i_{st\to s}, q^j_{st\to t})\Big]_-\end{aligned} \tag{8}$$

subject to the flow conservation constraints Eq. 7. Using Nesterov's smoothing approach for non-smooth functions [Nes05] (where we use a quadratic prox-function), we obtain a smooth approximation of $E^*_{\text{DC-MRF}}$ (still subject to linear constraints), which can be optimized e.g. by L-BFGS or FISTA [BT09]. We quickly discarded subgradient methods, since (in contrast to discrete MRFs) inference for joint discrete-continuous problems cannot efficiently performed on trees. Using only inference over edges as subproblems leads to very poor convergence.

4 Improving the DC-MRF Relaxation

4.1 The Impact of Reparametrizations

The presentation of $E^{\text{orig}}_{\text{DC-MRF}}$ (recall Eq. 1) follows the usual presentation of standard relaxations for discrete MRFs by using unary and pairwise potentials. For the standard discrete relaxation, unary potentials can be

freely moved to the pairwise ones without affecting the strength of the relaxation. This fact is exactly the basis for reparametrization approaches for discrete inference i.e. the foundation for many message-passing algorithms. In the following we will show that $E^{\text{orig}}_{\text{DC-MRF}}$ as written in Eq. 1 is weaker than necessary, and (potentially) stronger relaxations can be obtained by moving unary potentials closer to the pairwise ones. *Thus, by appropriate reparametrization the strength of the convex relaxation can be improved.*

In order to analyze the impact of reparametrization of the potentials we consider the following two expressions: first we write the unary potentials (for $t \in \text{out}(s)$ and $r \in \text{in}(s)$)

$$\begin{aligned} U^i_s \overset{\text{def}}{=} (f^i_s)_\oslash(x^i_s, y^i_s) &= (f^i_s)_\oslash\left(\sum_j x^{ij}_{st}, \sum_j y^{ij}_{st\to s}\right) \\ &= (f^i_s)_\oslash\left(\sum_j x^{ji}_{rs}, \sum_j y^{ji}_{rs\to s}\right) \end{aligned}$$

(where we made use of the marginalization constraints), and second, we introduce a reparametrized version,

$$\begin{aligned} V^i_s(\lambda) \overset{\text{def}}{=} &\sum_{t:(s,t)\in\mathcal{E}} \lambda_{st} \sum_j (f^i_s)_\oslash(x^{ij}_{st}, y^{ij}_{st\to s}) \\ &+ \sum_{r:(r,s)\in\mathcal{E}} \lambda_{rs} \sum_j (f^i_s)_\oslash(x^{ji}_{rs}, y^{ji}_{rs\to s}), \end{aligned}$$

where λ is restricted to the unit simplex (of appropriate dimension). We need the constraint that $\sum_t \lambda_{st} + \sum_r \lambda_{rs} = 1$ in order to conserve the overall cost, and $\lambda \geq 0$ (element-wise) to preserve the convexity of $\lambda_{st}(f^i_s)_\oslash$ and $\lambda_{rs}(f^i_s)_\oslash$.[2]

We have the following fact:

Proposition 1. *For all $\lambda \in \Delta$ it holds that*

$$U^i_s \leq V^i_s(\lambda).$$

Proof. As a perspective $(f^i_s)_\oslash$ is positively 1-homogeneous, i.e. $(f^i_s)_\oslash(kx, ky) = k(f^i_s)_\oslash(x, y)$ for $k \geq 0$. For any convex 1-positively homogeneous function ϕ we have

$$\begin{aligned} \phi\left(\sum_i \mathbf{x}_i\right) &\overset{\text{pos. 1-hom.}}{=} N\phi\left(\frac{\sum_i \mathbf{x}_i}{N}\right) \\ &\overset{\text{Jensen}}{\leq} \frac{N}{N}\sum_i \phi(\mathbf{x}_i) = \sum_i \phi(\mathbf{x}_i). \end{aligned}$$

[2] In discrete MRFs we have f^i_s are constant functions with value θ^i_s, hence the non-negativity constraint on λ can be dropped.

