Inaugural-Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht-Karls-Universität Heidelberg

vorgelegt von
Dipl.-Math. Dipl.-Inform. Jan Lellmann
aus Haan

Tag der mündlichen Prüfung: 12. Juli 2011

Revidierte Fassung vom 14. September 2011

Nonsmooth Convex Variational Approaches to Image Analysis

Gutachter: **Prof. Dr. Christoph Schnörr**
Prof. Dr. Dr. h.c. Hans Georg Bock

Zusammenfassung

Variationsmethoden bilden in vielen Gebieten der Bildverarbeitung die Grundlage für die Formulierung von Modellen sowie für deren tieferes Verständnis. In dieser Arbeit betrachten wir einen Variationsansatz für konvexe Relaxierungen des Mehrklassen-Segmentierungsproblems, formuliert auf kontinuierlichen Bildgebieten. Wir stellen mehrere zugehörige Relaxierungen für längenbasierte Regularisierer vor, die sich in der Mächtigkeit, aber auch in der numerischen Komplexität, unterscheiden. Durch die Formulierung im Rahmen der geometrischen Maßtheorie werden Diskretisierungsartefakte, die bei graphenbasierten kombinatorischen Verfahren aufgrund der frühzeitigen Diskretisierung auftreten, so weit wie möglich vermieden. Zur numerischen Lösung des zugehörigen nichtglatten Optimierungsproblems untersuchen wir Optimierungsmethoden erster Ordnung, basierend auf kontrollierter Glättung und Operator Splitting. Wir formulieren eine randomisierte Rundungsmethode für Mehrklassen-Segmentierungsprobleme auf kontinuierlichen Gebieten und zeigen, dass auf diese Weise ganzzahlige Lösungen mit einer *a priori*-Schranke für die Optimalität gefunden werden können. Weiterhin stellen wir eine "Sparse Representation"-basierte Methode vor, die es erlaubt, zusätzliches Vorwissen über die Objektform in Variationsansätze zu integrieren.

Abstract

Variational models constitute a foundation for the formulation and understanding of models in many areas of image processing and analysis. In this work, we consider a generic variational framework for convex relaxations of multiclass labeling problems, formulated on continuous domains. We propose several relaxations for length-based regularizers, with varying expressiveness and computational cost. In contrast to graph-based, combinatorial approaches, we rely on a geometric measure theory-based formulation, which avoids artifacts caused by an early discretization in theory as well as in practice. We investigate and compare numerical first-order approaches for solving the associated nonsmooth discretized problem, based on controlled smoothing and operator splitting. In order to obtain integral solutions, we propose a randomized rounding technique formulated in the spatially continuous setting, and prove that it allows to obtain solutions with an *a priori* optimality bound. Furthermore, we present a method for introducing more advanced prior shape knowledge into labeling problems, based on the sparse representation framework.

Acknowledgments

There are many people that have directly and indirectly contributed to this thesis. First and foremost, I'd like to thank my advisor, Prof. Christoph Schnörr, who introduced me to this fascinating subject. He not only provided continuous encouragement, but also deeply impressed me by his openness and genuine interest in promoting students and introducing them into the scientific community.

I am also very thankful to my present and former colleagues at the Heidelberg Collaboratory for Image Processing, many of whom have become personal friends. Special thanks go to Dirk Breitenreicher, whose open-mindedness and straightforward working style I greatly appreciate.

I also like to mention Florian Becker, Jörg Kappes, Frank Lenzen, Bogdan Savchynskyy, Stefan Schmidt, Fabian Rathke, and Bernhard Schmitzer, many of whom contributed to this thesis through helpful comments or proof-reading, and who made working at the Image and Pattern Analysis group such an enjoyable experience. I also warmly thank Evelyn Wilhelm, Barbara Werner, and Karin Kruljac for invaluable help in slaying the bureaucratic dragons and always being there with an open ear for our everyday problems.

During the time at the Image and Pattern Analysis group, I had the privilege of being able to meet many interesting people at conferences and workshops all over the world. I would like to particularly thank Prof. Gabi Steidl, Simon Setzer, and Tanja Teuber for many fruitful and open discussions.

Finally, I am eternally grateful to my family, and especially my parents, who unconditionally supported me at all times. In difficult times you always encouraged me and set me back on track. But most importantly, you taught me to always be curious and never to give up, no matter how intimidating a problem may seem at first, but to keep struggling until at some point you discover that all obstacles lie behind you. For this and everything else I cannot ever thank you enough.

