Continuous Multiclass Labeling Approaches and Algorithms

J. Lellmann and C. Schnörr
{lellmann,schnoerr}@math.uni-heidelberg.de

Image and Pattern Analysis & HCI
Dept. of Mathematics and Computer Science
University of Heidelberg
ipa.iwr.uni-heidelberg.de

February 2010

Abstract

We study convex relaxations of the image labeling problem on a continuous domain with regularizers based on metric interaction potentials. The generic framework ensures existence of minimizers and covers a wide range of relaxations of the originally combinatorial problem. We focus on two specific relaxations that differ in flexibility and simplicity – one can be used to tightly relax any metric interaction potential, while the other one only covers Euclidean metrics but requires less computational effort. For solving the nonsmooth discretized problem, we propose a globally convergent Douglas-Rachford scheme, and show that a sequence of dual iterates can be recovered in order to provide a posteriori optimality bounds. In a quantitative comparison to two other first-order methods, the approach shows competitive performance on synthetical and real-world images. By combining the method with an improved binarization technique for nonstandard potentials, we were able to routinely recover discrete solutions within 1%–5% of the global optimum for the combinatorial image labeling problem.

1 Problem Formulation

The multi-class image labeling problem consists in finding, for each pixel x in the image domain $\Omega \subseteq \mathbb{R}^d$ a label $\ell(x) \in \{1, \dots, l\}$ which assigns one of l class labels to x so that the labeling function ℓ adheres to some local data fidelity as well as nonlocal spatial coherency constraints.

This problem class occurs in many applications, such as segmentation, multiview reconstruction, stitching, and inpainting [PCF06]. We consider the variational formulation

$$\inf_{\ell:\Omega\to\{1,\dots,l\}} f(u),\ f(u) := \underbrace{\int_\Omega s(x,\ell(x))dx}_{\text{data term}} + \underbrace{J(\ell).}_{\text{regularizer}} \tag{1}$$

The *data term* assigns to each possible label $\ell(x)$ a *local cost* $s(x, \ell(x))$, while the *regularizer* J enforces the desired spatial coherency. We will in particular be interested in regularizers that penalize the weighted length of boundaries between regions

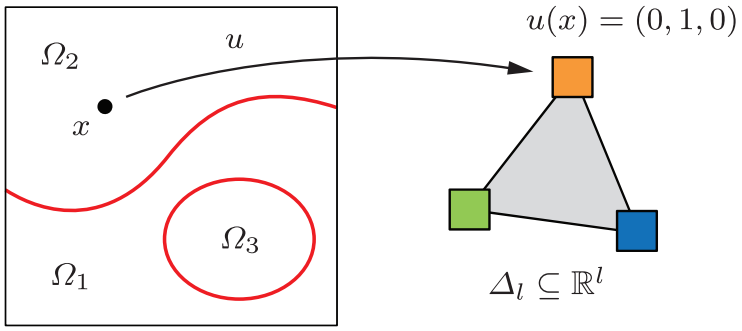

Figure 1: Convex relaxation of the multiclass labeling problem. The assignment of one unique label to each point in the image domain Ω is represented by a vector-valued function $u : \Omega \to \mathbb{R}^l$. Ideally, u partitions the image into l sets by assuming one of the unit vectors $\{e^1, \dots, e^l\}$ everywhere. By relaxing this set to the unit simplex Δ_l, the originally combinatorial problem can be treated in a convex framework.

of constant labeling. Minimizing f is an inherently combinatorial problem, as there is a discrete decision to be made for each point in the image.

In the fully discrete setting, the problem can be expressed in terms of a Markov Random Field [Win06] with a discrete state space, where the data and regularization terms can be thought of as unary and binary potentials, respectively. For graph-based discretizations of J, the resulting objective only contains terms that depend on the labels at one or two points, and the problem can be approached with fast graph cut-based methods. Unfortunately, this scheme introduces an anisotropy [Boy03] and thus does not represent isotropic regularizers well. Using ternary or higher-order terms, the metrication error can be greatly reduced, but graph-based methods then cannot be directly applied.

However, it can be shown that even in the graph-based representation the problem is NP-hard for relatively simple J [BVZ01], so we cannot expect to easily derive fast solvers for this problem. This is in part caused by the discrete nature of the feasible set. In the following, we will relax this set. This allows to solve the problem in a globally optimal way using convex optimization methods. On the downside, we cannot expect the relaxation to be exact for any problem instance, i.e. we might get non-discrete (or discrete but suboptimal) solutions.

There are several choices for the relaxation method, of which in our opinion the following is the most transparent (Fig. 1): Identify label i with the i-th unit vector $e^i \in \mathbb{R}^l$, set $E := \{e^1, \dots, e^l\}$, and solve

$$\inf_{u:\Omega \to E} f(u)\,, \quad f(u) := \int_\Omega \langle u(x), s_i(x) \rangle dx + J(u)\,. \tag{2}$$

The data term is now linear in u and fully described by the vector

$$s(x) := (s_1(x), \dots, s_l(x))^\top := (s(x, 1), \dots, s(x, l))^\top . \tag{3}$$

Due to the linearization, the local costs s may be arbitrarily complex, possibly derived from a probabilistic model, without affecting the overall problem class.

In this form, we *relax* the label set by allowing u to take "intermediate" values in

the convex hull of the original label set. This is just the unit simplex Δ_l,

$$\begin{aligned} \Delta_l &:= \operatorname{conv}\{e^1, \dots, e^l\} \\ &= \{a \in \mathbb{R}^l | a \geqslant 0, \sum_{i=1}^{l} a_i = 1\}. \end{aligned} \tag{4}$$

The problem is then considered on the relaxed feasible set $\mathcal{C}$,

$$\mathcal{C} := \{u \in \mathrm{BV}(\Omega)^l, u(x) \in \Delta_l \text{ a.e.}\}. \tag{5}$$

Assuming we can extend the regularizer J to the whole relaxed set $\mathcal{C}$, we get the relaxed problem

$$\inf_{u \in \mathcal{C}} f(u), \ f(u) := \int_\Omega \langle u(x), s_i(x) \rangle dx + J(u). \tag{6}$$

If J can be made convex, the overall problem is convex as well, and there is a good chance of it being computationally tractable. In addition, J should ideally have a closed expression, or at least lead to a computationally tractable problem.

Whether these points are satisfied strongly depends on the way a given regularizer is *extended* to the relaxed set. The prototypical example for such a regularizer is the *total variation*,

$$\mathrm{TV}(u) = \int_\Omega \|Du\|. \tag{7}$$

Note that u may be discontinuous, so the gradient Du has to be understood in a distributional sense. Much of the popularity of TV stems from the fact that it allows to include *boundary-length* terms: The total variation of the indicator function $\chi_\mathcal{S}$ of a set $\mathcal{S}$ is just the boundary length (or *perimeter*) of $\mathcal{S}$,

$$\mathrm{TV}(u) = \mathrm{Per}(\mathcal{S}). \tag{8}$$

In this paper, we will in more generality consider ways to construct regularizers which penalize interfaces between two adjoining regions with labels $i \neq j$ according to the *perimeter* (i.e. length or area) of the interface weighted by an *interaction potential* $d(i, j)$ depending on the labels. The simplest case is the "Potts" model, where $d(i, j) = \chi_{\{i=j\}}$, i.e. the regularizer corresponds to the total interface length, as seen above for the total variation.

As a prime motivation for our work, consider the two-class case $l = 2$ with $J = \mathrm{TV}$. As here $u_2 = 1 - u_1$, we may pose the relaxed problem in the form

$$\min_{u' \in \mathrm{BV}(\Omega), 0 \leqslant u' \leqslant 1} \int_\Omega \langle u'(x), s_1(x) - s_2(x) \rangle dx + 2\,\mathrm{TV}(u'), \tag{9}$$

i.e. $u'(x)$ is a scalar. This formulation is also referred to as *continuous cut* in analogy to graph cut methods. It can be shown [CEN06] that while there may be non-discrete solutions of the relaxed problem, a *discrete* – i.e. $u'(x) \in \{0, 1\}$ – global optimal solution can be recovered from *any* solution of the relaxed problem. We can thus reduce the *combinatorial* problem to a *convex* problem. While there are reasons to believe that this procedure cannot be extended to the multi-class case, we may still hope for "nearly" discrete solutions.

In view of the mathematical background sketched below in Sect. 2, the reader may ask why this paper focuses on *continuous* labeling approaches. The reason is that

solutions to the labeling problem provide a transition from the continuous world of visual signals to a (primitive) symbolic representation in terms of labels, and it seems somewhat myopic to ignore intrinsically continuous concepts like signal geometry. An obvious example is the length of level lines of a multidimensional signal which translates to well-defined measures through the coarea formula (see below). It seems hard to understand and to come up with reasonable discrete representations without resorting to such continuous concepts.

1.1 Related Work

Formulations of the labeling problem can be categorized based on

1. whether they tackle the binary (two-class) or the much harder multiclass problem, and

2. whether they rely on a graph representation or are formulated in a continuous framework.

An early analysis of a variant of the *binary continuous* cut problem and the associated dual *maximal flow* problem was done by Strang [Str83]. Chan et al. [CEN06] pointed out that by thresholding a nonbinary result of the relaxed, convex problem at almost any threshold one can generate binary solutions of the original, combinatorial problem (this can be carried over to *any* threshold and to slightly more general regularizers [Ber09] [ZNF09]). The proof heavily relies on the coarea formula [FR60], which unfortunately does not transfer to the multiclass case. The binary continuous case is also closely related to the thoroughly analyzed Mumford-Shah [MS89] and Rudin-Osher-Fatemi [ROF92] problems, where a quadratic data term is used.

For the *graph-based* discretization, the binary case can be formulated as a minimum-cut problem on a grid graph, which allows to solve the problem exactly and efficiently for a large class of metrics using graph cuts [KB05, BK04]. Graph cut algorithms have become a working horse in many applications as they are very fast for medium sized problems. Unfortunately they offer hardly any potential for parallelization. As mentioned in the introduction, the graph representation invariably introduces a grid bias [Boy03] (Fig. 2). While it is possible to reduce the resulting artifacts by using larger neighborhoods, this greatly increases graph size as well as processing time.

Prominent methods to handle the graph-based *multiclass* case rely on finding a local minimum by solving a sequence of binary graph cuts [BVZ01] (see [KT07] for a recent generalization). These methods have recently been extended to the continuous formulation [TPCB08] with similar theoretical performance [Ols09]. Our results can be seen as a continuous analogon to [Ish03], where it was shown that potentials of the form $d(i,j) = |i-j|$ can be exactly formulated as a cut on a multi-layered graph. An early analysis can be found in [KT99], where the authors also derive suboptimality bounds of a linear programming relaxation for metric distances. All these methods rely on the graph representation with pairwise potentials.

In this paper, we will focus on the *continuous multiclass* setting with higher order potentials in the discretization. Closely related to our approach is [CCP08]. In contrast to approaches that rely on a linear ordering of the labels [Ish03, BT09], the authors represent labels in a higher-dimensional space, which allows to formulate interaction potentials of the form $d(i,j) = \sigma(|i-j|)$ with nondecreasing, positive, concave σ. They provide a thorough analysis of the continuous model and propose a relaxation

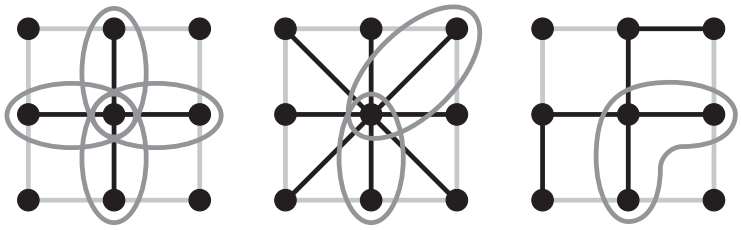

Figure 2: Discretization schemes. **Left to right:** Graph-based with 4- and 8-neighborhood; higher order potentials. In graph-based approaches the regularizer is discretized using terms depending on the labels of at most two neighboring points. This leads to artifacts as isotropic metrics are not approximated well. Using higher-order terms the discrete functional more closely approximates the continuous functional (Fig. 3).

Figure 3: Effect of the discretization on an inpainting problem. **Left to right:** Input image with unknown black region; graph-based method using pairwise potentials (α-β-swap code from [SZS$^+$06]); proposed method using higher order potentials. Due to the introduced anisotropy, the graph-based method shows a bias towards axis-parallel edges.

based on the convex envelope, which gives almost discrete solutions in many cases. We will extend this approach to a more general class of regularizers.

Recently the same authors proposed a "Fast Primal-Dual" algorithm with proven convergence to solve the associated saddle point problem [PCBC09a]. By lifting the objective to a higher dimensional space, it turns out that the same method can be used to solve many problems also with nonlinear data term [PCBC09b].

Our approach is a generalization of [ZGFN08, LKY$^+$09], where a similar linearization is used with the regularizer restricted to the Potts distance, and with less strong convergence results. These methods have also been extended to segmentation on manifolds [DFPH09].

Regarding the optimization methods, using Nesterov's method [Nes04] for the labeling problem was proposed in [LKY$^+$09]. An earlier analysis of the method in the context of ℓ_1-norm and TV minimization can be found in [WABF07]. In [BBC09] the method is applied to a class of general ℓ_1 regularized problems. In [GBO09] a predecessor of the proposed Douglas-Rachford approach was presented in the Split Bregman framework [Set09a] and restricted to the two-class case. We provide an extension to the multi-class case, together with complete characterization of the convergence and a sound stopping criterion.

1.2 Contribution

Our main contributions will be the following:

1. We formulate natural requirements on the regularizer J and show their implications on the choice of the interaction potential d (Sect. 3). In particular, d must necessarily be a metric under these requirements (Prop. 1).

2. Given such a metric, we study two possible approaches to extend regularizers J on the relaxed set (Sect. 4):

   - The "envelope" approach, which is a generalization of the method recently suggested by Chambolle et al. (Sect. 4.3). While there is no simple closed form expression, we show that it can be used to construct a true extension for *any* metric d (Prop. 5).
   - The "Euclidean distance" approach (Sect. 4.4), which yields exact extensions for Euclidean metrics d only but has a closed form expression. We review some methods for the approximation of non-Euclidean metrics.

   We provide a unified continuous framework and show existence of a minimizer.

3. Both approaches lead to non-smooth convex problems, which can be studied in a very general saddle point formulation (Sect. 5). Within this framework, we propose an improved binarization technique for nonstandard potentials to recover approximate solutions for the non-relaxed problem (Sect. 5.6).

4. We provide and analyze two different methods that are capable of minimizing the specific class of saddle point problems (Sect. 6):

   - A specialization of a method for nonsmooth optimization as suggested by Nesterov (Sect. 6.2). The method is virtually parameter-free and provides explicit a priori and a posteriori optimality bounds.
   - A Douglas-Rachford splitting approach (Sect. 6.3). We show that the approach allows to compute a sequence of dual iterates that provide an optimality bound and stopping criterion in form of the primal-dual gap.

   Both methods are highly parallelizable and are shown to converge.

5. Finally, we illustrate and compare our methods with the primal-dual technique from [PCBC09a] under a wide range of conditions and demonstrate the applicability on real-world problems (Sect. 7).

In contrast to existing graph-based methods, we provide a continuous and isotropic formulation, while in comparison with existing continuous approaches, we provide a unified framework for arbitrary, non-uniform metrics d. The Euclidean metric method and Nesterov optimization have been announced in less generality in [LKY+09, LBS09].

2 Mathematical Preliminaries

In the following sections we provide a reference of the notation used and a brief introduction to the concept of functions of bounded variation.

2.1 Notation

In the following, superscripts v^i denote a collection of vectors or columns of a matrix, while subscripts v_k denote vector components or rows of a matrix, i.e. for $A \in \mathbb{R}^{d\times l}$ we denote

$$A \quad = \quad (a^1|\dots|a^l) = (a_1|\dots|a_d)^\top . \tag{10}$$

We will frequently make use of the Kronecker product $\otimes : \mathbb{R}^{n_1\times m_1} \times \mathbb{R}^{n_2,m_2} \to \mathbb{R}^{(n_1n_2)\times(m_1m_2)}$ in order to formulate all results for arbitrary dimensions. We denote by $\Delta_l := \{x \in \mathbb{R}^l | x \geqslant 0, e^\top x = 1\}$ the unit simplex in $\mathbb{R}^l$, where $e := (1,\dots,1)^\top \in \mathbb{R}^l$. I_n is the identity matrix in $\mathbb{R}^n$ and $\|\cdot\|$ the usual Euclidean norm for vectors resp. the Frobenius norm for matrices. $\mathcal{B}_r(x)$ denotes the ball of radius r in x, and S^{d-1} the set of $x \in \mathbb{R}^d$ with $\|x\| = 1$. The indicator function $\chi_{\mathcal{S}}(x)$ of a set $\mathcal{S}$ is defined as $\chi_{\mathcal{S}}(x) = 1$ iff $x \in S$ and $\chi_{\mathcal{S}}(x) = 0$ otherwise. By $\delta_{\mathcal{S}} := +\infty \cdot \chi_{\mathcal{S}}$ we denote the corresponding characteristic function. For a convex set $\mathcal{C}$, $\sigma_{\mathcal{C}}(u) := \sup_{v\in\mathcal{C}} \langle u, v\rangle$ is the support function from convex analysis. $\mathcal{J}(u)$ denotes the classical Jacobian of u.

$C_c^1(\Omega)$ is the space of continuously differentiable functions on Ω with compact support. As usual $\mathcal{L}^d$ denotes the d-dimensional Lebesgue measure, while $\mathcal{H}^k$ denotes the k-dimensional Hausdorff measure.

2.2 Total Variation and BV

The total variation will be our main tool to construct the regularizer J. For a differentiable scalar-valued function u, the total variation is simply the integral over the norm of its gradient:

$$\mathrm{TV}(u) \quad = \quad \int_\Omega \|\nabla u\| dx\,. \tag{11}$$

As u is the designated labeling function, which is ideally should be piecewise constant, the differentiability and continuity assumptions have to be dropped. In the following we will shortly review the general definition of the total variation and its properties.

We require the image domain $\Omega \subseteq \mathbb{R}^d$ to be a bounded open domain with compact Lipschitz boundary, that is the boundary can locally be represented as the graph of a Lipschitz function.For simplicity, we will assume in the following that $\Omega = (0,1)^d$.

We consider general vector valued functions $u : \Omega \to \mathbb{R}^l$ which are locally absolutely integrable, i.e. $u \in L^1_{\mathrm{loc}}(\Omega)^l$. As Ω is bounded this is equivalent to $u \in L^1(\Omega)^l$.