We can write U^i_s as

$$\begin{aligned} U^i_s = &\sum_{t:(s,t)\in\mathcal{E}} \lambda_{st} \underbrace{(f^i_s)_\oslash\left(\sum_j x^{ij}_{st}, \sum_j y^{ij}_{st\to s}\right)}_{=(f^i_s)_\oslash(x^i_s, y^i_s)} \\ &+ \sum_{r:(r,s)\in\mathcal{E}} \lambda_{rs} \underbrace{(f^i_s)_\oslash\left(\sum_j x^{ji}_{rs}, \sum_j y^{ji}_{rs\to s}\right)}_{=(f^i_s)_\oslash(x^i_s, y^i_s)}. \\ \leq &\sum_{t:(s,t)\in\mathcal{E}} \lambda_{st} \sum_j (f^i_s)_\oslash(x^{ij}_{st}, y^{ij}_{st\to s}) \\ &+ \sum_{r:(r,s)\in\mathcal{E}} \lambda_{rs} \sum_j (f^i_s)_\oslash(x^{ji}_{rs}, y^{ji}_{rs\to s}) = V^i_s(\lambda), \end{aligned}$$

after applying the above inequality. □

Any reparametrization induced by λ does not change our original objective, since U^i_s are $V^i_s(\lambda)$ are the same if all x^{ij}_{st} have integral (i.e. either 0 or 1) values. Nevertheless, the strength of the relaxation after dropping the integrality constraints may depend on the choice of λ. Since we are aiming for the tightest relaxation, i.e. we maximize $V^i_s(\lambda)$ with respect to $\lambda \in \Delta$, we are only interested in the largest term:

$$\begin{aligned} \max_{\lambda\in\Delta} V^i_s(\lambda) = \max\Bigg\{ &\max_{t:(s,t)\in\mathcal{E}} \Big\{ \sum_j (f^i_s)_\oslash(x^{ij}_{st}, y^{ij}_{st\to s}) \Big\}, \\ &\max_{r:(r,s)\in\mathcal{E}} \Big\{ \sum_j (f^i_s)_\oslash(x^{ji}_{rs}, y^{ji}_{rs\to s}) \Big\} \Bigg\}. \end{aligned}$$

In order to reduce notational clutter in the following we introduce the set of neighboring nodes

$$N(s) \overset{\text{def}}{=} \{t \in V : (s,t) \in \mathcal{E} \vee (t,s) \in \mathcal{E}\},$$

and replicate variables as neccessary, e.g. $x^{ij}_{st} = x^{ij}_{ts}$ and $y^{ij}_{st\to s} = y^{ji}_{ts\to s}$. Consequently, we can write more compactly

$$\max_{\lambda\in\Delta} V^i_s(\lambda) = \max_{t\in N(s)} \left\{ \sum_j (f^i_s)_\oslash(x^{ij}_{st}, y^{ij}_{st\to s}) \right\}$$

Plugging this as a replacement for the unary potentials in Eq. 1 finally yields an improved relaxation,

$$\begin{aligned} E^{\text{tight}}_{\text{DC-MRF}}(\mathbf{x},\mathbf{y}) = &\overbrace{\sum_{s,i} \max_{t\in N(s)} \left\{ \sum_j (f^i_s)_\oslash(x^{ij}_{st}, y^{ij}_{st\to s}) \right\}}^{\overset{\text{def}}{=} E_{\text{unary}}(\mathbf{x},\mathbf{y})} \\ &+ \sum_{s\sim t} \sum_{i,j} (f^{ij}_{st})_\oslash(x^{ij}_{st}, y^{ij}_{st\to s}, y^{ij}_{st\to t}) \end{aligned} \tag{9}$$

subject to the normalization and marginalization constraints. Unfortunately, in many cases $E^{\text{tight}}_{\text{DC-MRF}}$ is much more difficult to optimize, especially if the f^i_s are non-linear. In the important setting of piece-wise linear potentials tightening the relaxation is easy to achieve (see Section 4.2 below). In other cases we already observed significantly stronger relaxations (compared to the model $E^{\text{orig}}_{\text{DC-MRF}}$) by evenly distributing the unary potentials to adjacent edges. The implementation complexity for such an approach is comparable to the one for the original model with node-based unary potentials, $E^{\text{orig}}_{\text{DC-MRF}}$.

