

Discrete and Continuous Models for Partitioning Problems

Jan Lellmann¹, Björn Lellmann², Florian Widmann³, and Christoph Schnörr⁴

¹ Department of Applied Mathematics and Theoretical Physics
University of Cambridge

² Department of Computing, Imperial College London

³ Image and Pattern Analysis Group & HCI
Department of Mathematics and Computer Science, University of Heidelberg

Abstract. Recently, variational relaxation techniques for approximating solutions of partitioning problems on continuous image domains have received considerable attention, since they introduce significantly less artifacts than established graph cut-based techniques. This work is concerned with the sources of such artifacts. We discuss the importance of differentiating between artifacts caused by *discretization* and those caused by *relaxation* and provide supporting numerical examples. Moreover, we consider in depth the consequences of a recent theoretical result concerning the optimality of solutions obtained using a particular relaxation method. Since the employed regularizer is quite tight, the considered relaxation generally involves a large computational cost. We propose a method to significantly reduce these costs in a fully automatic way for a large class of metrics including tree metrics, thus generalizing a method recently proposed by Strekalovskiy and Cremers [63].

1 Introduction and Overview

The issue of how to formulate, solve, and approximate labeling problems has a long-standing history as one of the classical problems in image processing, and occurs in many applications, such as segmentation, multi-view reconstruction, stitching, and inpainting [56]. In the past decade, and in particular with the introduction of maximum flow-based methods into image analysis, a very popular approach has been to first cast the problem into the form of a Markov Random Field with a finite number of states, i.e., to minimize an energy function depending on a finite number of labels associated with a finite number of points in the image domain.

By doing so, one obtains a finite-dimensional problem defined over a discrete set of possible configurations. Therefore the resulting problem can be treated

This document is an author-created copy of the article originally published in International Journal of Computer Vision, April 2013. The final publication is available at <http://www.springerlink.com>

and analyzed as a combinatorial problem, for which a large machinery of solvers is available, most notably graph cut-based algorithms such as α -expansion.

As an alternative, in the past few years it has become more and more common to postpone the discretization until the very end, and instead to work in the function space setting as long as possible.

In particular, we will consider the following formulation, which can be seen as a continuous analogue of the finite-dimensional pairwise MRF energy. A predecessor can be found in [51], while the model was proposed independently in [66, 45, 57].

The task is to assign to each point x in the image domain $\Omega \subseteq \mathbb{R}^d$ an *integral* label $\ell(x) \in \mathcal{I} := \{1, \dots, l\}$, so that the *label assignment* (or *labeling*) function ℓ minimizes the functional

$$\inf_{\ell: \Omega \rightarrow \mathcal{I}} f(\ell), \quad f(\ell) := \int_{\Omega} s(x, \ell(x)) dx + J(\ell), \quad (1)$$

where the left “data” term is strongly application-driven and defines *local* information obtained from the input image, and the right “regularization” term penalizes labeling functions ℓ according to some *nonlocal* irregularity measure, i.e., *prior information*.

Instead of considering (1) directly, one then embeds the labels into \mathbb{R}^l in the following way: Identify label i from the label set $\mathcal{I} := \{1, \dots, l\}$ with the i -th unit vector $e^i \in \mathbb{R}^l$, set $\mathcal{E} := \{e^1, \dots, e^l\}$, and solve

$$\inf_{u \in \mathcal{C}_{\mathcal{E}}} f(u), \quad f(u) := \int_{\Omega} \langle u(x), s(x) \rangle dx + \int_{\Omega} \Psi(Du), \quad (2)$$

$$\mathcal{C}_{\mathcal{E}} := \text{BV}(\Omega, \mathcal{E}) = \{u \in \text{BV}(\Omega)^l \mid u(x) \in \mathcal{E} \text{ a.e. } x \in \Omega\}.$$

The labels are thus embedded into a higher-dimensional space. The function space $\text{BV}(\Omega, \mathcal{E}) \subset (L^1)^l$ of functions of bounded variation guarantees a minimal regularity of the discontinuities of u , we refer to [1] for the mathematical background. The integrand Ψ is positively homogeneous and convex, and defines the regularizer.

In this work we will focus on a specific generic regularizer proposed in [57] and later extended in [48]: For some given metric $d : \{1, \dots, l\}^2 \rightarrow \mathbb{R}_{\geq 0}$, define

$$\Psi(z) := \Psi_d(z) := \sup\{\langle z, v \rangle \mid z \in \mathcal{D}_{\text{loc}}^d\}, \quad (3)$$

$$\mathcal{D}_{\text{loc}}^d := \bigcap_{i \neq j} \{v = (v^1, \dots, v^l) \in \mathbb{R}^{d \times l} \mid \dots\} \quad (4)$$

$$\|v^i - v^j\|_2 \leq d(i, j), \sum_k v^k = 0\}. \quad (5)$$

The idea behind this definition is that at points where u locally jumps from label e^j to label e^i along a normal ν , the gradient Du is the $(d-1)$ -dimensional Hausdorff measure multiplied by $\nu(e^i - e^j)^{\top}$. The function Ψ is constructed such that

$$\Psi(\nu(e^i - e^j)^{\top}) = d(i, j), \quad (6)$$

i.e., jumps are penalized by their length multiplied by $d(i, j)$. This can be seen as an infinite-dimensional equivalent of the *metric labeling problem* [37].

The slightly convoluted look of (3)–(5) originates from the wish that Ψ should be the *largest convex* function on $\mathbb{R}^{d \times l}$ that satisfies (6), which anticipates that ultimately we would like to consider a *convex relaxation* of (2), as will be discussed below. It is also worth mentioning that the equality constraint in (5) is not necessary, but does not change the value of Ψ on gradients of u , and ensures that $\mathcal{D}_{\text{loc}}^d$ is bounded.

As a simple example, consider the *uniform* metric

$$d_u(i, j) := 1_{\{i \neq j\}}(i, j), \quad (7)$$

where $1_S(x) = 1$ if $x \in S$ and 0 otherwise denotes the characteristic function of a set S . Under this metric, the regularizer simply penalizes the total interface length.

By definition, Ψ is rotationally invariant, i.e. $\Psi(Rz) = \Psi(z)$ for any rotation matrix $R \in \text{SO}(d)$. Therefore the regularizer is *isotropic* and *homogeneous*, in the sense that it is invariant under rotation and translation of the coordinates. By using a different norm in (5) and making Ψ depend on x , anisotropic and inhomogeneous variants can also be perceived, but will not be discussed here.

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

$$s(x) := (s_1(x), \dots, s_l(x))^{\top} \quad (8)$$

$$:= (s(x, 1), \dots, s(x, l))^{\top} \in \mathbb{R}^l. \quad (9)$$

Due to the linear structure, the local costs s may be arbitrarily complicated, possibly derived from a probabilistic model (Sect. 2.1), without affecting the overall problem class. We generally assume $s \geq 0$, however any problem with (possibly negative) s bounded from below can be equivalently transformed into this form by adding a sufficiently large constant to s .

In this form, the label set is then *relaxed* by allowing u to take intermediate (fractional) values in the convex hull $\text{conv } \mathcal{E}$ of the original label set. This is just the unit simplex Δ_l ,

$$\Delta_l := \text{conv}\{e^1, \dots, e^l\} = \{a \in \mathbb{R}^l \mid a \geq 0, \sum_{i=1}^l a_i = 1\}. \quad (10)$$

The problem is then considered on the *relaxed* constraint set \mathcal{C} ,

$$\mathcal{C} := \text{BV}(\Omega, \Delta_l) \quad (11)$$

$$= \{u \in \text{BV}(\Omega)^l \mid u(x) \in \Delta_l \text{ for a.e. } x \in \Omega\}, \quad (12)$$

and one obtains the *relaxed problem*

$$\inf_{u \in \mathcal{C}} f(u), \quad f(u) := \int_{\Omega} \langle u(x), s(x) \rangle dx + \int_{\Omega} \Psi(Du). \quad (13)$$

It can be shown that for any metric d , (13) admits a (possibly non-unique) minimizer. Since this functional is convex, a globally optimal solution can be found using convex optimization methods.

Ideally, one would hope for *integral* solutions, i.e., solutions that only assume values in \mathcal{E} . In practice, solving (13) leads to very good solutions of the labeling problem, with fewer artifacts than in MRF-based methods. Here the construction of Ψ comes into play: while other choices could be imagined that fulfill (6), the definitions (3)–(5) lead to the tightest relaxation that can be written in the integral form (13), which minimizes the amount of unwanted solutions [13, 48].

However, when solving (13) numerically, one still observes *fractional* solutions, i.e., solutions that assume values in $\Delta_l \setminus \mathcal{E}$ in some points.

In this survey, we focus on the possible sources of such fractional labels. In part, such labels can be attributed to the relaxation: by enlarging the constraint set, one introduces artificial solutions. Then, in order to obtain an *integral* solution that is close to the – unknown – *best integral* solution, one requires a rounding scheme. However, this introduces several questions: How should one round? Should the scheme be formulated on function spaces or on the discretized problem?

On the other hand, there is an important reason for allowing intermediate values: *Fractional values are actually required in order to provide a good approximation of the true spatially continuous solution.*

This aspect is often overlooked in related literature, and it raises the slightly subtle question of whether computing an *integral* solution of the *discretized* problem is the optimal approach.

Contribution. This manuscript is structured in three parts. First, we provide a short synopsis of the variational multiclass approach (13) and relate it to existing *continuous* (function-space) approaches (Sect. 2). We then review related *discrete* (finite-dimensional) methods and possible discretizations, and illustrate and compare them on several numerical examples (Sect. 3).

We feel that it is important to formulate these methods in a unified framework in order to point out the close connections. In particular, we hope to make more obvious the often overlooked necessity of fractional labels, even when the relaxation is tight.

After discussing the *discretization* aspect, in the second part (Sect. 4) we focus on the second source of fractional labels, the *relaxation* aspect. We provide an extended review on approaches in the literature and relate them to a recently proposed rounding approach for recovering integral from fractional solutions. For illustration we supply several numerical examples.

Finally, we propose a method for reducing the computational effort involved in solving the relaxed problem (13) by several orders of magnitude for larger number of labels (Sect. 5).

We close with an extended discussion on the justification of using combinatorial or relaxed problem formulations (Sect. 6).

For the technical mathematical details and proofs we refer to the thesis [43]. The proofs for the rounding method were announced in conference proceedings

form in [46] and have been extended to the more general regularizer (5) in [47]. We additionally supply the background and numerical evidence.

The present manuscript should be understood as an extended survey of the field with a specific focus on the interplay of *discretization* and *relaxation* in connection with *discrete* and *continuous* approaches to image partitioning.

Given that vast amount of literature on segmentation models and algorithms it seems almost impossible to give a complete account, and we make no attempt at doing so. Instead we deliberately focus on the model (2) and aim to illustrate the main connections and conclusions by means of this – necessarily rather specific – example.

2 Continuous Models

2.1 Variational Models

Problem (13) belongs to the class of *variational* problems, where the output is defined as the minimizer of an *objective function* f ,

$$u^* := \arg \min_{u \in \mathcal{C}} f(u), \quad (14)$$

\mathcal{C} is some subset of a space of functions that are defined on the continuous domain $\Omega \subseteq \mathbb{R}^d$, and f is a functional depending on the input data. The objective f is typically composed of a *data term* $H(u)$ and a *regularizer* $J(u)$,

$$f(u) = H(u|I) + J(u). \quad (15)$$

The data term depends on the input data I – such as color values of a recorded image, depth measurements, or other features – and promotes a good fit of the minimizer to the input data. In order to cope with noise and extract higher-order information from low-level image features, it is generally necessary to incorporate additional prior knowledge about the “typical” appearance of the desired output, which is the purpose of the regularizer. We refer to [60] for a general overview of variational methods in image processing.

The distinction between data term and regularizer in (15) often has a statistical background: Consider the problem of finding the best estimate of the unknown quantity u , given some observation (input) I which is assumed to be susceptible to noise, i.e. I is sampled from a random variable. Then, the configuration u with the highest probability can be inferred from the observation by maximizing the probability

$$u^* = \arg \max_u \mathbb{P}(u|I). \quad (16)$$

The modeling process consists in specifying the conditional probability. The conditional probability distribution can be either estimated directly, as in *discriminative* models, or it can be deduced from the Bayes theorem: Problem (16)

is equivalent to

$$u^* = \arg \max_u \mathbb{P}(I|u)\mathbb{P}(u) \quad (17)$$

$$= \arg \min_u \{-\log(\mathbb{P}(I|u)) - \log(\mathbb{P}(u))\}. \quad (18)$$

This approach requires to define a *generative* model, i.e. the joint distribution of the observation I and the unknown u . The right summand in (18) encodes prior knowledge about the (a priori) likelihood of a particular configuration u , while the conditional probability on the left relates possible u with the observation I – which could be color, texture, or any other observable quantity –, and can therefore be seen as the data term.

In fact, if one makes the common conditional independence assumption that the conditional probability in (17) factorizes on a per-point basis, one obtains (for finite Ω , i.e. after discretization):

$$-\log \mathbb{P}(I|u) = -\log \prod_{x \in \Omega} \mathbb{P}(I(x)|u(x)) \quad (19)$$

$$= \sum_{x \in \Omega} -\log \mathbb{P}(I(x)|u(x)). \quad (20)$$

For normally distributed noise, this corresponds directly to the classical ℓ^2 distance between u and I .

Some possible choices for s include:

- For color segmentation – which can be seen as a form of denoising with a *finite* number of color values –, each label k is associated with a prototypical color value c^k . Then $s(x, k)$ could be set to a distance measure between the color $I(x)$ in the input image at point x , such as $\|I(x) - c^k\|_2$ (corresponding to a Gaussian distribution), the more robust $\|I(x) - c^k\|_1$, or many other variants such as robust p -norms with $p < 1$.
- For general foreground-background segmentation, one usually estimates parametrized statistical models of the foreground and background, based on a range of features such as color, edges, texture, and scale computed at each point. The parameters and weights of the individual features are then determined in a learning step.
- For depth estimation from stereo image pairs, the labels correspond to possible point correspondences between the two involved images. For a calibrated stereo camera system, these are physically restricted to a one-dimensional subspace along the epipolar lines [34, 28]. The data term then describes how well the hypothesis of a certain depth at a certain point is supported by the observed images, i.e. how similar the corresponding image patches are.

Generally, the local cost term has a strong dependence on the input data, and its structure typically cannot be reliably predicted beforehand.

However, since s only appears in linear form in (13), the complexity of the original model does not affect the complexity of the functional f . In particular, f is always *convex* and can therefore be solved to a global optimum, clearly separating modeling and optimization aspects.

2.2 Contour-Based and Level-Set Methods

The model (1) represents one of the most recent developments in a long history of approaches for image segmentation on continuous domains. The first approaches were *contour-based*, restricting themselves to the two-dimensional, two-class case and parametrizing the interface between “class regions” $\Omega_1 := \ell^{-1}(\{1\})$ and $\Omega_2 := \ell^{-1}(\{2\})$ explicitly in the form of a closed curve $C : [0, 1] \rightarrow \mathbb{R}^2$, such as the *snake* [36] and *Geodesic Active Contours* models [11].

The major drawback with these methods is that they do not easily cope with different topologies. A major improvement consists in replacing the explicit parametrization with the *level set* technique [11], i.e., representing C as the *zero set* of some function $\phi : \Omega \rightarrow \mathbb{R}$ with the convention $C([0, 1]) = \phi^{-1}(\{0\})$. This allows to propagate curves of arbitrary topology (and number) using a single, fixed discretization, e.g. on a grid. The boundary curve C can be extracted with sub-pixel accuracy from ϕ , and higher-dimensional cases $d \geq 3$ can be transparently handled.

In the famous *Active Contours Without Edges* (Chan-Vese) approach [17], this method is applied to a *region-based* energy, formulated in terms of the *interior* and *exterior* of the region P_C described by C ,

$$\begin{aligned} f(C, c_1, c_2) = & \int_{\text{int}(P_C)} \|I - c_1\|_2^2 dx + \\ & \int_{\text{ext}(P_C)} \|I - c_2\|_2^2 dx + \\ & \mu \int_0^1 \|\nabla C\|_2 dp \end{aligned} \quad (21)$$

This formulation can be seen as a continuous version of the ferromagnetic Ising model [35] and is a special case of the more general model (1): in the two-class case, $\Omega_1 = \Omega \setminus \Omega_2$ and $\partial\Omega_1 = \partial\Omega_2$. Therefore we may set $s(\cdot, j) = \|I - c_j\|_2^2$ for the labels $j \in \{1, 2\}$, and $J(\ell) = \mu \mathcal{H}^{d-1}(\partial\Omega_1)$, where $\mathcal{H}^{d-1}(\partial\Omega_1)$ denotes the $(d-1)$ -dimensional Hausdorff measure, i.e. the length or area, of the boundary of Ω_1 .

These classical methods share several drawbacks:

- Level set-based methods suffer from the non-uniqueness of ϕ , which complicates optimization.
- Edge-based approaches usually do not have a plausible statistical explanation.
- Most importantly, the methods are inherently local, since they rely on an incremental curve evolution. Therefore a good initialization is mandatory, and model and optimization effects cannot be clearly separated.

Probably the most influential region-based model is the *Mumford-Shah* model [53]. Motivated by the Gibbs field [29] and weak membrane energy [6] methods, it can be seen as a first approach of explicitly introducing the possibility of discontinuous solutions into a spatially continuous framework.

It can be formulated as minimizing, for some constants $\lambda, \mu, \nu > 0$, the functional

$$f(u) = \lambda \int_{\Omega} (u - I)^2 dx + \mu \int_{\Omega} \|\nabla u\|_2^2 dx + \nu \mathcal{H}^{d-1}(S_u), \quad (22)$$

where $u \in \text{SBV}(\Omega)$ is a *special function of bounded variation*. It is allowed to be discontinuous on a non-empty set S_u , and the gradient can be defined almost everywhere (see [1] for the precise definitions).