For any such function u, the *total variation* $\mathrm{TV}(u)$ is defined in a dual way as [AFP00, (3.4)]

$$\begin{aligned}
\mathrm{TV}(u) \quad &:= \quad \sup_{v\in C_c^1(\Omega)^{d\times l}, \|v(x)\|\leqslant 1 \text{ a.e.}} \sum_{j=1}^{l} \int_\Omega u_j \operatorname{div} v_j dx \\
&= \quad \sup_{v\in C_c^1(\Omega)^{d\times l}, \|v(x)\|\leqslant 1 \text{ a.e.}} \int_\Omega \langle u, \operatorname{Div} v\rangle dx\,, \qquad (12)\\
\operatorname{Div} v \quad &:= \quad \left(\operatorname{div} v^1, \dots, \operatorname{div} v^l\right)^\top ,
\end{aligned}$$

where $\|v\|_\infty = \inf\{c \in [0,\infty] | \|v(x)\| \leqslant c, \mathcal{L}^d - \text{a.e. in}\Omega\}$ [AFP00, 1.16]. This definition can be derived for continuously differentiable u by extending (11) to vector

valued u,

$$\mathrm{TV}(u) \quad = \quad \int_\Omega \|\mathcal{J}(u)\| dx\,, \tag{13}$$

replacing the norm by its dual formulation and applying Gauss' theorem. If $\mathrm{TV}(u) < \infty$, u is said to be of *bounded variation*. The vector space of all such functions is denoted by $\mathrm{BV}(\Omega)^l$:

$$\mathrm{BV}(\Omega)^l \quad = \quad \left\{ u \in \left(L^1(\Omega)\right)^l \,|\, \mathrm{TV}(u) < \infty \right\}. \tag{14}$$

Equivalently, $u \in \left(\mathrm{BV}(\Omega)\right)^l$ iff $u \in L^1(\Omega)^l$ and its distributional derivative corresponds to a finite Radon measure; i.e. $u_j \in L^1(\Omega)$ and there exist $\mathbb{R}^d$-valued measures $Du_j = (D_1 u_j, \ldots, D_d u_j)$ on the Borel subsets $\mathcal{B}(\Omega)$ of Ω such that [AFP00, p.118]

$$\begin{aligned} \sum_{j=1}^{l} \int_\Omega u_j \,\mathrm{div}\, v^j dx = \ldots \\ - \sum_{j=1}^{l} \sum_{i=1}^{d} \int_\Omega v_i^j dD_i u_j \;\forall v \in (C_c^1(\Omega))^{d\times l}\,. \end{aligned} \tag{15}$$

These measures form the distributional gradient Du, which is again a measure that vanishes on any $\mathcal{H}^{(d-1)}$-negligible set. If $u \in \mathrm{BV}(\Omega)$ we have $|Du|(\Omega) = \mathrm{TV}(u)$, where $|Du|$ is the total variation of Du in the measure-theoretic sense [AFP00, 3.6].

The total variation of characteristic functions has an intuitive geometrical interpretation: For a Lebesgue-measurable subset $\mathcal{S}$ of $\mathbb{R}^l$, its *perimeter* is defined as the total variation of its characteristic function,

$$\mathrm{Per}(\mathcal{S}) \quad := \quad \mathrm{TV}(\chi_\mathcal{S})\,. \tag{16}$$

Assuming the boundary $\partial\mathcal{S}$ is sufficiently regular, $\mathrm{Per}(\mathcal{S})$ is just the classical length ($d = 2$) or area ($d = 3$) of the boundary.

2.3 Properties of TV and Compactness

We will review the most important ingredients for proving existence of minimizers for variational problems involving TV and BV.

Convexity. As TV is the pointwise supremum of a family of linear functions, it is *convex* and *positively homogeneous*, i.e. $\mathrm{TV}(\alpha u) = \alpha\,\mathrm{TV}(u)$ for $\alpha > 0$.

Lower Semicontinuity. A functional J is said to be *lower semicontinuous* with respect to some topology, if for any $u_k \to u$,

$$\liminf_{k\to\infty} J(u_k) \quad \geqslant \quad J(u)\,. \tag{17}$$

It can be shown that for fixed Ω, the total variation TV is well-defined on $L^1_{\mathrm{loc}}(\Omega)^l$ and lower semicontinuous in $\mathrm{BV}(\Omega)^l$ w.r.t. the $L^1_{\mathrm{loc}}(\Omega)^l$ topology [AFP00, 3.5,3.6]; hence also in $L^1(\Omega)^l$ due to the boundedness of Ω.

Compactness. For BV, instead of the norm topology

$$\|u\|_{\mathrm{BV}} \quad := \quad \int_\Omega \|u\| dx + \mathrm{TV}(u)\,, \tag{18}$$

which makes $\mathrm{BV}(\Omega)^l$ a Banach space but is often too strong, one frequently uses the weak* convergence: Define $u^k \to u$ weakly* iff

- $u, u^k \in \mathrm{BV}(\Omega)^l \, \forall k \in \mathbb{N}$,
- $(u^k) \to u$ in $L^1(\Omega)$ and
- $(Du^k) \to Du$ weakly* in measure, i.e.

$$\forall v \in C_0(\Omega) : \lim_{k\to\infty} \int_\Omega v dDu^k = \int_\Omega v dDu \,. \tag{19}$$

For $u, u^k \in \mathrm{BV}(\Omega)^l$ this is equivalent to $u^k \to u$ in $L^1(\Omega)^l$ and u^k bounded in $\mathrm{BV}(\Omega)^l$ [AFP00, 3.13]. For the weak* topology in BV, a compactness result holds [AFP00, 3.23]: If $(u^k) \subset \mathrm{BV}(\Omega)^l$ and $\|u^k\|_{\mathrm{BV}} \leqslant C < \infty$ uniformly in $\mathrm{BV}(\Omega)^l$, then u_k contains a subsequence weakly*-converging to some $u \in \mathrm{BV}(\Omega)^l$.

2.4 General Functionals on BV

We will now review how general functionals depending on the distributional gradient Du can be defined. Recall that for any $u \in \mathrm{BV}(\Omega)^l$ the distributional gradient Du is a measure and thus can be uniquely decomposed into three mutually singular measures

$$Du \quad = \quad D^a u + D^j u + D^c u \,, \tag{20}$$

that is: An *absolutely continuous* part D^a, the *jump* part D^j, and the *Cantor* part D^c. We will give a short intuitive explanation, see [AFP00, 3.91] for the full definitions.

The D^a part is absolutely continuous with respect to the d-dimensional Lebesgue measure $\mathcal{L}^d$, i.e. it vanishes on any $\mathcal{L}^d$-negligible set. It captures the "smooth" variations of u: in any neighborhood where u has a (weak) Jacobian $\mathcal{J}(u)$, the jump and Cantor parts vanish and

$$Du = D^a u = \mathcal{J}(u)\mathcal{L}^d \,. \tag{21}$$

The jump part D^j is concentrated on the set of points where locally u jumps between two values u^- and u^+ along a $(d-1)$-dimensional surface with normal $\nu_u \in S^{d-1}$ (unique up to a change of sign). In fact, there exist Borel functions $u^+, u^-, \nu_u : J_u \to \mathbb{R}^l \times \mathbb{R}^l \times S^{d-1}$ such that [AFP00, 3.78, 3.90]

$$D^j u = Du \llcorner J_u = (u^+ - u^-) \otimes \nu_u \mathcal{H}^{d-1} \llcorner J_u \,, \tag{22}$$

where $\mathcal{H}^{d-1} \llcorner J_u$ denotes the restriction of the $(d-1)$-dimensional Hausdorff measure on the jump set J_u.

The Cantor part D^c captures anything that is left. This is for example necessary if u jumps on a set with fractional Hausdorff dimension $s \in (d-1, d)$.

As an important consequence of the mutual singularity, the total variation decomposes into $|Du| = |D^a u| + |D^j u| + |D^c u|$. Using this idea, one can define functionals depending on the distributional gradient Du. Let $u \in \mathrm{BV}(\Omega)^l$ and define, for some $\Psi : \mathbb{R}^{d\times l} \to \mathbb{R}$,

$$\begin{aligned} J(u) \quad := \quad & \int_\Omega \Psi(Du) := \int_\Omega \Psi(\mathcal{J}(u)(x)) dx + \dots \\ & \int_{J_u} \Psi_\infty \left(\nu_u(x) \left(u^+(x) - u^-(x) \right)^\top \right) d\mathcal{H}^{d-1} + \dots \\ & \int_\Omega \Psi_\infty \left(\frac{D^c u}{|D^c u|} \right) d|D^c u| \,. \end{aligned} \tag{23}$$

Here Ψ_∞ is the recession function $\Psi_\infty(x) = \lim_{t\to\infty} \frac{\Psi(tx)}{t}$ of Ψ. We can see how the meaning of Ψ acting on the *Jacobian* of u is extended to the jump set as acting on the *difference* of the left and right side limits of u at the discontinuity. This a key point: by switching to the measure formulation, one can handle noncontinuous functions as well.

3 Necessary Properties of the Interaction Potential

Before applying the above methods to the labeling problem, we start with some basic considerations about the regularizer and the interaction potential d. We begin by formalizing the requirements on the regularizer of the relaxed problem as mentioned in the introduction. Let us assume we are given a general interaction potential $d : \{1, \dots, l\}^2 \to \mathbb{R}$. Intuitively, $d(i, j)$ assigns a weight to switching between label i and label j. We require

$$d(i,j) \quad > \quad 0, i \neq j\,, \tag{24}$$

but no other metric properties (i.e. symmetry or triangle inequality) for now. For $u \in \mathcal{C}$, we postulate that the regularizer should satisfy

(P1) J is convex and positively homogeneous on the relaxed set $\mathcal{C}$ as defined in (5).

(P2) $J(u) = 0$ for any constant u, i.e. there is no penalty for constant labelings.

(P3) For any partition (S, S^c) of Ω into two sets with finite perimeter $\mathrm{Per}(S) < \infty$, and any $i, j \in \{1, \dots, l\}$,

$$J(e^i \chi_S + e^j \chi_{S^c}) \quad = \quad d(i,j)\,\mathrm{Per}(S)\,. \tag{25}$$

That is, a change from label i to label j gets penalized proportional to $d(i, j)$ as well as the perimeter of the interface. Note that this implies that J is isotropic (i.e. rotationally invariant).

Convexity is required in (P1) in order to render global optimization tractable. Indeed, if J is convex, the overall objective function (6) will be convex as well due to the linearization of the data term. Positive homogeneity is included as it allows J to be represented as a support function (i.e. its convex conjugate is an indicator function and $J = \sigma_{\mathcal{D}}$ for some closed convex set $\mathcal{D}$), which will be exploited by our optimization methods.

Requirements (P3) and (P2) formalize the principle that the multilabeling problem should reduce to the classical continuous cut (9) in the two-class case. This allows to include boundary length-based terms in the regularizer that can additionally be weighted depending on the labels of the adjoining region (Fig. 4). Together, these requirements pose a natural restriction on d:

Proposition 1 *If (J, d) satisfy (P1) – (P3), it follows that d must be a metric, i.e. for all $i, j, k \in \{1, \dots, l\}$,*

1. $d(i, i) = 0$,
2. $d(i, j) = d(j, i)\ \forall i \neq j$,
3. *d is subadditive:* $d(i, k) \leqslant d(i, j) + d(j, k)$.

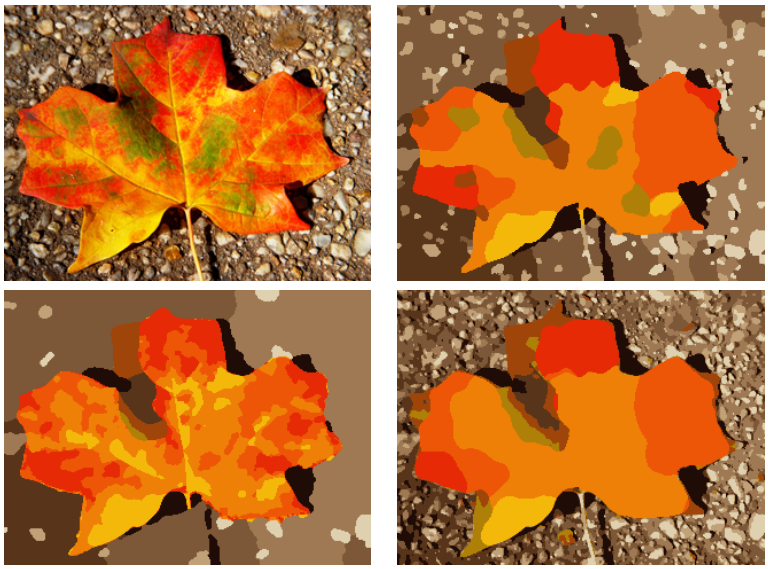

Figure 4: Effect of choosing different interaction potentials. **Top row:** The original image (left) is segmented into 12 regions corresponding to prototypical colors vectors. The Potts interaction potential penalizes the boundary length independently of the labels (right), which leads to a uniformly smooth segmentation. **Bottom row:** By modifying the interaction potential, the regularization strength is selectively adjusted to suppress background (left) or foreground (right) structures while allowing for fine details in the other regions.

Proof 1. follows from (P2) and (P3) by choosing $i = j$ and S with $\mathrm{Per}(S) > 0$. Symmetry in 2. is obtained from (P3) by replacing S with S^c, as $\mathrm{Per}(S) = \mathrm{Per}(S^c)$. To show 3., fix any set S with perimeter

$$c := \mathrm{Per}(S) > 0\,, \tag{26}$$

then

$$\begin{aligned} cd(i,k) &= cJ\left(e^i\chi_S + e^k\chi_{S^c}\right) && (27)\\ &\leqslant cJ\left(e^i\chi_S + e^j\chi_{S^c}\right) + J\left(e^k\chi_{S^c} + e^j\chi_S\right) && (28)\\ &= c\left(d\left(i,j\right) + d(j,k)\right) && (29) \end{aligned}$$

due to (P1). □

Note that if the requirement (24) is dropped, it is easy to show that if $d(i,j) = 0$ for some $i \neq j$, then $d(i,k) = d(j,k)$ for *any* k. Thus classes i and j can be collapsed into a single class as far as the regularizer is concerned. The decision between label i and j is then completely local, i.e. depends only on the data term and can be postponed to a post-processing step by modifying the data term to

$$s_i'(x) := s_j'(x) \quad := \quad \min\{s_i(x), s_j(x)\}\,. \tag{30}$$

Thus (24) is not a real limitation and can be always assured.

As a side note, it can be shown that, under some assumptions and in the space of piecewise constant functions, the subadditivity of d already follows if J is required to be lower semicontinuous [Bra02, p.88].

Proposition 1 implies that for non-metric d, we generally cannot expect to find a regularizer satisfying (P1)–(P3). Note that here J is not required to be of any particular form. In the following sections, we will show that on the other hand, if d is metric as in Proposition 1, such a regularizer can always be constructed. This implies that the interaction potentials allowing for a regularizer that satisfies (P1)–(P3) are exactly the metrics.

4 Constructing Regularizers from the Interaction Potential

We study next how to construct regularizers on $\mathrm{BV}(\Omega)^l$ satisfying (P1)–(P3). As in (23) we set

$$J(u) := \int_\Omega \Psi(Du)\,. \tag{31}$$

We additionally require $\Psi : \mathbb{R}^{d\times l} \to \mathbb{R}^+$ as in (23) to be a support function, i.e.

$$\Psi(z) = \sigma_{\mathcal{D}_{\mathrm{loc}}}(z) = \sup_{v(x)\in\mathcal{D}_{\mathrm{loc}}} \langle z, v(x)\rangle \tag{32}$$

for some closed, convex $\mathcal{D}_{\text{loc}} \subseteq \mathbb{R}^{d\times l}$. As a support function, Ψ coincides with its recession function Ψ_∞, so

$$\begin{aligned} J(u) \quad = \quad & \int_\Omega \Psi(\mathcal{J}(u)(x))dx + \dots \\ & \int_{J_u} \Psi\left(\nu_u(x)\left(u^+(x) - u^-(x)\right)^\top\right) d\mathcal{H}^{d-1} + \dots \\ & \int_\Omega \Psi\left(\frac{D^c u}{|D^c u|}\right) d|D^c u| \,. \end{aligned} \tag{33}$$

Also, we have an equivalent dual formulation in analogy to (12),

$$\begin{aligned} J(u) = \sup\{\int_\Omega \quad & \langle u \,, \operatorname{Div} v\rangle dx | v \in C_c^\infty(\Omega)^{d\times k} \,, \\ & v(x) \in \mathcal{D}_{\text{loc}} \text{ a.e. in } \Omega\} \,. \end{aligned} \tag{34}$$

For simplicity we will also assume that $\mathcal{D}_{\text{loc}}$ is rotationally invariant along the image dimensions, i.e. for any rotation matrix $R \in \mathbb{R}^{d\times d}$,

$$v = (v^1, \dots, v^l) \in \mathcal{D}_{\text{loc}} \quad \Leftrightarrow \quad (Rv^1, \dots, Rv^l) \in \mathcal{D}_{\text{loc}} \,. \tag{35}$$

This is equivalent to J being isotropic.

We will now show under which circumstances a minimizer exists in BV, and then see how the approach can be used to construct regularizers for specific interaction potentials.

4.1 Existence of Minimizers

The complete problem considered here is of the form (cf. (6) and (33))

$$\inf_{u\in\mathcal{C}} f(u) \,, \quad f(u) := \int_\Omega \langle u, s\rangle dx + J(u) \tag{36}$$

and $J(u)$ as in (23) with

$$\Psi(z) = \sup_{v(x)\in\mathcal{D}_{\text{loc}}} \langle z, v(x)\rangle \tag{37}$$

for some closed convex $\mathcal{D}_{\text{loc}}$. Note that f is convex, as it is the pointwise supremum of affine functions, cf. (34). Again for simplicity we set $\Omega = (0,1)^d$. Then we have the following

Proposition 2 *Let $\mathcal{D}_{\text{loc}}$ be closed convex, $\Psi = \sigma_{\mathcal{D}_{\text{loc}}}$, $s \in L^1(\Omega)^l$, and*

$$f(u) = \int_\Omega \langle u, s\rangle dx + \int_\Omega \Psi(Du) \,. \tag{38}$$

Additionally assume that $\mathcal{D}_{\text{loc}} \subseteq \mathcal{B}_{\rho_u}(0)$ for some $0 < \rho_u$. Then f is lower semicontinuous in $\operatorname{BV}(\Omega)^l$ with respect to L^1 convergence.