The lack of equivalence of reparamatrization also means, that introducing higher-order clique variables without associated potentials ("zero-constraints" [Wer07, SMG+08]) will in general only weakly strengthen the relaxation. In order to make better use of higher-order unknowns, lower-order potentials need to be incorporated into higher order ones, either by using a generalization of Eq. 9 beyond pairwise cliques, or by fixing the reparametrization weights in advance (e.g. using a uniform weighting).

4.2 Piece-wise Linear Unary Potentials

In this section we derive $E^{\text{tight}}_{\text{DC-MRF}}$ (Eq. 9) for unary potentials, that are piecewise linear and have bounded domain. Assume that w.l.o.g. we want to assign one continuous label $t_s \in [0,1]$ at each node s, and the respective linear potentials read as

$$U_s(t_s) = \sum_{i=1}^{L_s} \left(a^i_s + b^i_s t_s + \imath_{[l^i_s, u^i_s]}(t_s) \right), \qquad (10)$$

where a^i_s and b^i_s induce the linear cost, and $l^i_s \in [0,1]$ and $u^i_s \in [0,1]$ define the domain of the particular segment. L_s is the number of linear segments at node s. W.l.og. we will assume that $\{[l^i_s, u^i_s]\}_i$ is a partition of $[0,1]$, i.e. the unary potentials are piece-wise linear functions of t with domain $[0,1]$. The terms corresponding to the unaries in $E^{\text{tight}}_{\text{DC-MRF}}$ read as (and after verifying that $f_\oslash(x,y) = ax + by + \imath_{[lx,ux]}(y)$ for $f(t) = a + bt + \imath_{[l,u]}(t)$)

$$E_{\text{unary}}(\mathbf{x},\mathbf{y}) = \sum_{s\in\mathcal{V}} \sum_{i=1}^{L_s} \max_{t\in N(s)} \left\{ \sum_{j=1}^{L_t} \left(a^i_s x^{ij}_{st} + b^i_s y^{ij}_{st\to s} \right) + \sum_{j=1}^{L_t} \imath \left\{ y^{ij}_{st\to s} \in [l^i_s x^{ij}_{st}, u^i_s x^{ij}_{st}] \right\} \right\}$$

subject to the marginalization constraints. But for the inner maximization problem,

$$\hat{U}^i_s \stackrel{\text{def}}{=} \max_{t\in N(s)} \sum_{j=1}^{L_t} \left(a^i_s x^{ij}_{st} + b^i_s y^{ij}_{st\to s} + \imath_{[l^i_s x^{ij}_{st}, u^i_s x^{ij}_{st}]}(y^{ij}_{st\to s}) \right)$$

$$\stackrel{(1)}{=} \max_{t\in N(s)} \sum_{j=1}^{L_t} \left(a^i_s x^{ij}_{st} + b^i_s y^{ij}_{st\to s} \right) + \sum_{j=1}^{L_t} \imath_{[l^i_s x^{ij}_{st}, u^i_s x^{ij}_{st}]}(y^{ij}_{st\to s})$$

$$\stackrel{(2)}{=} \max_{t\in N(s)} \left\{ a^i_s x^i_s + b^i_s y^i_s \right\} + \sum_{j=1}^{L_t} \imath \left\{ y^{ij}_{st\to s} \in [l^i_s x^{ij}_{st}, u^i_s x^{ij}_{st}] \right\}$$