The Mumford-Shah approach is not a pure labeling method in the strict sense, but rather a case of *simultaneous* labeling (inference) and model parameter optimization: it divides the image domain into an – a priori unknown – number of connected components, on each of which the observation I is explained by a smoothed version. This fact reflects in the non-convexity of the functional, which is caused by the Hausdorff term.

If one considers the limit $\mu \rightarrow +\infty$, restricts the model to two regions, and assumes that the optimal constant values c_1, c_2 for each are known, one obtains the Chan-Vese model (21), see also [53, 52] for many other limiting and special cases. However, directly transferring optimization methods for the Mumford-Shah functional to the labeling formulation is difficult:

- While the model (1) can be viewed as a generalization of the inference part in the Mumford-Shah model, the latter is restricted to quadratic distances.
- Optimization methods for the Mumford-Shah model are generally tailored to the main difficulty – the non-convex regularizer – and do not allow to find globally optimal solutions of the *labeling* problem.

2.3 Convex Variational Approaches

The model (13) was originally motivated by the work of Chan, Esedoğlu, and Nikolova [16], who considered the objective

$$f(u') = \lambda \int_{\Omega} ((1 - u')(c_1 - I)^2 + u'(c_2 - I)^2) dx + \nu \int_{\Omega} \|Du'\|_2, \quad (23)$$

formulated on *indicator* functions $u' : \Omega \rightarrow \{0, 1\}$ representing the foreground set, in order to solve the *geometry denoising* problem [16].

The total variation term in the rightmost integral evaluates exactly to the length of $\partial(u^{-1}(\{0\})) = \partial\Omega_1$, corresponding to the uniform metric $d(i, j) = 1_{\{i \neq j\}}$. Upon setting $u_1 := u'$, $u_2 := 1 - u'$, and

$$s(x) = (c_2 - I)^2 - (c_1 - I)^2, \quad (24)$$

this shows that the multi-class model (13) can be seen as a true extension of the two-class formulation (23).

Convex approaches such as (23) and (13) have several major advantages:

- The functionals are region-based, do not require an explicit parametrization of the boundary, and therefore can deal with partitions of arbitrary topology.
- As already noted, the data term is always linear in u , independent of its original form.
- The functional is *convex* on the *relaxed* set of functions $u' \rightarrow [0, 1]$, which allows to find global minimizers.

Such problems have become known as *continuous cut* problems, in analogy to combinatorial graph-cut techniques. For the two-class case, the dual problem has been formulated in a maximal-flow setting in [62] and numerically solved in [2]. It can be shown that the functional satisfies a “coarea-like” formula

$$f(u') = \int_0^1 f(1_{\{u' > \alpha\}}) d\alpha. \quad (25)$$

From this in turn it can be shown that solutions of the *relaxed* problem, with $u'(x) \in [0, 1]$, can be thresholded at almost any threshold α such that the resulting *integral* $u' := 1_{\{u' > \alpha\}}$ constitutes a solution of the original problem [16].

This stands in close analogy to discrete min-cut/max-flow problems, where the optimal solution can be obtained in polynomial time. Such finite-dimensional methods have been tremendously popular in image processing [8], however they are inherently formulated on graphs, i.e. they are formulated on the *discretized* problem, which often introduces an anisotropy, and prohibits rotational invariance in a strict sense (Sect. 3).

Several slightly different formulations of the multi-class formulation (13) have been proposed: The works of [66] and [45] mainly differ in the way the regularizer is relaxed, one using the “decoupling” regularizer $\Psi(z) = \sum_{i=1}^l \|z^i\|_2$, and one using the standard vectorial total variation $\Psi(z) = (\sum_{i=1}^l \|z^i\|_2^2)^{1/2}$. A systematic treatment was published by Chambolle et al. in [13], based on [58]. Their approach differs from (13) in the sense that instead of embedding into \mathbb{R}^l via $\mathcal{E} = \{e^1, \dots, e^l\}$, they embed the labels into \mathbb{R}^{l-1} , effectively reparametrizing the unit simplex. Also, they already propose to use the tight regularizer (5) in the slightly more restricted setting where $d(i, j) = \gamma(|i - j|)$ for some nondecreasing, positive, concave function γ .

In a sense, the piecewise constant Mumford-Shah formulation in [51] can be seen as a predecessor of these approaches. The authors represent the label assignment ℓ using a piecewise-constant real-valued function, and parametrize this function using a set of l polynomial basis functions, which can be regarded as the individual components of u .

3 Finite-Dimensional Models and Discretization

In this section we consider discretization strategies for the labeling problem. In order to point out the differences between the approaches we consider the general

setting: Assume the goal is to find an optimal labeling $\ell^* : \Omega \rightarrow \mathcal{I} = \{1, \dots, l\}$ based on input data I modeled as $s : \Omega \rightarrow \mathbb{R}^l$, and assume that there is some function space \mathcal{U} and an objective function f depending on s , whose minimizer

$$u^* = \arg \min_{u \in \mathcal{U}} f(u) \quad (26)$$

provides some information about ℓ^* . In our setting, $\mathcal{U} = \text{BV}(\Omega, \mathcal{E})$ corresponds to the combinatorial problem (2), and $\mathcal{U} = \text{BV}(\Omega, \Delta_I)$ to the relaxed problem (13).

In order to represent the problem in finite memory, one has to consider a *discretized* problem, i.e. the goal is to find a finite-dimensional approximation $u^{h,*}$ (where h denotes the scale, e.g. grid spacing) that approximates ℓ^* in some sense, and can be computed by solving a finite-dimensional problem,

$$u^{h,*} = \arg \min_{u^h \in \mathcal{U}^h} f^h(u^h). \quad (27)$$

Several important questions arise:

- How should \mathcal{U}^h be chosen? In particular, should $u^{h,*}$ be restricted to the same values as u^* , i.e. \mathcal{E} or Δ_I ?
- What are the semantics of $u^{h,*}$, i.e. in what sense does it provide information about the original u^* ?
- Do the discretized functionals f^h converge to the original, spatially continuous functional f ? Moreover, is it possible to reconstruct u^* from $u^{h,*}$ for infinite resolution, i.e. do the minimizers of the discretized problems converge to a minimizer of the original problem?

In the following, we consider several approaches and how they compare with respect to these questions. In particular, we argue that it may be better *not* to pose the finite-dimensional problem as a combinatorial problem, even if integral solutions are required.

An important concept for such an analysis is Γ -convergence. This type of convergence for sequences of functionals was introduced by De Giorgi [24] and can be interpreted as set-convergence of the epigraphs [23, Chap. 4]; we refer to [24, 23, 9] for an overview.

Two conditions have to be fulfilled for a sequence of functionals (f^h) to Γ -converge to a functional f : a “lim inf”-condition, i.e.,

$$f(u) \leq \liminf_{h \rightarrow 0} f^h(u^h) \quad (28)$$

for any u and any converging sequence $(u^h) \rightarrow u$ in the domain of f , and a “lim sup”-condition, i.e., for all u in the domain of f there must be a converging sequence $(u^h) \rightarrow u$ such that

$$f(u) \geq \limsup_{h \rightarrow 0} f^h(u^h). \quad (29)$$

The central feature of Γ -converging sequences used in the following sections is that if a sequence of functionals f^h Γ -converges to some functional f (and is

equicoercive), then the minimizers of these functionals converge to a minimizer of f [23, Def. 7.10, Thm. 7.23].

In particular, if it can be shown that a sequence of discretized functionals f^h converge to a continuous energy f as the grid spacing h approaches zero, this implies that the minimizers of the *discretized* energies converge to the minimizer of the original *continuous* energy (Sect. 3.2).

3.1 Combinatorial Finite-Dimensional Schemes

The classical approach for discretizing labeling problems is to fix, for some given grid spacing h , a set of *points* $\{x^{\bar{i}} \in \mathbb{R}^d | \bar{i} \in \mathcal{J}\}$ on the image domain, where $\mathcal{J} \subseteq \mathbb{Z}^d$ is a set of multi-indices, and to approximate the values $\ell(x^{\bar{i}})$. For simplicity we assume that the original continuous domain Ω is the unit box $(0, 1)^d$ and set $\mathcal{J} = \{0, \dots, k-1\}^d$, where k is the grid size and $h = 1/k$ is the scale/grid spacing. We consider the regular grid

$$\Omega^h = \{x^{\bar{i}} = (\bar{i} + e/2)h | \bar{i} \in \mathcal{J}\}. \quad (30)$$

With any $\ell : \Omega \rightarrow \mathcal{I}$, we associate its discretization $\ell^h : \Omega^h \rightarrow \mathcal{I}$, $\ell_i^h := \ell(x^{\bar{i}})$.

Classically, a discretized *combinatorial* energy $f^h : \mathcal{I}^n \rightarrow \mathbb{R}$ is then constructed such that ideally $f^h(\ell^h)$ approximates the energy $f(\ell)$. Minimizing f^h , one obtains

$$\ell^{h,*} = \arg \min_{\ell^h : \Omega^h \rightarrow \mathcal{I}} f^h(\ell^h). \quad (31)$$

In contrast to the continuous models (Sect. 2), the resulting problem is combinatorial in nature in the sense that it is formulated on a discrete, finite set. However, since this set is very large, the problem is in general very hard to solve without further knowledge about the structure of f^h .

Markov Random Fields A common approach to model such structure is to represent f^h in terms of a *Markov Random Field* (MRF), also known as an *undirected graphical model*. The MRF consists of an undirected graph $G = (V, E)$, where each vertex $v \in V$ is associated with the variable $\ell^h(v)$. For the uniform grid as above, we have $V = \Omega^h$ and $v = x^{\bar{i}}$ for some $\bar{i} \in \mathcal{J}$. Note that the common convention in graphical model literature is to denote by x^v or x^i the random variables, and by v^i (or just i) the vertices in V . This clashes with the continuous formulation, where x is the spatial variable. For the sake of consistency, we therefore denote by x or $x^{\bar{i}}$ the vertices, and by $\ell_i^h = \ell^h(x^{\bar{i}})$ the labels. In the MRF approach, f^h is (non-uniquely) written as

$$f^h(\ell^h) = \sum_{C \in \text{cl}(G)} \psi_C(\ell_C^h), \quad (32)$$

where the sum is taken over all sets $\text{cl}(G)$ of *cliques* of G , $\ell_C^h \in \mathcal{I}^{|C|}$ denotes the *restriction* of ℓ to the vertices in the clique, and $\psi_C : \mathcal{I}^{|C|} \rightarrow \mathbb{R}$ are the individual *factors* of f .

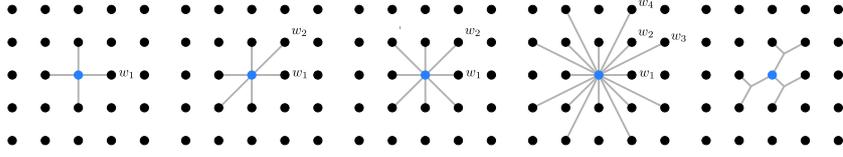


Fig. 1. Graph-based discretization on a grid. **Left to right:** Pairwise terms with 4-, 6-, 8- and 16-neighborhood; higher-order discretization with ternary terms. The dots correspond to the vertices of the graph, the lines indicate factors in the representation (32) and weights in the pairwise representation (38).

The structure of (32) is directly related to the factorization of the corresponding MRF distribution, that in turn is guaranteed by the Hammersley-Clifford theorem and the conditional independence model related to set separation in the underlying graph [42, 39]. Minimizing (32) amounts to computing the most likely configuration for a given observation (MAP inference).

An important special case occurs when only *unary* ($|C| = 1$) and *pairwise* ($|C| = 2$) terms exist, i.e. the graph G contains no higher-order cliques. In this case f^h can be written in pairwise form,

$$f^h(\ell^h) = \sum_{x \in V} \psi_x(\ell^h(x)) + \sum_{(x,y) \in E} \psi_{x,y}(\ell^h(x), \ell^h(y)). \quad (33)$$

This is the most popular approach for discretizing labeling problems with spatial regularizers. In order to represent a local data term together with a regularizer implementing the uniform metric (see the introduction and (7)), a classical choice is to set

$$\psi_x(\ell^h(x)) = s(x, \ell^h(x)) = \mathbb{P}(I(x) | \ell^h(x)), \quad (34)$$

$$\psi_{x,y}(p, q) = \begin{cases} w_{x,y}, & p \neq q, \\ 0, & \text{otherwise} \end{cases} \quad (35)$$

for some $w_{x,y} \geq 0$, and choose E such that each vertex in the grid is connected to its four neighboring vertices. This principle can be generalized by adding terms for a larger neighborhood, such as 8 or 16 neighbors, or by adding higher-order terms, i.e., terms that depend on three or more labels (Fig. 1).

Graph Cuts and Metrics For the two-class case, symmetric pairwise potentials can be considered as edges in the grid graph. By adding some constant to the overall energy, they can be normalized to

$$\psi_{x,y}(1, 1) = \psi_{x,y}(2, 2) = 0, \quad (36)$$

$$\psi_{x,y}(1, 2) = \psi_{x,y}(2, 1) = w_{x,y}, \quad (37)$$

where $w_{x,y} \in \mathbb{R}$ is some weight, i.e. $\psi_{x,y}(p,q) = w_{x,y}1_{\{p \neq q\}}$. Then, discarding the constant which is irrelevant for the optimization,

$$f^h(\ell^h) = \sum_{x \in V} \psi_x(\ell^h(x)) + \sum_{(x,y) \in E} w_{x,y}1_{\{\ell^h(x) \neq \ell^h(y)\}}. \quad (38)$$

This indicates that each edge yields a certain cost when the edge between x and y is *cut* by the interface separating the two class regions. Minimizing the energy (38) therefore amounts to computing a partition of the nodes into two subsets that minimizes the total sum of the weights of the edges that are *cut by the interface* between the partitions. The unary potentials can be included by adding edges to designated “source” and “sink” outside of the image plane.

For nonnegative $w_{x,y}$ such problems with pairwise terms can be solved in polynomial time by min cut/max flow algorithms [5]. Therefore a central question – which we will discuss in the next paragraphs – is how to discretize some given *spatially continuous* functional f in a way such that the discretized energy $f^h(\ell^h)$ approximates the continuous energy $f(\ell)$ if one sets $\ell_i^h = \ell(x_i^{\bar{}})$.

In [7], this question was answered for two-class problems. They consider regularizers that are formulated in terms of the length of the interface between the two classes, measured by a Riemannian metric (see [40] for a generalization to a larger class of metrics). Specifically, let $C : [0, |C|] \rightarrow \mathbb{R}^2$ denote some curve parametrized by arc length with tangent τ_C , and define its anisotropic length $|C|_{\mathcal{R}}$ by

$$|C|_{\mathcal{R}} := \int_0^{|C|} \|A(C(s)) \tau_C(s)\|_2 ds, \quad (39)$$

where $A : \mathbb{R}^2 \rightarrow \text{GL}_2(\mathbb{R})$ defines the Riemannian metric. Such metrics can be used as regularizers for two-class labeling problems by setting $C = \partial\Omega_1 = \partial\{x \in \mathbb{R}^d | \ell(x) = 1\}$. In our framework (1) this corresponds, for some suitable $A' : \mathbb{R}^2 \rightarrow \text{GL}_2(\mathbb{R})$, to the length-based regularizer

$$J(\ell) = \int_{\mathbb{R}^d} \left\| A' \frac{D1_{P^1}}{|D1_{P^1}|} \right\|_2 d|D1_{P^1}|. \quad (40)$$

The mathematically precise formulation (40) in terms of measure-valued derivatives of discontinuous functions can indeed be intuitively interpreted as contour integrals penalizing boundaries of a partition [1, 43]. For $A = A' = I$ we obtain the classical isotropic length of $\partial\Omega_1$.

We denote by $\mathcal{N} = \mathcal{N}(x) = \{y^1, \dots, y^{|\mathcal{N}|}\}$ the neighborhood system of x , i.e. vertices connected to x via an edge, and assume that such a system is given. Examples are 4-, 6-, 8- or 16-neighborhoods as shown in Fig. 1. Then, for some fixed x , the vectors

$$g^m := y^m - x, \quad m = 1, \dots, |\mathcal{N}| \quad (41)$$

denote the offsets between some point x and its neighbors. Assuming that the g^m are in increasing order with respect to their angle α_m relative to g^1 , Boykov et al. [7] construct a regularizer J^h with pairwise potentials as in (38) by choosing the weights according to

$$\begin{aligned} \psi_{x,y^m}(p,q) &= w_{x,y^m} 1_{\{p \neq q\}}, \\ w_{x,y^m} &= \frac{h^2 \|g^m\|_2^2 (\alpha_{m+1} - \alpha_m) \det(A(x))}{2 \|A(x) g^m\|_2^3}. \end{aligned} \quad (42)$$

Under some regularity assumptions on ℓ , they show that $J^h(\ell^h)$ then converges to $J(\ell)$ if

$$h \rightarrow 0, \quad \sup_m |g^m| \rightarrow 0 \quad \text{and} \quad \sup_m |\alpha_m| \rightarrow 0. \quad (43)$$

This establishes a consistency result for the representation of metrics using pairwise terms. However it also has some drawbacks:

- The result only holds for two-class problems and on two-dimensional grids.
- Convergence of the energy is only shown pointwise. There is no indication how *minimizers* of functionals involving J^h relate to minimizers of the associated spatially continuous problem.
- While the choice of weights is good enough to give the desired result in the limit, it does not necessarily provide an (in some sense) optimal representation for a *given* connectivity.
- The last condition in (43) implies that the neighborhood size must approach infinity in order to obtain a consistent scheme.

In particular the last point is troublesome, as it means that the number of pairwise potentials must grow faster than the number of vertices for $h \rightarrow 0$. Since solvers for such problems usually rely on solving the dual “max flow” problem on the edges of the graph, the increasing connectivity multiplies the problem size and greatly slows down optimization. This problem becomes potentially worse in higher dimensions due to the larger neighborhood.