Proof As the data term is continuous, it suffices to show that the regularizer is lower semicontinuous. This is an application of Thm. 5.47 in [AFP00], proven by Ambrosio/Dal Maso. In fact, the theorem shows that f is the relaxation of $\tilde{f} : C^1(\Omega)^l \to \mathbb{R}$,

$$\tilde{f}(u) := \int_\Omega \langle u, s\rangle dx + \int_\Omega \Psi(\mathcal{J}u(x))dx \,, \tag{39}$$

on $\mathrm{BV}(\Omega)^l$ w.r.t. L^1_{loc} (thus here L^1) convergence and thus lower semicontinuous in $\mathrm{BV}(\Omega)^l$. To apply the theorem, we have to show that Ψ is quasiconvex in the sense of [AFP00, 5.25], which holds as it is convex by construction. The other precondition is (at most) linear growth of Ψ, which holds with $0 \leqslant \Psi(x) \leqslant \rho_u \|x\|$. □

Proposition 3 *Let f, Ψ, s as in Prop. 2 and additionally assume that*

$$\mathcal{B}_{\rho_l}(0) \cap \{(v^1, \ldots, v^l) | \sum_i v^i = 0)\} \subseteq \mathcal{D}_{\mathrm{loc}} \subseteq \mathcal{B}_{\rho_u}(0) \tag{40}$$

for some $0 < \rho_l \leqslant \rho_u$. Then the problem

$$\min_{u \in \mathcal{C}} f(u) \tag{41}$$

has at least one minimizer.

Proof From the inner and outer bounds it holds that $\rho_l \|x\| \leqslant \Psi(x) \leqslant \rho_u \|x\|$ for any x with $x^1 + \ldots + x^l = 0$. In particular this applies to Du as $Du^1 + \ldots + Du^l = 0$ due to $u \in \mathcal{C}$. Combining this with the dual representations of TV and J we get

$$0 \leqslant \rho_l \,\mathrm{TV}(u_k) \leqslant J(u_k) \leqslant \rho_u \,\mathrm{TV}(u_k)\,. \tag{42}$$

From

$$\int_\Omega \langle u, s \rangle dx \geqslant - \int_\Omega \|u(x)\|_\infty \|s(x)\|_1 dx\,, \tag{43}$$

$s \in L^1(\Omega)^l$ and boundedness of Ω and Δ_l it follows that the data term is bounded from below on $\mathcal{C}$.

We now show coercivity of f with respect to the BV norm: Let $(u_k) \in \mathcal{C}$ with $\|u_k\|_1 + \mathrm{TV}(u_k) \to \infty$. As the data term is bounded from below, and since $J(u_k) \geqslant \rho_l \,\mathrm{TV}(u_k)$, it follows that $f(u_k) \to +\infty$. Thus f is coercive.

On the other hand, the above considerations also show that f is bounded from below. Thus we can choose a minimizing sequence (u_k) with $f(u_k) \searrow 0$. Due to the coercivity, $\|u_k\|_1 + \mathrm{TV}(u_k)$ must then be bounded from above.

From this and [AFP00, Thm. 3.23] we conclude that there is a subsequence of u_k weakly*- (and thus L^1-) converging to some $u \in \mathrm{BV}(\Omega)^l$. With the lower semicontinuity from Prop. 2 and closedness of $\mathcal{C}$ with respect to L^1 convergence, existence of a minimizer follows. □

4.2 Relation to the Interaction Potential

To relate such J to the labeling problem in view of (P3), we have the following

Proposition 4 *Let $\Psi = \sigma_{\mathcal{D}_{\mathrm{loc}}}$ and $J(u) = \int_\Omega \Psi(Du)$ as above. Let $u(x) = (1 - u'(x))a + u'(x)b$ for some $u' \in \mathrm{BV}(\Omega)$ and vectors $a, b \in \Delta_l$. Then for any vector $y \in \mathbb{R}^d$ with $\|y\| = 1$,*

$$\begin{aligned} J(u) &= \left(\sup_{v \in \mathcal{D}_{\mathrm{loc}}} \|v\,(b-a)\,\| \right) \mathrm{TV}(u') \\ &= \Psi\left(y(b-a)^\top\right) \mathrm{TV}(u')\,. \end{aligned} \tag{44}$$

In particular, if $\Psi\left(y(e^i - e^j)^\top\right) = d(i,j)$ for some y with $\|y\| = 1$, then J fulfills (P3).

Proof We start with the dual formulation (34),

$$J(u) \quad = \quad \sup_{v \in \mathcal{D}} \int_\Omega \langle (1 - u'(x))b + u'(x)a, \operatorname{Div} v(x) \rangle dx \,, \tag{45}$$

where $\mathcal{D} := \{v \in C_c^\infty(\Omega)^{d \times l}, v(x) \in \mathcal{D}_{\mathrm{loc}} \text{ a.e.}\}$. Due to the compact support of v the constant part vanishes,

$$J(u) \quad = \quad \sup_{v \in \mathcal{D}} \int_\Omega u'(x) \langle b - a, \operatorname{Div} v(x) \rangle dx \,. \tag{46}$$

From the linearity of div, we get

$$J(u) \quad = \quad \sup_{v \in \mathcal{D}} \int_\Omega u'(x) \operatorname{div}(\underbrace{v(x)(b-a)}_{=:w(x)}) dx \,, \tag{47}$$

where $v(x)(b-a)$ has to be understood as a matrix-vector product. Denote by $\mathcal{D}'_{\mathrm{loc}}$ the set of all possible $w(x)$, i.e. $\mathcal{D}'_{\mathrm{loc}} = \{w \in \mathbb{R}^d | w = v(b-a), v \in \mathcal{D}_{\mathrm{loc}}\}$. As $\mathcal{D}_{\mathrm{loc}}$ is rotationally invariant, $\mathcal{D}'_{\mathrm{loc}}$ is as well and thus must be a ball, $\mathcal{D}'_{\mathrm{loc}} = \mathcal{B}_r(0)$, with

$$r \quad := \quad \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \|v(b-a)\| \,. \tag{48}$$

Together with the vector space property of $C_c^\infty(\Omega)$, we get

$$\begin{aligned} J(u) \quad &= \quad \sup_{w \in C_c^\infty(\Omega)^d, \|w(x)\| \leqslant r \text{ a.e.}} \int_\Omega u'(x) \operatorname{div} w(x) dx \\ &= \quad r \operatorname{TV}(u') \,, \end{aligned} \tag{49}$$

which proves the first equality of (44). The second equality follows again from the rotational invariance of $\mathcal{D}_{\mathrm{loc}}$ via

$$r \quad = \quad \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \sup_{z \in \mathbb{R}^d, \|z\| \leqslant 1} \langle z, v(b-a) \rangle \tag{50}$$

$$= \quad \sup_{z \in \mathbb{R}^d, \|z\| \leqslant 1} \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \langle z, v(b-a) \rangle \tag{51}$$

$$= \quad \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \langle y, v(b-a) \rangle = \Psi(y(b-a)^\top) \,. \tag{52}$$

□

As a consequence, if the relaxed multiclass formulation is restricted to two classes by parametrizing $u = (1-u')a + u'b$ for some $u' \in [0,1]$, it essentially reduces to the scalar continuous cut problem: Solving

$$\min_{\substack{u' \in \mathrm{BV}(\Omega), \\ u'(x) \in [0,1]}} \int_\Omega \langle (1-u')a + u'b, s \rangle dx + J(u) \tag{53}$$

is equivalent to solving

$$\min_{\substack{u' \in \mathrm{BV}(\Omega), \\ u'(x) \in [0,1]}} \int_\Omega u'(b-a) dx + \Psi(y(b-a)^\top) \operatorname{TV}(u') \tag{54}$$

which is just the classical binary continuous cut approach with data $(b-a)$ and regularizer weight $\Psi(y(b-a)^\top)$. For the multiclass case, assume that

$$u = u_P \quad = \quad e^1 \chi_{P^1} + \ldots + e^l \chi_{P^l} \tag{55}$$

for some partition $P^1 \cup \ldots \cup P^l = \Omega$ with sufficiently regular boundaries such that the Cantor parts of the χ_{P^i} vanish. Then as u_P is locally constant also the absolute part vanishes, and we are left with the jump part

$$J(u_P) \quad = \quad \int_{S_{u_P}} \Psi\left(\nu_{u_P} \left(u_{P+} - u_{P-}\right)^\top\right) d\mathcal{H}^{d-1}, \tag{56}$$

where $S_{u_P} = \bigcup_{i=1,\ldots,l} \partial P^i$ is the union of the interfaces between regions. Define $i(x), j(x)$ s.t. $u_{P+}(x) = e^{i(x)}$ and $u_{P-}(x) = e^{j(x)}$. Then

$$J\left(u_P\right) \quad = \quad \int_{S_{u_P}} \Psi\left(\nu_{u_P} \left(e^{i(x)} - e^{j(x)}\right)^\top\right) d\mathcal{H}^{d-1}. \tag{57}$$

By rotational invariance,

$$J\left(u_P\right) \quad = \quad \int_{S_{u_P}} \Psi\left(y \left(e^{i(x)} - e^{j(x)}\right)^\top\right) d\mathcal{H}^{d-1}. \tag{58}$$

for some y with $\|y\| = 1$. Thus the regularizer locally penalizes the jump with the boundary length between regions with labels i and j, multiplied by the factor $\Psi(y(e^i - e^j))$ depending on the labels of the adjoining regions.

The question is now how to choose the dual set $\mathcal{D}_{\text{loc}}$ such that $\Psi(y(e^i - e^j)^\top) = d(i,j)$ for a given interaction potential d. We will consider two approaches which differ with respect to expressiveness and simplicity of use: In the *local envelope* approach, $\mathcal{D}_{\text{loc}}$ is chosen as large as possible. In turn, J is as large as possible with the integral formulation, which prevents introducing artificial minima generated by the relaxation and potentially keeps minimizers of the relaxed problem close to minimizers of the discrete problem. However, $\mathcal{D}_{\text{loc}}$ is only implicitly defined, which complicates optimization. In contrast, in the *embedding* approach, $\mathcal{D}_{\text{loc}}$ is considerably simpler, at the cost of being able to represent only a subset of all metric potentials d. For an illustration of the two approaches, see Fig. 5.

4.3 Local Envelope Method

Chambolle et al. [CCP08] proposed an interesting approach for potentials d of the form $d(i,j) = \sigma(|i-j|)$ for a positive concave function σ. The approach is derived by specifying the value of J on discrete u only and then constructing an approximation of the convex envelope by pulling the convexification into the integral.

This potentially generates tight extensions and thus we may hope that the convexification process does not generate too many artificial non-discrete solutions.

We propose to extend this approach to arbitrary metric d by setting (cf. Fig. 5)

$$\begin{aligned} \mathcal{D}^d_{\text{loc}} \quad := \bigcap\nolimits_{i \neq j} \{ \quad & v = \left(v^1, \ldots, v^l\right) \in \mathbb{R}^{d \times l} | \ldots \\ & \|v^i - v^j\| \leqslant d(i,j), \sum_k v^k = 0 \} \end{aligned} \tag{59}$$

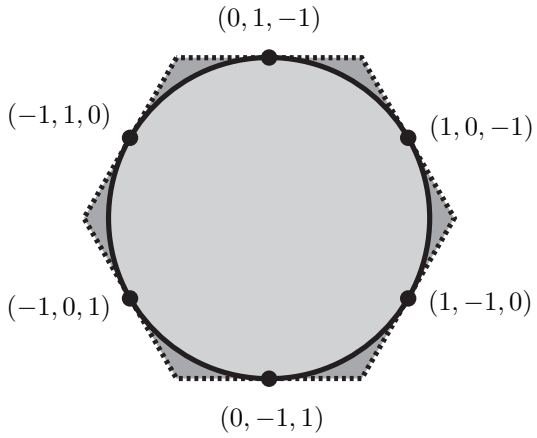

Figure 5: Illustration of the set $\mathcal{D}_{\mathrm{loc}}$ used to build the regularizer for the uniform distance $d(i,j) = \chi_{\{i \neq j\}}$ for $l = 3$ in $d = 1$ – dimensional space. Shown is a cut through the $z^1 + z^2 + z^3 = 0$ plane; the labels correspond to the points $e^i - e^j$ with $i \neq j$. The local envelope method leads to a larger set $\mathcal{D}_{\mathrm{loc}}$ (dashed) than the Euclidean metric method (solid). This improves the tightness of the relaxation, but requires more expensive projection steps.

for some given interface potential $d(i,j)$. By definition $\mathcal{D}^d_{\mathrm{loc}}$ is rotationally invariant. By the considerations in Sect. 3 we may assume d to be a metric. Then the inner and outer bound assumptions required for existence of a minimizer in Prop. 3 are satisfied. Moreover, d can be reconstructed from $\mathcal{D}^d_{\mathrm{loc}}$:

Proposition 5 *Let $d : \{1,\dots,l\}^2 \to \mathbb{R}^+_0$ be a metric. Then for any i, j,*

$$\sup_{v \in \mathcal{D}^d_{\mathrm{loc}}} \left(\left(v^i\right)_1 - \left(v^j\right)_1 \right) \quad = \quad d(i,j)\,. \tag{60}$$

Proof "$\leqslant$" follows from the definition. "$\geqslant$" can be shown using a network flow argument: We have

$$\begin{aligned}
& \sup_{v \in \mathcal{D}^d_{\mathrm{loc}}} \left(\left(v^i\right)_1 - \left(v^j\right)_1 \right) && (61)\\
\geqslant \quad & \sup\{p_i - p_j \mid p \in \mathbb{R}^l, \sum_k p_k = 0, \dots \\
& \forall i', j' : p_{i'} - p_{j'} \leqslant d(i',j')\} && (62)\\
= \quad & \sup\{p_i - p_j \mid p \in \mathbb{R}^l, \dots \\
& \forall i', j' : p_{i'} - p_{j'} \leqslant d(i',j')\} && (63)\\
\overset{\text{see text}}{=} \quad & d(i,j)\,. && (64)
\end{aligned}$$

The last equality follows from [Mur03, 5.1] with $\gamma = d$ (and, as d is a metric, $\bar{\gamma} = d$ since the triangle inequality implies that the length of the shortest path from i to j is always $d(i,j)$). □

The final result of this section is the following:

Proposition 6 *Let $d : \mathbb{R}^{l\times l} \to \mathbb{R}_0^+$ be a metric. Define $\mathcal{D}_{\mathrm{loc}} := \mathcal{D}^d_{\mathrm{loc}}$ as in (59), $\Psi_d := \sigma_\Omega$ and $J_d := \int_\Omega \Psi_d(Du)$ as in (33). Then J_d satisfies (P1)–(P3).*

Proof (P1) and (P2) are clear from the definition of J_d. (P3) follows directly from Prop. 5 and Prop. 4 with $y = e^1$. □

Defining $\mathcal{D}^d_{\mathrm{loc}}$ as in (59) provides us with a way to extend the desired regularizer for *any* metric d to non-discrete $u \in \mathcal{C}$ via (33). The price to pay is that there is no simple closed expression for Ψ and thus for J_d, which potentially complicates optimization.

Also note that in order to define $\mathcal{D}^d_{\mathrm{loc}}$, d does not have to be a metric. However Prop. 5 then does not hold in general, so J is not a true extension of the desired regularizer.

4.4 Euclidean Metric Method

In this section, we consider a regularizer which is less powerful but much easier to evaluate. Recall the classical total variation for vector-valued u as defined in (12),

$$\mathrm{TV}(u) = \int_\Omega \|Du\| \,. \tag{65}$$

This classical definition has also been used in color denoising and is also referred to as MTV [SR96, DAV08]. We propose to extend this definition by choosing an *embedding matrix* $A \in \mathbb{R}^{k\times l}$, and defining

$$J_A(u) \quad := \quad \mathrm{TV}(Au) \,. \tag{66}$$

This corresponds to substituting the Frobenius matrix norm on the distributional gradient with a linearly weighted variant. In the framework of (33), it amounts to setting $\mathcal{D}_{\mathrm{loc}} = \mathcal{D}^A_{\mathrm{loc}}$ (cf. Fig. 5) with

$$\mathcal{D}^A_{\mathrm{loc}} \quad := \quad \{v'A | v' \in \mathbb{R}^{d\times k}, \|v'\| \leqslant 1\} = \mathcal{B}_1(0)A \tag{67}$$

In fact,

$$\begin{aligned} \Psi(z) \quad &= \quad \sigma_{\mathcal{D}^A_{\mathrm{loc}}}(z) = \sup_{v'\in\mathcal{B}_1(0)A} \langle z, v'\rangle = \sup_{v\in\mathcal{B}_1(0)} \langle z, vA\rangle && (68)\\ &= \quad \sup_{v\in\mathcal{B}_1(0)} \langle zA^\top, v\rangle = \|zA^\top\| \,. && (69) \end{aligned}$$

In particular, we formally have

$$\Psi(Du) \quad = \quad \|(Du)A^\top\| = \|D(Au)\| \,, \tag{70}$$

as $u \mapsto Du$ is linear in u. To further clarify the definition, we may rewrite this to

$$\mathrm{TV}_A(u) \quad = \quad \int_\Omega \sqrt{\|D_1 u\|_A^2 + \ldots + \|D_d u\|_A^2} \,, \tag{71}$$

where $\|v\|_A := (v^\top A^\top A v)^{1/2}$. The approach can thus be understood as replacing the Euclidean norm by a linearly weighted variant.

It remains to show for which interaction potentials d assumption (P3) can be satisfied. The next proposition shows that this is possible for the class of *Euclidean* metrics.

Proposition 7 *Let d be an Euclidean metric, i.e. there is a $k \in \mathbb{N}$ and $z^1, \ldots, z^l \in \mathbb{R}^k$ such that $d(i,j) = \|z^i - z^j\|$. Then for $A = (z^1, ..., z^l)$, $J_A := \mathrm{TV}_A$ satisfies (P1)–(P3).*

Proof (P1) and (P2) are clearly satisfied. In order to show (P3) we assume $\|y\| = 1$ and apply (69) to get

$$\begin{aligned} \Psi(y(e^i - e^j)^\top) &= \|y(e^i - e^j)^\top A^\top\| \\ &= \|y(z^i - z^j)^\top\| = \|z^i - z^j\| \,. \end{aligned} \tag{72}$$

□

The class of Euclidean metrics comprises some important special cases:

- The *uniform*, *discrete* or *Potts* metric as also considered in [ZGFN08, LKY+09] and as a special case in [KT99, KT07]. Here $d(i,j) = 0$ iff $i = j$ and $d(i,j) = 1$ in any other case, which corresponds to $A = (1/\sqrt{2})I$.
- The *linear* (label) metric, $d(i,j) = c|i - j|$, with $A = (c, 2c, \ldots, lc)$. This regularizer is suitable to problems where the labels can be naturally ordered, e.g. depth from stereo or grayscale image denoising.
- More generally, if label i corresponds to a prototypical vector z^i in k-dimensional feature space, and the Euclidean norm is an appropriate metric on the features, it is natural to set $d(i,j) = \|z^i - z^j\|$, which is Euclidean by construction. This corresponds to a regularization in feature space, rather than in "label space".

Note that the boundedness assumption required for the proof of existence of a minimizer in Prop. 3 is only fulfilled if $k \geqslant l$ and A is not rank-deficient. Thus if d is a degenerate Euclidean metric which can be represented by an embedding into a lower-dimensional space, it has to be made regular first, e.g. by adding ϵI_k to A before solving the problem. This case occurs for example in the case of the linear metric.

Non-metric or non-Euclidean d, such as the *truncated linear metric*, $d(i,j) = \min\{2, |i - j|\}$, cannot be represented exactly by TV_A. We will now see how to construct approximations for these cases.