$$= a^i_s x^i_s + b^i_s y^i_s + \sum_{j=1}^{L_t} \imath \left\{ y^{ij}_{st\to s} \in [l^i_s x^{ij}_{st}, u^i_s x^{ij}_{st}] \right\}.$$

where (1) holds since we assume that any minimizer has finite cost, and (2) follows from the marginalization constraint. This means, that we only need to add respective bounds constraints to the pairwise unknowns. Plugging this into E_{unary} we obtain

$$E_{\text{unary}}(\mathbf{x},\mathbf{y}) = \sum_{s\in\mathcal{V}} \sum_{i=1}^{L_s} \left(a^i_s x^i_s + b^i_s y^i_s \right) + \sum_{s,t} \sum_{i,j} \imath \left\{ y^{ij}_{st\to s} \in [l^i_s x^{ij}_{st}, u^i_s x^{ij}_{st}] \right\}$$

subject to the marginalization constraints. By replacing

$$\sum_{s\in\mathcal{V}} \sum_{i=1}^{L_s} \left(a^i_s x^i_s + b^i_s y^i_s + \imath \left\{ y^i_s \in [l^i_s x^i_s, u^i_s x^i_s] \right\} \right)$$

with E_{unary} we have strengthened the relaxation, since $y^{ij}_{st\to s} \in [l^i_s x^{ij}_{st}, u^i_s x^{ij}_{st}]$ for all j implies $y^i_s \in [l^i_s x^i_s, u^i_s x^i_s]$ (via the marginalization constraints) but not vice versa. Overall. for piece-wise linear unary potentials it is very straightforward to obtain a stronger relaxation without having a negative impact on implementation complexity.

4.3 Numerical Illustration

Strengthening of the relaxation as described in Section 4.1 can have a huge impact in practice. We illustrate the behavior of different relaxation using a robustified, non-convex TV-L^1 energy for image denoising,

$$E_{\text{TV-}L^1}(\mathbf{z};\mathbf{f}) = \sum_s \lambda \min\{|z_s - f_s|, T\} + \sum_{s\sim t} |z_s - z_t|$$

where $\mathbf{f}$ is the given source image and T is an inlier threshold. $\lambda > 0$ is a weighting parameter. A straightforward convex relaxation would introduce two unary potential functions (since the unary cost is defined by two convex alternatives),

$$f^0_s(z_s) = \lambda |z_s - f_s| \qquad \text{and} \qquad f^1_s(z_s) = \lambda T$$

and convexify

$$\min_{\mathbf{x},\mathbf{z}} \sum_{s,i} x^i_s f^i_s(z^i_s) + \sum_{s\sim t} |z_s - z_t| \qquad (11)$$

(a) (b) (c) (d) (e)

Figure 1: The impact of reparametrization on the convex relaxation. (a) Result of the weak relaxation Eq. 12. (b) corresponding variables x_s^0 indicating whether a pixel s in an inlier or an outlier. (c) Result of the relaxation Eq. 13 and corresponding "inlier status" of pixels. Notice that (d) is largely a binary image (stronger relaxation), and (b) contains many fractional pixels (weak relaxation). (e) Baseline result using discrete inference after discretization of the continuous image intensity into 256 states.

(subject to $x_s = x_s^0 + x_s^1$, and $x_s^i \geq 0$), leading to the convex program

$$\min_{\mathbf{x},\mathbf{y}} \sum_{s,i} (f_s^i)_\oslash(x_s^i, y_s^i) + \sum_{s\sim t} |y_s - y_t| \tag{12}$$

subject to the same constraints on x and $y_s = y_s^0 + y_s^1$. This relaxation turns out to be extremely weak (and therefore useless, see Fig. 1(a,b)). A significantly stronger relaxation can be obtained by introducing four states per node (corresponding to the number of segments in the piece-linear unary potential $z_s \mapsto \min\{|z_s - f_s|, T\}$), and to move the unary potentials to the pairwise ones (according to Section 4.2), resulting in a convex program of the shape