Pairwise Functionals and LP Relaxation One distinct advantage of the pairwise energy (38) is that it has a tight natural linear programming (LP) relaxation in the two-class case: Associate ℓ^h with $u^h : \Omega^h \rightarrow \{0, 1\}$ in the sense that $u^h(x) = 0 \Leftrightarrow \ell^h(x) = 1$, and consider the relaxed problem

$$\min_{u^h \in [0,1]^n} f_{\text{LP}}^h(u^h), \quad (44)$$

$$\begin{aligned} f_{\text{LP}}^h(u^h) &:= \sum_{x \in V} (\psi_x(1) - \psi_x(2)) u^h(x) + \\ &\quad \sum_{(x,y) \in E} w_{x,y} |u^h(x) - u^h(y)|. \end{aligned} \quad (45)$$

Such formulations are very versatile and, by varying the graph structure and the weights, can also be adapted to many other problems besides segmentation, see [33] for a recent overview.

Problem (44) can clearly be solved as a linear program. One can also show that for $w_{x,y} \geq 0$, *integral* solutions of the pairwise energy f^h from (38) can be found by minimizing the LP relaxation f_{LP}^h and thresholding (cf. [14] and Sect. 4).

By setting $u^h(x) := u(x)$, the LP energy (44) can be extended to the set of relaxed functions $u : \Omega \rightarrow [0, 1]$ as follows:

$$f_{\text{LP}}(u) := \sum_{x \in V} (\psi_x(1) - \psi_x(2))u(x) + \sum_{(x,y) \in E} w_{x,y} |u(x) - u(y)|. \quad (46)$$

Although formulated on functions defined on continuous domains, energies such as (46) are formulated in a *nonlocal* way, i.e. using a sum of pairwise differences instead of local properties such as ∇u . A comprehensive analysis can be found in [30], where the authors consider energies of the form

$$J_{G,h}(u) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \eta(g) \varphi_{h\|g\|_2} \left(\frac{|u(x+hg) - u(x)|}{h\|g\|_2} \right) dg dx, \quad (47)$$

where $u \in L^1_{\text{loc}}(\mathbb{R}^d)$, $\eta \in L^1(\mathbb{R}^d)$ with $0 \neq \eta \geq 0$, and for each h , $\varphi_h : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is continuous, nondecreasing and either convex, concave, or pieced together from a convex and a concave part. Moreover, let ψ and φ be the Γ -limits for $h \rightarrow 0$ of $h\varphi_h(z/h)$ and $\varphi_h(z)$, respectively, and define

$$J_G(u) := \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \varphi(|\langle \nabla u, g \rangle / \|g\|_2|) \eta(g) dg dx + \|\eta\|_{L^1} \int_{S_u} \psi(|u^+ - u^-|) d\mathcal{H}^{d-1}. \quad (48)$$

Then, under some technical assumptions, the functionals $J_{G,h}$ Γ -converge to J_G in L^1_{loc} as $h \rightarrow 0$ [30, Thm. 4.3]. As a consequence, one obtains pointwise convergence of the functionals, as well as convergence of their minimizers.

This convergence result can be seen as an extended variant of the result in [7] (cf. (43)), formulated for nonlocal functionals in terms of (possibly non-integral) functions u . However, in formulation (47) the finite sum of pairwise terms (33) has been replaced by the convolution with weights specified by φ_h . Therefore, it cannot be formulated on finite-dimensional representations of u , since it depends on the values of u on all of \mathbb{R}^d .

A formulation closer to (33) has been considered in [12], in order to approximate the nonconvex part of the Mumford-Shah energy,

$$J_{\text{MS}}(u) = \lambda \int_{\mathbb{R}^d} \|\nabla u\|_2^2 dx + \mu \mathcal{H}^{d-1}(S_u). \quad (49)$$

The considered nonlocal energies are of the form

$$J_N^h(u^h) = \sum_{\bar{i} \in \mathcal{J}} h^d \sum_{\bar{j} \in \mathbb{Z}^d} \frac{1}{h} \varphi_h \left(\frac{\left(u^h(x^{\bar{i}+\bar{j}}) - u^h(x^{\bar{i}}) \right)^2}{h} \right) \eta(\bar{j}), \quad (50)$$

i.e. if one again sets $u^h(x^{\bar{i}}) = u(x^{\bar{i}})$, then

$$J_N^h(u^h) = \sum_{\bar{i} \in \mathcal{J}} h^d \sum_{\bar{j} \in \mathbb{Z}^d} \frac{1}{h} \varphi_h \left(\frac{\left(u(x^{\bar{i}+h\bar{j}}) - u(x^{\bar{i}}) \right)^2}{h} \right) \eta(\bar{j}). \quad (51)$$

Under some technical assumptions, these functionals can be shown to Γ -converge to

$$J_N(u) = \int_{\Omega} \sum_{\bar{j} \in \mathbb{Z}^d} \eta(\bar{j}) \alpha_{\bar{j}} |\langle \nabla u, \bar{j} \rangle|^2 dx + \int_{S_u} \sum_{\bar{j} \in \mathbb{Z}^d} \eta(\bar{j}) \beta_{\bar{j}} |\langle \nu_u, \bar{j} \rangle| d\mathcal{H}^{d-1} \quad (52)$$

for a collection of scalar weights $\alpha_{\bar{j}}$ and $\beta_{\bar{j}}$.

As a special case assume that u is integral, i.e. $u : \Omega \rightarrow \{0, 1\}$. Then the absolutely continuous part of Du vanishes, i.e. $\nabla u = 0$, and $J_N(u)$ is precisely the length of the discontinuity set S_u as defined by the right-hand integral in (52). Clearly, in order to obtain an isotropic regularizer in the limit, there must be infinitely many $\eta(\bar{j}) \neq 0$: the image domain and the connectivity needs to be infinitely large. This parallels the result of Boykov et al. in the graph cut setting.

The above results show that *it is possible to approximate energies involving length-based terms using sums of pairwise terms*, even for non-integral u . However, as for the graph cut approach, a finite neighborhood size invariably introduces an anisotropy.

Note that all these results are formulated on scalar u , and therefore apply directly only to the two-class case. However, they still provide an indication on what issues can be expected when applying similar techniques to multiclass labeling problems with vector-valued u . A prototypical finite-dimensional extension to the multi-class case is the LP relaxation [37, 41]

$$f_{\text{MLP}}^h(u^h) := \sum_{x \in V} \sum_{j=1}^l \psi_x(j) (u^h(x))_j + \frac{1}{2} \sum_{(x,y) \in E} w_{x,y} \|u^h(x) - u^h(y)\|_1, \quad (53)$$

where f_{MLP}^h is minimized over all $u^h : \Omega^h \rightarrow \Delta_l$. Unfortunately, unlike the two-class case, integral solutions cannot trivially be obtained using thresholding due to the constraints.

Note that (53) is essentially a non-local, discretized version of the main model (2), where $s_j(x)$ is defined by $\psi_x(j)$ and Ψ is replaced by a non-local, weighted integral defined by the $w_{x,y}$.

It is also worth mentioning that recent works [61, 20] have led to a non-linear generalization of the basic LP relaxation, where the energy (46) is replaced by

$$\begin{aligned}
 f_{\text{PW}}^h(u^h) := & \sum_{x \in V} \psi_x(1)^p |1 - u^h(x)|^q + \\
 & \sum_{x \in V} \psi_x(2)^p |u^h(x)|^q + \\
 & \sum_{(x,y) \in E} (w_{x,y})^p |u^h(x) - u^h(y)|^q, \tag{54}
 \end{aligned}$$

for some $p, q \in [0, \infty]$ and the resulting u is thresholded at $1/2$ to obtain an integral result. By varying p and q one obtains – in a limit sense in case $p = \infty$ or $q = \infty$ – variants of the Voronoi, Graph Cut, Random Walker and Shortest Path Forest segmentations.

A special case termed *Power Watersheds* is the limit $p \rightarrow \infty$ with $1 < q < \infty$. The solutions can be approximated efficiently and show very little geometric artifacts [20]; however we are not aware of any rigorous convergence proofs for infinitesimal grid spacing comparable to those available for the LP formulation.

3.2 Finite-Differences Schemes and Convergence

The above-mentioned methods all have in common that in order to achieve isotropy, even in the limit, they require either an infinite number of terms or an adaptive discretization. In this section we apply a finite-differences scheme from [13] which has a fixed neighborhood but still provides isotropy in the limit. This allows to keep the number of artifacts low without introducing nonlocal terms that would greatly increase the computational cost.

The original work [13] already contains informal versions of the convergence theorems; we provide a rigorous formulation in our framework and strengthen them to also obtain equicoercivity. We confine ourselves here to detailing the discretization and the main results. For technical details we refer to [43].

Again we represent multidimensional functions $u : V \rightarrow \mathbb{R}^l$ on Ω by a matrix $u^h = (u^{h,1} | \dots | u^{h,l}) \in \mathbb{R}^{n \times l}$. The l -dimensional vector associated with \bar{i} or $x^{\bar{i}}$ is denoted by $u_{\bar{i}}^h = u^h(x^{\bar{i}}) \in \mathbb{R}^l$. The standard forward-differences approximation

for $\nabla u(x^{\bar{i}})$ is

$$\nabla_{\bar{i}} u^h = \frac{1}{h} \begin{pmatrix} (u_{\bar{i}+e^1}^h - u_{\bar{i}}^h)^\top \\ \vdots \\ (u_{\bar{i}+e^d}^h - u_{\bar{i}}^h)^\top \end{pmatrix}, \quad (55)$$

with the convention $u_{\bar{i}+e^j}^h := u_{\bar{i}}^h$ if \bar{i} corresponds to a point on the right boundary, i.e. $\bar{i}+e^j \geq n_j = k$. For some $s \in L^\infty(\Omega)$ we compute the discrete approximation s^h by

$$s_{\bar{i}}^h := s^h(x^{\bar{i}}) := \frac{1}{h^d} \int_{C_{\bar{i}}^h} s(x) dx, \quad (56)$$

where $C_{\bar{i}}^h$ denotes the box corresponding to the \bar{i} -th pixel in the image,

$$C_{\bar{i}}^h := x^{\bar{i}} + \frac{1}{2}(-h, h)^d \quad (57)$$

$$= (i_1 h, (i_1 + 1)h) \times \cdots \times (i_d h, (i_d + 1)h). \quad (58)$$

Then, for the relaxed multiclass labeling functional (13), we define the discretization

$$f^h(u^h) := \sum_{\bar{i} \in \mathcal{J}} h^d \langle u_{\bar{i}}^h, s_{\bar{i}}^h \rangle + \sum_{\bar{i} \in \mathcal{J}} h^d \Psi(\nabla_{\bar{i}} u^h). \quad (59)$$

The forward-differences scheme introduces a slight asymmetry. Although this has no effect on the Γ -convergence as shown below, it can be somewhat reduced by taking the mean over variants that use backward- and mixed forward-backward differences.

In order to properly define consistency and convergence of minimizers in a common function space, we identify each discretized function

$$u^h \in \mathcal{U}^h := \{u^h : \Omega^h \rightarrow \Delta_l\} = (\Delta_l)^n \quad (60)$$

with the piecewise constant function $\tilde{u}^h \in \text{BV}(\Omega)^l$,

$$\tilde{u}^h(x) = u^h(x^{\bar{i}}) \in \mathbb{R}^l, \mathcal{L}^d\text{-a.e. } x \in C_{\bar{i}}^h. \quad (61)$$

For each h , we denote by $\tilde{\mathcal{U}}^h$ the space of such piecewise constant functions \tilde{u}^h ,

$$\tilde{\mathcal{U}}^h = \left\{ u \in \text{BV}(\Omega)^l \mid \exists u^h \in \mathcal{U}^h : u = \tilde{u}^h \right\}. \quad (62)$$

Likewise, we extend some functional $f^h : \mathcal{U}^h \rightarrow \mathbb{R}$ to $\text{BV}(\Omega)^l$ by setting

$$\tilde{f}^h : \text{BV}(\Omega)^l \rightarrow \bar{\mathbb{R}} \quad (63)$$

$$\tilde{f}^h(u') := \begin{cases} f^h(u^h), & \exists u^h \in \mathcal{U}^h : u' = \tilde{u}^h, \\ +\infty, & \text{otherwise.} \end{cases} \quad (64)$$

In the following, we will see that the *discretized* functionals \tilde{f}^h Γ -converge, for $h \rightarrow 0$, to the *true* constrained functional $f_{\mathcal{C}}$,

$$f_{\mathcal{C}} : \text{BV}(\Omega) \rightarrow \mathbb{R}, \quad (65)$$

$$f_{\mathcal{C}}(u) := \int_{\Omega} \langle u, s \rangle dx + \int_{\Omega} d\Psi(Du) + \delta_{\mathcal{C}}(u). \quad (66)$$

Note that the following propositions do not require Ψ to be isotropic.

Proposition 1. *Let $\Psi : \mathbb{R}^{d \times l} \rightarrow \mathbb{R}$ be continuous, convex and positively homogeneous with $\rho_l \|z\|_2 \leq \Psi(z) \leq \rho_u \|z\|_2 \forall z \in \mathbb{R}^{d \times l}$ for some $0 < \rho_l \leq \rho_u$, and define*

$$J_{\mathcal{C}}(u) := \int_{\Omega} d\Psi(Du) + \delta_{\mathcal{C}}(u). \quad (67)$$

Denote

$$\tilde{J}^h(u') := \begin{cases} \sum_{\bar{i} \in \mathcal{J}} h^d \Psi(\nabla_{\bar{i}} u^h), & \exists u^h \in \mathcal{U}^h : u' = \tilde{u}^h, \\ +\infty, & \text{otherwise.} \end{cases} \quad (68)$$

Then \tilde{J}^h Γ -converges (with respect to L^1 -convergence) to $J_{\mathcal{C}}$ for $h \rightarrow 0$.

Proof. An outline can be found in [13, Prop. 3.1]. The “lim inf” part is shown by proving that

$$\liminf_{h \rightarrow 0} \tilde{J}^h(\tilde{u}^h) \geq - \int_{\Omega} \langle u, \text{Div } v \rangle dx, \quad (69)$$

which is technical but straightforward after substituting the exact definition of \tilde{J}^h . Showing the “lim sup” inequality amounts to finding, for arbitrary but fixed $u \in \text{BV}(\Omega)^l$, a sequence (\tilde{u}^h) converging to u in L^1 s.t.

$$\limsup_{h \rightarrow 0} \tilde{J}^h(\tilde{u}^h) \leq J_{\mathcal{C}}(u). \quad (70)$$

This is achieved by choosing a sequence of smooth functions $(u^{(j)}) \subseteq \{u : \Omega \rightarrow \Delta_l | u \in C^\infty(\Omega)^l\}$ converging to u in L^1 and satisfying $\text{TV}(u^{(j)}) \rightarrow \text{TV}(u)$, which is possible due to [1, Thm. 3.9]. It is important to note that these functions are constructed by mollification, i.e., they are generally *not* integral even if u is. Then it is possible to show that by evaluating $u^{(j)}$ at suitable points and restriction to a subsequence, one can find a sequence \tilde{u}^i that fulfills (70). We refer to [43] for the full derivations.

Theorem 1. *Let $f^h, f_{\mathcal{C}}$ as defined in (59), (65) for $s \in L^\infty(\Omega), s \geq 0$. Then \tilde{f}^h as defined in (63) Γ -converges with respect to L^1 -convergence to the constrained functional $f_{\mathcal{C}}$ for $h \rightarrow 0$ and is equicoercive with respect to the L^1 -topology.*

Proof. Denote the data term by

$$g^h(u^h) := \sum_{\tilde{i} \in \mathcal{J}} h^d \langle u_{\tilde{i}}^h, s_{\tilde{i}}^h \rangle, \quad (71)$$

then $f^h = g^h + J^h$. For any $\tilde{u}^h \in \tilde{\mathcal{U}}^h$, one can show that $\tilde{g}^h(\tilde{u}^h) = \int_{\Omega} \langle \tilde{u}^h, s \rangle dx$. Therefore, since $\tilde{J}^h(u) = +\infty$ for all $u \notin \tilde{\mathcal{U}}^h$, \tilde{f}^h can be represented as

$$\tilde{f}^h(u) = \int_{\Omega} \langle u, s \rangle dx + \tilde{J}^h(u) \quad (72)$$

for any $u \in \text{BV}(\Omega)^l$, with the data term independent of h . Since $s \in L^\infty(\Omega)$, the linear term is continuous with respect to L^1 -convergence. Therefore, since Γ -convergence is stable under continuous perturbations [9, Rem. 1.7], and \tilde{J}^h Γ -converges to $J_{\mathcal{C}}$ due to Prop. 1, \tilde{f}^h Γ -converges to $f_{\mathcal{C}}$.

In order to show equicoercivity, by [23, Def. 7.6, Prop. 7.7] it suffices to provide a lower semicontinuous, coercive function f' with $\tilde{f}^h \geq f'$ uniformly for all h . This can be achieved by considering the spatially separable $\Psi' : \mathbb{R}^{d \times l} \rightarrow \mathbb{R}$,

$$\Psi' \left(\begin{pmatrix} (z^1)^\top \\ \vdots \\ (z^d)^\top \end{pmatrix} \right) := \frac{\rho l}{\sqrt{d}} \sum_{j=1}^d \|z^j\|_2 \quad (73)$$

and the corresponding functional $f'_{\mathcal{C}}$ and discretization \tilde{f}^h . Again we refer to [43] for the details.

Remark 1. In view of [23, Def. 7.10, Thm. 7.23], Thm. 1 shows that from a sequence (u^h) of minimizers of the *discretized* problems f^h , a (piecewise constant) sequence (\tilde{u}^h) of functions on the continuous domain Ω can be constructed that converge to a minimizer of the original, isotropic energy f .

Thm. 1 can also be in part applied to pairwise energies for multiclass problems. In particular, consider the multiclass LP relaxation (53) with a simple 4-neighborhood, setting $\psi_x(j) = h^d \left(s^h \left(x^{\tilde{i}} \right) \right)_j$ and $w_{x,y} = h^{d-1}$. This energy coincides with f^h as in (59) for the integrand

$$\Psi(z) = \sum_{i=1}^d \sum_{j=1}^l |(z^j)_i|. \quad (74)$$

Therefore Thm. 1 shows that the (properly weighted) LP relaxation objective Γ -converges to the *anisotropic* objective

$$f_1(u) = \int_{\Omega} \langle u, s \rangle dx + \sum_{i=1}^d \int_{\Omega} d \|D_i u\|_1 + \delta_{\mathcal{C}}(u), \quad (75)$$

i.e. it with the exception of the constraints it is separable.