Assume that d is a general metric with squared matrix representation $D \in \mathbb{R}^{l \times l}$, $D_{ij} = d(i,j)^2$. Then it is known [BG05] (Chap.) that d is Euclidean iff for $C := I - \frac{1}{l} ee^\top$, the matrix $T := -\frac{1}{2} CDC$ is positive semidefinite. In this case, A can be found by factorizing $T = A^\top A$. If T is not positive semidefinite, setting the nonnegative eigenvalues in T to zero yields an Euclidean approximation. This method is known as *classical scaling* [BG05] and does not necessarily give good absolute error bounds.

More generally, for some non-metric, nonnegative d, we can formulate the problem of finding the "closest" Euclidean distance matrix E as the minimization problem of a matrix norm $\|E - D\|_M$ over all $E \in \mathcal{E}_l$, the set of $l \times l$ Euclidean distance matrices. Fortunately, there is a linear bijection $B : \mathcal{P}_{l-1} \to \mathcal{E}_l$ between $\mathcal{E}_l$ and the space of positive semidefinite $(l-1) \times (l-1)$ matrices $\mathcal{P}_{l-1}$ [Gow85, JT95]. This allows us to rewrite our problem as a *semidefinite program* [WSV00, (p.534–541)]

$$\inf_{S \in \mathcal{P}_{l-1}} \|B(S) - D\|_M \,. \tag{73}$$

The resulting problem can be solved using available numerical solvers. Then $E = B(S) \in \mathcal{E}_l$, and A can be extracted by factorizing $-\frac{1}{2} CEC$. Since E and D are explicitly known, we can compute an a posteriori bound on the maximum distance error,

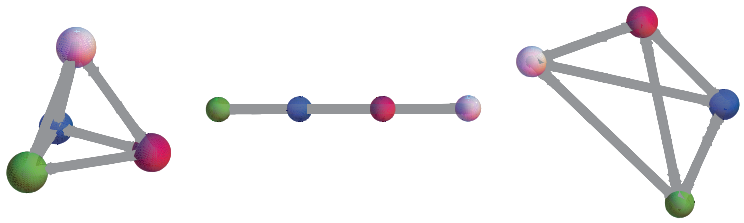

Figure 6: Euclidean embeddings into $\mathbb{R}^3$ for several interaction potentials with four classes. **Left to right:** Potts; linear metric; non-Euclidean truncated linear metric. The vertices correspond to the columns $z^1, \dots, z^l$ of the embedding matrix A. For the truncated linear metric an optimal approximate embedding was computed as outlined in Sect. 4.4 with $\|X\|_M := \max_{i,j} |X_{ij}|$.

$\varepsilon_E := \max_{i,j} |(E_{ij})^{1/2} - (D_{ij})^{1/2}|$. Fig. 6 shows a visualization of some embeddings for a four-class problem. In many cases, in particular when the number of classes is large, the Euclidean embedding provides a good approximation for non-Euclidean metrics (Fig. 7).

Based on the embedding matrices computed in this way, the Euclidean distance approach can be used to solve approximations of the labeling problem with arbitrary metric interaction potentials, with the advantage of having a closed expression for the regularizer.

5 Discretized Problem

5.1 Saddle Point Formulation

We now turn to solving the discretization of (6). In order to formulate generic algorithms, we study the bilinear saddle point problem,

$$\begin{aligned} &\min_{u \in \mathcal{C}} \max_{v \in \mathcal{D}} g\,(u, v)\ , \\ &g(u, v) := \langle u, s \rangle + \langle Lu, v \rangle - \langle b, v \rangle\,. \end{aligned} \tag{74}$$

As will be shown in Sect. 5.2–5.4, this covers both J_d and J_A as well as many other – even non-uniform and non-isotropic – regularizers.

In a slight abuse of notation, we will denote by $u, s \in \mathbb{R}^n$ also the discretizations of u and s on a uniform grid. We have a linear operator $L \in \mathbb{R}^{m \times n}$, a vector $b \in \mathbb{R}^m$ for some $m, n \in \mathbb{N}$, and bounded closed convex sets $\mathcal{C} \subseteq \mathbb{R}^n, \mathcal{D} \subseteq \mathbb{R}^m$. Intuitively, L discretizes the gradient operator and $\mathcal{D}$ corresponds to $\mathcal{D}_{\mathrm{loc}}$, i.e. specifies Ψ in a dual formulation. The primal and dual objectives are

$$f(u) := \max_{v \in \mathcal{D}} g\,(u, v) \quad \text{and} \quad f_d(v) := \min_{u \in \mathcal{C}} g(u, v)\,, \tag{75}$$

respectively. The dual problem then consists of maximizing $f_d(v)$ over $\mathcal{D}$. As $\mathcal{C}$ and $\mathcal{D}$ are bounded, it follows from [Roc70, Cor. 37.6.2] that a saddle point (u^*, v^*) of g exists. With [Roc70, Lemma 36.2], this implies strong duality, i.e.

$$\min_{u \in \mathcal{C}} f(u) = f(u^*) = g(u^*, v^*) = f_d(v^*) = \max_{v \in \mathcal{D}} f_d(v)\,. \tag{76}$$

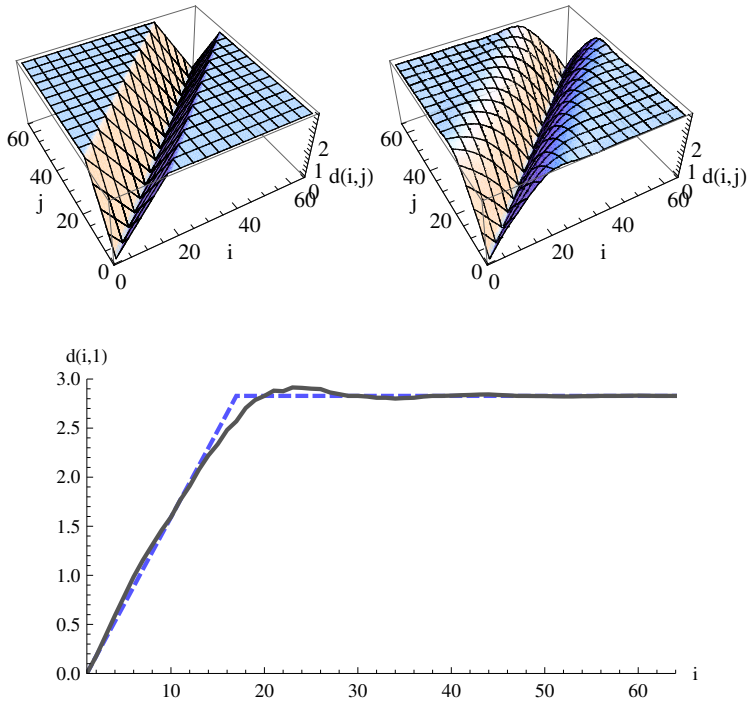

Figure 7: Euclidean approximation of the Non-Euclidean truncated linear metric $d(i,j) = \sqrt{2}/8 \min\{|i-j|, 16\}$. **Left to Right:** Original potential for 64 classes; potential after Euclidean approximation. **Bottom:** cross section of the original (dashed) and approximated (solid) metric at $i = 1$. The approximation was computed using semidefinite programming as outlined in Sect. 4.4. It represents the closest Euclidean metric with respect to the matrix norm $\|X - Y\|_M := \sum_{i,j} |X_{ij} - Y_{ij}|$. The maximal error with respect to the original potential is $\varepsilon_E = 0.2720$.

In our case, $\mathcal{C}, \mathcal{D}$ exhibit a product structure, which allows to compute f_d as well as the orthogonal projections $\Pi_{\mathcal{C}}$ and $\Pi_{\mathcal{D}}$ efficiently, a fact that will prove important in the algorithmic part. The evaluation of f can be more difficult depending on the definition of $\mathcal{D}_{\mathrm{loc}}$, i.e. $\mathcal{D}$, but is not required by the optimizer.

Note that in the discrete framework, we may easily substitute non-homogeneous, spatially varying or even nonlocal regularizers by choosing L and b appropriately.

5.2 Discretization

We discretize Ω by a regular grid, $\Omega = \{1, \dots, n_1\} \times \cdots \times \{1, \dots, n_d\} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, consisting of $n := |\Omega|$ pixels, and u by its (vectorial) values $u_i = (u_i^1, \dots, u_i^l)$ for a lexicographical ordering of the pixels in Ω, i.e. $u \in \mathbb{R}^{n \times l}$. The multidimensional image space $\mathbb{R}^{n \times l}$ is equipped with the Euclidean inner product $\langle \cdot, \cdot \rangle$ over the vectorized elements. We naturally identify $u = (u^1, \dots, u^l) \in \mathbb{R}^{n \times l}$ with $((u^1)^\top \cdots (u^l)^\top)^\top \in \mathbb{R}^{nl}$.

Let $\mathrm{grad} := (\mathrm{grad}_1^\top, \dots, \mathrm{grad}_d^\top)^\top$ be the d-dimensional forward difference gradient operator for Neumann boundary conditions. Accordingly, $\mathrm{div} := -\mathrm{grad}^\top$ is the backward difference divergence operator for Dirichlet boundary conditions. These operators extend to $\mathbb{R}^{n \times l}$ via $\mathrm{Grad} := (I_l \otimes \mathrm{grad})$, $\mathrm{Div} := (I_l \otimes \mathrm{div})$.

We define for some $k \geqslant 1$ the convex sets

$$\mathcal{C} := \{u \in \mathbb{R}^{n \times l} | u_i \in \Delta_l, i = 1, \dots, n\}, \tag{77}$$

$$\mathcal{D} := \prod_{x \in \Omega} \mathcal{D}_{loc} \subseteq \mathbb{R}^{n \times d \times k}. \tag{78}$$

The set $\mathcal{D}_{\mathrm{loc}}$ and the operator L depend on the chosen regularizer. Note that for $L := \mathrm{Grad}$, $k = l$ and

$$\mathcal{D}_{loc} := \mathcal{D}_{\mathrm{loc}}^I := \left\{p = (p^1, \dots, p^l) \in \mathbb{R}^{d \times l} | \|p\| \leqslant 1\right\}, \tag{79}$$

the primal objective in the saddle point formulation discretizes the classical vector-valued total variation,

$$\mathrm{TV}(u) := \sigma_{\mathrm{Div}\,\mathcal{D}}(u) = \sigma_{\mathcal{D}}(\mathrm{Grad}\, u) = \sum_{x \in \Omega} \|G_x u\|, \tag{80}$$

where G_x is an $(ld) \times n$ matrix composed of rows of (Grad) s.t. $G_x u$ constitutes the vectorized discrete Jacobian of u at the point x.

Projections on the set $\mathcal{C}$ are highly separable and can be computed exactly in a finite number of steps [Mic86]. The dual objective

$$f_d(v) = \min_{u \in \mathcal{C}} \langle u, L^\top v + s \rangle \tag{81}$$

can in any case be evaluated by summing the pointwise minimum over the components of $L^\top v + s$,

$$f_d(v) = \sum_{i \in \Omega} \mathrm{vecmin}(L^\top v + s). \tag{82}$$

This fact is helpful if evaluating the primal objective is costly because of a complicated dual set $\mathcal{D}_{\mathrm{loc}}$.

5.3 Specialization for the Local Envelope Method

We identify a given metric d with the matrix $d \in \mathbb{R}^{l\times l}, d_{ij} = d(i,j)$, set $k := l$ and

$$\begin{aligned} L &:= \operatorname{Grad}, \\ \mathcal{D}_{\mathrm{loc}} &:= \mathcal{D}^d_{\mathrm{loc}} = \bigcap_{i\neq j} \{v = \left(v^1, \ldots, v^l\right) \in \mathbb{R}^{d\times l} | \ldots \\ &\qquad \|v^i - v^j\| \leqslant d(i,j)\} \end{aligned} \tag{83}$$

as in (59). We arrive at the saddle point form (74) with $\mathcal{C}$, $\mathcal{D}$, and L defined as above, $m = nl$ and $b = 0$. However due to the generality of the regularizer, the primal objective f cannot be easily computed anymore (for the special case of 3 labels there is a derivation in [CCP08]). Projections on $\mathcal{D}$ can be computed for all $x \in \Omega$ in parallel. Projections on the individual sets $\mathcal{D}^d_{\mathrm{loc}}$ can be computed by the Dykstra algorithm, cf. [CCP08] and Sect. 6.4.

5.4 Specialization for the Euclidean Metric Method

For $A \in \mathbb{R}^{k\times l}$ as in (66), define

$$\begin{aligned} L &:= (\operatorname{Grad})(A \otimes I_n), \\ \mathcal{D}_{\mathrm{loc}} &:= \mathcal{D}^I_{\mathrm{loc}} = \left\{v = (v^1, \ldots, v^k) \in \mathbb{R}^{d\times k} | \|v\| \leqslant 1\right\}. \end{aligned} \tag{84}$$

Departing from the definition (67) of $\mathcal{D}^A_{\mathrm{loc}}$, A is merged into L, as the optimization method will rely on projections on $\mathcal{D}_{\mathrm{loc}}$. Including A into $\mathcal{D}_{\mathrm{loc}}$, i.e. by setting

$$\mathcal{D}_{\mathrm{loc}} := (A \otimes I_n)^\top \mathcal{D}^I_{\mathrm{loc}} \tag{85}$$

and $L := \operatorname{Grad}$, would prevent computing the projection in closed form. Projecting on the unit ball $\mathcal{D}^I_{\mathrm{loc}}$ on the other hand is trivial. The discretized regularizer can be explicitly evaluated, as

$$\Psi(z) \quad = \quad \|zA^\top\|. \tag{86}$$

A comparison to (80) yields

$$\mathrm{J}_A(u) \quad := \quad \mathrm{TV}((A \otimes I_n)u). \tag{87}$$

We finally arrive at the form (74) with $\mathcal{C}$, $\mathcal{D}$, and L defined as above, $m = nk$ and $b = 0$. Projections $\mathcal{D}$ are highly separable and thus can be computed easily. The primal objective can be evaluated in closed form using (87).

5.5 Optimality

If f_d and f can be explicitly computed, any $v \in \mathcal{D}$ gives an optimality bound on the primal objective via the numerical *primal-dual gap* $f(u) - f_d(v)$,

$$0 \leqslant f(u) - f(u^*) \leqslant f(u) - f_d(v). \tag{88}$$

Assuming f and f_d can be evaluated, the gap is a convenient stopping criterion. To improve the scale invariance, it is often practical to stop on the *relative gap*

$$\frac{f(u) - f_d(v)}{f_d(v)}. \tag{89}$$

instead, which gives a similar bound. However convergence in the objective alone does not necessarily imply convergence in u: The minimizer of the original problem (74) is generally not unique, in contrast to the well-known ROF-type problems [ROF92] where a quadratic data term is used.

For some applications, after solving the relaxed problem a discrete solution – or "hard" labeling – still needs to be recovered, i.e. the relaxed solution needs to be *binarized*. For the continuous two-class case with the classical TV regularizer, [CEN06] showed that an exact solution can be obtained by thresholding at almost any threshold. However, their results do not obviously transfer to the multi-class case.

Another difficulty lies in the discretization: In order to apply the thresholding theorem, a crucial "coarea"-like property must hold for the *discretized* problem [CD09], which holds for the graph-based pairwise ℓ_1-, but not the higher order ℓ_2-discretization of the TV. Thus, even in the two-class case, simple thresholding may lead to a suboptimal discrete solution.

Currently we are not aware of an a priori bound on the error introduced by the binarization step in the general case. In practice, any dual feasible point together with (88) yields an a posteriori optimality bound: Let (u^N, v^N) be a pair of primal resp. dual feasible iterates, $\bar{u}^N$ the result of the binarization step applied to u^N, and $\bar{u}^*$ the optimal discrete solution. Then $\bar{u}^N$ is primal feasible, and its suboptimality is bounded from above by

$$\begin{aligned} f(\bar{u}^N) - f(\bar{u}^*) &\leqslant f(\bar{u}^N) - f(u^*) && (90) \\ &\leqslant f(\bar{u}^N) - f_d(v^N)\,. && (91) \end{aligned}$$

Computation of f_d, and in many cases also f, is efficient as outlined in Sect. 5.2.

5.6 Improving the Binarization

There seems to be no obvious best choice for the binarization step. The simplest choice is the *first-max* approach: the label $\ell(x)$ is set to the index of the first maximal component of the relaxed solution $\bar{u}^N(x)$. However, this might lead to undesired side effects: Consider the segmentation of a grayscale image with the 3 labels $1, 2, 3$ corresponding to the gray level intensity, together with the linear distance $d(i,j) = |i-j|$, and assume there is a region where $u(x) = (1/3+\delta(x), 1/3, 1/3-\delta(x))$ for some small $\delta(x) \in \mathbb{R}$. The most "natural" choice given the interpretation as grayscale values is the constant labeling $\ell(x) = 2$. The first-max approach gives $\ell(x) \in \{1, 3\}$, depending on the sign of $\delta(x)$, which leads to a noisy resulting image.

By closer inspection, the first-max approach – which works well for the Potts distance – corresponds to choosing

$$\ell(x) \;=\; \arg \min_{\ell \in \{1,\dots,l\}} \|u(x) - e^l\|\,, \tag{92}$$

with the smallest l chosen in case of ambiguity. We propose to extend this to non-uniform distances by setting

$$\begin{aligned} \ell(x) \;=\; & \arg \min_{\ell \in \{1,\dots,l\}} \bar{\Psi}\left(u(x) - e^\ell\right)\,, && (93) \\ & \bar{\Psi} : \mathbb{R}^l \to \mathbb{R}, \bar{\Psi}(z) := \Psi\left(e^1 \otimes z^\top\right)\,. \end{aligned}$$

That is, we select the label corresponding to the nearest unit vector *with respect to* $\bar{\Psi}$ (note that instead of e^1 we could choose any normalized vector as Ψ is rotationally

Algorithm 1 FPD Multi-Class Labeling

Choose $\bar{u}^0 \in \mathbb{R}^{n\times l}, v^0 \in \mathbb{R}^{n\times d\times l}$.
Choose $\tau_p > 0, \tau_d > 0, N \in \mathbb{N}$.
for $k = 0, \dots, N-1$ **do**
  $v^{k+1} \leftarrow \Pi_{\mathcal{D}}\left(v^k + \tau_d\left(L\bar{u}^k - b\right)\right)$.
  $u^{k+1} \leftarrow \Pi_{\mathcal{C}}\left(u^k - \tau_p\left(L^\top v^{k+1} + s\right)\right)$.
  $\bar{u}^{k+1} \leftarrow 2u^{k+1} - u^k$.
end for

invariant). In fact, for the linear distance example above we have $\bar{\Psi}(z) = |-z_1 + z_3|$. Thus

$$\begin{aligned} \bar{\Psi}(u(x) - e^1) &= |1 - 2\delta(x)|, \\ \bar{\Psi}(u(x) - e^2) &= |2\delta(x)|, \\ \bar{\Psi}(u(x) - e^3) &= |1 + 2\delta(x)|, \end{aligned} \tag{94}$$

and for small δ we will get the stable and semantically correct choice $\ell(x) = 2$.