Table of contents

List of Symbols xiii

List of Publications xvii

1 Introduction and Overview 1

1.1 Nonsmooth Variational Models 1
1.2 Variational Multiclass Labeling 4
1.3 Contribution 10
1.4 Notation 11

2 A Variational Convex Formulation for Multi-Class Labeling 13

2.1 Introduction and Overview 13
2.2 Related Work 17
2.3 Properties of the Interaction Potential 18
2.4 Existence of Minimizers 20
2.5 Regularizers for Specific Interaction Potentials 21
2.5.1 Relaxation Based on the Local Envelope 25
2.5.2 Relaxation Based on Embeddings 27
2.5.3 Relaxation with Emphasized Uncertainty 31
2.6 Optimality 32
2.6.1 The Two-Class Case 32
2.6.2 Generalized Coarea Formulas 34
2.6.3 Higher Codimensions 34
2.6.4 Separable Regularizer 35
2.7 Relation to other Approaches and Extensions 36
2.7.1 Isotropic Regularizers 36
2.7.2 Anisotropic and Inhomogeneous Regularizers 37
2.7.3 Linearly Ordered Label Set 39
2.7.4 Other Extensions 43
2.8 Summary and Further Work 43

3 Discretization of Functionals with Length-Based Regularizers 45

3.1 Introduction and Overview 45
3.2 Related Work 47
3.3 Convergent Finite-Differences Approximation of the Relaxed Problem 53
3.4 Experimental Comparison 62
3.5 Discussion: Integral and Fractional Models 67
3.6 Summary and Further Work 72

4 Nonsmooth Optimization . . . 73

4.1 Introduction and Overview . . . 73
4.2 Related Work . . . 76
4.3 First-Order Schemes for Multiclass Labeling . . . 79
4.3.1 Nesterov Approach . . . 80
4.3.2 Douglas-Rachford Splitting . . . 84
4.3.2.1 Primal Constraint Splitting . . . 86
4.3.2.2 Auxiliary Variables . . . 87
4.3.2.3 Multiple-Constraint Dual Variables . . . 90
4.4 Implementation Details . . . 92
4.5 Experimental Comparison . . . 95
4.5.1 General Observations . . . 95
4.5.2 Varying Problem Size and Parameters . . . 97
4.5.3 Breaking Points . . . 99
4.5.4 Choice of the Relaxation . . . 102
4.5.5 Dual Multiple-Constraint Douglas-Rachford . . . 105
4.6 Summary and Further Work . . . 108

5 Optimality and Rounding . . . 111

5.1 Introduction and Overview . . . 111
5.2 Related Work . . . 113
5.2.1 Isolation Heuristic and α-Expansion . . . 113
5.2.2 Continuous Binary Fusion . . . 115
5.2.3 LP Relaxation with Derandomization . . . 116
5.3 Improved Deterministic Schemes . . . 117
5.3.1 First-Max . . . 117
5.3.2 Modified First-Max . . . 118
5.4 Coarea Formula and Probabilistic Rounding . . . 118
5.5 A Priori Bounds . . . 120
5.5.1 Probabilistic Rounding for Multiclass Image Partitions . . . 120
5.5.2 Termination Properties . . . 121
5.5.3 Intermediate Results . . . 125
5.5.4 A Probabilistic A Priori Optimality Bound . . . 130
5.6 Experimental Comparison . . . 136
5.7 Summary and Further Work . . . 140

6 Sparse Representation of Shape . . . 143

6.1 Introduction and Overview . . . 143
6.2 Related Work . . . 145
6.3 A Convex Model for Sparse Shape . . . 147
6.4 Optimization . . . 148
6.5 Experimental Evaluation . . . 149
6.6 Summary and Further Work . . . 152

7 Conclusion . . . 155

Appendix A Mathematical Preliminaries . . . 157

A.1 Functions of Bounded Variation 157
A.1.1 Total Variation and BV 157
A.1.2 Properties of TV and Compactness 159
A.1.3 Decomposition and general functionals on BV 160
A.1.4 The Coarea Formula 164
A.2 Γ-Convergence 167
A.3 Set-Valued Operators and Proximal Steps 168

Bibliography 173

List of Symbols

Basic Conventions

$\mathbb{N}$, $\mathbb{N}_0$	natural numbers $\{1, 2, ...\}$, natural numbers including 0
$\bar{\mathbb{R}}$	extended real line, $\mathbb{R} \cup \{\pm\infty\}$
I	identity operator/matrix
e	all-one vector $(1, ..., 1)$
e^i	i-th unit vector
$\langle x, y \rangle$	standard inner product in $\mathbb{R}^n$, L^2
$A \otimes B$	Kronecker product of matrices A, B
$\mathcal{B}_h(x)$	ℓ^2 ball with radius h, centered in x
S^{d-1}	unit sphere in $\mathbb{R}^d$
2^A	set of subsets of a set A
$\operatorname{vecmax}(A)$	row-wise maximum of matrix or vector A
$A \odot B$	elementwise (Hadamard) product
$\mathrm{SO}(d)$	special orthogonal group, rotations on $\mathbb{R}^d$

Convex Analysis and Operator Splitting

χ	characteristic function, $\chi(x) \in \{0, 1\}$
δ	indicator function, $\delta(x) \in \{0, +\infty\}$
$T\colon X \rightrightarrows Y$, T^{-1}	set-valued mapping $T\colon X \to 2^Y$, inverse
$\operatorname{dom} T$, $\operatorname{range} T$	domain, range of a set-valued mapping T
$J_{\lambda T}$	resolvent of an operator $T\colon X \rightrightarrows X$, $J_{\lambda T} = (I + \lambda T)^{-1}$
$P_{\tau f}(x)$	proximal step of f in x with step size τ
$\operatorname{dom} f$	domain of a function f
∂f, $\partial f(x)$	subdifferential mapping of f, set of subgradients of f in x
f^*, $(f^*)^*$	Legendre-Fenchel conjugate, biconjugate of a function f
$\operatorname{ri} C$	relative interior of a set C
$\Pi_{\mathcal{C}}$	projection onto $\mathcal{C}$
N_C	normal cone of a set C
σ_C	support function for a set C
$\operatorname{conv} A$	convex hull of a set A
$\operatorname{hyp} u$	hypograph of u
Ψ_∞	recession function of Ψ