While formulated for the forward differences discretization (55), these results can be generalized to most related schemes, such as backward- and centered differences, and up-/downwind schemes as used in [15].

Remark 2. Finite-differences discretizations generally do *not* fulfill a coarea-like formula such as (25) even if the continuous problem they were derived from has this property, as in the case of the two-class continuous cut. Therefore it is not trivial to obtain integral minimizers, in contrast to pairwise energies. However, the finite-differences energy approximates isotropic regularizers without requiring an infinitely large neighborhood in the limit. Moreover, in the following sections we will argue that computing an integral minimizer may actually not be the optimal approach.

3.3 Dual Approaches

Recently several authors have also proposed to use *dual* discretization approaches in the context of image segmentation [21, 3]. These approaches rely on a dual representation of the regularizer, and can be applied to (13) using the primal-dual equivalency

$$\inf_{u \in \mathcal{C}} \int_{\Omega} \langle u(x), s(x) \rangle dx + \int_{\Omega} \Psi(Du) \quad (76)$$

$$= \sup_{v \in \mathcal{D}} \int_{\Omega} \min\{0, s - \text{Div } v\} dx, \quad (77)$$

$$\mathcal{D} := \{v \in (C_c^\infty(\Omega))^{d \times l} \mid \Phi(v(x)) \leq 1, \forall x \in \Omega\}, \quad (78)$$

where Φ refers to the *dual norm* of Ψ . Then v is discretized on the *edges* of a 4- or higher-connected graph (cf. Fig. 1), and the constraint $\Phi(v(x)) \leq 1$ is discretized on the *vertices* [21] or *cell centers* [3].

The key is that the local constraints that define \mathcal{D} do not have to be (point-wise) separable after discretization, and in fact non-separable choices have been observed to give good results. The main difficulty is that under these discretizations the regularizer in (76) cannot be evaluated easily, therefore the *dual* problem must be solved instead. A primal solution u can then be recovered using the multipliers of an interior-point scheme [21] or a combination of dual smoothing and thresholding [3].

3.4 Experimental Comparison

In order to evaluate the practical consequences of the above theoretical results, we compared the different approaches on several two-class labeling problems, i.e. continuous cut problems.

Due to the coarea-like property (25), the original continuous problem then admits an integral minimizer. In particular this implies that if the solution is unique then it is integral. Therefore, by restricting ourselves to the two-class case, we make sure that any fractional solutions of the discretized problems are purely caused by the discretization.

We compared the following energies:

discretization	w_1	w_2	w_3	w_4
4-neighborhood	$\frac{\pi}{4}$			
8-neighborhood	$\frac{\pi}{8}$	$\frac{\pi}{8\sqrt{2}}$		
16-neighborhood	$\frac{\arctan(1/2)}{2}$	$\frac{\arctan(2) - \arctan(1/2)}{2\sqrt{2}}$	$\frac{\pi/4 - \arctan(1/2)}{2\sqrt{5}}$	$\frac{\pi/2 - \arctan(2)}{2\sqrt{5}}$
6-nb/fd-fw (comb.)	$\frac{\sqrt{2}}{2}$	$\frac{2 - \sqrt{2}}{2}$		
8-nb/fd-sym (comb.)	$\frac{\sqrt{2}}{2}$	$\frac{2 - \sqrt{2}}{4}$		

Table 1. Weights used for the graph-based pairwise discretizations (cf. Fig. 1). The weights in the last two rows were chosen to match the finite-differences energies (fd-fw and fd-sym) on integral labelings.

- Classical pairwise energies with 4-, 8-, and 16- neighborhood (44) as depicted in Fig. 1 (n_4, n_8, n_{16}). The weights were chosen according to (42) and are listed in Table 1. For non-integral labelings, the pairwise LP relaxation (46) was employed.
- The “isotropic” forward-differences scheme as outlined in Sect. 3.2 (*fd-fw*), an averaged variant that also involves backward differences in order to make it more symmetric (*fd-sym*), and centered differences evaluated on the cell centers (*center*).
- First- and second-order [59] upwind discretizations (*up1, up2*). We also implemented the up-winding approach of [15] and found that the energies were generally so close to *up1* as to be visually indistinguishable, therefore we omit them from the figures.
- The “combinatorial continuous maximal flow” [21] and “mimetic” [3] dual discretizations from Sect. 3.3 (*ccmf-d, mi-d*) and their primal counterparts (*ccmf-p, mi-p*).

Anisotropy of the Discretization In order to get a quantitative impression on the anisotropy induced by the various methods, we evaluated the energies on a labeling u^h rotated by different angles. The rotated source labelings were generated at a resolution of 512×512 pixels and downsampled to 128×128 pixels in order to reduce artifacts (Fig. 2).

We first evaluated the functionals on *integral* labelings (Fig. 3). Since the images were artificially generated, the true length is known to be $\pi + 2 = 5.14$ for the half disc with radius normalized to 1. The anisotropies of the 4-, 8-, and 16-neighborhood are clearly visible with the number of bumps increasing and the magnitude of the anisotropy decreasing for larger neighborhoods. The isotropic finite-differences and dual energies seemingly do not work very well in comparison: the energy is overestimated, and they show large oscillations.

However, when the edges of the shape are slightly blurred, the picture changes (Fig. 4): for a light 4-pixel Gaussian blur, the variation of the finite-differences and dual energies over all rotations is already close to one pixel width (an energy difference of 0.024 at this scale), with the *center*, *ccmf-d*, and *mi-d* energies showing the least amount of anisotropy. For a 10-pixel blur, the anisotropy fur-



Fig. 2. Artificial labelings u^h used for the tests in Fig. 3 and Fig. 4. **First row:** The integral labelings were downscaled from a larger 512×512 pixel image, rotated by different angles. **Second and third row:** Non-integral labelings were similarly obtained by smoothing the source image using a Gaussian filter with increasing variance and subsequent downscaling.

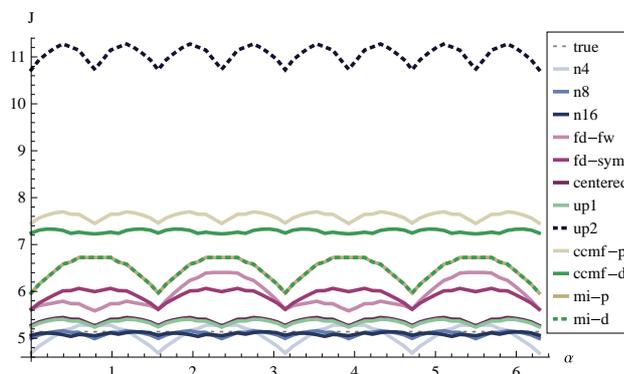


Fig. 3. Energy of the rotated integral labelings in Fig. 2, first row, vs. rotation angle. The pairwise discretizations with 4-, 8-, and 16-neighborhoods exhibit a distinct anisotropy, which decreases as the neighborhood size increases. On such integral labelings the isotropic finite-differences- and dual energies overestimate the true length.

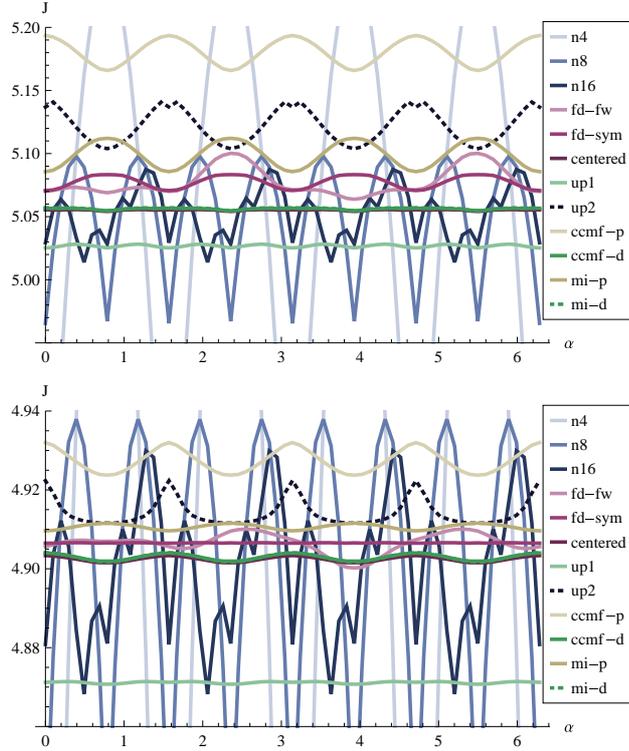


Fig. 4. Energies for the fractional labelings in Fig. 2, second and third row (note the different scales). By allowing a certain amount of fractional labels, the isotropy of the finite-differences- and dual energies is greatly improved. Note that a ground truth is not available since the total variation is affected by the smoothing process; this also explains the overall lower energies compared to Fig. 3.

ther decreases. The important point is that this is only possible because we allow a moderate amount of fractional labels.

In contrast, the LP relaxations of the graph cut energies show no reduction in the discretization-induced anisotropy, with a length variation equivalent to 3 ($n16$), 6 ($n8$) and 25 ($n4$) pixels, clearly preferring some directions over others.

3.5 Consistency of the Discretization

As noted in Sect. 3.1, pairwise energies can be shown to approximate the true length for infinitesimal grid spacing, but only if the neighborhood size simultaneously grows to infinity. To investigate whether this is a problem in practice, we first computed a large half-disc shaped template with a size of 2048×2048 pixels. From this template we generated a range of downscaled copies with resolutions down to 32×32 (Fig. 5). For each resolution we computed the energies using the different regularizers.

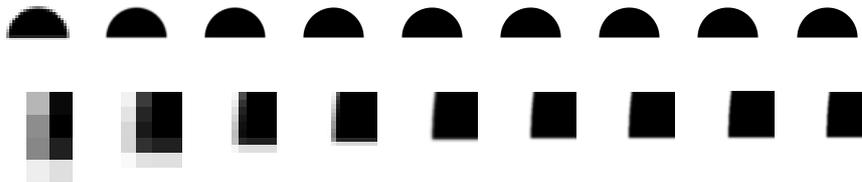


Fig. 5. Labelings used for comparing the isotropy of the regularizer at different resolutions (Fig. 6). **Top row:** Discretization of labeling functions for grid sizes between 32×32 and 2048×2048 . **Bottom row:** Detail (lower left corner).

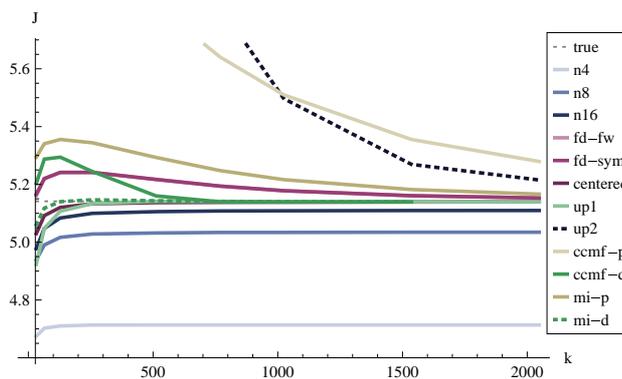


Fig. 6. Energy comparison for different resolutions. Shown are the energies for the templates in Fig. 5 vs. the horizontal grid size $k \in \{32, \dots, 2048\}$. For a fixed neighborhood, the graph-based energies ($n4$, $n8$, $n16$) exhibit a systematic anisotropy, while the finite-differences and dual discretizations converge to the true energy as the resolution increases (note that $fd-fw$ and $fd-sym$ coincide in this example). Fastest convergence is achieved by the centered differences ($center$), upwind ($up1$), and dual discretizations ($ccmf-d, mi-d$).

It becomes apparent that the LP relaxation-/graph cut-based energies exhibit the same anisotropy over all scales if the neighborhood size is kept the same, and systematically underestimate the true length (Fig. 6). This is in accordance with the observations in Sect. 3.1 (see also (75)): for a fixed neighborhood size, the discretized functionals Γ -converge to an *anisotropic* spatially continuous functional. In contrast, the length estimated by the finite-differences schemes converges to the true length as the resolution increases as predicted by Prop. 1.

When the template is rotated by 45 degrees, the 4-neighborhood energy increases beyond the ground truth, while the finite-differences energies again converge to the true length (Fig. 7). The fastest convergence is achieved by the centered differences ($center$), upwind ($up1$), and dual discretizations ($ccmf-d, mi-d$).

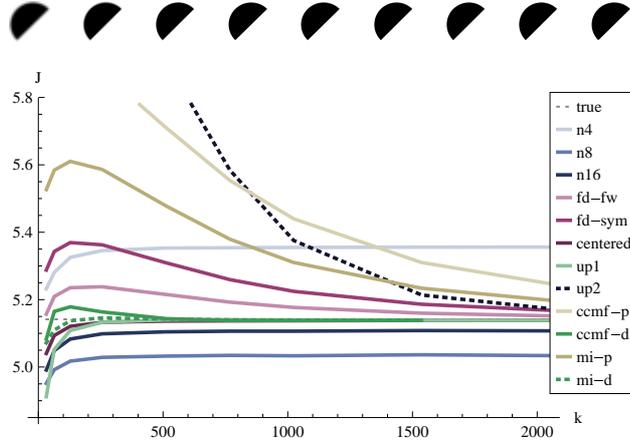


Fig. 7. Variant of the experiment in Fig. 6 for a rotated template. Again, the graph-based discretizations exhibit a systematic error. Due to the rotation, the true length is now overestimated by the 4-neighborhood pairwise energy (n_4). The finite-differences and dual energies converge to the true isotropic length.

integrality	n4	n8	n16	fd-fw-c	fd-sym-c	center	ccmf-d
$\beta = 0.05$	100.00	100.00	100.00	100.00	100.00	97.19	96.66
$\beta = 0.01$	100.00	100.00	100.00	100.00	100.00	96.14	95.14

integrality	up1	mi-d	fd-fw	up2	fd-sym	mi-p	ccmf-p
$\beta = 0.05$	96.44	96.24	94.71	94.17	93.33	92.02	89.21
$\beta = 0.01$	95.12	94.73	92.79	90.67	90.80	88.99	84.64

Table 2. Percentage of points in which $\min(|u(x)|, |1 - u(x)|) < \beta$ for the different regularizers, sorted in ascending order. Minimizers of the combinatorial energies are fully integral. In this example, centered differences (*center*, highlighted red) achieve the least amount of fractional values amongst the isotropic energies. Isotropic energies that use larger neighborhoods (*ccmf-p*, *mi-p*, also *fd-sym*, *up2* to some extent) perform worse as the larger stencil decreases resolution.

3.6 Integral and Fractional Minimizers

In order to see how the choice of the discretization affects the minimizer, consider the problem in Fig. 8, first row. The input consists of an image with two slightly rotated circular segments. The gray regions are “uncertain” and have to be filled in by the regularizer. We added subtle Gaussian noise in order to render the minimizer unique.

The minimization problems were solved to a high accuracy (relative gap 10^{-8}) in order to avoid effects caused by suboptimal solutions. We used the commercial MOSEK solver, which employs an interior-point method that is well suited to solving small problems to high accuracy. The results of the graph-based energies are integral and thus global minimizers of their energies over the set of *integral* labelings.

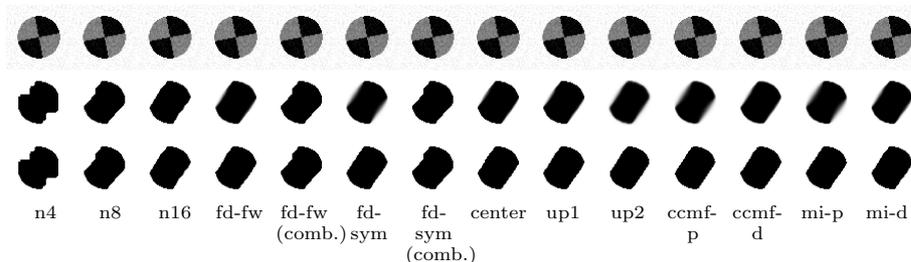


Fig. 8. Segmentation results for different discretizations. **First row:** Input image. **Second row:** Minimizers for the different discretizations. The graph-based discretizations generate artifacts in the solution. The finite-differences energies show significantly less artifacts and generate *fractional* labelings at the boundaries. **Third row:** Thresholded solutions. The *thresholded fractional* solutions of the finite-differences and dual energies result in a better approximation of the desired shape than the *integral* solutions of the graph-based pairwise energies.

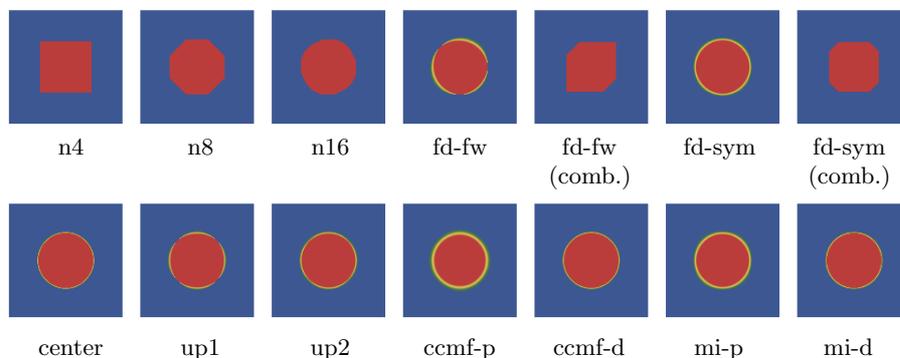


Fig. 9. Minimizers for the artificial test function $s(x) = 1/\|x\| - 3/2$. The artificial coloring highlights non-integral values (yellow). The combinatorial energies have fully integral minimizers but a visible anisotropy. Of the isotropic energies, centered differences (*center*) and the dual energies (*ccmf-d, mi-d*) perform best with respect to the integrality of the solution (see also Table 2).