This method proved to work well in practice, and considerably reduced the suboptimality introduced by the binarization step (Sect. 7.5). In case there is no closed form expression of Ψ, it can be numerically approximated as outlined in Sect. 6.5.

6 Optimization

When optimizing the saddle point problem (74), one must take care of its large-scale nature and the inherent nonsmoothness of the objective. While interior point solvers are known to be very fast for small to medium sized problems, they are not particularly suited well for massively parallel computation, such as on the upcoming GPU platforms, due to the expensive inner Newton iterations.

We will instead focus on *first order* methods involving only evaluations of L and $L^\top$ and projections on $\mathcal{C}$ and $\mathcal{D}$, as these operations are can be highly parallelized due to their local nature. The first two methods are stated here for comparison, the third one improves on these and is new.

6.1 Fast Primal-Dual Method

One of the most straightforward approaches for optimizing (74) is to fix small primal and dual step sizes τ_p resp. τ_d, and alternatingly apply projected gradient descent resp. ascent on the primal resp. dual variables. This *Arrow-Hurwicz* approach [AHU64] was proposed in a PDE framework for solving the two-class labeling problem in [AT06] and recently used in [CCP08]. An application to denoising problems can be found in [ZC08]. However it seems nontrivial to derive sufficient conditions for convergence. Because of this, in [PCBC09a] the authors propose the *Fast Primal-Dual* (FPD) method, a variant of the Popov's saddle point method [Pop80]. The algorithm is summarized in Alg. 1.

Due to the explicit steps involved, there is an upper bound condition on the step size to assure convergence, which can be shown to be $\tau_p\tau_d < 1/\|L\|^2$ [PCBC09a]. For the two presented methods, the operator norm can be bounded according to

$$\|L\| \leqslant \|\operatorname{Grad}\| \leqslant 2\sqrt{d} \tag{95}$$

Algorithm 2 Nesterov Multi-Class Labeling

Let $c_1 \in \mathcal{C}$, $c_2 \in \mathcal{D}$ and $r_1, r_2 \in \mathbb{R}$ s.t. $\mathcal{C} \subseteq \mathcal{B}_{r_1}(c_1)$ and $\mathcal{D} \subseteq \mathcal{B}_{r_2}(c_2)$; $C \geqslant \|L\|$.
Choose $x^0 \in \mathcal{C}$ and $N \in \mathbb{N}$.
Let $\mu \leftarrow \frac{2\|L\|}{N+1} \frac{r_1}{r_2}$.
Set $G^{(-1)} = 0, w^{(-1)} = 0$.
for $k = 0, \ldots, N$ **do**
  $V \leftarrow \Pi_{\mathcal{D}} \left(c_2 + \frac{1}{\mu} \left(L x^k - b \right) \right)$.
  $w^k \leftarrow w^{k-1} + (k+1)V$.
  $v^k \leftarrow \frac{2}{(k+1)(k+2)} w^k$.
  $G \leftarrow s + L^\top V$.
  $G^k \leftarrow G^{k-1} + \frac{k+1}{2} G$.
  $u^k \leftarrow \Pi_{\mathcal{C}} \left(x^k - \frac{\mu}{\|L\|^2} G \right)$.
  $z^k \leftarrow \Pi_{\mathcal{C}} \left(c_1 - \frac{\mu}{\|L\|^2} G^k \right)$.
  $x^{k+1} \leftarrow \frac{2}{k+3} z^k + \left(1 - \frac{2}{k+3} \right) u^k$.
end for

for the envelope method, and

$$\|L\| \leqslant \| \operatorname{Grad} \| \|A\| \leqslant 2\sqrt{d} \|A\| \,. \tag{96}$$

for the Euclidean metric method. As both the primal and dual iterates are always feasible, a stopping criterion based on the primal-dual gap as outlined in Sect. 5.5 can be employed.

6.2 Nesterov Method

We will provide a short summary of the application of Nesterov's multi-step method [Nes04] to the saddle point problem (74) as proposed in [LBS09]. In contrast to the FPD method, it treats the nonsmoothness by first applying a smoothing step and then using a smooth constrained optimization method. The amount of smoothing is balanced in such a way that the overall number of iterations to produce a solution with a specific accuracy is minimized.

The algorithm has a theoretical complexity of $O(1/\varepsilon)$ for finding an ε-optimal solution, and has been shown to give accurate results for denoising [Auj08] and general ℓ_1-norm based problems [WABF07] [BBC09]. Besides the desired accuracy, no other parameters have to be provided. The complete algorithm for our saddle point formulation is shown in Alg. 2.

The only expensive operations are the projections $\Pi_{\mathcal{C}}$ and $\Pi_{\mathcal{D}}$, which are efficiently computable as shown above. The algorithm converges in any case and provides explicit suboptimality bounds:

Proposition 8 *In Alg. 2, iterates u^k,v^k are primal resp. dual feasible, i.e. $u^k \in \mathcal{C}$ and $v^k \in \mathcal{D}$. Moreover, for any solution u^* of the relaxed problem (74), the relation*

$$f(u^N) - f(u^*) \leqslant f(u^N) - f_d(v^N) \leqslant \frac{2 r_1 r_2 C}{(N+1)} \tag{97}$$

holds for the the final iterates u^N,v^N.

Proof Apply [Nes04, Thm. 3] with $\hat{f}(u) = \langle u, s\rangle$, $A = L$, $\hat{\phi}(v) = \langle b, v\rangle$, $d_1(u) := \frac{1}{2}\|u - c_1\|^2$, $d_2(v) := \frac{1}{2}\|v - c_2\|^2$, $D_1 = \frac{1}{2}r_1^2$, $D_2 = \frac{1}{2}r_2^2$, $\sigma_1 = \sigma_2 = 1$, $M = 0$. □

Corollary 1 *For given $\varepsilon > 0$, applying Alg. 2 with*

$$N \quad = \quad \left\lceil 2r_1 r_2 C\varepsilon^{-1} - 1 \right\rceil \tag{98}$$

yields an ε-optimal solution of (74), i.e. $f(u^N) - f(u^*) \leqslant \varepsilon$.

For the discretization in Sect. 5.2, we may choose $c_1 = \frac{1}{l}e$ and $r_1 = \sqrt{n(l-1)/l}$, which leads to the following complexity bounds to u^N with respect to the suboptimality ε.

- *Envelope method (83).* From the previous remarks, $C = 2\sqrt{d}$. Moreover $c_2 = 0$ and by Prop. 10 (see Appendix), we have $\mathcal{D}_{\text{loc}}^d \subseteq \mathcal{B}_{\alpha_d}(0)$ with

$$\alpha_d \quad := \quad \min_i \left(\sum_j d(i,j)^2 \right)^{1/2} , \tag{99}$$

  and thus $r_2 = \alpha_d\sqrt{n}$. The total complexity in terms of the number of iterations is thus

$$O(\varepsilon^{-1} n\sqrt{d}\alpha_d) . \tag{100}$$

- *Euclidean metric method (84).* Again, $C = 2\sqrt{d}\|A\|$. We may set $c_2 = 0$ and $r_2 = \sqrt{n}$ for a total complexity of

$$O(\varepsilon^{-1} n\sqrt{d}\|A\|) . \tag{101}$$

We arrive at a parameter-free algorithm, with the exception of the desired suboptimality bound. From the sequence (u^k, v^k) we may again compute the current primal-dual gap at each iteration. As a unique feature, the number of required iterations can be determined a priori and independently of the variables in the data term, which could prove useful in real-time applications.

6.3 Douglas-Rachford Method

Here we demonstrate how to apply and evaluate the Douglas-Rachford splitting approach [DR56] to our problem. By introducing auxiliary variables, we can again reduce the inner steps to projections on the sets $\mathcal{C}$ and $\mathcal{D}$. This is in contrast to a more straightforward splitting approach such as [LKY+09], where the inner steps require to solve ROF-type problems.

Minimization of a proper, convex, lower-semicontinuous (*lsc*) function $f : X \to \mathbb{R} \cup \{-\infty, +\infty\}$ over a finite dimensional vector space $X := \mathbb{R}^n$ can be regarded as finding a zero of its – necessarily maximal monotone [RW04] – subdifferential operator $T := \partial f : X \to 2^X$. In the operator splitting framework, ∂f is assumed to be decomposable into the sum of two "simple" operators, $T = A + B$, of which forward and backward steps can practically be computed. Here, we consider the (tight) *Douglas-Rachford-Splitting* iteration [DR56, LM79] with the fixpoint step

$$\bar{u}^{k+1} \quad \in \quad (J_{\tau A}(2J_{\tau B} - I) + (I - J_{\tau B}))(\bar{u}^k) , \tag{102}$$

Algorithm 3 Douglas-Rachford Multi-Class Labeling

1: Choose $\bar{u}^0 \in \mathbb{R}^{n\times d}$, $\bar{w}^0 \in \mathbb{R}^{n\times d\times l}$ (or set $\bar{w}^0 = L\bar{u}^0$).
2: Choose $\tau > 0$.
3: $k \leftarrow 0$.
4: **while** (not converged) **do**
5: $\quad u^k \leftarrow \Pi_{\mathcal{C}}\left(\bar{u}^k - \tau s\right)$.
6: $\quad w''^k \leftarrow \Pi_{\mathcal{D}}\left(\frac{1}{\tau}(\bar{w}^k - b)\right)$.
7: $\quad u'^k \leftarrow (I + L^\top L)^{-1}\left((2u^k - \bar{u}^k) + L^\top(\bar{w}^k - 2\tau w''^k)\right)$.
8: $\quad w'^k \leftarrow Lu'^k$.
9: $\quad \bar{u}^{k+1} \leftarrow \bar{u}^k + u'^k - u^k$.
10: $\quad \bar{w}^{k+1} \leftarrow w'^k + \tau w''^k$.
11: $\quad k \leftarrow k + 1$.
12: **end while**

where $J_{\tau T} := (I + \tau T)^{-1}$ is the *resolvent* of T. Under the very general constraint that A and B are maximal monotone and $A + B$ has at least one zero, the sequence $(\bar{u}^k)$ will converge to a point $\bar{u}$, with the additional property that $u := J_{\tau B}(\bar{u})$ is a zero of T [Eck89, Thm. 3.15, Prop. 3.20, Prop. 3.19].

In particular if $f = f_1 + f_2$ for proper, convex, lsc f_i such that the relative interiors of their domains have a nonempty intersection

$$\operatorname{ri}(\operatorname{dom} f_1) \cap \operatorname{ri}(\operatorname{dom} f_2) \neq \emptyset\,, \tag{103}$$

it can be shown [RW04, Cor. 10.9] that $\partial f = \partial f_1 + \partial f_2$, and $A := \partial f_1$ as well as $B := \partial f_2$ are maximal monotone. As $x \in J_{\tau\partial f_i}(y) \Leftrightarrow x = \operatorname{argmin}\{(2\tau)^{-1}\|x - y\|_2^2 + f_i(x)\}$, the computation of the resolvents reduces to proximal point optimization problems involving only the f_i.

The idea is now to add auxiliary variables before splitting the objective in order to simplify the individual steps of the algorithm [EB92] [Set09b]. We introduce $w = Au$ and split according to

$$\begin{aligned} & \min_{u\in\mathcal{C}} \max_{v\in\mathcal{D}} \langle u, s\rangle + \langle Lu, v\rangle - \langle b, v\rangle && (104)\\ = \; & \min_{u} \langle u, s\rangle + \sigma_{\mathcal{D}}(Lu - b) + \delta_{\mathcal{C}}(u) && (105)\\ = \; & \min_{u,w} \underbrace{\delta_{Au=w}(u, w)}_{f_1} + \underbrace{\langle u, s\rangle + \delta_{\mathcal{C}}(u) + \sigma_{\mathcal{D}}(w - b)}_{f_2}\,. && (106) \end{aligned}$$

Applying the tight Douglas-Rachford iteration to this formulation yields Alg. 3.

In addition to the usual projections on $\mathcal{C}$ and $\mathcal{D}$, the algorithms requires to solve a linear equation system in step 7. However, in many cases this can be greatly accelerated by exploiting the fact that under the forward difference discretization with Neumann boundary conditions, $\operatorname{grad}^\top \operatorname{grad}$ diagonalizes with respect to the basis of the discrete cosine transform (DCT-2) [Str99] [LKY+09]:

$$\operatorname{grad}^\top \operatorname{grad} = B^{-1} \operatorname{diag}(e) B \tag{107}$$

where B is the orthogonal transformation matrix of the DCT and e is the set of eigenvalues of the discrete Laplacian. In both approaches presented above, L is of the form $L = A \otimes \operatorname{grad}$ for some (possibly identity) matrix $A \in \mathbb{R}^{k\times l}, k \leqslant l$. We first compute the singular value decomposition $A = U \operatorname{diag}(d) V^\top$ with orthogonal matrices

$U \in \mathbb{R}^{k\times k}$, $V \in \mathbb{R}^{k\times l}$ and a vector $d \in \mathbb{R}^k$. Then

$$\begin{aligned} & (I + L^\top L)^{-1} && (108)\\ = \;& I + (V \otimes I_n)(-I + (I_k \otimes B^{-1}) \cdot \ldots && (109)\\ & (I + \operatorname{diag}(d) \otimes \operatorname{diag}(e))^{-1}(I_k \otimes B))(V^\top \otimes I_n)\,. \end{aligned}$$

Thus step 7 can be achieved fast and accurately through matrix multiplications with V, discrete cosine transforms and one $O(nk)$ product for inverting the inner diagonal matrix.

Interestingly, in addition to convergence of the primal iterates (u^k), we can show that the sequence (w''^k) from Alg. 3 actually converges to a solution of the dual problem:

Proposition 9 *Let $\mathcal{C}$, $\mathcal{D}$ bounded closed convex sets with $\operatorname{ri}(\mathcal{C}) \neq \emptyset$ and $\operatorname{ri}(\mathcal{D}) \neq \emptyset$. Then Alg. 3 generates a sequence of primal/dual feasible pairs $(u^k, w''^k) \in \mathcal{C} \times \mathcal{D}$ such that, for any saddle point (u^*, v^*) of the relaxed problem (74),*

$$f(u^k) \overset{k\to\infty}{\to} f(u^*) = f_d(v^*)\,, \tag{110}$$

$$f_d(w''^k) \overset{k\to\infty}{\to} f_d(v^*) = f(u^*)\,. \tag{111}$$

In addition,

$$f_d\left(w''^k\right) \leqslant f_d(v^*) = f(u^*) \leqslant f(u^k)\,. \tag{112}$$

holds. Therefore

$$f(u^k) - f(u^*) \;\leqslant\; f(u^k) - f_d(w''^k)\,, \tag{113}$$

provides an upper bound for the suboptimality of the current iterate.

Proof See Appendix. □

Thus the Douglas-Rachford approach also allows to use the primal-dual gap

$$f(u^k) - f_d(w''^k) \tag{114}$$

as a stopping criterion.

6.4 Projection on the Dual Constraint Set

For the Euclidean metric approach, projection on $\mathcal{D}^I_{\text{loc}}$ and thus on $\mathcal{D}$ is trivial:

$$\Pi_{\mathcal{D}^I_{\text{loc}}}(v) \;=\; \begin{cases} v, & \|v\| \leqslant 1\,, \\ \frac{v}{\|v\|}\,, & \text{otherwise}\,. \end{cases} \tag{115}$$

Projection on $\mathcal{D}^d_{\text{loc}}$ for the general metric case is more involved. We represent $\mathcal{D}^d_{\text{loc}}$ as the intersection of convex sets,

$$\mathcal{D}^d_{\text{loc}} = \mathcal{R} \cap \mathcal{S}\,,\ \mathcal{R} := \{v \in \mathbb{R}^{d\times l} | \sum_i v^i = 0\}\,, \tag{116}$$

$$\mathcal{S} := \bigcap_{i<j} \mathcal{S}^{i,j}\,,\ \mathcal{S}^{i,j} := \{v \in \mathbb{R}^{d\times l} | \|v^i - v^j\| \leqslant d(i,j)\}\,.$$

Algorithm 4 Dykstra's Method for Projecting onto the Intersection of Convex Sets

Associate with each $(i,j), i<j$ a unique index $t \in \{1,\dots,k\}$, $k := l(l-1)$.
$x \leftarrow v \in \mathbb{R}^{d\times l}$.
$y^1,\dots,y^k \leftarrow 0 \in \mathbb{R}^{d\times l}$.
while (x not converged) **do**
  for $t = 1,\dots,k$ **do**
    $x' \leftarrow \Pi_{S^{(i_t,j_t)}}(x + y^t)$.
    $y^t \leftarrow x + y^t - x'$.
    $x \leftarrow x'$.
  end for
end while

Since $\mathcal{R}$ and $\mathcal{S}$ are orthogonal, the projection decouples:

$$\Pi_{\mathcal{D}^d_{\mathrm{loc}}}(v) = \Pi_{\mathcal{R}}(\Pi_{\mathcal{S}}(v))\,. \tag{117}$$

We then follow the idea of [CCP08] to use Dykstra's method [BD86] for iteratively computing an approximation of $\Pi_{\mathcal{S}}$ using only projections on the individual sets $\mathcal{S}^{i,j}$. In our case, $\Pi_{\mathcal{S}^{i,j}}$ can be stated in closed form:

$$\Pi_{\mathcal{S}_{i,j}}(v) \quad = \quad \begin{cases} v, & \|v^i - v^j\| \leqslant d(i,j)\,, \\ (w^1,\dots,w^l), & \text{otherwise}\,, \end{cases} \tag{118}$$

where

$$w^k \quad = \quad \begin{cases} v^k, & k \notin \{i,j\}\,, \\ v^i - \frac{\|v^i-v^j\|-d(i,j)}{2}\frac{v^i-v^j}{\|v^i-v^j\|}, & k = i\,, \\ v^j + \frac{\|v^i-v^j\|-d(i,j)}{2}\frac{v^i-v^j}{\|v^i-v^j\|}, & k = j\,. \end{cases} \tag{119}$$

The complete method is outlined in Alg. 4. While the sequence of y may be unbounded, x converges to $\Pi_{\mathcal{S}}(v)$ (see [GM89] [Xu00] for a convergence analysis).