Measures and Functions of Bounded Variation

$\mathcal{L}^d$	d-dimensional Lebesgue measure
$\mathcal{H}^{d-1}$	$(d-1)$-dimensional Hausdorff measure
C_c^∞	smooth functions with compact support
C_c^k	k-times continuously differentiable functions with compact support
C_0	completion of C_c^0
$BV(\Omega, \mathcal{K})$	functions of bounded variation assuming values in a set $\mathcal{K}$
$\lvert Du\rvert$	total variation measure of u
$\mu/\lvert\mu\rvert$	polar decomposition of a measure μ
$\mu \llcorner A$	restriction of a measure μ on a set A
$\nu \ll \mu$	ν is absolutely continuous with respect to μ
$\nu \perp \mu$	ν, μ are mutually singular measures
$\lvert A\rvert$	Lebesgue content of a set A
∂A	classical boundary of a set A
$\mathrm{int}(A)$, $\mathrm{ext}(A)$	classical interior, exterior of a set A
$(E)^t$	set of points with density $t \in [0,1]$ with respect to the set E
$(E)^1, (E)^0$	measure-theoretic interior, exterior of a set E
$\mathrm{Per}(E)$	perimeter of a set E, $\mathrm{TV}(\chi_E)$
$\mathcal{F}E$	reduced boundary of a set E
ν_E	generalized inner normal to a set E
$D^a u, D^j u, D^c u$	absolutely continuous part, jump part, Cantor part of Du
$D^s u$	singular part of Du, $D^s u = D^j u + D^c u$
Du	distributive derivative, $\mathbb{R}^{d\times l}$-valued measure for $u\colon \Omega \to \mathbb{R}^l$
$D_i u$	distributive directional derivative in the direction $i \in \{1, \ldots, d\}$
S_u	approximate discontinuity set of u
J_u	set of approximate jump points of u
ν_u, u^+, u^-	normal and one-sided limits of u on the approximate jump set J_u
$u^+_{\mathcal{F}E}, u^-_{\mathcal{F}E}$	approximate one-sided limits of u on the reduced boundary of a set E
df_x	Fréchet differential of f in x

Optimality and Rounding

Γ	space of sequences of thresholding parameters, $\Gamma = (\mathcal{I} \times [0,1])^{\mathbb{N}}$
$\mathbb{P}(X)$, $\mathbb{P}(X\lvert Y)$	probability of the event X, conditional probability
$\mathbb{E}f = \mathbb{E}_x f(x)$	expectation of f applied to the random variable x
$\bar{u}$, $\bar{u}_\gamma$, $\bar{u}_\alpha$	result of (parametrized) rounding operation applied to u

Problem Specification

$\Omega \subseteq \mathbb{R}^d$	image domain
d	dimension of image domain
l	number of labels/classes
s	local potentials, data term
b	affine part of regularizer
$\mathcal{I} = \{1, ..., l\}$	set of labels
$\mathcal{E} = \{e^1, ..., e^l\}$	embedded set of labels
Δ_l	l-dimensional unit simplex, convex hull of $\mathcal{E}$
f, f_D	primal, dual objective
$f_{\mathcal{C}}$	constrained functional $f + \delta_{\mathcal{C}}$
g	saddle-point function
$\mathcal{C}$	convex constraint set for u, primal constraint set
$\mathcal{C}_{\mathcal{E}}$	constraint set for *integral* u
$\mathcal{D}$	constraint set for support function, dual constraint set
$\mathcal{D}_{\text{loc}}$	local constraints pointwise defining the dual constraint set
J	positively homogeneous, convex regularizer
$\Psi\colon \mathbb{R}^{d\times l} \to \mathbb{R}_{\geqslant 0}$	integrand defining the regularizer
Ψ_d, Ψ_A, Ψ_u	integrand for envelope-, embedding-, uncertainty-based regularizer
$d(i, j)$	interaction potential, metric
Div, Grad	multidimensional divergence, gradient
ρ_l, ρ_u	lower, upper bounds for u
u^*	optimal relaxed solution
$u^*_{\mathcal{E}}$	optimal integral solution
$\ell\colon \Omega \to \mathcal{I}$	labeling function
ℓ^*	optimal labeling function