For this experiment we additionally included two regularizers proposed in [38]. They correspond to a restriction of the finite-differences energies to *combinatorial* objectives, i.e., they coincide with the finite-differences energy on *integral* labelings. The observation made in [38] was that these regularizers can be represented using submodular ternary potentials and therefore be globally optimized. In fact, it turns out that both regularizers can be implemented using *pairwise terms only* by adding diagonal edges, resulting in 6- and 8- neighborhoods (*fd-fw (comb.)*, *fd-sym (comb.)* in Fig. 1, Table 1). By construction, minimizers of these energies minimize the finite-differences objective *on the set of integral labelings*.

From the results it becomes clear that all graph-based pairwise energies exhibit visible artifacts due to the anisotropy (Fig. 8, second row). Enlarging the neighborhood reduces the artifacts, however they cannot be completely avoided. In contrast, the isotropic energies generate solutions that are much closer to an approximation of the true, continuous solution, with a small amount of fractional labels at the slanted edges.

As mentioned in the discussion, there may be cases where an integral output is required. We therefore thresholded the output of the finite-differences methods at $1/2$ (Fig. 8, last row). Again, the solutions obtained by solving the finite-differences approximation of the relaxed problem and subsequent thresholding are visually superior to the solutions obtained by minimizing the pairwise energies.

In order to test the amount of fractional labels introduced by the discretizations, we computed the minimizers for an artificial rotationally symmetric input (Fig. 9). The pairwise energies generate fully integral solutions, but the anisotropy is clearly visible; this is also the case for some non-symmetric finite-differences energies (*fd-fw, up1*). The primal energies *ccmf-p, mi-p* clearly suffer from their larger stencils. The centered and dual energies (*center, ccmf-d, mi-d*) perform best, see also the quantitative comparison (Table 2).

Naturally the question arises how fractional minimizers of the finite-differences energy compare to integral minimizers, i.e. whether it makes sense to minimize the finite-differences energy in a combinatorial setting. From Fig. 10 it becomes clear that this cannot be recommended:

Integral minimizers of the finite-differences energy are visually clearly inferior to those obtained by rounding a fractional solution.

In fact, the latter are *not* integral minimizers, as can be seen by comparing the energies (Table. 3). This effect has also been observed in [38], and may seem quite counter-intuitive at first. We will discuss its implications in Sect. 6.

In particular, it indicates that isotropic energies over the set of integral labelings may not be an optimal approach. For the same reason, comparing energies of solutions obtained by rounding and solutions obtained by combinatorial optimization has only very limited value, since the energy does not necessarily provide an indication which solution better approximates the spatially continuous solution.

In order to focus on the quality of the discretization, all of the above results were computed on the two-class problem. In the multi-class setting, additional

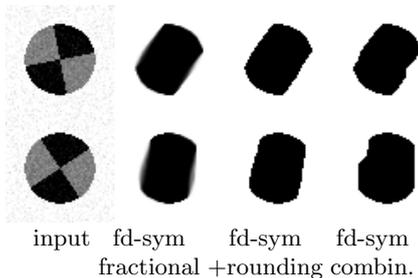


Fig. 10. Minimizing energies over fractional vs. integral labelings. Finding a minimizer (second from left) of the finite-differences discretization and subsequent rounding (second from right) results in less artifacts than solving the combinatorial problem of minimizing *exactly the same* energy over the set of integral labelings (right).

problem	fractional	rounded	combinat.
1st row Fig. 10	-2207.13	-2164.41	-2166.86
2nd row Fig. 10	-2213.08	-2177.46	-2181.04

Table 3. Energies for the solutions in Fig. 10. The rounded fractional solution is *not* a combinatorial (integral) minimizer of the energy, but it is nevertheless visually clearly superior. This is an example where it is not reasonable to minimize the energy under an integrality constraint.

uncertainties may be introduced by the relaxation of the continuous problem. Therefore, fractional solutions cannot be unambiguously classified as caused by either the discretization or the relaxation. However, in principle the above considerations also apply – with less theoretical justification – to the multi-class case.

4 Optimality and Rounding

As noted above, the multiclass relaxation (13) is convex and can thus be solved globally optimal. However, the minimizer u^* of the relaxed problem may not lie in $\mathcal{C}_{\mathcal{E}}$, i.e., it is not necessarily integral. Therefore, in applications that require a true partition of Ω , some rounding process is needed in order to generate an integral labeling \bar{u}^* . Unlike in the two-class case, this may increase the objective, and lead to a suboptimal solution of the original problem (2).

Note that this behavior is independent of the effects discussed in the previous section, where we considered the occurrence of fractional labels due to the process of switching from the continuous problem formulation to a *discrete*, finite-dimensional one, and explicitly ruled out relaxation effects by exclusively considering the two-class case. In this section, we consider the *second* source of fractional solutions, which is the *relaxation* of the original combinatorial *multi-class* problem to a convex problem, and rule out discretization effects by working completely in the spatially continuous setting.

An important question is whether it is possible to obtain, using the convex relaxation (13), *integral* solutions with an upper bound on the objective. Specifically, we concentrate on inequalities of the form

$$f(\bar{u}^*) \leq (1 + \varepsilon)f(u_\varepsilon^*) \quad (79)$$

for some constant $\varepsilon > 0$, which provide an *upper bound on the objective* of the *rounded* integral solution \bar{u}^* with respect to the objective of the (unknown) *optimal* integral solution u_ε^* of (2). The specific form (79) can be attributed to the alternative interpretation

$$\frac{f(\bar{u}^*) - f(u_\varepsilon^*)}{f(u_\varepsilon^*)} \leq \varepsilon, \quad (80)$$

which provides a bound for the relative gap to the optimal objective of the combinatorial problem. Such ε can be obtained *a posteriori* by actually computing (or approximating) \bar{u}^* and a dual feasible point: Assume that a feasible primal-dual pair $(u, v) \in \mathcal{C} \times \mathcal{D}$ is known, where u approximates u^* , and assume that some integral feasible $\bar{u} \in \mathcal{C}_\varepsilon$ has been obtained from u by a rounding process. Then the pair (\bar{u}, v) is feasible as well since $\mathcal{C}_\varepsilon \subseteq \mathcal{C}$, and from the considerations in Sect. 4 we obtain an *a posteriori* optimality bound of the form (80) with respect to the optimal *integral* solution u_ε^* :

$$\frac{f(\bar{u}) - f_D(u_\varepsilon^*)}{f_D(u_\varepsilon^*)} \leq \frac{f(\bar{u}) - f_D(u_\varepsilon^*)}{f_D(v)} \leq \frac{f(\bar{u}) - f_D(v)}{f_D(v)} =: \varepsilon'. \quad (81)$$

However, this requires that f and f_D can be accurately evaluated, and requires to compute a minimizer of the problem for the specific input data, which is generally difficult, especially in the spatially continuous formulation.

In contrast, true *a priori* bounds do not require knowledge of a solution and apply uniformly to all problems of a class, irrespective of the particular input.

When considering rounding methods, one generally has to discriminate between

- *deterministic* vs. *probabilistic* methods, and
- *spatially discrete (finite-dimensional)* vs. *spatially continuous* methods.

Most known *a priori* approximation results only hold in the finite-dimensional setting, and are usually proven using graph-based pairwise formulations. In contrast, in this section we will again focus on an “optimize first” perspective. From the remarks in Sect. 2.3, we see that in the spatially continuous setting the *two-class* problem admits the trivial rounding approach with

$$\bar{u}_\alpha^* := e^1 1_{\{u_1^* > \alpha\}} + e^2 1_{\{u_1^* \leq \alpha\}} \quad (82)$$

for almost every $\alpha > 0$ due to the coarea formula. In view of the ε -optimality bound (79), this amounts to $f(\bar{u}^*) = f(u_\varepsilon^*)$, i.e. $\varepsilon = 0$. In the finite-dimensional multiclass case, the most prominent approaches for finding integral combinatorial minimizers are the α -expansion approach, more general move-making methods such as continuous binary fusion, and LP relaxations. In the following sections, we give an overview of the various connections.

Algorithm 1. Graph-Based α -Expansion

- 1: Choose $\ell^{(0)} : V \rightarrow \mathcal{I}$.
 - 2: $k \leftarrow 0$.
 - 3: **repeat**
 - 4: **for** all $j \in \mathcal{I}$ **do**
 - 5: For $V' \subseteq V$, let $\ell_{j,V'}(x) := \begin{cases} j, & x \in V', \\ \ell^{(k)}(x), & x \notin V'. \end{cases}$
 - 6: Find V' s.t.

$$\ell' := \ell_{j,V'} = \arg \min_{V' \supseteq \{x \in V \mid \ell^{(k)}(x) = j\}} f(\ell_{j,V'}). \quad (84)$$
 - 7: $\ell^{(k+1)} \leftarrow \begin{cases} \ell', & f(\ell') < f(\ell^{(k)}), \\ \ell^{(k)}, & \text{otherwise.} \end{cases}$
 - 8: $k \leftarrow k + 1$.
 - 9: **end for**
 - 10: **until** $f(\ell^{(k)})$ did not decrease in at least one of the inner iterations.
 - 11: **Output:** $\ell^+ := \ell^{(k)}$.
-

4.1 Isolation Heuristic and α -Expansion

Probably the best-known bound for obtaining solutions of the multiclass labeling problem on graphs with pairwise terms is provided in the original “graph cut” paper by Boykov et al. [8], and is based on the α -expansion method.

The α -expansion method provides a way to reduce multiclass labeling problems to a sequence of two-class problems, which can then be solved globally optimal, for instance using graph cuts.

Denote by $G = (V, E)$ the (undirected) graph representation of the problem, where the energy for a labeling $\ell : V \rightarrow \mathcal{I} := \{1, \dots, l\}$ is

$$f(\ell) = \sum_{x \in V} s_x(\ell(x)) + \sum_{e=(x^1, x^2) \in E} d(\ell(x^1), \ell(x^2)) \quad (83)$$

for nonnegative $s_x : \mathcal{I} \rightarrow \mathbb{R}_{\geq 0}$ and metric $d : \mathcal{I}^2 \rightarrow \mathbb{R}_{\geq 0}$.

An early idea for generating integral solutions from the solution of the relaxed problem was provided by [22] in a multiterminal cut framework, which corresponds to the case where d is the uniform metric. It uses an *isolation heuristic*, which consists in computing l individual cuts (i.e. two-class segmentations), where each label in turn is segmented against all others. The multiterminal cut is then constructed as the union of the $l - 1$ best cuts. Using this approach, a bound of $\varepsilon = 1 - 2/l$ was proven in [22] for the finite-dimensional problem (83).

The α -expansion method can be seen as a repeated, sequential application of the steps in the isolation heuristic, extending it to general metrics. It proceeds in a number of outer iterations, as shown in Alg. 1:

In each step, one label j is selected, and $\ell_{j,V'}$ is constructed from $\ell^{(k)}$ so that each vertex either keeps its current label or switches to label j . Thus, during one

step the set of points which carry label j may only expand. Therefore the steps are referred to as α -*expansion moves*, with α referring to the selected label j in the original work [8]. The inner problem (84) is a *two-class* labeling problem, and, under the assumptions on the discretization and on d , contains semi-metric pairwise terms, and can thus be solved exactly using graph cut techniques.

The output ℓ^+ can be considered as a *local minimum* with respect to expansion moves, as there cannot be an expansion move starting from ℓ^+ that decreases the energy. The authors then show the following proposition:

Proposition 2. [8, Thm. 6.1] *Let ℓ^+ be a local minimum with respect to expansion moves, and ℓ^* be a global minimizer of (83). Then*

$$f(\ell^+) \leq 2cf(\ell^*), \quad (85)$$

where

$$c := \frac{\max_{i \neq i'} d(i, i')}{\min_{i \neq i'} d(i, i')} \geq 1. \quad (86)$$

From (85) we therefore obtain $\varepsilon = 2c - 1$ for the α -expansion method. The principle of reducing multi-class problems to a sequence of two-class problems such as (84) is also the basis for the α - β -swap technique from the same authors, which can handle the case of semi-metric d and in practice often leads to a smaller energy, but provides no bound similar to (85). A generalization can be found in [49, 50], where the authors view the problem of finding the optimal expansion step (84) as the decision between two solutions: identifying V' with its characteristic function $u' := 1_{V'} : V \rightarrow \{0, 1\}$, the expansion step becomes

$$u' = \arg \min_{u' : V \rightarrow \{0, 1\}} f \left((1 - u')\ell^{(k)} + u'j \right), \quad (87)$$

This can be seen as a “binary fusion” between two candidate solutions: the current iterate $\ell^{(k)}$ and the constant solution $\ell \equiv j$. As these problems correspond to two-class labeling, they can be solved globally optimal, e.g. using graph cuts.

4.2 Continuous Binary Fusion

The finite-dimensional approach (87) can be generalized to the *spatially continuous* case by essentially replacing V with Ω . This was proposed in [65] in an informal way, without specifying the actual function spaces and assumptions on the functionals. In [54, 55], the authors argue that Prop. 2 similarly holds for functionals of the form (2), with the separable (but anisotropic) regularizer

$$\Psi(z) = \sum_{j=1}^l \|Az^j\|_2, \quad z \in \mathbb{R}^{d \times l}, \quad (88)$$

for some $A \in \mathbb{R}^{d \times d}$. However, their proof seems to be insufficient in several aspects. The authors do not specify the function spaces, and use a pointwise

argument much as the vertex- and edge-wise argument used in Prop. 2. This requires additional justification when dealing with BV functions, which are defined only almost everywhere. Also, there are some technical problems regarding the use of the classical definition for the interior and exterior of a set, which affects well-definedness in connection with functionals on BV involving the total variation.

Apart from these problems, it is nontrivial to show that the continuous analogue to Alg. 1 actually terminates. This is not an issue in the finite-dimensional setting: since there is only a finite number of configurations, there can only be a finite number of iterations until the energy does not decrease anymore, and the algorithm stops.

4.3 LP Relaxation with Derandomization

In [10], the authors consider an LP relaxation of the multiway cut and provide a randomized approximation algorithm with $\varepsilon = \frac{1}{2} - \frac{1}{l}$. In the multiclass labeling setting, their formulation corresponds to the graph-based discretization (83) with (locally weighted) uniform metric regularizer $d(i, j) = 1_{\{i \neq j\}}$. As seen in the previous chapter, for grid graphs this corresponds to $\Psi = \|\cdot\|_1$.

In order to cope with general metrics, [37] adapted a variant of the LP formulation (Sect. 3.1), which raises the bound to $\varepsilon = 1$ for the uniform metric. For the uniform metric, the LP relaxation has the form

$$\min_{y,z} \sum_{x \in V} \sum_{j \in \mathcal{I}} s_x(j) y_{x,j} + \sum_{e \in E} w_e z_e \quad (89)$$

$$s.t. \quad y_{x,\cdot} \in \Delta_l, \quad x \in V, \quad (90)$$

$$z_e = \frac{1}{2} \sum_{j \in \mathcal{I}} z_{e,j}, \quad (91)$$

$$z_{e,j} \geq y_{x^1,j} - y_{x^2,j}, \quad (x^1, x^2) \in E, \quad (92)$$

$$z_{e,j} \geq y_{x^2,j} - y_{x^1,j}, \quad (x^1, x^2) \in E. \quad (93)$$

The variables $y_{x,j}$ correspond to $u_j(x)$, i.e. semantically $y_{x,j} = 1$ iff $\ell(x) = j$, and the scalars w_e constitute edge weights to allow for non-homogeneous regularizers. Without the slack variables, the LP amounts to

$$\min_{y \in (\Delta_l)^n} \left\{ \sum_{x \in V} \sum_{j \in \mathcal{I}} s_x(j) y_{x,j} + \frac{1}{2} \sum_{e=(x^1,x^2) \in E} w_e \sum_{j \in \mathcal{I}} |y_{x^1,j} - y_{x^2,j}| \right\}. \quad (94)$$

Assuming $w_e \equiv 1$, this is equivalent to a graph-based discretization of (83) and closely resembles (13), which motivates to adapt the proof for the spatially continuous setting.

For finite-dimensional problems, the bound of $\varepsilon = 1$ is proven in [37] by first considering a randomized rounding method and then showing that it can be derandomized in polynomial time.

The proof strongly relies on the vertex/edge representation, and therefore cannot be trivially transferred to the spatially continuous setting. In particular, several steps directly involve the slack variables z_e and $z_{e,j}$, which do not have a direct analogue in the continuous setting.

An interesting question is how to transfer the derandomization to the continuous setting – they crucially depend on fixing single labels, which is not available in a well-defined sense in the continuous setting. To our knowledge, there have been no real advances in this direction.

4.4 Probabilistic Multi-Class Results

A first approximation result for the full spatially continuous model (2) was proven in [47]. It relies on proving a generalized *approximate* coarea formula,

$$(1 + \varepsilon)f(u) \geq \int_{\Gamma} f(\bar{u}_\gamma) d\mu(\gamma). \quad (95)$$

The right-hand side encompasses a *parameter space* Γ for a “rounding parameter” γ with an associated probability distribution μ , and a *parametrized rounding method* $(u, \gamma) \mapsto \bar{u}_\gamma$. With this notation, (95) can be interpreted as the expectation of the objective function applied to u after rounding according to a random rounding parameter γ , i.e.,

$$(1 + \varepsilon)f(u) \geq \int_{\Gamma} f(\bar{u}_\gamma) d\mu(\gamma) = \mathbb{E}_\gamma f(\bar{u}_\gamma). \quad (96)$$

If such a relation is shown, it implies that for any minimizer u^* of the relaxed problem (13),

$$\mathbb{E}_\gamma f(\bar{u}_\gamma^*) \leq (1 + \varepsilon)f(u^*) \leq (1 + \varepsilon)f(u_\mathcal{E}^*), \quad (97)$$

holds, bounding the ratio between the objective of the *rounded relaxed solution* and the *optimal integral solution* – in a probabilistic sense – by $(1 + \varepsilon)$.

In [48] it is shown that if one chooses Γ as the space of *sequences* γ of pairs $\gamma^k = (i^k, \alpha^k)$ with the uniform distribution, it is possible to construct such a bound.