6.5 Computing the Objective

For computing the primal-dual gap and the binarization step, it is necessary to evaluate the objective for a given u. Unfortunately, in the general case this is nontrivial as the objective's integrand Ψ is defined implicitly as

$$\Psi(z) \quad = \quad \max_{v\in\mathcal{D}^d_{\mathrm{loc}}} \langle z, v\rangle\,. \tag{120}$$

We exploit that $\Pi_{\mathcal{D}}$ can be computed and use an iterative gradient-projection method,

$$v^{k+1} \quad \leftarrow \quad \Pi_{\mathcal{D}}(v^k + \tau z) \tag{121}$$

for some $\tau > 0$, starting with $v^0 = z$. As the objective in (120) is linear and Ψ is bounded from above as $\mathcal{D}^d_{\mathrm{loc}}$ is bounded, convergence follows for any given step size τ [Gol64] [LP66] [WX04]:

$$\lim_{k\to\infty} \langle z, v^k\rangle = \Psi(z)\,. \tag{122}$$

There is a trade-off in choosing the step size, as large τ lead to a smaller number of outer iterations, but an increased number of nontrivial operations in the projection. We chose $\tau = 2$, which worked well for all examples. It is also possible to use any of the other nonsmooth optimization methods presented above.

7 Experiments

Regarding the practical performance of the presented approaches, we focused on two main issues: convergence speed and tightness of the relaxation. We will first quantitatively compare the presented algorithms, and then provide some results on the effect of the Euclidean metric vs. the envelope regularization.

The algorithms were implemented in Matlab with some core functions, such as the computation of the gradient and the projections on the dual sets, implemented in C++. We used Matlab's built-in FFTW interface for computing the DCT for the Douglas-Rachford approach. All experiments were conducted on an Intel Core2 Duo 2.66 GHz with 4 GB of RAM and 64-bit Matlab 2009a.

7.1 Relative Performance

To compare the convergence speed of the three different optimization methods, we computed the relative primal-dual gap at each iteration as outlined in Sect. 6. As it bounds the suboptimality of the current iterate (see Sect. 5.5), it constitutes a reliable and convenient criterion for performance comparison.

Unfortunately the gap is not available for the envelope method, as the primal objective must be evaluatable. Using an iterative approximation as in Sect. 6.5 is not an option, as these methods can only provide a *lower* bound for the objective. This would lead to an underestimation of the gap, which is critical as one is interested in the behavior when the gap is very close to zero. Therefore we restricted the core performance tests to the Euclidean metric regularizer.

In order to make a fair comparison we generally analyzed the progression of the gap with respect to computation time, rather than the number of iterations.

For the first tests we used the synthetical 256×256 "four colors" input image (Fig. 8). It represents a typical segmentation problem with several objects featuring sharp corners and round structures above a uniform background. The label set consists of three classes for the foreground and one class for the background. The image was overlaid with iid Gaussian noise with $\sigma = 1$ and truncated to $[0, 1]$ on all RGB channels. We used a simple ℓ^1 data term, $s_i(x) = \|g(x) - c^i\|_1$, where $g(x) \in [0, 1]^3$ are the RGB color values of the input image in x, and c^i is the prototypical color for the i-th class.

The time-based analysis shows that FPD and Douglas-Rachford perform very similar, while the Nesterov method falls behind in both the primal and the dual objective by a large margin (Fig. 9).

The picture changes when considering the gap with respect to the number of iterations rather than time, eliminating influences of the implementation and system architecture. To achieve the same accuracy, Douglas-Rachford requires only one third of the iterations compared to FPD (Fig.10). This advantage does not fully translate to the time-based analysis as the DCT steps increase the per-step computational cost significantly. However in this example the projections on the sets $\mathcal{C}$ and $\mathcal{D}$ were relatively cheap compared to the DCT. In situations where the projections dominate the time per step, the reduced iteration count can be expected to lead to an almost equal reduction in computation time.

One could ask if these relations are typical to the synthetical data used. However we found them confirmed on a large majority of the problems tested. As one example of a real-world example, consider the "leaf" image (Fig. 4). We computed a segmentation

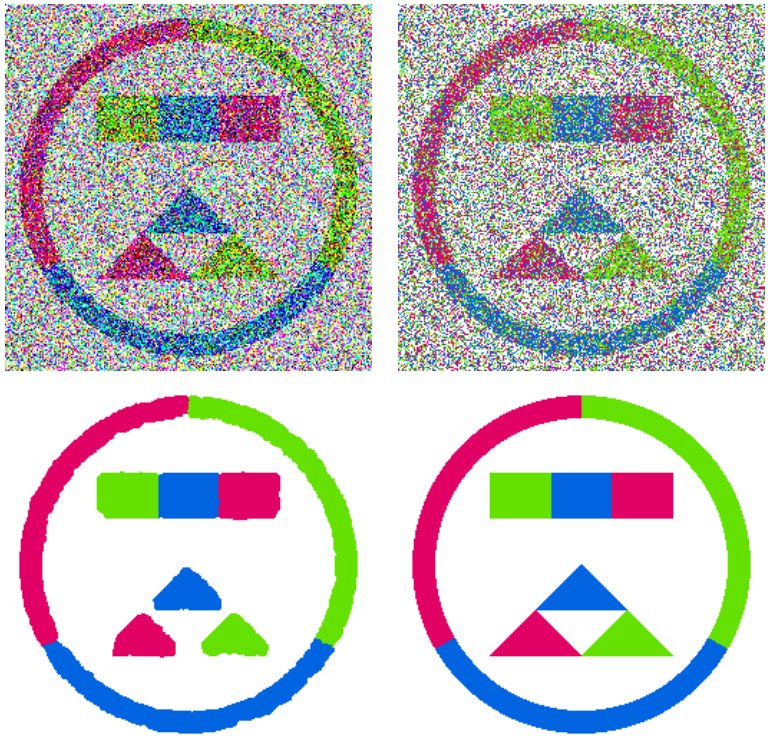

Figure 8: Synthetical "four colors" input image for the performance tests. **Top row:** Input image with Gaussian noise, $\sigma = 1$; local labeling without regularizer. **Bottom row:** Result with the Potts regularizer and Douglas-Rachford optimization; ground truth.

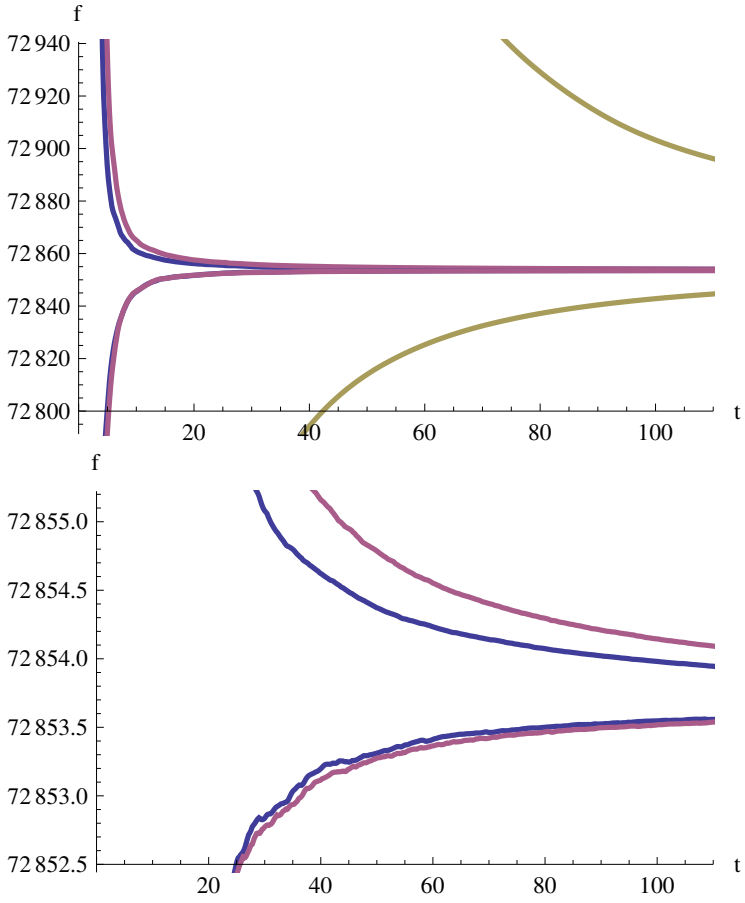

Figure 9: Convergence speed for the "four colors" image in Fig. 8. **Top:** Primal (upper) and dual (lower) objectives vs. computation time for the (from top to bottom) Nesterov, Fast Primal-Dual (FPD) and Douglas-Rachford methods. **Bottom:** Detailed view of the FPD and DR methods. The primal and dual objectives provide upper and lower bounds for the objective of the true optimum. Douglas-Rachford and FPD perform equally, while the Nesterov method falls behind.

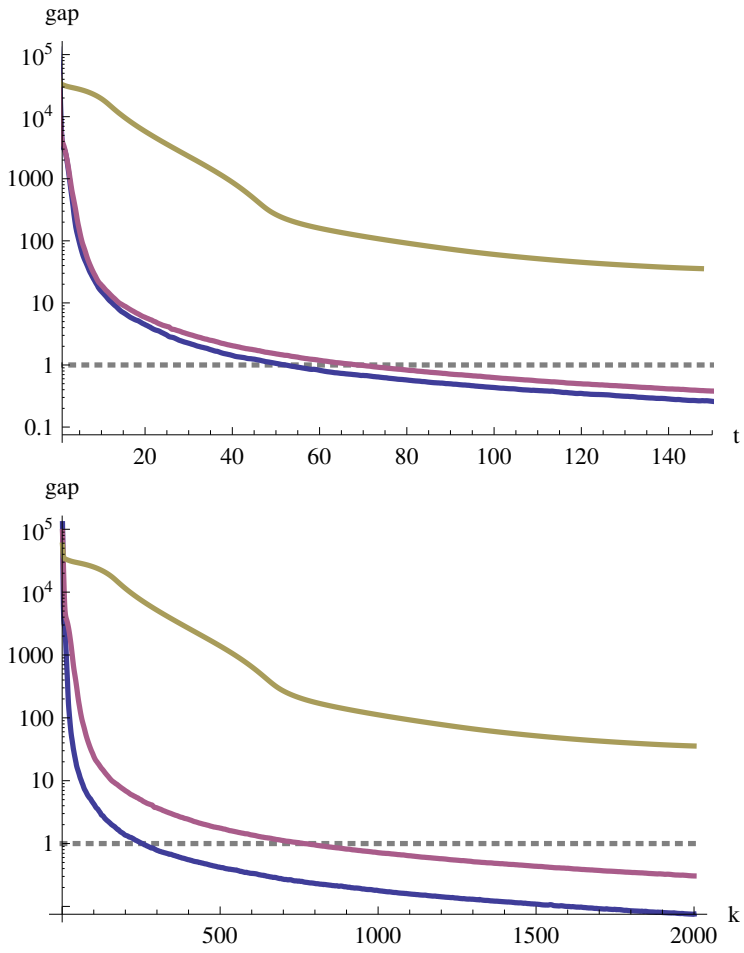

Figure 10: Primal-Dual gap for Fig. 9 with respect to time and number of iterations. **Top:** Primal-Dual gap vs. time and number of iterations. The Nesterov method (top) again falls behind, while FPD (center) and Douglas-Rachford (bottom) are equally fast. **Bottom:** Primal-Dual gap vs. number of iterations. The Douglas-Rachford method requires only one third of the FPD iterations, which makes it suitable for problems with expensive inner steps.

into 12 classes with Potts regularizer, again based on the ℓ^1 distances for the data term, with very similar relative performance as for the "four colors" image (Fig. 11).

7.2 Number of Variables and Regularization Strength

To see how the presented methods scale with increasing image size, we evaluated the "four colors" image at 20 different scales ranging from 16×16 to 512×512. Note that for a constant grid spacing of 1, the regularizer weights must be scaled proportionally to the image width resp. height in order to obtain structurally comparable results, and not mix up effects of the problem size and of the regularization strength.

In order to compensate for the increasing number of variables, the stopping criterion was based on the relative gap. The algorithms terminated as soon as the relative gap fell below 10^{-4}. The Nesterov method consistently produced gaps in the 10^{-3} range and never achieved the threshold within the limit of 2000 iterations. Douglas-Rachford and FPD scale only slightly superlinearly with the problem size, which is quite a good result for such comparatively simple first-order methods (Fig. 12).

While we deliberately excluded influences of the regularizer in the previous experiment, it is also interesting to see how algorithms cope with varying regularization strength. We fixed a resolution of 256×256 and performed the same analysis as above, scaling the regularization term by an increasing sequence of λ in the $[0.1, 5]$ range (Fig. 13).

Interestingly, for low regularization, where much of the noise remains in the solution, FPD clearly takes the lead. For scenarios with large structural changes, Douglas-Rachford performs better. We observed two peaks in the runtime plot which we cannot completely explain. However we found that at the first peak, structures in the image did not disappear at in parallel but rather one after the other, which might contribute to the slower convergence. Again, the Nesterov method never achieved the required accuracy.

7.3 Breaking Points

We have no clear explanation why the Nesterov method appears to almost always fall behind. However it is possible to compare its behavior with the a priori bound from Prop. 8. By inspecting the derivations in the original work [Nes04], it can be seen that exactly one half of the the final bound comes from the smoothing step, while the other half is caused by the finite number of iterations:

$$\delta_{\mathrm{total}} = \delta_{\mathrm{smooth}} + \delta_{\mathrm{iter}}, \quad \text{where} \quad \delta_{\mathrm{smooth}} = \delta_{\mathrm{iter}}\,. \tag{123}$$

Moreover, δ_{iter} decreases with $1/(k+1)^2$, which gives an a priori per-iteration suboptimality bound of

$$\delta^k_{\mathrm{total}} = \delta_{\mathrm{smooth}} + \left(\frac{N+1}{k+1}\right)^2 \delta_{\mathrm{iter}}\,. \tag{124}$$

On the "four colors" image, the actual gap stays just below $\delta^k_{\mathrm{total}}$ in the beginning (Fig. 14). This implies that the theoretical bound can hardly be improved, e.g. by choosing constants more precisely. Unfortunately, the bound is rather large at $\delta_{\mathrm{total}} = 256.8476$ for 2000 iterations. While the Nesterov method outperforms the theoretical bound δ_{total} by a factor of 2 to 10 and even drops well below the worst-case smoothing error δ_{smooth}, it still cannot compete with the other methods, which achieve a gap of 0.3052 (FPD) resp. 0.0754 (Douglas-Rachford).

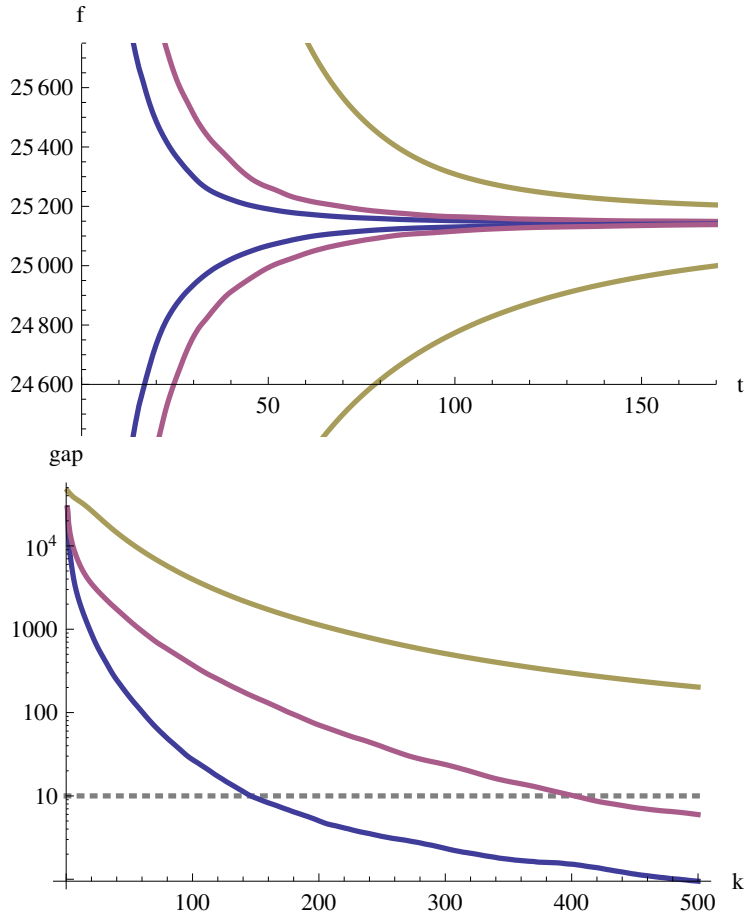

Figure 11: Objectives and primal-dual gap for the real-world leaf image in Fig. 4 with 12 classes and Potts potential. **Top:** Primal (upper) and dual (lower) objectives vs. time. The Nesterov method (top) falls behind the FPD (center) and Douglas-Rachford (bottom) methods. **Bottom:** Gap vs. number of iterations. As with the synthetical four-colors image (Fig. 10), the Douglas-Rachford approach reduces the number of required iterations by approximately a factor of 3.

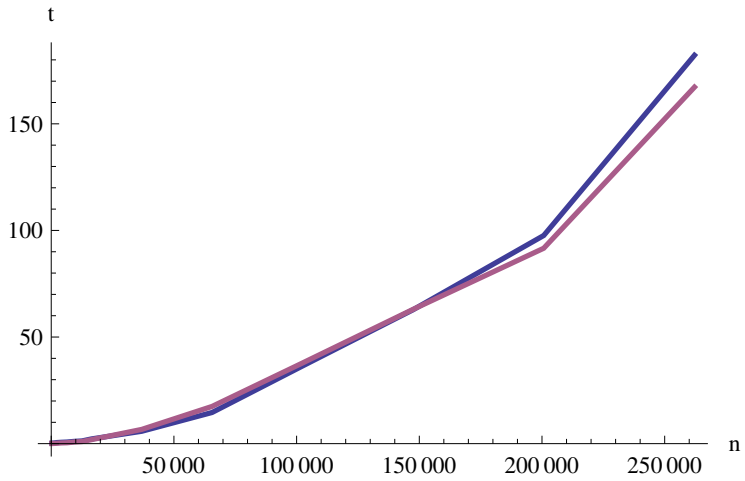

Figure 12: Computation time for increasing problem size for the Douglas-Rachford (top, dark) and FPD (bottom, light) methods. Shown is the time in seconds required to achieve a relative gap of 10^{-4}. The computational effort scales slightly superlinearly with the number of pixels. The Nesterov method never converged to the required accuracy within the limit of 2000 iterations.

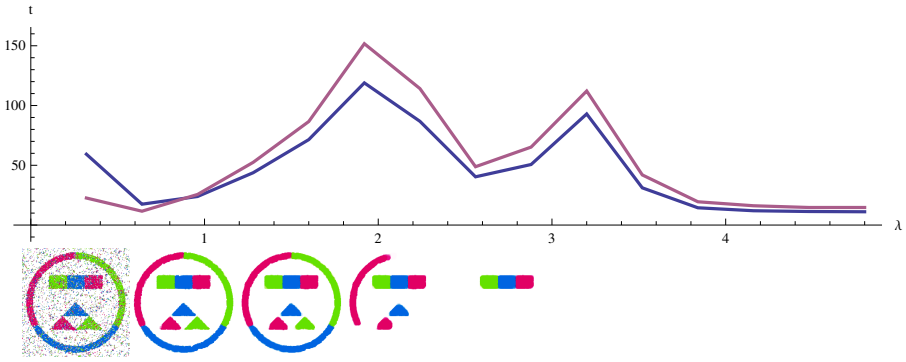

Figure 13: Computation time for varying regularization strength λ for the Douglas-Rachford (top, dark) and FPG (bottom, light) methods. The images at the bottom show the final result for the λ above. FPD is strong on low regularization problems, while Douglas-Rachford is better suited for problems with large structural changes. The Nesterov method never achieved the relative gap of 10^{-5} within 2000 iterations.