$$E(\mathbf{x},\mathbf{y}) = \sum_{s\sim t} \sum_{i,j\in\{1,\dots,4\}} (f_{st}^{ij})_\oslash(x_{st}^{ij}, y_{st\to s}^{ij}, y_{st\to t}^{ij}) \tag{13}$$

subject to the standard discrete-continuous marginalization constraints Eq. 2. A minimizer of this energy is illustrated in Fig. 1(c,d), together with a baseline solution using a fine discretization of the continuous label space in Fig. 1(e).

5 Layered Depth Denoising

Assume we are given noisy and partially missing depth (or disparity) observations $\hat{z} : V \to \mathbb{R}$, and we want to denoise the map and jointly to segment the image domain into layers based on the observations. We assume that layers also extend to occluded image regions, where another layer is closer to the depth sensor (see e.g. [Wei97]). We derive the energy for a background layer (with unknown depth $z^b : \Omega \to \mathbb{R}_0^+$) and a foreground layer (with depth $z^f : \Omega \to \mathbb{R}_0^+$). The discrete-continuous labeling problem is now

$$E(\mathbf{x},\mathbf{z}) = \sum_s \sum_i x_s^i f_s^i(z_s^b, z_s^f) + \sum_{s\sim t} \sum_{i,j} x_{st}^{ij} f^{ij}(z_s^b, z_t^b, z_s^f, z_t^f). \tag{14}$$

Note that the continuous unknown is a 2-vector per pixel and not just a single scalar. The discrete choice i at each pixel s is either that background is observed directly ($i = 0$, which implies that foreground depth is not existent) or foreground occludes background ($i = 1$, which means that both z^b and z^f are defined for this pixel). Given the quantized nature of disparities reported by the Kinect depth sensor we model f_s^i as "capped" L^1-penalty,

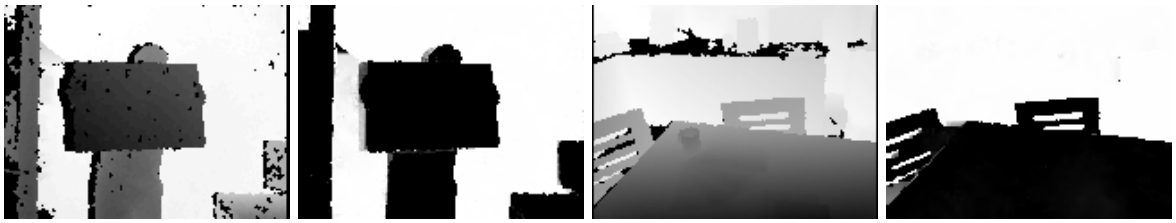
$$f_s^0(z_s^b, z_s^f) = m_s \alpha \left[|z_s^b - \hat{z}_s| - \delta \right]_+$$
$$f_s^1(z_s^b, z_s^f) = m_s \alpha \left[|z_s^f - \hat{z}_s| - \delta \right]_+ + \imath\{z_s^b \leq z_s^f\},$$

where δ defines the quantization level. If foreground is directly visible we enforce that the background depth is behind the foreground. α is a weighting parameter for data fidelity, and $m_s \in \{0, 1\}$ is a mask indicating whether a depth value $\hat{z}_s$ is available for pixel s. We essentially assume smoothly varying depth for each layer and utilize a homogeneous (quadratic) regularizer to penalize spatial discontinuities. Further, we assume smooth segmentation boundaries between foregound and background, leading to pairwise potentials

$$f^{00}(z_s^b, z_t^b, z_s^f, z_t^f) = \left(z_s^b - z_t^b\right)^2$$
$$f^{01}(z_s^b, z_t^b, z_s^f, z_t^f) = f^{10}(z_s^b, z_t^b, z_s^f, z_t^f) = \left(z_s^b - z_t^b\right)^2 + \beta$$
$$f^{11}(z_s^b, z_t^b, z_s^f, z_t^f) = \left(z_s^b - z_t^b\right)^2 + \left(z_s^f - z_t^f\right)^2.$$

$\beta > 0$ is the cost paid to switch between foreground and background and corresponds to the smoothness parameter in standard binary segmentation.