Discretization

n	number of pixels/vertices in the discretization of Ω
u^h, $u^h_{\bar{i}}$	discretization of u, vector corresponding to pixel $\bar{i} \in \mathcal{J}$
$u^{h,*}$	minimizer of the discretized problem
$\nabla_{\bar{i}}$	discretized gradient at pixel $\bar{i}$
Ω^h	discretized domain for grid size/scale h
$\mathcal{J}$	set of multiindices $\bar{i}$ indexing the pixels $x^{\bar{i}}$ in Ω
f^h, Ω^h	discretized functional
$\psi_C, \psi_x, \psi_{x,y}$	factor corresponding to clique C, unary term, pairwise term
$\tilde{u}^h$, $u^h_{\bar{i}}$, $C^h_{\bar{i}}$	L^1 representation of u^h, $\tilde{u}^h$ assumes the value $u^h_{\bar{i}}$ in $C^h_{\bar{i}}$
$\mathcal{U}^h$	set of discretized u^h
$\tilde{\mathcal{U}}^h$	set of piecewise constant functions defined by $u^h \in \mathcal{U}^h$
$\tilde{f}^h$	discretized functional f^h extended to piecewise constant functions in $\tilde{\mathcal{U}}^h$
$G = (V, E)$	graph with vertex set V and edge set E

List of Publications

1. J. Lellmann, J. Kappes, J. Yuan, F. Becker, and C. Schnörr: *Convex Multi-Class Image Labeling by Simplex-Constrained Total Variation.* IWR Preprint, University of Heidelberg, November 2008.
2. J. Lellmann, J. Kappes, J. Yuan, F. Becker, and C. Schnörr: *Convex Multi-Class Image Labeling by Simplex-Constrained Total Variation.* In Scale Space and Variational Methods in Computer Vision, volume 5567 of LNCS, pages 150–162, 2009. Springer.
3. J. Lellmann, F. Becker, and C. Schnörr: *Convex Optimization for Multi-Class Image Labeling with a Novel Family of Total Variation Based Regularizers.* In Proceedings of the 12th IEEE International Conference on Computer Vision, pages 646–653, 2009.
4. J. Lellmann, D. Breitenreicher, and C. Schnörr: *Fast and Exact Primal-Dual Iterations for Variational Problems in Computer Vision.* In Proceedings of the 11th European Conference on Computer Vision, volume 6312 of LNCS, pages 494–505, 2010. Springer.
5. D. Breitenreicher, J. Lellmann and C. Schnörr: *Sparse Template-Based Variational Image Segmentation.* In Advances in Adaptive Data Analysis, Special Issue on Sparse Representation of Multiscale Data and Images, 2011. World Scientific. In Press.
6. F. Lenzen, F. Becker, J. Lellmann, S. Petra, and C. Schnörr: *Variational Image Denoising with Adaptive Constraint Sets.* In Proceedings of the 3nd International Conference on Scale Space and Variational Methods in Computer Vision 2011. Springer. In Press.
7. J. Lellmann and C. Schnörr. *Continuous Multiclass Labeling Approaches and Algorithms.* SIAM Journal on Imaging Sciences, 2011. In Press. Preprint available on arXiv:1102.5448v2 [cs.CV].
8. J. Lellmann, F. Lenzen, and C. Schnörr: *Optimality Bounds for a Variational Relaxation of the Image Partitioning Problem.* Accepted as oral for the 8th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, 2011. In Press.

Chapter 1
Introduction and Overview

1.1 Nonsmooth Variational Models

Variational Methods. In this work, we will be concerned with a class of *variational* problems used in image processing and analysis. The output of a variational method is defined as the minimizer of an *objective function* f,

$$u^* := \arg\min_{u \in \mathcal{C}} f(u) \,, \tag{1.1}$$

where $\mathcal{C}$ 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 interpretation of u is governed by the problem to be solved: For the prototypical example of color denoising, $u\colon \Omega \to \mathbb{R}^3$ could directly describe the colors of the output image on the image domain Ω, while for segmentation problems, $u\colon \Omega \to \{0, 1\}$ could assign each point to the foreground ($u(x) = 1$) or background ($u(x) = 0$) class. We will in particular consider the case where the range of u is continuous and multi-dimensional, i.e. u is vector-valued.

Usually the objective f is composed of a *data term* $H(u)$ and a *regularizer* $J(u)$,

$$f(u) \;=\; H(u|I) + J(u). \tag{1.2}$$

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. However, 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 [SGG+09] for a general overview of variational methods in image processing.

The distinction between data term and regularizer in (1.2) also 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

$$u^* \;=\; \arg\max_u \mathbb{P}(u|I). \tag{1.3}$$

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 (1.3) is equivalent to

$$\begin{aligned} u^* &= \arg\max_u \mathbb{P}(I|u)\,\mathbb{P}(u) && (1.4)\\ &= \arg\min_u \{-\log(\mathbb{P}(I|u)) - \log(\mathbb{P}(u))\}. && (1.5) \end{aligned}$$

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 (1.5) 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 (simplifying) assumption that the conditional probability in (1.4) decouples on a per-point basis, one obtains (for finite Ω, i.e. after discretization):

$$\begin{aligned} -\log \mathbb{P}(I|u) &= -\log \prod_{x\in\Omega} \mathbb{P}(I(x)|u(x)) && (1.6)\\ &= \sum_{x\in\Omega} -\log \mathbb{P}(I(x)|u(x)). && (1.7) \end{aligned}$$

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

In contrast to this finite-dimensional example, we will generally formulate models on *continuous* domains Ω, i.e. contiguous subsets of $\mathbb{R}^d$, following the "analyze/optimize first" paradigm. Compared to "discretize first" approaches, this allows to get a deeper insight into the underlying problem, and to abstract from inaccuracies caused by the discretization.