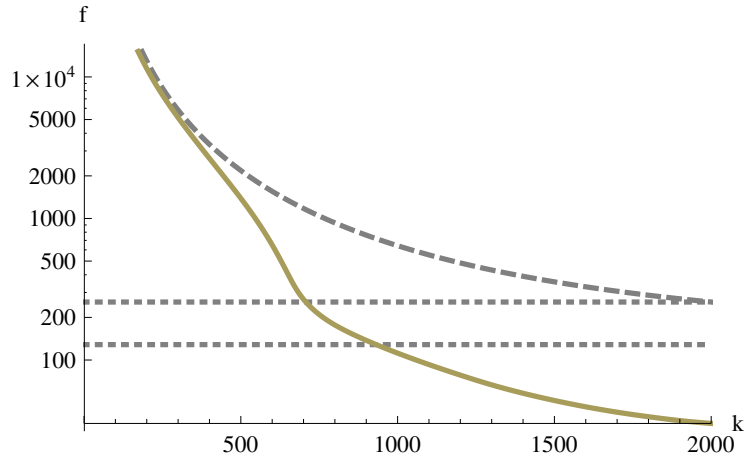

Figure 14: Theoretical vs. practical performance of the Nesterov method for Fig. 9. As expected, the method stays below the theoretical per-iteration bound δ^k_{total} (dashed). At the final iteration, the worst-case total bound δ_{total} (dotted, top) is outperformed by a factor of 7, which implicates that the error introduced by smoothing is also well below its worst-case bound δ_{smooth} (dotted, bottom).

There is an interesting extreme case where the Nesterov method seems to come to full strength. Consider the noise-free "triple point" inpainting problem (15). The triple junction in the center can only be reconstructed by the Potts regularizer, as the ℓ^1 data term has been blanked out around the center. By just reversing the sign of the data term, one gets the "inverse triple point" problem, an extreme case that has also been studied in [CCP08] and shown to be an example where the relaxation leads to a strictly nonbinary solution.

On the inverse problem the Nesterov method catches up and even surpasses FPD. This stands in contrast with the regular triple point problem, where all methods perform as usual. We conjecture that this sudden strength comes from the inherent averaging over all previous gradients (step 7 in Alg. 2): in fact, on the inverse problem Douglas-Rachford and FPD display a pronounced oscillation in the primal and dual objectives, which is accompanied by slow convergence. In contrast, the Nesterov method consistently shows a monotone and smooth convergence.

7.4 Performance for the Envelope Relaxation

Undoubtedly, the difficulty when using the envelope based regularizer comes from the slow and inexact projection steps which have to be approximated iteratively. Therefore we re-evaluated the "four colors" benchmark image with the envelope regularizer. The iterative Dykstra projection (Alg. 4) was stopped when the iterates differed by at most $\delta = 10^{-2}$, with an additional limit of 50 iterations. While the gap cannot be computed in this case, the dual objective can still be evaluated and provides an indicator for the convergence speed.

We found that in comparison to the Euclidean metric regularizer from the previous examples, the margin between FPD and Douglas-Rachford increases significantly. This is consistent with the remarks in Sect. 7.1: the lower iteration count of the Douglas-Rachford method becomes more important, as the projections dominate

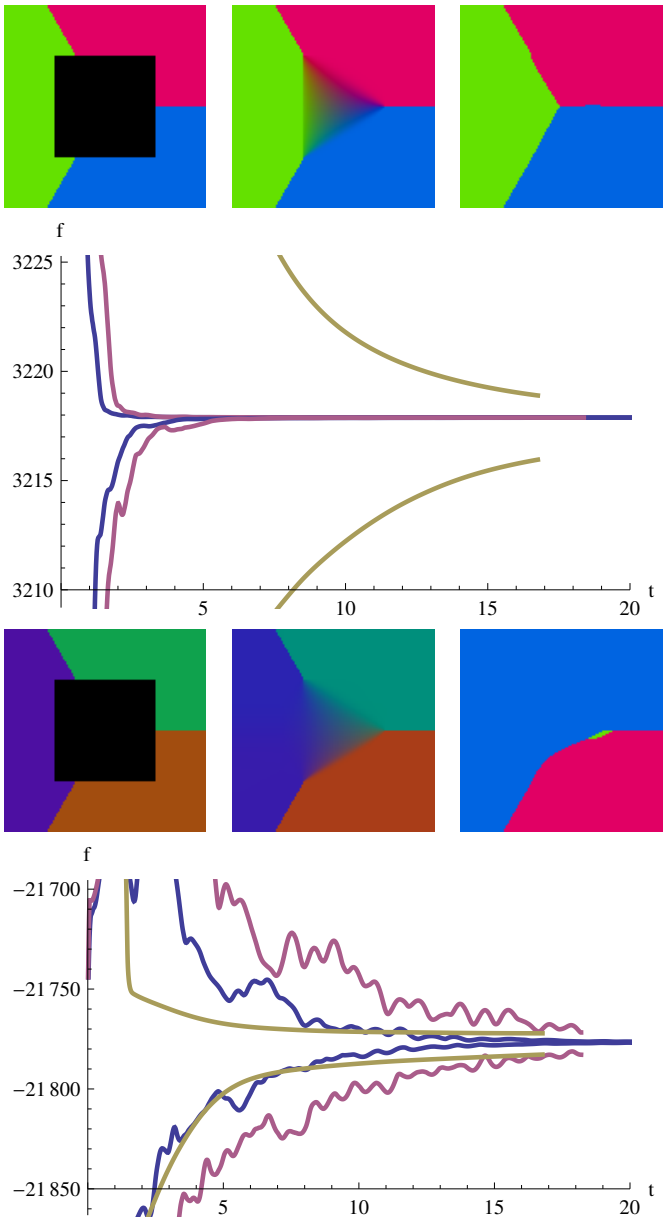

Figure 15: Primal and dual objectives for the triple point (top) and inverse triple point (bottom) inpainting problems. **Left to right**: Input image with zeroed-out region around the center; relaxed solution; binarized solution. The triple junction in the center has to be reconstructed solely by the Potts regularizer. The inverse triple point problem exhibits a strictly nonbinary relaxed solution. Shown below are the primal (top) and dual (bottom) objectives. For the inverse triple point, Douglas-Rachford (bottom) and FPD (center) show an oscillatory behavior which slows down convergence. The Nesterov approach (top) does not suffer from oscillation due to the inherent averaging, and surpasses FPD on the inverse problem.

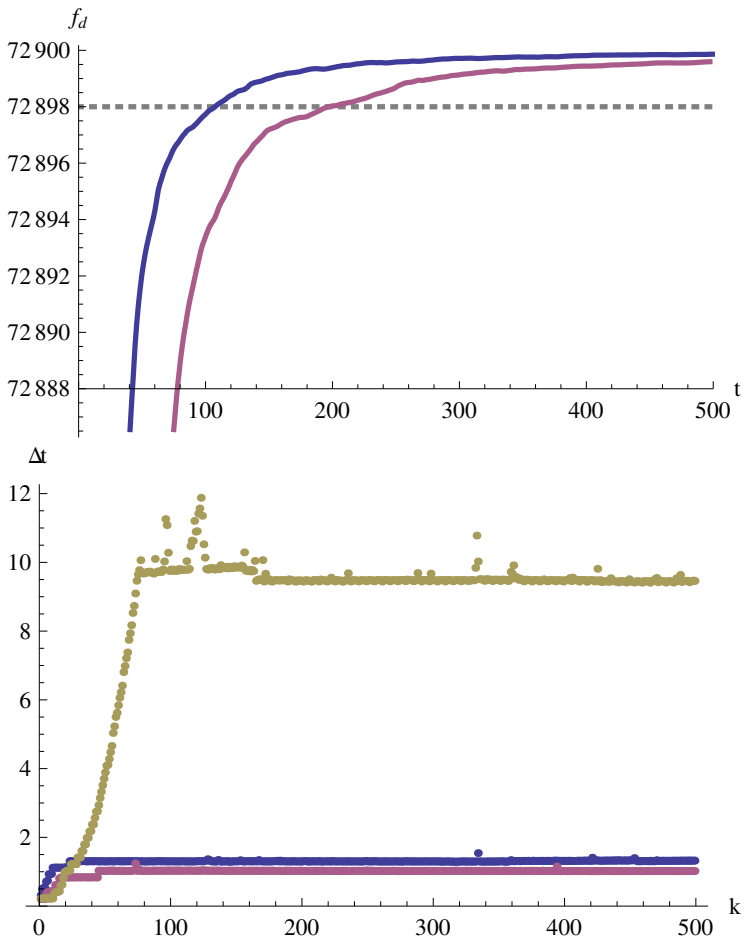

Figure 16: Performance on the "four colors" image with Potts interaction potential and the envelope regularizer. **Top:** Dual objectives of for Douglas-Rachford (top) and FPD (bottom) vs. time. The reduced iteration count of the Douglas-Rachford method becomes more apparent in the time plot as the time per iteration is now dominated by the projection rather than the DCT. **Bottom:** Time per iteration for Nesterov (top), Douglas-Rachford (center) and FPD (bottom). The Nesterov method fails to converge as it accumulates errors from the approximate projections, which in turn leads to slower and more inexact projections.

the per-iteration runtime (Fig. 16).

Surprisingly the Nesterov method did not converge at all. On inspecting the per-iteration runtime, we found that after a few steps the iterative projections became very slow and eventually terminated after the limit of 50 iteration with δ remaining between 2 and 5. In contrast, 20 Dykstra iterations were usually sufficient to obtain $\delta = 10^{-9}$ (Douglas-Rachford) or even 10^{-11} (FPD).

We again attribute this to the averaging property of the Nesterov method: as it accumulates the results of the previous projections, errors from the inexact projections build up. This is accelerated by the dual variables quickly becoming infeasible with increasing distance to the dual feasible set, which in turn puts higher demands on the iterative projections. Douglas-Rachford and FPD did not display this behavior and consistently required 5 to 6 Dykstra iterations from the first to the last iteration.

7.5 Tightness of the Relaxations

Besides the properties of the optimization methods, it is interesting to study the effect of the relaxation – i.e. Euclidean metric or envelope type – on the relaxed and binarized solutions.

To get an insight into the tightness of the relaxations, we used the Douglas-Rachford method to repeat the "triple point" inpainting experiment from [CCP08] with both relaxations (Fig. 17). Despite the inaccuracies in the projections, the envelope regularizer generates a nearly binary solution: 97.6% of all pixels were assigned "almost binary" labels with an ℓ^∞ distance of less than 0.05 to one of the unit vectors $\{e^1, \dots, e^l\}$. For the Euclidean metric method, this constraint was only satisfied at 88.6% of the pixels. The result for the envelope relaxation is very close to the sharp triple junction one would expect from the continuous formulation, and shows that the envelope relaxation is tighter than the Euclidean metric method.

However, after binarization both approaches generate almost identical discrete results. The Euclidean metric method was more than four times faster, with 41.1 seconds per 1000 iterations vs. 172.16 seconds for the envelope relaxation, which used 8–11 Dykstra steps per iteration.

While the triple point is a problem specifically designed to challenge the regularizer, real-world images usually contain more structure as well as noise, while the data term is available for over most or all of the image. To see if the above results also hold under these conditions, we repeated the previous experiment with the "sailing" image and four classes (Fig. 18). The improved tightness of the envelope relaxation was also noticeable, with 96.2% vs. 90.6% of "almost binary" pixels. On the downside, due to the larger number of labels and the larger image size of 360×240, run times increased to 4253 (envelope) vs. 420 (Euclidean metric) seconds.

The relaxed as well as the binarized solutions show some differences but are hard to distinguish visually. It is difficult to pinpoint if these differences are caused by the tighter relaxation or by numerical issues: while the Douglas-Rachford method applied to the Euclidean metric relaxation converged to a final relative gap of $1.5 \cdot 10^{-6}$, no such bound is available to estimate the accuracy of the solution for the envelope relaxation, due to the inexact projections and the intractable primal objective.

7.6 Binarization and Global Optimality

As an example for a problem with a large number of classes, we analyzed the "penguin" problem from [SZS+06]. We chose 64 labels corresponding to 64 equally spaced

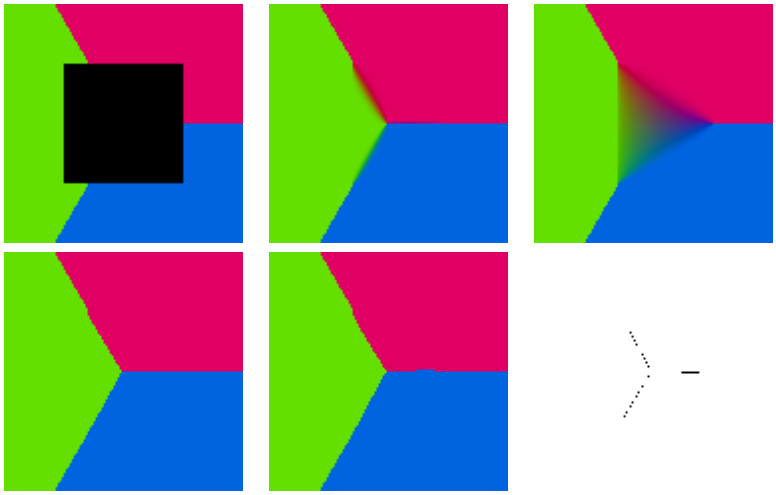

Figure 17: Tightness of the relaxation. **Top row:** In the input (left), the data term was blanked out in a quadratic region. All structure within the region is generated purely by the regularizer with a standard Potts interface potential. The envelope relaxation is tighter and generates a much more "binary" solution (center) than the Euclidean metric method (right). **Bottom row:** After binarization of the relaxed solutions, the envelope (left) and Euclidean metric (center) methods generate essentially the same solution, as can be seen in the difference image (right). The Euclidean metric method performed more than four times faster due to the inexpensive projections.

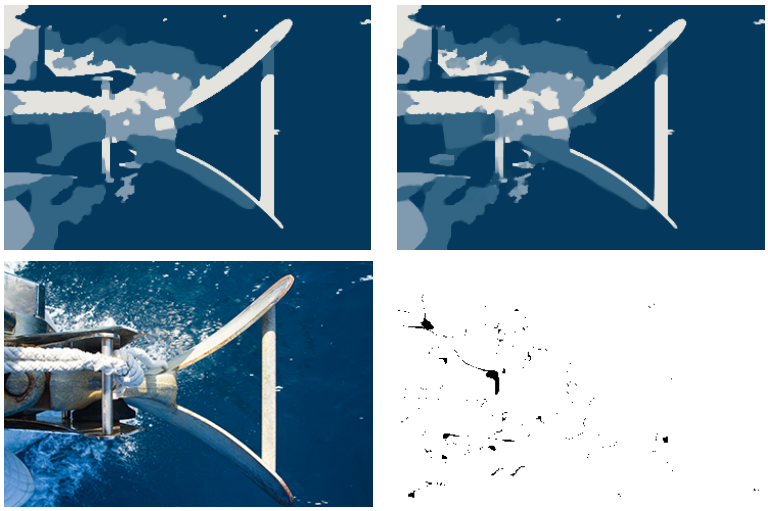

Figure 18: Effect of the choice of relaxation method on the real-world "sailing" image (image courtesy of F. Becker). **Top row:** Four-class segmentation using envelope (left) and Euclidean metric (right) methods. Shown are the solutions of the relaxed problem. **Bottom row:** Original image (left); difference image of the discretized solutions (right). While the envelope relaxation leads to substantially more "almost discrete" values in the relaxed solution, it also runs more than 10 times slower and does not provide a suboptimality bound. The generated solutions are visually almost identical.

gray values. The input image contains a region where the image must be inpainted in addition to removing the considerable noise. Again the data term was generated by the ℓ^1 distance, which reduces here to the absolute difference of the gray values. In order to remove noise but not overly penalize hard contrasts, such as between the black wing and the white front, we chose a regularizer based on the truncated linear potential as introduced in Sect. 4.4.

Due to the large number of labels, this problem constitutes an example where the Euclidean metric approach is very useful. As the complexity of the projections for the envelope relaxation grows quadratically with the number of labels, computation time becomes prohibitively long for a moderate amount of classes. In contrast, the Euclidean metric method requires considerably less computational effort and still approximate the potential function to a reasonable accuracy (Fig. 7).

In the practical evaluation, the Douglas-Rachford method converged in 1000 iterations to a relative gap of $8.3 \cdot 10^{-4}$, and recovered both smooth details near the beak, and hard edges in the inpainting region (Fig. 19).

This example also clearly demonstrates the superiority of the improved binarization scheme proposed in Sect. 5.6. Opposed to the first-max scheme, it generated considerably less noise. The energy increased only by 2.78% compared to 15.78% for the first-max approach.

The low energy increase is directly related to global optimality for the discrete problem: as the relaxed solution is provably nearly optimal, we conclude that the energy of the *binarized* solution must lie within 2.78% of the optimal energy for the original *combinatorial* problem (1). Similar results were obtained for the other images: 5.64% for the "four colors" demo, 1.02% for the "leaf" image and 0.98% for the "triple point" problem.

These numbers indicate that the relaxation seems to be quite tight in many cases, and allows to recover good approximations for the solution of the discrete combinatorial labeling problem by solving the convex relaxed problem.

8 Conclusion and Further Work

We hope that the present work provides a reference and framework for continuous multilabeling approaches. The presented algorithms are robust and fast and are suited for massive parallelization. From the experiments it became clear that solving the convex relaxed problem allows to recover very good solutions for the original combinatorial problem in many cases.

The performance evaluations showed that the Douglas-Rachford method consistently requires about one third of the iterations compared to the Fast Primal-Dual method. For low regularization and fast projections, FPD outperforms the Douglas-Rachford method. In all other cases, Douglas-Rachford performs equally or better, with a speedup of 2-3 if the projections are expensive. Overall, the proposed Douglas-Rachford method approach appears to be a solid all-round method that also handles extreme cases well.