Given a sequence $\gamma \in (\mathcal{I} \times [0, 1])^\mathbb{N}$, the rounding step $R_\gamma : u \mapsto \bar{u}_\gamma$ can be recursively defined as

$$R_\gamma(u) := R_{\gamma'} \left(e^{i^1} 1_{\{u_{i^1} > \alpha^1\}} + u 1_{\{u_{i^1} \leq \alpha^1\}} \right), \quad (98)$$

where $\gamma' := (\gamma^2, \gamma^3, \dots)$.

Intuitively, the recursion “terminates” as soon as almost every point in Ω has been assigned a “hard” label in \mathcal{E} . More precisely, it can be shown that its expectation $\mathbb{E}_\gamma f(\bar{u}_\gamma)$ is well-defined on $\text{BV}(\Omega, \Delta_l)$ in the sense that for almost every γ it generates $\bar{u}_\gamma \in \text{BV}(\Omega, \mathcal{E})$.

Then, assuming that there exists a lower bound λ_l for the integrand Ψ such that

$$\begin{aligned} \Psi(z = (z^1, \dots, z^l)) &\geq \\ \lambda_l \frac{1}{2} \sum_{i=1}^l \|z^i\|_2 \quad \forall z \in \mathbb{R}^{d \times l}, \sum_{i=1}^l z^i &= 0, \end{aligned} \quad (99)$$

and an upper bound $\lambda_u < \infty$ such that

$$\begin{aligned} \Psi(y(e^i - e^j)^\top) &\leq \\ \lambda_u \quad \forall i, j \in \{1, \dots, l\}, y \in \mathbb{R}^d, \|y\|_2 &= 1, \end{aligned} \quad (100)$$

it is possible to show the following theorem:

Theorem 2. [47] *Assume $s \in L^\infty(\Omega)^l$, $s \geq 0$. Furthermore, let $\Psi : \mathbb{R}^{d \times l} \rightarrow \mathbb{R}_{\geq 0}$ be positively homogeneous, convex and continuous, and $u \in \mathcal{C}$. Assume that there exist λ_l, λ_u such that the lower- and upper-boundedness conditions (99) and (100) are satisfied. Then the method (98) generates an integral labeling $\bar{u} \in \mathcal{C}_{\mathcal{E}}$ almost surely, and*

$$\mathbb{E}f(\bar{u}) \leq 2 \frac{\lambda_u}{\lambda_l} f(u). \quad (101)$$

The theorem shows that by randomly rounding a minimizer u^* of the relaxed problem (13), one obtains an *integral* solution \bar{u}^* with a guaranteed bound

$$\mathbb{E}f(\bar{u}^*) \leq 2 \frac{\lambda_u}{\lambda_l} f(u_{\mathcal{E}}^*) \quad (102)$$

with respect to the *optimal* integral solution $u_{\mathcal{E}}^*$. For the metric regularizer in (5) with given metric d , it is possible to specialize the result to

$$2 \frac{\lambda_u}{\lambda_l} = 2 \frac{\max_{i,j} d(i,j)}{\min_{i \neq j} d(i,j)}, \quad (103)$$

paralleling the above-mentioned results for the finite-dimensional LP relaxation and α -expansion methods.

One may ask the question if this bound, and equally the one in Prop. 2, is actually relevant, i.e., if it the obtained solution is better than just assigning a constant label to the whole image. To see why the latter method may perform very bad in terms of the approximation factor, consider a problem where $s(x) = e - e^i(x)$ for some function $i : \Omega \rightarrow \mathcal{E}$ that partitions the image domain into l regions of non-zero area and finite perimeter, and d is the weighted uniform distance $d = \delta d_u$ with δ (infinitesimally) small.

Setting $u(x) = e^{i(x)}$ yields an upper bound on the optimal energy that is arbitrarily close to zero, while any constant labeling has an energy larger than the area of the smallest region, $\min_{l \in \mathcal{E}} |i^{-1}(\{l\})|$, and therefore much worse than the factor of 2 achieved by (103). For the finite-dimensional case, one can find examples where the bound (103) is tight, at least asymptotically for an infinite number of labels [37]. However, in the continuous setting we are not aware of such a result.

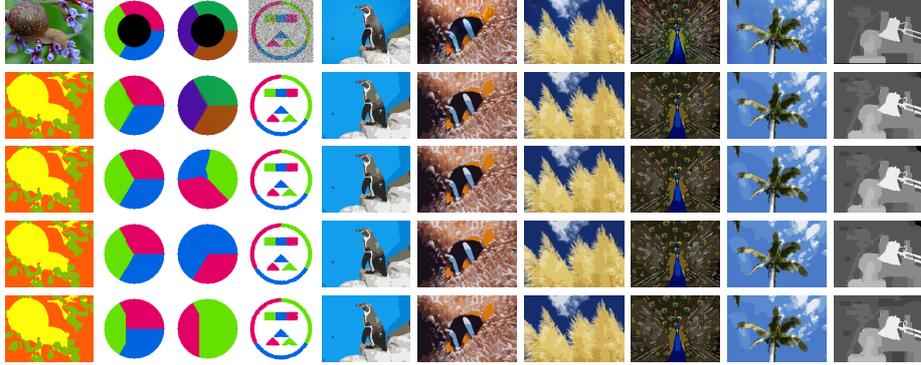


Fig. 11. Top to bottom: Problems 1 – 10 of the test set. **Top to bottom:** Original input; relaxed (fractional) solution; integral solutions obtained by deterministic rounding; result of probabilistic rounding; result of α -expansion (for comparison). In specially crafted situations, the probabilistic method may perform slightly worse (second column) or better (third column) than the deterministic approaches. On real-world data, results are very similar. In contrast to the deterministic approaches, the probabilistic method provides true *a priori* optimality bounds.

4.5 Experimental Performance

In this section we show some prototypical examples of what can be expected from the different rounding strategies. It is important to keep in mind that for the *discretized* problem, analog bounds to those provided by Thm. 2 are only valid if the discretization respects the coarea formula. However, for these energies the original finite-dimensional proof [37] already applies. For the finite-differences discretization, a comparison of the *a posteriori* bounds computed via the primal-dual gap must be taken with care, since the gap is caused by the relaxation as well as the discretization. However, unlike in the two-class case, a comparison makes at least limited sense in the multiclass case, since a large *a posteriori* bound suggests that it is not only caused by the discretization, and that the underlying integral solution may be suboptimal due to the relaxation. With this in mind, the observations in the following subsections should be seen only as indicators of what results can be expected qualitatively.

A Priori and A Posteriori Bounds In order to evaluate the tightness of the bound (102) in Thm. 2 in practice, we selected 10 prototypical multiclass labeling problems with 3 – 16 labels each. For each we computed the relaxed solution u^* and the mean as well as the best objective of the rounded solution \bar{u}^* after 20 runs of the probabilistic method (98), see Fig. 11 for some exemplary results. In order to compute the objectives with the sophisticated regularizer (5), we used the commercial MOSEK solver with an accuracy of 10^{-8} . The primal-dual optimization approach provides an (approximate) *a posteriori* bound ε' , in contrast to the theoretical *a priori* upper bound $\varepsilon = 2\lambda_u/\lambda_l - 1$. In practice,

problem	1	2	3	4	5	6	7	8	9	10
# points	76800	14400	14400	129240	76800	86400	86400	76800	86400	110592
# labels	3	4	4	4	8	12	12	12	12	16
mean # iterations	7.27	7.9	8.05	10.79	31.85	49.1	49.4	49.4	49.7	66.1
<i>a priori</i> ε	1.	1.	1.	1.	1.	1.	1.	1.	1.	2.6332
<i>a posteriori</i>										
- first-max	0.0007	0.0231	0.2360	0.0030	0.0099	0.0102	0.0090	0.0101	0.0183	0.0209
- prob. best	0.0010	0.0314	0.1073	0.0045	0.0177	0.0195	0.0174	0.0219	0.0309	0.0487
- prob. mean	0.0007	0.0231	0.0547	0.0029	0.0138	0.0152	0.0134	0.0155	0.0247	0.0292

Table 4. Number of pixels N , number of labels l , mean number of iterations k , predicted *a priori* bound $\varepsilon = 2\lambda_u/\lambda_l - 1$, *a posteriori* bounds for the different rounding methods. The *a posteriori* bound for the probabilistic method is well below the worst-case bound predicted by Thm. 2.

the *a posteriori* bound stayed well below the theoretical bound (Table 4), which is consistent with the good practical performance of α -expansion, which has a similar *a priori* bound.

Also note that the theoretical bounds do not directly apply to the discretized problem, and cannot be directly used as an indicator for the quality of the result, see Sect. 3. However, a large energy increase indicates that the increase might at least partially be caused by the relaxation, rather than the discretization, and therefore the finite-dimensional solution likely does not represent the spatially continuous solution well.

Deterministic and Probabilistic Methods Compared to a simple deterministic rounding to the closest unit vector, the probabilistic method (98) usually leads to a slightly larger energy increase (Table. 4).

However, for problems that are inherently difficult for convex relaxation approaches, we found that the probabilistic approach often generated better solutions. An example is the “inverse triple junction” inpainting problem (second row in Fig. 11), which has at least 3 distinct integral solutions. A variant of this problem, formulated on graphs, was used as a worst case to show the tightness of the LP relaxation bound in [37].

4.6 Future Directions

From these experiments it becomes clear that the *a priori* bounds are generally outperformed by a large margin in practice. One possible strategy to achieve even tighter bounds is to adapt further arguments from [37]: For general metrics, one may consider a variant of the linear program that incorporates an approximation of the metric by *r-hierarchically well-separated tree metrics*. Such metrics are shortest-path metrics generated by weighted graphs with tree structure [4, Def. 6], with the additional property that the edge weights decrease by at least a factor of r on any path from the root to a leaf. For such metrics with $r > 2$, the authors of [37] provide a derandomized algorithm with $\varepsilon = 1 + 4/(r - 2)$.

A probabilistic result [4, Thm. 9], later derandomized in [18], shows that for any metric d , an *r-hierarchically well-separated tree metric approximation* d_r

can be constructed such that

$$d(i, j) \leq d_r(i, j) \leq \alpha d(i, j), \quad (104)$$

with a bound of $O(r \log l \log \log l)$ for the approximation quality α . If the requirement of well-separability is dropped, a tight bound of $O(\log l)$ holds [27, Thm. 1].

Using these techniques, a bound close to the above-mentioned $\varepsilon = 1+4/(r-2)$ should be feasible for the spatially continuous case. Another open question is how to construct worst-case examples in order to prove tightness of the bounds.

5 Numerical Speedup Through Metric Reduction

In order to solve nonsmooth convex problems such as the relaxed problem (13), recently first-order primal-dual methods have gained a lot of popularity (see [26] for an overview), since they

- are relatively straightforward to implement and well-analyzed,
- allow to exploit sparse problem structure in a particularly straightforward way,
- can also be formulated in general Hilbert spaces [19],
- provide a good stopping-criterion in the form of the primal-dual gap,
- involve only basic operations that can be easily parallelized due to their local nature, and are therefore amenable to the massive parallelization available on the upcoming GPU platforms [66], which is much more difficult for the combinatorial graph cut-based methods [32, 25].

While asymptotically they usually do not guarantee linear convergence, they are often very fast for the moderate accuracy required in image processing even when compared to highly optimized commercial interior-point solvers (Fig. 12 and 13; an accuracy of 10^{-4} is usually sufficient for typical computer vision applications).

Primal-dual methods usually start from a saddle-point representation of the problem,

$$\min_{u \in \mathcal{C}} \max_{v \in \mathcal{D}} \{ \langle u, s \rangle + \langle Du, v \rangle - \langle b, v \rangle \}, \quad (105)$$

and generate a sequence of primal-dual iterates (u^k, v^k) based on evaluation of the gradient operator D and projections on the sets \mathcal{C} and \mathcal{D} . Unfortunately, for the tight regularizer (5), the latter is difficult, since there is no known way to project onto the set \mathcal{D} in closed form.

Such projections are then either computed approximately [13], or they are avoided by introducing a suitable number of auxiliary variables [44]. In both cases, the computational effort is at least proportional to the number of constraints in \mathcal{D}_{loc} as defined in (5):

$$\mathcal{D}_{\text{loc}}^d = \{ v \mid \|v^i - v^j\|_2 \leq d(i, j) \forall i \neq j, \sum_k v^k = 0 \}. \quad (106)$$

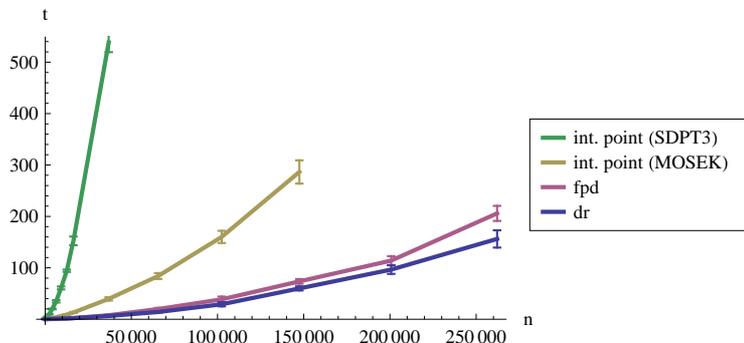


Fig. 12. Performance of first-order and interior-point methods for increasing problem size (number of pixels) n . Shown is the mean time in seconds over 10 problems with different noise, with error bars at 2σ . For a moderate accuracy of 10^{-4} , the primal-dual methods (shown as FPD and DR) outperform both the non-commercial SDPT3 and the commercial MOSEK interior-point solvers, which additionally exceeded the available memory at image sizes of 192×192 and 384×384 , respectively.

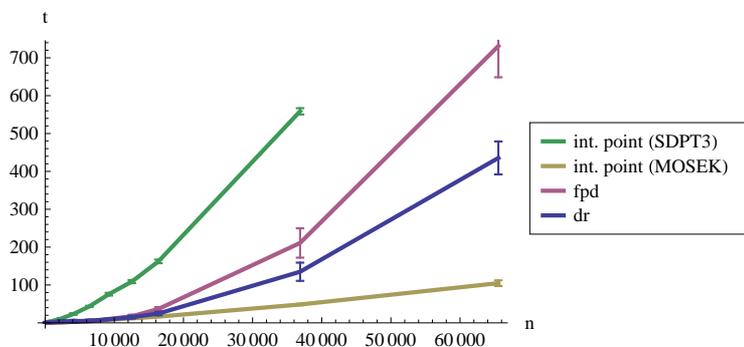


Fig. 13. Performance of first-order and interior-point methods for an accuracy of 10^{-6} , cf. Fig. 12 (note the different scales). For higher requested accuracy, the better asymptotical convergence rate of the interior-point methods becomes advantageous. However, the first-order methods still outperform the SDPT3 solver.

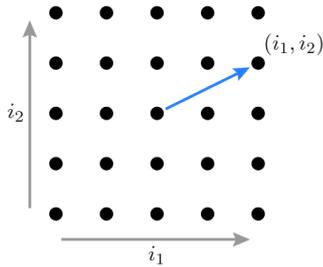


Fig. 14. High label counts are often required when dealing with quantized multi-dimensional properties such as velocity vectors in optical flow estimation. Each point represents a label associated with a combination of velocities in x - and y -direction, therefore the number of labels grows quadratically with the accuracy of the quantization.

For general metrics the number of constraints scales quadratically with the number of labels, which renders the approach unsuitable for larger labels counts. Unfortunately such high label counts occur naturally in problems where the labels correspond to a quantized *multi-dimensional* quantity, such as the two- or three-dimensional velocity vectors in optical flow estimation (Fig. 14).

There have been some suggestions on how to overcome this problem: In [31] and [64], the authors show that for applications where the labels represent vector-valued quantities with n steps in each component, it is possible to work with $2n$ labels instead of n^2 , reducing the number of constraints from $O(n^4)$ to $O(n^2)$. However, this requires the regularizer to be separable in the components, and slightly weakens the relaxation. Another variant is [63], where the authors consider the special case of “cyclic” total variation for angular values, and show how to reduce the number of constraints from $O(n^2)$ to $O(n)$.

Since both of these settings are quite restricted, it is an interesting question whether one can apply such complexity reduction to the general metric regularizer (106). In the following, we outline a generic approach for reducing the number of constraints that covers these special cases and also applies to many other interesting metrics.

5.1 Reducing Metrics

We denote

$$\mathcal{D}_{ij} = \{v \mid \|v^i - v^j\|_2 \leq d(i, j)\}, \quad (107)$$

then

$$\mathcal{D}_{\text{loc}}^d = \bigcap_{i \neq j} \mathcal{D}_{ij} \cap \left\{ v \mid \sum_k v^k = 0 \right\}. \quad (108)$$

As a motivation, assume we can find (i, j, k) with pairwise different i, j, k such that

$$d(i, j) \geq d(i, k) + d(k, j). \quad (109)$$

In fact, since d is a metric and satisfies the triangle inequality, this implies equality. Then

$$\|v^i - v^j\| \leq \|v^i - v^k\| + \|v^k - v^j\| \quad (110)$$

$$\leq d(i, k) + d(k, j) \leq d(i, j). \quad (111)$$

Therefore, if we can find such (i, j, k) satisfying (109), then $(\mathcal{D}_{ik} \cap \mathcal{D}_{kj}) \subseteq \mathcal{D}_{ij}$, and removing the constraint for (i, j) does not change $\mathcal{D}_{\text{loc}}^d$.

This raises the question in which order one should continue removing constraints. In the following we will show that the order is actually irrelevant, and that it suffices to check condition (109) on the original metric only.

In order to simplify notation, instead of metrics we consider mappings m that may assume the value $+\infty$:

$$m \in \mathcal{M} := \{m : \{1, \dots, l\}^2 \rightarrow \mathbb{R}_{\geq 0} \cup \{+\infty\}, \\ m(i, j) > 0 \forall i \neq j\}. \quad (112)$$

Such mappings do not have to satisfy the triangle inequality, but it is still possible to define the associated dual constraint set $\mathcal{D}_{\text{loc}}^m$ analogously to (5). A value of $m(i, j) = +\infty$ indicates that the corresponding constraint may just be skipped. Therefore the goal is to find, for a given metric d , a mapping m with as few finite values as possible such that $\mathcal{D}_{\text{loc}}^m = \mathcal{D}_{\text{loc}}^d$.