Applying variational approaches generally requires *two* steps:

- choosing a suitable *model*,
- and providing a numerical *solver* for the associated discretized problem.

More intricate models usually complicate the optimization process, therefore choosing a model always involves a trade-off between modeling accuracy and numerical tractability. A particular difficulty concerns the evaluation and comparability of models: for moderately complex models, the associated problem can usually only be solved locally optimal, making it difficult to pinpoint whether an unexpected result should be attributed to the model or to the solver.

In this work, we will mainly consider *convex* models. Since these can generally be solved to a global optimum, modeling and optimization aspects are clearly separated. Certainly, this comes at the price of reduced modeling accuracy; the prominent question is therefore how to construct sufficiently simple convex approximations to difficult problems.

In many typical imaging problems, the data term is little problematic, and can be modeled fairly well using a convex local (pointwise) term as in (1.7). However, the prior knowledge encoded in the regularizer is usually of a much more "nonlocal" type, and finding suitable – ideally convex – regularizers is a difficult problem.

In this work, we will focus on *nonsmooth* regularizers, i.e. we do not assume differentiability. Such regularizers have become very popular in the field of image processing and computer vision in the last two decades. In many cases, introducing simple nonsmooth terms in the regularizer has intriguing effects.

Variational Denoising. As an example, consider the problem of removing noise from an image, in order to improve its visual quality or as a preprocessing step for further feature extraction. The most basic, classical example is Gaussian L^2-L^2 denoising: For an input image $I\colon \Omega \to \mathbb{R}^l$ and regularization weight $\lambda > 0$, minimize

$$f(u) \;=\; \frac{1}{2}\int_\Omega \|u - I\|_2^2 \, d\,x + \frac{\lambda}{2}\int_\Omega \|\nabla\, u\|_2^2 \, d\,x \tag{1.8}$$

over $u\colon \Omega \to \mathbb{R}^l$. Note that both the data term and the regularizer exhibit quadratic growth.

Problem (1.8) is convex, and after discretization can be solved globally optimal as a linear equation system. While the approach removes Gaussian noise very well, it tends to smear hard edges in the image. This is caused by the quadratic growth of the regularizer, which makes it susceptible to "outliers" – i.e. large gradients – in the form of hard edges.

Many approaches have been proposed to circumvent this problem. Most prominently, the *anisotropic diffusion* approach consists in solving (1.8) using gradient descent, at each step locally modifying the norm in the regularizer to reduce smoothing across directions where the current iterate has a large gradient, i.e. across potential edges. While this is widely used and gives good results in many cases, the output cannot be characterized in the variational way as the minimizer of a certain functional. A more one-step approach is the seminal work of Rudin-Osher-Fatemi [ROF92], who introduce the total variation into image processing and formulate the $L^2-\mathrm{TV}$ (*ROF*) model

$$f(u) \;=\; \frac{1}{2}\int_\Omega \|u - I\|_2^2 \, d\,x + \lambda \int_\Omega d|D\,u|. \tag{1.9}$$

The integral on the right-hand side involving the distributional derivative Du is known as the *total variation* (TV) of u, and is a generalization over the integral over $\|\nabla u\|_2$ for discontinuous u. The key difference to (1.8) is that, while the data term still has quadratic growth, the regularizer only grows linearly.

In practice this seemingly simple change partly solves the problem of dealing with hard edges: Gaussian noise is generally removed from regions with smooth gradient, while hard edges are retained. However, while still convex, the model is nonsmooth due to the missing exponent in the regularizer. Moreover, the theoretical analysis is much more involved since it requires to consider discontinuous u with appropriate generalizations of the derivatives. However, these issues can be rigorously addressed in the framework of *functions of bounded variation*, we refer to Appendix A.1 for an overview.

For non-Gaussian noise such as salt-and-pepper, (1.9) is suboptimal, as it is quite sensitive to outliers in the input image I. Also, ROF denoising invariably leads to a reduction in contrast. By going one step further, these drawbacks can also be addressed: Consider the $L^1-\mathrm{TV}$ model (see [Nik01] for an overview)

$$f(u) \;=\; \int_\Omega \|u - I\|_1 d\,x + \lambda \int_\Omega d|D\,u|. \tag{1.10}$$

Here both the data term as well as the regularizer exhibit linear growth. Existence and well-posedness of (1.9) and (1.10) can be shown in a precise sense within the class of functions of bounded variation [AFP00, AMT91].

Functionals such as (1.10) are extremely tolerant to noise. The downside is that they also tend to generate a "staircasing" effect on smooth gradients, i.e. the solution tends to be piecewise constant. We refer to [DAG09] and the references therein for a detailed analysis. For image denoising this is certainly not desirable, therefore some effort has been put into reducing staircasing while preserving robustness, mostly by including higher-order derivatives (see e.g. [CL97, Sch98, CMM00, LT06, BKP10]). However, in some applications staircasing is explicitly *desired*. One such application is image segmentation or more generally multiclass labeling, which will be our main interest.