Note that f_s^i is already piecewise linear, if one neglects the constraint $z_s^b \leq z_s^f$ in f_s^1. In order to strengthen the convex relaxation, we move the unary constraints to the pairwise potentials as indicated in Section 4.2.



(a) (b) (c) (d)

Figure 2: (a, c) Input depth maps (black pixels indicating missing values). (b,d) "Background only" marginals $x_s^0 = \sum_k x_s^{0,k}$.

Since we have 3 branches in the piecewise linear function $z \mapsto \left[|z_s^b - \hat{z}_s| - \delta\right]_+$, we obtain $2 \cdot 3 = 6$ discrete states per pixel s. We use a combined index (i,k) (or (j,l)), where i (respectively j) indicates the absence ($i = 0$) or presence ($i = 1$) of foreground. After moving the nonlinearities (constraints) from the unary potentials to the pairwise terms, we obtain the following new pairwise potentials for an edge st (dropping the explicit arguments $(z_s^b, z_t^b, z_s^f, z_t^f)$ to $f_{st}^{ij,kl}$ for brevity)

$$\begin{aligned}
f_{st}^{00,kl} &= \left(z_s^b - z_t^b\right)^2 + \imath_{R_{st}^{kl}}(z_s^b, z_t^b) \\
f_{st}^{01,kl} &= \left(z_s^b - z_t^b\right)^2 + \beta + \imath\{z_t^f \leq z_t^b\} + \imath_{R_{st}^{kl}}(z_s^b, z_t^f) \\
f_{st}^{10,kl} &= \left(z_s^b - z_t^b\right)^2 + \beta + \imath\{z_s^f \leq z_s^b\} + \imath_{R_{st}^{kl}}(z_s^f, z_t^b) \\
f_{st}^{11,kl} &= \left(z_s^b - z_t^b\right)^2 + \left(z_s^f - z_t^f\right)^2 \\
&\quad + \imath\left\{z_s^f \leq z_s^b,\, z_t^f \leq z_t^b\right\} + \imath_{R_{st}^{kl}}(z_s^f, z_t^f),
\end{aligned}$$

where

$$R_{st}^{kl} \stackrel{\text{def}}{=} [l_s^k, u_s^k] \times [l_t^l, u_t^l]$$

is a feasible rectangle in 2D. We have $l_s^0 = 0$, $u_s^0 = l_s^1 = \hat{z}_s - \delta$, $u_s^1 = l_s^2 = \hat{z}_s + \delta$, and $u_s^2 = \infty$ (or some maximal depth) for all s. Finally,

$$\begin{aligned}
f_s^{0,0}(z_s^b, z_t^f) &= \alpha(\hat{z}_s - z_s^b) & f_s^{1,0}(z_s^b, z_t^f) &= \alpha(\hat{z}_s - z_s^f) \\
f_s^{0,1}(z_s^b, z_t^f) &= 0 & f_s^{1,1}(z_s^b, z_t^f) &= 0 \\
f_s^{0,2}(z_s^b, z_t^f) &= \alpha(z_s^b - \hat{z}_s) & f_s^{1,2}(z_s^b, z_t^f) &= \alpha(z_s^f - \hat{z}_s).
\end{aligned}$$