With each $m \in \mathcal{M}$ we associate the complete directed graph \mathcal{G}^m with nodes $\{1, \dots, l\}$ and edge weights $m(i, j)$. We denote its shortest-path metric by \bar{m} , i.e., $\bar{m}(i, j)$ is the length of the shortest path from i to j in the graph associated with m . Note that $m = \bar{m}$ in case m is a metric due to the triangle inequality.

Two such mappings induce the same dual constraint set if their hulls coincide:

Proposition 3. *Assume $m, m' \in \mathcal{M}$ satisfy $\bar{m} = \bar{m}'$. Then, with the definition (5) for \mathcal{D}_{loc} ,*

$$\mathcal{D}_{\text{loc}}^m = \mathcal{D}_{\text{loc}}^{m'}. \quad (113)$$

Proof. See appendix. \square

A mapping m is said to *represent* the original metric d if $\bar{m} = d$. We now show that it is possible to “remove” a pair (i, j) satisfying (109) from such an m by setting it to $+\infty$, so that the modified mapping still represents the original metric d .

Proposition 4. *Let d be a metric and $m \in \mathcal{M}$ such that m represents d , i.e., $\bar{m} = d$. Assume there exist pairwise different indices $i, j, k \in \{1, \dots, l\}$ such that*

$$d(i, j) \geq d(i, k) + d(k, j), \quad (114)$$

and define $m' \in \mathcal{M}$ by

$$m'(i', j') := \begin{cases} +\infty, & (i', j') = (i, j) \\ m(i', j'), & \text{otherwise.} \end{cases} \quad (115)$$

Then m' also represents d :

$$\bar{m}' = d. \quad (116)$$

Proof. See appendix. \square

These observations suggest a way of removing constraints from the constraint set $\mathcal{D}_{\text{loc}}^d$:

Proposition 5. *Let d be a metric. Define*

$$\begin{aligned} \mathcal{R} := & \{(i, j) \in \{1, \dots, l\} \mid i \neq j, \\ & \forall k \in \{1, \dots, l\} \setminus \{i, j\} : d(i, j) < d(i, k) + d(k, j)\}. \end{aligned} \quad (117)$$

Then

$$\mathcal{D}_{\text{loc}}^d = \bigcap_{(i,j) \in \mathcal{R}} \mathcal{D}_{ij} \cap \left\{ v \mid \sum_k v^k = 0 \right\}. \quad (118)$$

Proof. See appendix. \square

Prop. 5 provides a straightforward way of eliminating exactly all redundant constraints by checking condition (117), leaving only a “skeleton” of the original metric. This process requires $O(l^3)$ but can be performed in a precomputation step if the metric is known beforehand.

5.2 Experimental Results

We provide here several examples of metrics and their reduced form. The performance is quantified by the number of constraints remaining in the dual constraint set. In conjunction with the FPD primal-dual method we observed that the runtime is dominated by the projections on $\mathcal{D}_{\text{loc}}^d$, and is therefore almost proportional to the number of constraints. Therefore the number of remaining constraints provides a good indicator for the speed-up that can be expected.

The first thing to notice is that for the well-known *uniform (Potts) metric*, $d(i, j) = 1_{\{i \neq j\}}$, the method does not provide any reduction (Fig. 15). However, this is to be expected since the metric does not contain any “linear” structure.

In contrast, for the *linear metric*, $d(i, j) = |i - j|$, we observe a full reduction from $l(l - 1)/2$ to $(l - 1)$ constraints, effectively reducing the complexity from $O(l^2)$ to $O(l)$ (Fig. 16).

The linear metric is a special case of the class of *tree metrics*, i.e., shortest-path metrics on tree graphs (Fig. 17). Any such metric can be fully reduced to $l - 1$ constraints, which is the minimal number required if d is not allowed to assume $+\infty$. This makes the class of tree metrics very appealing, in particular in the light of the comments in Sect. 4.6.

The most straightforward non-tree metric is the *circular metric*, $d(i, j) = \min\{|i - j|, l - |i - j|\}$, which describes the geodesic distance on a circle (Fig. 18). This metric was proposed in [63] as a regularizer for angular data such as orientation. It is interesting to see that the required constraints are exactly the same as for the linear metric, with one additional constraint “connecting” labels 1 and l . This directly relates to the simplified representation in [63], but the

experiments shows that it can be derived in a fully automatic way, reducing the number of constraints from $l(l-1)/2$ to l .

In applications where there may be sharp contrasts between objects, such as depth from stereo, the *truncated linear* metric, $d(i, j) = \min\{|i - j|, c\}$, is often useful, since it linearly penalizes small variations but does not overly penalize large jumps (Fig. 16). As a hybrid between the linear ($c = l$) and Potts metric ($c = 0$), the reduction lies in between, depending on the choice of c .

In order to evaluate the method on unstructured graphs, we generated a graph by randomly assigning a weight between 1 and 5 to 60% of the edges and setting all other edge weights to $+\infty$ (Fig. 20). The method correctly recovers a subset of the original graph and reduces the number of constraints from 66 to 26.

A very interesting application is the reduction of metrics arising from the quantization of *vector-valued* data, such as optical flow vectors. We denote by n the number of quantization steps *per dimension*. For two-dimensional data, this results in $l = n^2$ and n^4 constraints, rendering the approach impractical already for very small n .

In such cases the regularizer often is separable, i.e., $d((i_1, i_2), (j_1, j_2)) = d_1(i_1, j_1) + d_2(i_2, j_2)$. For the linear metric $d_1(i, j) = d_2(i, j) = |i - j|$ this results in a reduction from $n^2(n^2 - 1)/2$ to $2n(n - 1)$, i.e., from $O(n^4)$ to $O(n^2)$ (Fig. 21).

But even for the case of the sum of two uniform metrics, $d_1(i, j) = d_2(i, j) = 1_{\{i \neq j\}}$, each of one is not reducible as seen previously in Fig. 15, a certain reduction can be obtained (Fig. 22): the number of constraints reduces from $n^2(n^2 - 1)/2$ to $2n^2(n - 1)/2$, i.e., from $O(n^4)$ to $O(n^3)$.

6 Integral or Fractional Models?

The experiments in the previous sections, in particular Sect. 3 lead to the conclusion that in order to obtain the visually best results, it is better to minimize over the set of relaxed labelings using finite differences, and threshold if necessary, than to solve a combinatorial problem directly. In this section we will discuss several aspects of this conclusion.

6.1 Point- and Region-Based Interpretation

A fundamental decision implied when using combinatorial methods such as graph cuts is that the result of a labeling method should consist of a vector of integral labels. This is a consistent choice at first glance, but it immediately brings forward the question of the semantics of such a solution, i.e. how the discretization reflects properties of the optimal *spatially continuous* labeling u^* . When deriving graph-based pairwise energies, a strong focus lies on the correspondence between label variables and *points* in the image domain: the label $\ell^h(x^i)$ denotes *in which of the class regions $\Omega_1, \dots, \Omega_l$ the point x^i is contained*. This is clearly a combinatorial decision, and does not allow any intermediate values.

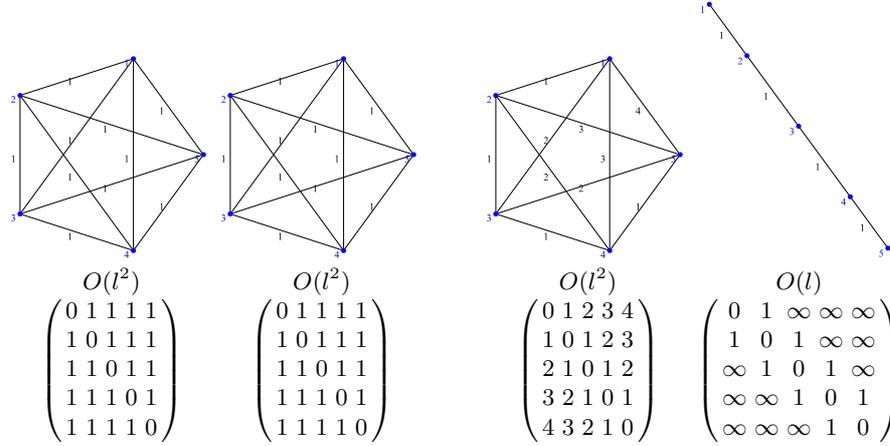


Fig. 15. Metric reduction for the uniform metric, $d(i, j) = 1_{\{i \neq j\}}$. In this case no reduction is possible, as the metric does not contain any linear structure.

Fig. 16. Metric reduction for the linear metric, $d(i, j) = |i - j|$. The linear structure of the metric can be fully exploited, yielding a reduction from $O(l^2)$ to $O(l)$ constraints.

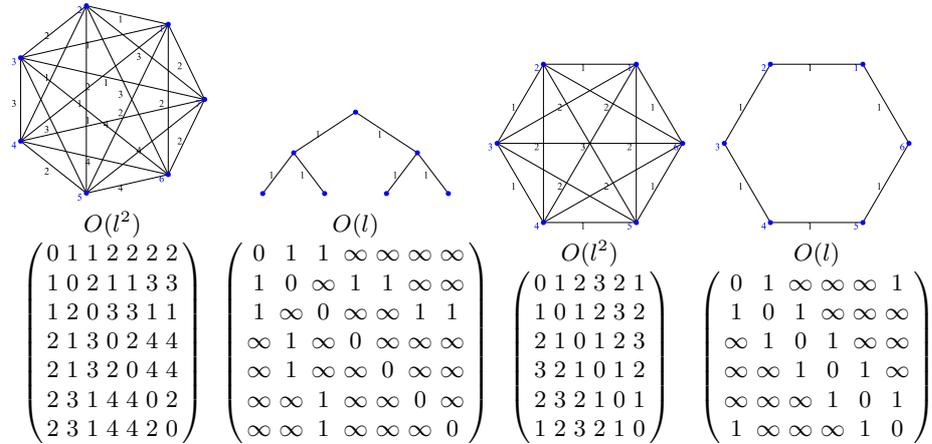


Fig. 17. Shortest-path metrics on trees are special in that they can be fully reduced to the minimal number of $(l - 1)$ constraints, a reduction from $O(l^2)$ to $O(l)$.

Fig. 18. The circular metric, $d(i, j) = \min\{|i - j|, l - |i - j|\}$. Such metrics are useful for regularizing angular/directional data, and also allow a full reduction from $l(l - 1)/2$ to l constraints. The constraints are similar to the ones proposed in [63], but can be recovered in a fully automatic way.

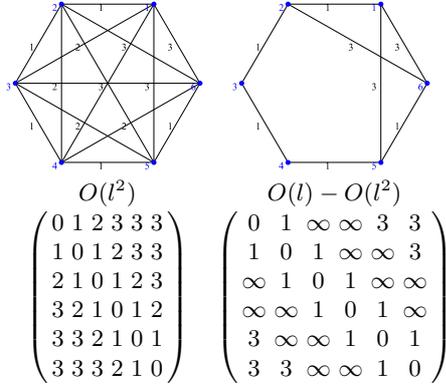


Fig. 19. Reduction for the truncated linear metric $d(i, j) = \min\{|i - j|, c\}$ with $c = 3$. Depending on the parameter c , the number of constraints varies from $O(l)$ ($c = l$, linear metric) to $O(l^2)$ ($c = 0$, uniform metric).

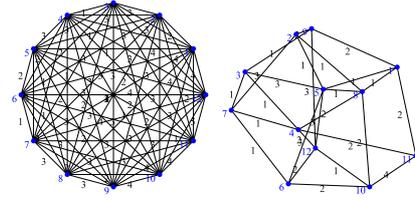


Fig. 20. Metric reduction of a shortest-path metric on an unstructured random graph. Since no particular structure is assumed a priori, the reduction method can be applied to any metric in a black-box fashion, and correctly identifies a subset of the original graph in order to form the constraint set.

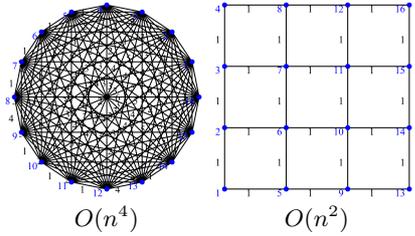


Fig. 21. Metric reduction on problems with vector-valued quantities (cf. Fig. 14), where each direction is quantized using n steps. The separable structure of the metric $d((i_1, i_2), (j_1, j_2)) = |i_1 - j_1| + |i_2 - j_2|$ is automatically fully exploited and permits a perfect reduction from $O(n^4)$ to $O(n^2)$ constraints.

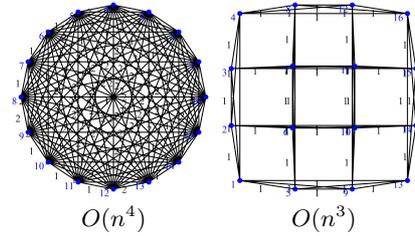


Fig. 22. In contrast to the pure uniform metric (Fig. 15), for the separable uniform regularizer $d((i_1, i_2), (j_1, j_2)) = 1_{\{i_1 \neq j_1\}} + 1_{\{i_2 \neq j_2\}}$, a reduction from $O(n^4)$ to $O(n^3)$ constraints is possible and automatically identified.

The derivations for the edge weights etc. all depend on this assumption: cutting an edge is semantically equivalent to separating two *points*.

Such a hard pointwise decision is fully justified when dealing with e.g. network problems, where the nodes of the graph correspond to finite entities in the real world, such as factory locations in production planning problems. However, in a sense it neglects the origin of imaging data, which usually comes from cameras or sensors that average the continuous input over a number of *pixel areas*, i.e. regions with nonzero area in the image plane. In contrast to points, and even in perfect camera models, pixel values always accumulate some statistics from their respective rectangular region. The same holds for higher-dimensional data such as voxels describing a section of the real world.

If one therefore associates each label $\ell^h(x^{\bar{i}})$ with a *pixel*, the interpretation changes in the sense that each label now represents *all* labels in the rectangular region associated with the pixel containing $x^{\bar{i}}$. In Thm. 1 this is quantified by assuming that the piecewise constant function \tilde{u}^h associated with some u^h should approximate the optimal *continuous* function u in the L^1 sense, similarly for \tilde{s}^h and s^h .

Using this interpretation, the decision to only allow integral labels becomes questionable. In fact, enforcing integral values then corresponds to approximating the true function u using integer-valued functions that are piecewise constant on the region associated to each pixel. However, such integral approximations can only have axis-parallel edges. If the energy respects this structure, as is the case for the 4-neighborhood LP relaxation scheme, the corresponding artifacts occur.

Therefore, for the region-(pixel-)based interpretation, a more adequate choice is to allow fractional values for u^h in order to better approximate the true continuous u using the piecewise constant \tilde{u}^h . This interpretation is very natural when dealing with images: Assume for a moment that an optimal two-class labeling $u^* : \Omega \rightarrow \{0, 1\}$ of some real-world image is known, and that we are given the task of finding a good approximation u^h of the labeling on a grid (we represent e^1 and e^2 with the scalar values $\{0, 1\}$ as in the two-class continuous cut). Possibly the most natural approach, and what is intuitively expected by humans, is to simulate the effect of a camera, i.e. to formally paint the scene in black and white according to the true labeling u^* and to *average* the values within the region for each pixel. This inevitably leads to fractional values if the pixel region is intersected by an interface.

The central message is that such fractional values are *not* introduced by a relaxation process, but by honoring the fact that each of the values $u^h(x^{\bar{i}})$ correspond to a whole *region* of labels. In other words, the fractional values occur only as an effect of approximating the true *continuous* solution on a *finite* grid.

Therefore we argue that the region-based interpretation should be preferred. In fact, an integral labeling is rarely ever actually required by subsequent image processing steps in the sense that they cannot be reformulated to account for the

region interpretation. Often, a certain smoothness at the boundaries is actually desirable, as in the case of segmentation for image manipulation.

6.2 Obtaining Integral Solutions

Nevertheless, let us assume that the user has a valid reason for restricting the solution to hard labels, and consider what would be an optimal segmentation $\bar{u}^h : \mathbb{R}^n \rightarrow \{0, 1\}$ if one had *perfect knowledge* about the optimal true segmentation u^* . Since with perfect knowledge there is no reason to infer any structure that is not contained in the known segmentation, again the most reasonable choice is to find an (integral) \bar{u}^h that best approximates the continuous segmentation,

$$\bar{u}^h = \arg \min_{u^h : \Omega^h \rightarrow \{0,1\}} \|u^* - \tilde{u}^h\|_{L^1}, \quad (119)$$

or a similar formulation with a different norm. For the usual L^1 -distance, this simply corresponds to setting $\bar{u}^h(x^{\bar{i}}) = 1$ if more than half of the labels corresponding to the \bar{i} -th pixel have label 1. Note that this is a purely local decision, since it only depends on labels for points inside the region associated with the pixel containing the point $x^{\bar{i}}$.

In reality, u^* cannot be represented in finite memory and is therefore only available as a finite-dimensional fractional approximation $u^{h,*}$ obtained by minimizing a functional f^h . The best guess to find \bar{u}^h is then to approximate $u^{h,*}$, i.e.

$$\bar{u}^{h,*} = \arg \min_{u^h : \Omega^h \rightarrow \{0,1\}} \|u^{h,*} - \tilde{u}^h\|_{L^1}, \quad (120)$$

$$u^{h,*} = \arg \min_{u^h : \Omega^h \rightarrow [0,1]} f^h(u^h). \quad (121)$$

This amounts exactly to *rounding* the fractional values of $u^{h,*}$ to integral values, and again is a purely local operation. The rounding is a direct consequence of the region-based interpretation when combined with the requirement for integral solutions.

A key point is that *this is not the same as minimizing f^h over the set of integral u^h* , i.e. solving the combinatorial problem

$$\bar{u}^h = \arg \min_{u^h : \Omega^h \rightarrow \{0,1\}} f^h(u^h). \quad (122)$$

For an illustration, see Fig. 23. This provides an explanation for the results observed in the experimental section:

1. It does not make sense to compute global *integral* minimizers of energies that are formulated with a region-based interpretation in mind. The proper way to generate integral approximations to the best continuous segmentation is to first compute the best *fractional* minimizer and then (locally) round.