1.2 Variational Multiclass Labeling

In this work, we focus on a particular class of variational problems that originate from the *multiclass labeling* problem, also known as *multiclass image segmentation*. We will first outline our generic model, and then relate it to the existing approaches.

The task is to assign to each point x in the image domain $\Omega \subseteq \mathbb{R}^d$ an *integral* label $\ell(x) \in \mathcal{I} := \{1, ..., l\}$, so that the *label assignment* (or *labeling*) function ℓ adheres to some local data fidelity as well as nonlocal spatial coherency constraints (Fig. 1.1). This problem class occurs in many applications such as segmentation, multiview reconstruction, stitching, and inpainting [PCF06]. We consider the generic variational formulation

$$\inf_{\ell:\Omega\to\mathcal{I}} f(\ell),\ f(\ell) := \underbrace{\int_\Omega s(x,\ell(x))\,dx}_{\text{data term}} + \underbrace{J(\ell)}_{\text{regularizer}}, \tag{1.11}$$

where we deliberately do not fix any function spaces yet, as we will settle on a slightly different form. Formulation (1.11) directly relates to the general variational approach (1.2), however we deliberately distinguish ℓ and u in order to emphasize the finite number of possible values at each point.

The *data term* assigns to each label $\ell(x)$ a *local cost* $s(x, \ell(x)) = s_I(x, \ell(x)) \in \mathbb{R}$ depending on the observation I. These costs are specific to the application and often derived from a statistical model: in view of (1.7), $s(x, k)$ can be interpreted as the negative log-probability $-\log \mathbb{P}(I(x)|\ell(x) = k)$, for any label $k \in \mathcal{I}$. 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 $\|I(x) - c^k\|_2$, 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 [HZ00, FL01]. 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 very strong dependence on the input data, and its structure typically cannot be reliably predicted beforehand.

As the input data is usually subject to noise or missing some parts, or generally not sufficient to extract the desired higher-order information, additional prior knowledge must be introduced through the regularizer J. Usually, at least some spatial coherency is desired. In this work we will in particular be interested in regularizers that penalize the weighted length of interfaces between regions of constant labeling. Note that the regularizer may generally involve terms depending on the observation I. While this somehow blurs the clear distinction between prior and posterior knowledge as in (1.5), it is common practice and often leads to better results.

Contour-Based Segmentation and Level Sets. The model (1.11) puts an emphasis on the *region-based* interpretation of the segmentation problem: Effectively, the labeling function ℓ partitions the image domain Ω into l regions $\Omega_1, \dots, \Omega_l$, where

$$\Omega_k \;:=\; \ell^{-1}(\{k\}). \tag{1.12}$$

The data term can be reformulated as region integrals,

$$\int_\Omega s(x, \ell(x))\, dx \;=\; \sum_{k=1}^{l} \int_{\Omega_k} s(x,k)\, dx. \tag{1.13}$$

However, for two-dimensional domains one could alternatively think of the labeling problem as the problem of finding a set of *contours*, i.e. closed curves $C^k\colon [0,1] \to \mathbb{R}^2$, describing the boundaries $\partial\Omega_k$. One of the earliest contour-based approaches is the *snake* model [KWT87], where one considers a single contour $C = C^k$ and minimizes for some weights $\beta, \lambda > 0$ the energy

$$f(C) \;=\; -\lambda \int_0^1 \|\nabla I(C(p))\|_2\, dp + \int_0^1 \{\|\nabla C(p)\|_2^2 + \beta\, \|\nabla^2 C(p)\|_2^2\}\, dp. \tag{1.14}$$

The left integral incorporates the observed grayscale image I by drawing the contour towards hard edges in the image. The right integral takes the role of a regularizer, penalizing curve length and curvature. This model has two important disadvantages: first, it is not parametrization-independent with respect to C, and second, it relies on an explicit parametrization of C, which requires much effort in order to cope with changing topology and multiple objects.

These issues are addressed by the well-known *Geodesic Active Contours* model [CKS97]. The parametrization invariance is obtained by defining, for some decreasing *edge detector* function $g\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{>0}$ and $h(x) := g(\|\nabla I(x)\|_2)$, the energy

$$f(C) \;=\; \int_0^1 h(C(p)) \|\nabla C(p)\|_2\, dp. \tag{1.15}$$

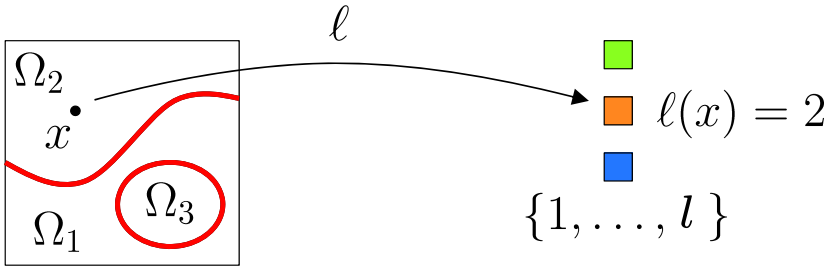

Figure 1.1. Multiclass labeling problem. The task is to find a label assignment function ℓ that partitions the image into l regions $\Omega_1, ..., \Omega_l$ such that the assignment fits to the observed input data as well as the prior knowledge about the structure of the regions.