From the viewpoint of numerics, in our evaluations we completely excluded the effect of choosing different step sizes for the Douglas-Rachford method. Also, it seems as if the smooth optimization step in the Nesterov method usually performs much better than its theoretical bound. Adjusting the strategy for choosing the smoothing parameter could yield a faster overall convergence and possibly render the method competitive. In this regard, it would also be interesting to include the inexactness of the projections

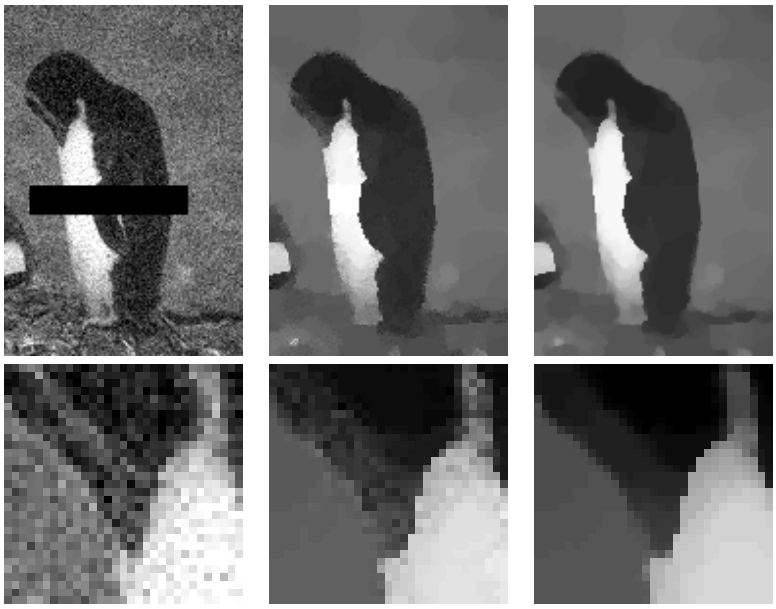

Figure 19: Denoising/inpainting problem with 64 classes and the nontrivial truncated linear potential approximated by an Euclidean metric. **Left to right:** Noisy input image with inpainting region marked black [SZS$^+$06]; result with first-max binarization; result with the proposed binarization scheme (93). The first-max method introduces noticeable noise in the binarization step. The proposed method takes into account the non-uniformity of the used "cut-linear" potential (Fig. 7), resulting in a clean labeling and an energy increase of only 2.78% vs. 15.78% for the first-max method. This shows that the obtained solution solves the originally combinatorial multiclass labeling problem to a suboptimality of 2.78%.

into the convergence analysis. There are also several theoretical questions left, such as how to include non-metric distances into the continuous formulation.

In any case we think that the present framework unites continuous and discrete worlds in an appealing way, and hopefully contributes to reducing the popularity gap compared to purely grid-based and graph cut methods.

Acknowledgments. The authors would like to thank Simon Setzer for the stimulating discussion.

9 Appendix

Proposition 10 *Let* $v = (v^1, \dots, v^l) \in \mathcal{D}^d_{\mathrm{loc}}$, *then*

$$\|v\| \leqslant \min_i \Big(\sum_j d(i,j)^2 \Big)^{\frac{1}{2}} . \tag{125}$$

Proof of Prop. 10. From the constraint $\sum_j v^j = 0$ we deduce, for any $i \in \{1, \dots, l\}$,

$$\begin{aligned} \sum_j \|v^j\|^2 &\leqslant \sum_j \|v^j\|^2 - 2\langle v^i, \sum_j v^j \rangle + l\|v^i\|^2 \\ &= \sum_j \big(\|v^j\|^2 - 2\langle v^i, v^j \rangle + \|v^i\|^2 \big) \\ &= \sum_j \|v^j - v^i\|^2 \leqslant \sum_j d(i,j)^2. \end{aligned}$$

As i was arbitrary this proves the assertion.

Proof of Prop. 9. The idea of the proof is to show that the sequence (z'^k) is exactly the minimizing sequence produced by the algorithm applied to the dual problem (75), with step size $1/\tau$. Thus, if the dual algorithm converges, (z'^k) converges to the solution of the dual problem, which proves the proposition.

First note that the formulation of the primal problem already covers the dual problem by the substitutions (replacing w with z for readability)

$$v \leftrightarrow u,\ \mathcal{C} \leftrightarrow \mathcal{D}\,,\ b \leftrightarrow s\,,\ w \leftrightarrow z\,,\ L \leftrightarrow -L^\top\,, \tag{126}$$

which lead to the dual algorithm, Alg. 5. We first show convergence of the primal and dual Douglas-Rachford algorithms, Alg. 3 and Alg. 5. As both parts of the objective are proper, convex and lsc, it suffices to show additivity of the subdifferentials [RW04, Cor. 10.9] [Eck89, Thm. 3.15] [Eck89, Prop. 3.20, Prop. 3.19] [EB92]. We denote

$$\begin{aligned} g(u,w) &:= \delta_{\mathcal{C}}(u) + \max_{v \in \mathcal{D}} \left(\langle u, s \rangle + \langle w, v \rangle - \langle v, b \rangle \right) \\ &= \delta_{\mathcal{C}}(u) + \langle u, s \rangle + \sigma_{\mathcal{D}}(w - b), \qquad (127) \\ h(v,z) &:= \delta_{\mathcal{D}}(v) + \max_{u \in \mathcal{C}} \left(\langle v, b \rangle + \langle z, v \rangle - \langle u, s \rangle \right) \\ &= \delta_{\mathcal{D}}(v) + \langle v, b \rangle + \sigma_{\mathcal{C}}(z - s)\,. \qquad (128) \end{aligned}$$

Algorithm 5 Dual Douglas-Rachford for Multi-Class Labeling

Choose $\bar{v}^0 \in \mathbb{R}^{n\times d\times l}$, $\bar{z}^0 \in \mathbb{R}^{n\times d}$.
Choose $\tau_d > 0$.
$k \leftarrow 0$.
while (not converged) **do**
  $v^k \leftarrow \Pi_{\mathcal{D}}\left(\bar{v}^k - \tau_d b\right)$.
  $z''^k \leftarrow \Pi_{\mathcal{C}}\left(\frac{1}{\tau_d}(\bar{z}^k - s)\right)$.
  $v'^k \leftarrow (I + LL^\top)^{-1}\left((2v^k - \bar{v}^k) + (-L)(\bar{z}^k - 2\tau_d z''^k)\right)$.
  $z'^k \leftarrow (-L^\top)v'^k$.
  $\bar{v}^{k+1} \leftarrow \bar{v}^k + v'^k - v^k$.
  $\bar{z}^{k+1} \leftarrow z'^k + \tau_d z''^k$.
  $k \leftarrow k + 1$.
end while

Then, due to the boundedness of $\mathcal{C}$ and $\mathcal{D}$,

$$\begin{aligned} & \mathrm{ri}\,(\mathrm{dom}\, g) \cap \mathrm{ri}\,(\mathrm{dom}\, \delta_{Lu=w}) \\ = \;& (\mathrm{ri}\,(\mathcal{C}) \times \mathbb{R}^p) \cap \{Lu = w\} && (129) \\ = \;& \{(u, Lu) | u \in \mathrm{ri}(\mathcal{C})\}, \\ & \mathrm{ri}\,(\mathrm{dom}\, h) \cap \mathrm{ri}\,(\mathrm{dom}\, \delta_{-L^\top v=z}) \\ = \;& (\mathrm{ri}(\mathcal{D}) \times \mathbb{R}^q) \cap \{-L^\top v = z\} && (130) \\ = \;& \{(v, -L^\top v) | v \in \mathrm{ri}(\mathcal{D})\}. \end{aligned}$$

Both of these sets are nonempty as $\mathrm{ri}(\mathcal{C}) \neq \emptyset \neq \mathrm{ri}(\mathcal{D})$. This implies additivity of the subdifferentials for the proposed objective [RW04, Cor. 10.9] and thus convergence of the tight Douglas-Rachford iteration.

We will now show that the primal and dual algorithms essentially generate the same iterates, i.e. from $\tau := \tau_p = 1/\tau_d$, $\bar{u}^k = \tau\bar{z}^k$ and $\bar{w}^k = \tau\bar{v}^k$, it follows that $\bar{u}^{k+1} = \tau\bar{z}^{k+1}$, $\bar{w}^{k+1} = \tau\bar{v}^{k+1}$ and $u^k = z''^k$, $v^k = w''^k$.

The last two equalities follow immediately. Also,

$$\begin{aligned} & \bar{v}^{k+1} = \bar{v}^k + v'^k - v^k && (131) \\ = \;& \bar{v}^k + (I + LL^\top)^{-1}((2v^k - \bar{v}^k) \\ & + (-L)(\bar{z}^k - 2\tau^{-1}z''^k)) - v^k && (132) \\ = \;& \tau^{-1}\bar{w}^k + (I + LL^\top)^{-1}((2w''^k - \tau^{-1}\bar{w}^k) \\ & + (-L)(\tau^{-1}\bar{u}^k - 2\tau^{-1}u^k)) - w''^k && (133) \\ = \;& \tau^{-1}\bar{w}^k + (I + LL^\top)^{-1}(2w''^k - \tau^{-1}\bar{w}^k) - w''^k \\ & -(I + LL^\top)^{-1}L(\tau^{-1}\bar{u}^k - 2\tau^{-1}u^k). && (134) \end{aligned}$$

By the Woodbury identity, $(I + LL^\top)^{-1} = I - L(I + L^\top L)^{-1}$ and in particular $(I + LL^\top)^{-1}L = L(I + L^\top L)^{-1}$. Consequently,

$$
\begin{aligned}
\bar{v}^{k+1} &= \tau^{-1}\bar{w}^k + 2w''^k - \tau^{-1}\bar{w}^k \\
&\quad -L(I + L^\top L)^{-1}L^\top(2w''^k - \tau^{-1}\bar{w}^k) - w''^k \\
&\quad -L(I + L^\top L)^{-1}(\tau^{-1}\bar{u}^k - 2\tau^{-1}u^k) \qquad (135)\\
&= w''^k - L(I + L^\top L)^{-1}(L^\top(2w''^k - \tau^{-1}\bar{w}^k) \\
&\quad +(\tau^{-1}\bar{u}^k - 2\tau^{-1}u^k)) \qquad (136)\\
&= \tau^{-1}L(I + L^\top L)^{-1}((2u^k - \bar{u}^k) \\
&\quad +L^\top(\bar{w}^k - 2\tau w''^k)) + w''^k \qquad (137)\\
&= \tau^{-1}(Lu'^k + \tau w''^k) = \tau^{-1}\bar{w}^{k+1}\,. \qquad (138)
\end{aligned}
$$

By primal-dual symmetry, the same proof shows that $\bar{u}^{k+1} = \tau\bar{z}^{k+1}$.

To conclude, we have shown that $w''^k = v^k$ for $\tau_d = 1/\tau$ and suitable initialization of the dual method. As the dual method was shown to converge, w''^k must be a maximizing sequence for f_d. Together with the convergence of u^k to a minimizer of f, this proves the proposition.

References

[AFP00] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Clarendon Press, 2000.

[AHU64] Kenneth Joseph Arrow, Leonid Hurwicz, and Hirofumi Uzawa. *Studies in linear and non-linear programming. With contributions by Hollis B. Chenery [and others]*. Stanford University Press, 1964.

[AT06] Ben Appleton and Hugues Talbot. Globally minimal surfaces by continuous maximal flows. *Patt. Anal. Mach. Intell.*, 28:106–118, 2006.

[Auj08] Jean-François Aujol. Some algorithms for total variation based image restoration. *CMLA Preprint*, (2008-05), 2008.

[BBC09] S. Becker, J. Bobin, and E. J. Cands. *NESTA: A Fast and Accurate First-order Method for Sparse Recovery*, April 2009.

[BD86] J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersections of convex sets in Hilbert spaces. *Lecture Notes in Statistics*, 37:28–47, 1986.

[Ber09] Benjamin Berkels. An unconstrained multiphase thresholding approach for image segmentation. In *Scale Space and Var. Meth. in Comp. Vis.*, volume 5567 of *Springer LNCS*, pages 26–37, 2009.

[BG05] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling*. Springer, 2nd edition, 2005.

[BK04] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Patt. Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

[Boy03] Y. Boykov. Computing geodesics and minimal surfaces via graph cuts. In *Int. Conf. Comp. Vis.*, pages 26–33, 2003.

[Bra02] A. Braides. *Gamma-convergence for Beginners*. Oxford Univ. Press, 2002.

[BT09] E. Bae and X.-C. Tai. Graph cut optimization for the piecewise constant level set method applied to multiphase image segmentation. In *Scale Space and Var. Meth.*, volume 5567 of *LNCS*, pages 1–13, 2009.

[BVZ01] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Patt. Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[CCP08] A. Chambolle, D. Cremers, and T. Pock. A convex approach for computing minimal partitions. Tech. Rep. 649, Ecole Polytechnique CMAP, 2008.

[CD09] A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parameteric maximum flows. *Int. J. Comp. Vis*, 84:288–307, 2009.

[CEN06] Tony F. Chan, Selim Esedoḡlu, and Mila Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *J. Appl. Math.*, 66(5):1632–1648, 2006.

[DAV08] Vincent Duval, Jean-François Aujol, and Luminita Vese. A projected gradient algorithm for color image decomposition. *CMLA Preprint*, (2008-21), 2008.

[DFPH09] A. Delaunoy, K. Fundana, E. Prados, and A. Heyden. Convex multi-region segmentation on manifolds. In *Int. Conf. Comp. Vis.*, 2009.

[DR56] Jim Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Trans. of the AMS*, 82(2):421–439, 1956.

[EB92] Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Prog.*, 55:293–318, 1992.

[Eck89] Jonathan Eckstein. *Splitting Methods for Monotone Operators with Application to Parallel Optimization*. PhD thesis, MIT, 1989.

[FR60] W. H. Fleming and R. Rishel. An integral formula for total gradient variation. *Archiv der Mathematik*, 11(1):218–222, 1960.

[GBO09] T. Goldstein, X. Bresson, and S. Osher. Geometric applications of the split bregman method: Segmentation and surface reconstruction. CAM Report 09-06, UCLA, 2009.

[GM89] Norbert Gaffke and Rudolf Mathar. A cyclic projection algorithm via duality. *Metrika*, 36(1):29–54, 1989.

[Gol64] A. A. Goldstein. Convex programming in hilbert space. *Bull. Amer. Math. Soc.*, 70:709–710, 1964.

[Gow85] J.C. Gower. Properties of Euclidean and non-Euclidean distance matrices. *Lin. Alg. and its Appl.*, 67:81–97, 1985.

[Ish03] Hiroshi Ishikawa. Exact optimization for Markov random fields with convex priors. *Patt. Anal. Mach. Intell.*, 25(10):1333–1336, 2003.

[JT95] C. R. Johnson and P. Tarazaga. Connections between the real positive semidefinite and distance matrix completion problems. *Lin. Alg. and its Appl.*, 223–224:375–391, 1995.

[KB05] V. Kolmogorov and Y. Boykov. What metrics can be approximated by geo-cuts, or global optimization of length/area and flux. *Int. Conf. Comp. Vis.*, 1:564–571, 2005.

[KT99] J.M. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. In *Found. Comp. Sci.*, pages 14–23, 1999.

[KT07] N. Komodakis and G. Tziritas. Approximate labeling via graph cuts based on linear programming. *Patt. Anal. Mach. Intell.*, 29(8):1436–1453, 2007.

[LBS09] Jan Lellmann, Florian Becker, and C. Schnörr. Convex optimization for multi-class image labeling with a novel family of total variation based regularizers. In *Int. Conf. Comp. Vis.*, 2009.

[LKY+09] Jan Lellmann, Jörg Kappes, Jing Yuan, Florian Becker, and Christoph Schnörr. Convex multi-class image labeling by simplex-constrained total variation. In *Scale Space and Var. Meth.*, volume 5567 of *LNCS*, pages 150–162, 2009.

[LM79] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Num. Anal.*, 16(6):964–979, 1979.

[LP66] E.S. Levitin and B.T. Polyak. Constrained minimization problems. *U.S.S.R. Comput. Math. Math. Phys.*, 6:1–50, 1966.

[Mic86] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *J. Optim. Theory and Appl.*, 50(1):195–200, 1986.

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

[Mur03] Kazuo Murota. *Discrete Convex Analysis*. SIAM, 2003.

[Nes04] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Prog.*, 103(1):127–152, 2004.

[Ols09] Carl Olsson. *Global Optimization in Computer Vision: Convexity, Cuts and Approximation Algorithms*. PhD thesis, Lund University, 2009.

[PCBC09a] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the mumford-shah functional. In *Int. Conf. Comp. Vis.*, 2009.

[PCBC09b] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. Technical report, Graz Univ. of Tech., 2009. Preprint.

[PCF06] N. Paragios, Y. Chen, and O. Faugeras, editors. *The Handbook of Mathematical Models in Computer Vision*. Springer, 2006.

[Pop80] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Math. Notes*, 28:845–848, 1980.

[Roc70] R.T. Rockafellar. *Convex Analysis*. Princeton UP, 1970.

[ROF92] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[RW04] R.T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2nd edition, 2004.

[Set09a] S. Setzer. Split Bregman algorithm, Douglas-Rachford splitting and frame shrinkage. In *Scale Space and Variational Methods in Computer Vision*, volume 5567 of *LCNS*, pages 464–476, 2009.

[Set09b] Simon Setzer. *Splitting Methods in Image Processing*. PhD thesis, University of Mannheim, September 2009.

[SR96] G. Sapiro and D. L. Ringach. Anisotropic diffusion of multi-valued images with applications to color filtering. In *Trans. Image Process.*, volume 5, pages 1582–1586, 1996.

[Str83] Gilbert Strang. Maximal flow through a domain. *Math. Prog.*, 26:123–143, 1983.

[Str99] Gilbert Strang. The discrete cosine transform. *SIAM Review*, 41(1):135–147, 1999.

[SZS+06] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields. In *Europ. Conf. Comp. Vis.*, volume 2, pages 19–26, 2006.

[TPCB08] W. Trobin, T. Pock, D. Cremers, and H. Bischof. Continuous energy minimization by repeated binary fusion. In *ECCV*, volume 4, pages 667–690, 2008.

[WABF07] Pierre Weiss, Gilles Aubert, and Laure Blanc-Féraud. Efficient schemes for total variation minimization under constraints in image processing. Tech. Rep. 6260, INRIA, 2007.

[Win06] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer, 2006.

[WSV00] H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors. *Handbook of Semidefinite Programming*. Kluwer Academic Publishers, 2000.

[WX04] Changyu Wang and Naihua Xiu. Convergence of the gradient projection method for generalized convex minimization. *Comp. Opt. Appl.*, 16:111–120, 2004.

[Xu00] Shusheng Xu. Estimation of the convergence rate of Dykstra's cyclic projections algorithm in polyhedral case. *Acta Mathematicae Applicatae Sinica*, 16(2):217–220, 2000.

[ZC08] Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. CAM Report 08-34, UCLA, 2008.

[ZGFN08] C. Zach, D. Gallup, J.-M. Frahm, and M. Niethammer. Fast global labeling for real-time stereo using multiple plane sweeps. In *Vis. Mod. Vis*, 2008.

[ZNF09] C. Zach, M. Niethammer, and J.-M. Frahm. Continuous maximal flows and Wulff shapes: Application to MRFs. In *Comp. Vis. Patt. Recogn.*, pages 1911–1918, 2009.