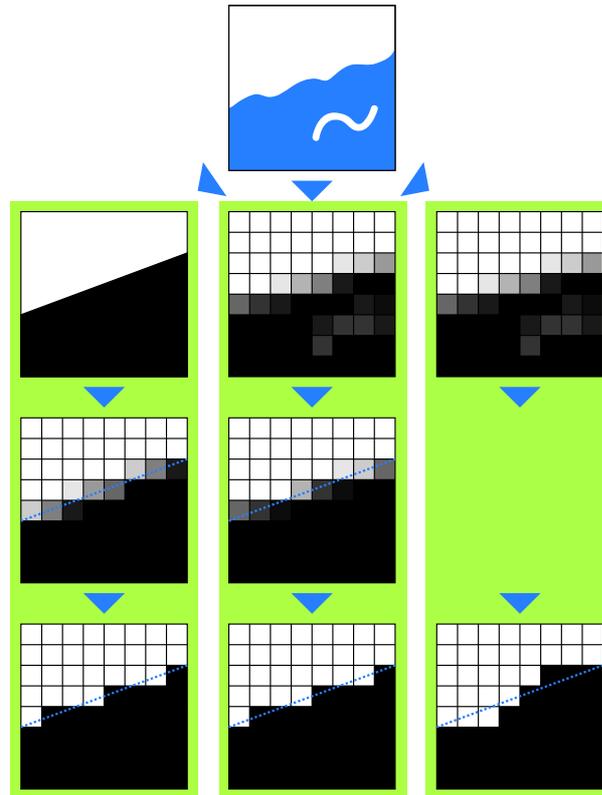


Fig. 23. Approximation of the optimal segmentation (left) by solving the discretized relaxed problem and rounding (center) vs. minimizing the same energy as a combinatorial problem. The relaxed approach respects the origin of the data from a continuous world, and returns an approximation to the *best fractional approximation* of the continuous segmentation in terms of a fractional solution, from which an integral approximation can be recovered. By allowing fractional values the continuous functional can be approximated fairly well. In contrast, for reasonably large neighborhoods, combinatorial approaches correspond to a crude approximation of the true functional, which introduces undesired minima (right).

2. Analogously, it is not reasonable to compare the energy of a rounded fractional solution to the energy of some other solution obtained via a combinatorial optimization method of the same energy. In particular, the rounding step must *not* be seen a way to approximate integral minimizers.

If one persists on using a combinatorial optimization method, the proper way would be to formulate a combinatorial energy f_c^h whose *integral* minimizers approximate u^* as well as possible, and solve

$$\bar{u}_c^{h,*} = \arg \min_{u^h: \Omega^h \rightarrow \{0,1\}} f_c^h(u^h). \quad (123)$$

The graph cut approach can be seen as an implementation of this idea (123), while the relaxed approach considered in this work conforms to (121). This constitutes a case where, if one actually requires integral labelings, it is much easier to construct *relaxed* energies f^h such that the rounded minimizer \bar{u}^h from (121) is a good approximation of u^* , than it is to formulate a *combinatorial* energy f_c^h such that the same holds for its integral minimizer $\bar{u}_c^{h,*}$.

We attribute this to the fact that f_c^h has much less degrees of freedom; in fact there is only a finite number of choices for f_c^h . In contrast, f^h contains much more information, since it describes a function defined on a continuum of values. In a sense, adding a single term that involves the Euclidean norm $\|\cdot\|_2$ to f^h conveys as much information as adding an infinite number of pairwise terms to f_c^h , cf. (52).

With this in mind, (121) can be seen as a convenient way of formulating combinatorial optimization problems involving an otherwise too complicated combinatorial objective f_c^h , that facilitates a very compact representation by introducing an intermediate *relaxed* problem.

6.3 Multi-Class Case

Note that the rounding process is not connected in any way to uncertainties related to the formulation of the minimization problem, but rather is the exact step to find the best integral approximation to a fractional image. Allowing intermediate might seem to be related to the process of switching from MAP to marginal estimation in graphical models, however this connection is deceiving: In the latter, marginal probabilities correspond to the uncertainty of choosing *one specific* label. In contrast, in our framework the intermediate values are required to accommodate for the *infinitely many* points within the pixel that should receive a label. Pixel labels are not the same as point labels, but in a sense statistics about the labels of *all* points associated with the pixel.

However, this view slightly changes for the multi-class problem where u may already assume fractional values due to the relaxation step. Here fractional values can additionally be caused by the relaxation, and it is important to develop relaxations that are as tight as possible. However, from the considerations above we see that even if the relaxations is perfectly tight, as in the two-class case, fractional labelings are still useful in order to better approximate the spatially

continuous solution, and provide the necessary freedom to precisely discretize the original objective.

Appendix

Proof (Prop. 3). Let $v \in \mathcal{D}_{\text{loc}}^m$ and $i \neq j$. Denote a shortest path in \mathcal{G}^m from i to j by $(k_1 = i, k_2, \dots, k_{n-1}, k_n = j)$ for some $n \geq 2$. Then, due to the triangle inequality of the norm,

$$\|v^i - v^j\|_2 \leq \|v^{k_1} - v^{k_2}\|_2 + \dots + \|v^{k_{n-1}} - v^{k_n}\|_2. \quad (124)$$

Since $v \in \mathcal{D}_{\text{loc}}^m$, the terms on the right can be bounded,

$$\|v^i - v^j\|_2 \leq m(k_1, k_2) + \dots + m(k_{n-1}, k_n). \quad (125)$$

By definition, the sum equals $\bar{m}(i, j)$ and therefore by assumption $\bar{m}'(i, j)$, thus

$$\|v^i - v^j\|_2 \leq \bar{m}(i, j) = \bar{m}'(i, j) \leq m'(i, j). \quad (126)$$

Consequently $v \in \mathcal{D}_{\text{loc}}^{m'}$, and, since v was arbitrary, $\mathcal{D}_{\text{loc}}^m \subseteq \mathcal{D}_{\text{loc}}^{m'}$. By symmetry the reverse inclusion holds as well, and therefore equality. \square

Proof (Prop. 4). Since d is a metric and thus satisfies the triangle inequality, (114) is equivalent to

$$d(i, k) + d(k, j) = d(i, j). \quad (127)$$

Since d is the shortest-path metric of m , eq. (127) states that there exists a shortest path from i to j in \mathcal{G}^m passing through k . This path consists of at least two edges which both have positive weight by the definition of \mathcal{M} . Therefore the path cannot contain the edge (i, j) , since $m(i, j) \geq d(i, j)$ which would contradict (127).

This means that the path is also contained in $\mathcal{G}^{m'}$. By construction of m' it must then also be a shortest path from i to j in m' , which shows that $\bar{m}'(i, j) = \bar{m}(i, j)$. Since $m'(i', j') = m(i', j')$ for all $(i', j') \neq (i, j)$, this immediately implies $m' = \bar{m}$ and therefore the assertion. \square

Proof (Prop. 5). Define

$$m(i, j) := \begin{cases} +\infty, & \exists k \neq i, j : d(i, j) \geq d(i, k) + d(k, j) \\ d(i, j), & \text{otherwise.} \end{cases}$$

An iterative application of Prop. 4 shows that $\bar{m} = d$. Prop. 3 then states that their associated dual constraint sets coincide,

$$\mathcal{D}_{\text{loc}}^m = \mathcal{D}_{\text{loc}}^d. \quad (128)$$

Since $\bigcap_{(i, j) \in \mathcal{R}} \mathcal{D}_{ij} = \mathcal{D}_{\text{loc}}^m$ by construction, this concludes the proof. \square

Acknowledgments. The second and third author were supported by Engineering and Physical Sciences Research Council (EPSRC)-Project EP/H016317/1. This publication is partly based on work supported by Award No. KUK-I1-007-43, made by King Abdullah University of Science and Technology (KAUST).

References

1. Ambrosio, L., Fusco, N., Pallara, D.: *Functions of Bounded Variation and Free Discontinuity Problems*. Clarendon Press (2000)
2. Appleton, B., Talbot, H.: Globally minimal surfaces by continuous maximal flows. *Patt. Anal. Mach. Intell.* **28**, 106–118 (2006)
3. Bae, E., Yuan, J., Tai, X.C.: Global minimization for continuous multiphase partitioning problems using a dual approach. *Int. J. Comp. Vis.* **92**, 112–129 (2011)
4. Bartal, Y.: On approximating arbitrary metrics by tree metrics. In: *ACM Symposium on Theory of Computing* (1998)
5. Bertsekas, D.P.: *Network Optimization: Continuous and Discrete Models*. Athena Scientific (1998)
6. Blake, A., Zisserman, A.: *Visual Reconstruction*. MIT Press (1987)
7. Boykov, Y.: Computing geodesics and minimal surfaces via graph cuts. In: *Int. Conf. Comp. Vis.*, pp. 26–33 (2003)
8. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *Patt. Anal. Mach. Intell.* **23**(11), 1222–1239 (2001)
9. Braides, A.: *Gamma-convergence for Beginners*. Oxford Univ. Press (2002)
10. Călinescu, G., Karloff, H., Rabani, Y.: An improved approximation algorithm for multiway cut. *J. Comp. System Sci.* **60**, 564–574 (2000)
11. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. *Int. J. Comp. Vis.* **22**, 61–79 (1997)
12. Chambolle, A.: Finite-differences discretizations of the Mumford-Shah functional. *Model. Math. et Anal. Num.* **33**, 261–288 (1999)
13. Chambolle, A., Cremers, D., Pock, T.: A convex approach to minimal partitions. *J. Imaging Sci.* **5**(4), 1113–1158 (2012)
14. Chambolle, A., Darbon, J.: On total variation minimization and surface evolution using parametric maximum flows. *Int. J. Comp. Vis.* **84**, 288–307 (2009)
15. Chambolle, A., Levine, S.E., Lucier, B.J.: An upwind finite-difference method for Total Variation-based image smoothing. *J. Imaging Sci.* **4**(1), 277–299 (2011)
16. Chan, T.F., Esedoğlu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. *J. Appl. Math.* **66**(5), 1632–1648 (2006)
17. Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Trans. Image Proc.* **10**(2), 266–277 (2001)
18. Charikar, M., Chekuri, C., Goel, A., Guha, S., Plotkin, S.: Approximating a finite metric by a small number of tree metrics. In: *ACM Found. Comp. Sci.*, pp. 379–388 (1998)
19. Combettes, P.L., Pesquet, J.C.: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chap. Proximal splitting methods in signal processing. Springer New York (2010)
20. Couprie, C., Grady, L., Najman, L., Talbot, H.: Power watersheds: A unifying graph-based optimization framework. *Patt. Anal. Mach. Intell.* **33**(7), 1384–1399 (2011)
21. Couprie, C., Grady, L., Talbot, H., Najman, L.: Combinatorial continuous maximum flow. *J. Imaging Sci.* **4**(3), 905–930 (2011)
22. Dahlhaus, E., Johnson, D.S., Papadimitriou, C.H., Seymour, P.D.: The complexity of multiterminal cuts. *J. Computing* **23**(4), 864–894 (1994)
23. Dal Maso, G.: *An Introduction to Γ -Convergence*. Birkhäuser (1993)

24. De Giorgi, E., Franzoni, T.: Su un tipo di convergenza variazionale. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.* **68**, 842–850 (1975)
25. Dixit, N., Keriven, R., Paragios, N.: GPU-cuts: Combinatorial optimisation, graphic processing units and adaptive object extraction. Tech. Rep. 05-07, CERTIS, ENPC (2005)
26. Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *J. Imaging Sci.* **3**, 1015–1046 (2010)
27. Fakcharoenphol, J., Rao, S., Talwar, K.: A tight bound on approximating arbitrary metrics by tree metrics. *J. Comp. System Sci.* **69**(3), 485–497 (2004)
28. Faugeras, O., Luong, Q.T.: *The Geometry of Multiple Images*. MIT Press (2001)
29. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.* **6**, 721–741 (1984)
30. Gobbino, M., Mora, M.G.: Finite-difference approximation of free-discontinuity problems. In: *Proc. Royal Soc. Edinburgh*, vol. 131, pp. 567–595 (2001)
31. Goldluecke, S., Cremers, D.: Convex relaxation for multilabel problems with product label spaces. In: *Europ. Conf. Comp. Vis.*, pp. 225–238 (2010)
32. Goldschlager, L.M., Shaw, R.A., Staples, J.: The maximum flow problem is log space complete for P. *Theor. Comp. Sci.* **21**, 105–111 (1982)
33. Grady, L.J., Polimeni, J.R.: *Discrete Calculus: Applied Analysis on Graphs for Computational Science*. Springer (2010)
34. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press (2000)
35. Ising, E.: Beitrag zur Theorie des Ferromagnetismus. *Zeitschr. f. Physik* **31**(1), 253–258 (1925)
36. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contours models. *Int. J. Comp. Vis.* **1**(4), 321–331 (1987)
37. Kleinberg, J.M., Tardos, E.: Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. In: *Found. Comp. Sci.*, pp. 14–23 (1999)
38. Klodt, M., Schoenemann, T., Kolev, K., Schikora, M., Cremers, D.: An experimental comparison of discrete and continuous shape optimization methods. In: *Europ. Conf. Comp. Vis.*, pp. 332–345. Marseille, France (2008)
39. Koller, D., Friedman, N.: *Probabilistic Graphical Models: Principles and Techniques*. MIT Press (2009)
40. Kolmogorov, V., Boykov, Y.: What metrics can be approximated by geo-cuts, or global optimization of length/area and flux. *Int. Conf. Comp. Vis.* **1**, 564–571 (2005)
41. Komodakis, N., Tziritas, G.: Approximate labeling via graph cuts based on linear programming. *Patt. Anal. Mach. Intell.* **29**(8), 1436–1453 (2007)
42. Lauritzen, S.: *Graphical Models*. Oxford Univ. Press (1996)
43. Lellmann, J.: *Nonsmooth convex variational approaches to image analysis*. Ph.D. thesis, University of Heidelberg (2011)
44. Lellmann, J., Breitenreicher, D., Schnörr, C.: Fast and exact primal-dual iterations for variational problems in computer vision. In: *Europ. Conf. Comp. Vis.*, *LNCS*, vol. 6312, pp. 494–505 (2010)
45. Lellmann, J., Kappes, J., Yuan, J., Becker, F., Schnörr, C.: Convex multi-class image labeling by simplex-constrained total variation. In: *Scale Space and Var. Meth.*, *LNCS*, vol. 5567, pp. 150–162 (2009)

46. Lellmann, J., Lenzen, F., Schnörr, C.: Optimality bounds for a variational relaxation of the image partitioning problem. In: *Energy Min. Meth. Comp. Vis. Patt. Recogn.* (2011)
47. Lellmann, J., Lenzen, F., Schnörr, C.: Optimality bounds for a variational relaxation of the image partitioning problem. *J. Math. Imaging Vis.* p. In press. (2012)
48. Lellmann, J., Schnörr, C.: Continuous multiclass labeling approaches and algorithms. *J. Imaging Sci.* **4**(4), 1049–1096 (2011)
49. Lempitsky, V., Rother, C., Blake, A.: Logcut – efficient graph cut optimization for Markov random fields. In: *Europ. Conf. Comp. Vis.*, pp. 1–8 (2007)
50. Lempitsky, V., Rother, C., Roth, S., Blake, A.: Fusion moves for Markov random field optimization. *Patt. Anal. Mach. Intell.* **32**(8), 1392–1405 (2010)
51. Lie, J., Lysaker, M., Tai, X.C.: A variant of the level set method and applications to image segmentation. *Math. Comp.* **75**, 1155–1174 (2006)
52. Morel, J.M., Solimini, S.: *Variational methods in image segmentation.* Birkhäuser (1995)
53. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.* **42**, 577–685 (1989)
54. Olsson, C.: *Global optimization in computer vision: Convexity, cuts and approximation algorithms.* Ph.D. thesis, Lund Univ. (2009)
55. Olsson, C., Byröd, M., Overgaard, N.C., Kahl, F.: Extending continuous cuts: Anisotropic metrics and expansion moves. In: *Int. Conf. Comp. Vis.* (2009)
56. Paragios, N., Chen, Y., Faugeras, O. (eds.): *The Handbook of Mathematical Models in Computer Vision.* Springer (2006)
57. Pock, T., Chambolle, A., Cremers, D., Bischof, H.: A convex relaxation approach for computing minimal partitions. In: *Comp. Vis. Patt. Recogn.*, pp. 810–817 (2009)
58. Pock, T., Schönemann, T., Graber, G., Bischof, H., Cremers, D.: A convex formulation of continuous multi-label problems. In: *Europ. Conf. Comp. Vis.*, vol. 3, pp. 792–805 (2008)
59. Rickett, J., Fomel, S.: *A second-order fast marching Eikonal solver.* Tech. Rep. 100, Stanford Expl. Project Report (1999)
60. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: *Variational Methods in Imaging, Applied Mathematical Sciences*, vol. 167. Springer (2009)
61. Sinop, A.K., Grady, L.: A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm. In: *Int. Conf. Comp. Vis.*, pp. 1–8 (2007)
62. Strang, G.: Maximal flow through a domain. *Math. Prog.* **26**, 123–143 (1983)
63. Strekalovskiy, E., Cremers, D.: Total variation for cyclic structures: Convex relaxation and efficient minimization. In: *Conf. Comp. Vis. Patt. Recogn.*, pp. 1905–1911 (2011)
64. Strekalovskiy, E., Goldluecke, B., Cremers, D.: Tight convex relaxations for vector-valued labeling problems. In: *Int. Conf. Comp. Vis.*, pp. 2328–2335 (2011)
65. Trobin, W., Pock, T., Cremers, D., Bischof, H.: Continuous energy minimization by repeated binary fusion. In: *Europ. Conf. Comp. Vis.*, vol. 4, pp. 667–690 (2008)
66. Zach, C., Gallup, D., Frahm, J.M., Niethammer, M.: Fast global labeling for real-time stereo using multiple plane sweeps. In: *Vis. Mod. Vis* (2008)