The integral can be interpreted as the curve length of C measured in the locally weighted Riemannian metric defined by h. Therefore, the (local) minimizers of f represent geodesics with respect to this metric. In higher-dimensional spaces a similar approach can be formulated in terms of minimal surfaces [CKSS96].

Due to the parametrization invariance of (1.15), by the usual technique of considering the first variation of (1.15) with respect to C, one obtains the "steepest-descent" partial differential equation (PDE)

$$\frac{\partial C}{\partial t} \;=\; (\kappa\, h - \langle \nabla h, \nu \rangle)\, \nu, \tag{1.16}$$

expressed solely in *intrinsic* properties of C, i.e. its *normal* ν and *curvature* κ. Starting from some arbitrary $C(\cdot, 0)$ at $t = 0$, (1.16) is then integrated with respect to the artificial time parameter t in order to obtain a local minimum of (1.15).

Instead of the explicit parametrization of C, a very popular approach also proposed in [CKS97] is to use *level sets* [OS88]: the curve C is represented as the *zero set* of some function $\phi\colon \Omega \to \mathbb{R}$, i.e. $C([0,1]) = \phi^{-1}(\{0\})$. A common convention is to require $\phi(x) < 0$ in the interior of C, and $\phi(x) > 0$ in the exterior. Then ν and κ can be readily expressed as

$$\nu \;=\; -\frac{\nabla \phi}{\|\nabla \phi\|_2}, \quad \kappa \;=\; \operatorname{div}\left(\frac{\nabla \phi}{\|\nabla \phi\|_2}\right), \tag{1.17}$$

and (1.16) amounts to

$$\frac{\partial \phi}{\partial t} \;=\; h\, \|\nabla \phi\|_2 \operatorname{div}\left(\frac{\nabla \phi}{\|\nabla \phi\|_2}\right) + \langle \nabla h, \nabla \phi \rangle. \tag{1.18}$$

Originally, (1.16) defines (1.18) only on C, i.e. on the zero set $\phi^{-1}(\{0\})$. The fundamental trick when employing level set methods is to integrate (1.18) also in all remaining $x \in \Omega$. This allows to propagate curves of arbitrary topology (and number) using a single, fixed discretization, e.g. on a grid. As soon as a steady state is achieved, the boundary curve C can be extracted with subpixel accuracy from the zero crossings of ϕ. Moreover, the higher-dimensional case $d \geqslant 3$ can be transparently handled.

In this sense, the level set approach is a *hybrid* method, encoding the originally contour-based functional in a region-based integral form. Consequently, Chan and Vese proposed to apply the level-set technique to a region-based formulation, also known as *Active Contours Without Edges* [CV01]. Here the *edge*-(gradient-)based model (1.15) is replaced by a *region*-based approach, formulated in terms of the *interior* and *exterior* of the region P_C described by C,

$$f(C, c_1, c_2) \;=\; \int_{\mathrm{int}(P_C)} \|I - c_1\|_2^2 \, dx + \int_{\mathrm{ext}(P_C)} \|I - c_2\|_2^2 \, dx + \mu \int_0^1 \|\nabla C\|_2 \, dp \qquad (1.19)$$

The functional can be rewritten in terms of a level set function ϕ by formally introducing the Heaviside function $H(\phi)$ in order to discriminate between the interior and exterior of P_C, and the common PDE-based flow can be computed.

However, all these methods share several important drawbacks:

- In general, the level set function ϕ is not unique. This can be avoided by postulating that ϕ should be a signed distance function, at the cost of complicating the optimization process.
- In contrast to region-based formulations, edge-based approaches do not have a plausible statistical explanation.
- For the Chan-Vese formulation, in order to obtain differentiability, a smoothed variant of the Heaviside function must be used, which requires a trade-off between accuracy and convergence speed.
- Most importantly, the methods are inherently local, since they rely on the steepest-descent PDE (1.16). Therefore a good initialization is mandatory, and model and optimization effects cannot be clearly separated.

Note that formulation (1.19) is directly covered by the general model (1.11): 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.

Region-Based Segmentation and Mumford-Shah. Probably the most influential region-based model is the *Mumford-Shah* model [MS89]. Motivated by the Gibbs field [GG84] and weak membrane energy [BZ87] methods, it can be seen as a first approach of explicitly introducing the possibility of discontinuous solutions into a spatially continuous framework.

It consists of minimizing, for some $\lambda, \mu, \nu > 0$, the functional

$$f(u, K) = \lambda \int_\Omega (u - I)^2 dx + \mu \int_{\Omega \setminus K} \|\nabla u\|_2^2 dx + \nu \, \mathcal{H}^{d-1}(K), \qquad (1.20)$$

where $K \subseteq \Omega$ is closed and u is differentiable outside of K, i.e. $u \in C^1(\Omega \setminus K)$. Essentially, this corresponds to the $L^2 - L^2$ denoising approach (1.8), with the exception that u is allowed to be discontinuous on a "boundary" set K that should be "small" as measured by the Hausdorff term.