We utilize the smoothing technique sketched in Section 3.2 and use both L-BFGS and FISTA for optimization. Figs. 2(a,c) illustrates two input depth maps (with missing data), and Figs. 2(b,d) display the "only background" pseudo-marginals, $x_s^0 = \sum_k x_s^{0,k}$. The results are largely binary indicating a strong relaxation. Convergence is rather slow for—what we assume–the following reasons: (i) only the sparse set of strong depth discontinuities determines the layer segmentation, (ii) computation of the gradients for the smooth approximation requires to solve a small-scale QP (which we solve by exhaustive case analysis), (iiI) the number of dual variables is $10L$ per node/pixel (in contrast to just $2L$ for discrete inference with the same number of discrete states), and (iv) the overall objective is nonlinear. Fig. 2 illustrate obtained results for example depth maps (normalized to a $[0,1]$ range) for $\alpha = 1$, $\beta = 4/1000$, and $\delta = 2/1000$.

Note that in this particular application it is non-trivial to initialize alternation-based methods: it is not obvious how to initialize the segmentation due to the lack of a per-pixel layer preference, and starting with an initial estimate for the depth is also challenging due to missing and occluded data in the input, which has to be filled in.

6 Conclusion

In this work we deepened the understanding of convex relaxations for joint discrete-continuous inference problems by deriving the relaxation via dual decomposition. Further, we obtained several insights how to properly strengthen the convex relaxations, which immediately points to fundamental differences between convex LP relaxations for inference problems with only discrete states and relaxations addressing discrete-continuous state spaces.

Given the improved understanding of the convex discrete-continuous formulation we aim for more efficient inference methods in the future based on the message passing principle. While standard message passing essentially leads to agreement of costs in overlapping cliques, an extended discrete-continuous message passing scheme will likely be based on joint agreement of costs and location of minimizers. Achieving this agreement between overlapping cliques in an efficient (e.g. closed-form) manner is ongoing research.

Acknowledgements

I am grateful to Andrew Fitzgibbon and Pushmeet Kohli for discussions, and to Nevena Lazic for detailed feedback on the manuscript.

References

[BR96] M. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *IJCV*, 19(1):57–92, 1996.

[BSA98] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *Proc. CVPR*, pages 434–441, 1998.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2:183–202, 2009.

[BZ87] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, 1987.

[CV01] T. F. Chan and L. Vese. Active contours without edges. *IEEE Trans. Image Processing*, 10(2):266–277, 2001.

[DLR77] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[For65] E. Forgey. Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics*, 21:768, 1965.

[GLY95] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14:211–226, 1995.

[Har58] H. Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14:174–194, 1958.

[IM09] A. Ihler and D. McAllester. Particle belief propagation. In *AISTATS*, pages 256–263, 2009.

[KPT11] N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):531–552, 2011.

[KS04] S.B. Kang and R. Szeliski. Extracting view-dependent depth maps from a collection of images. *IJCV*, 58(2):139–163, 2004.

[LRRB10] V. Lempitsky, C. Rother, S. Roth, and A. Blake. Fusion moves for Markov random field optimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(8):1392–1405, 2010.

[LSR+10] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *Proc. BMVC*, pages 104.1–11, 2010.

[MS89] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.

[Nes05] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Programming*, 103:127–152, 2005.

[PHMU11] J. Peng, T. Hazan, D. McAllester, and R. Urtasun. Convex max-product algorithms for continuous MRFs with applications to protein folding. In *Proc. ICML*, 2011.

[SGJ11] D. Sontag, A. Globerson, and T. Jaakkola. *Optimization for Machine Learning*, chapter Introduction to Dual Decomposition for Inference. MIT Press, 2011.

[SMG+08] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, 2008.

[SSB10] D. Sun, E. Sudderth, and M. Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In *NIPS*, pages 2226–2234, 2010.

[WA94] J.Y.A. Wang and E.H. Adelson. Representing moving images with layers. *IEEE Trans. Image Proc.*, 3(5):625–638, 1994.

[Wei97] Y. Weiss. Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *Proc. CVPR*, pages 520–527, 1997.

[Wer07] T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(7), 2007.

[ZK12] C. Zach and P. Kohli. A convex discrete-continuous approach for Markov random fields. In *Proc. ECCV*, 2012.