The initial idea of K being a set of piecewise smooth curves is difficult to treat analytically, therefore it has since been formalized using a *weak formulation* in the framework of functions of bounded variation: Replacing K by the discontinuity set S_u of some function u (see Appendix A.1 for the precise definitions), define

$$f(u) \;=\; \lambda \int_\Omega (u-I)^2\,dx + \mu \int_\Omega \|\nabla u\|_2^2\,dx + \nu\,\mathcal{H}^{d-1}(S_u). \tag{1.21}$$

With a proper redefinition of the gradient as an *approximate* gradient defined almost everywhere on Ω, and under a restriction of u to the set of *special functions of bounded variation* $\mathrm{SBV}(\Omega)$, a minimizer of (1.21) exists and allows to recover a minimizing pair (K,u) of the original functional (1.20).

Many important results concerning the structure of solutions to the Mumford-Shah problem have been derived in this framework, and are still under active research [AFP00, Dav05]. A large part of its popularity certainly stems from the fact that many region-based functionals can be considered as limiting or special cases of (1.20) as already derived in the original publication [MS89]. See also [AK00] for an overview.

While originally proposed for image segmentation, the Mumford-Shah functional is not a labeling approach in the sense that each point is assigned one of a fixed number of labels, each corresponding to a specific model. Instead, 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.

Therefore, in a sense the Mumford-Shah model is a case of *simultaneous* labeling and model parameter optimization, which is generally a much more difficult problem than pure labeling. Consequently it is not surprising that the functional is nonconvex: in the Hausdorff term, "jumps" of u along some boundary are always counted by the length of the boundary irregardless of the height of the jump, violating convexity.

An important connection to labeling problems occurs in the limit $\mu \to +\infty$. Here, the optimal u must be *piecewise constant*, i.e. it is constant on each connected component Q^i of $\Omega \setminus K$, necessarily assuming the mean of I in this region. For this special case, the weak formulation (1.21) reduces to

$$f(u) \;=\; \lambda \sum_i \int_{Q^i} \left(u - \frac{1}{|Q^i|} \int_{Q^i} I dy \right)^2 dx + \nu\,\mathcal{H}^{d-1}(S_u). \tag{1.22}$$

This formulation is also known as the *piecewise constant Mumford-Shah* model. In the labeling/model parameter estimation interpretation, this reduces the models that explain the individual regions to simple Gaussian models with a single fixed mean value. If one additionally restricts the number of clusters to a fixed integer l, and denotes the mean values by $c_1, \ldots, c_l$, the problem can be rewritten in the form

$$f(\ell) \;=\; \lambda \int_\Omega (c_{\ell(x)} - I(x))^2\,dx + \nu\,\mathcal{H}^{d-1}(S_\ell), \tag{1.23}$$

which amounts to the labeling problem (1.11) upon setting $s(x,\ell(x)) := \lambda\,(c_{\ell(x)} - I(x))^2$.

As seen above, for the two-class case this corresponds to the *Chan-Vese* model (1.19). Therefore, in a sense our general functional (1.11) constitutes the *labeling/inference* part of the simultaneous labeling and model parameter estimation performed by the Mumford-Shah and Chan-Vese models, i.e. the task of finding an optimal partition of the image domain if the optimal model parameters are known.

While (1.23) is a region-based formulation, it does not suffer from the non-uniqueness of the region-based active contours formulation: the labeling function ℓ describes the membership of a point to some region directly in terms of an *index*, rather than the *sign* of some level set function ϕ. This introduces the problem of coping with discontinuous ϕ, which is avoided in the level set formulation.

However, directly transferring optimization techniques for the Mumford-Shah functional to the labeling formulation is difficult:

- While the model (1.11) can be viewed as a generalization of the inference part in the Mumford-Shah model, the latter is restricted to quadratic distances.
- The complete (even piecewise constant) Mumford-Shah problem is inherently nonconvex. While many optimization methods have been suggested, such as Simulated Annealing, Graduated Nonconvexity [BZ87], Phase Field [AT90] and Level Set approaches, they are all tailored to the nonconvex regularizer, and do not generally allow to find global solutions of the *labeling* problem.

Convex Labeling. Recently, a third class of approaches for the two-class segmentation problem on continuous domains has emerged [CEN06]. It is based on the observation that if one replaces $\ell(x)$ in (1.23) with some function $u\colon \Omega \to \{0,\, 1\}$ such that $u(x) = 0$ iff $\ell(x) = 1$, the problem can be rewritten as

$$f(u) \;=\; \lambda \int_\Omega \left((1-u)\,(c_1 - I)^2 + u\,(c_2 - I)^2\right) dx + \nu \int_\Omega \|Du\|_2. \tag{1.24}$$

This is possible since the total variation term in the rightmost integral evaluates exactly to the length of $\partial(u^{-1}(\{0\})) = \partial\Omega_1$. In view of the $L^2 - L^2$, $L^2 - \mathrm{TV}$ and $L^1 - \mathrm{TV}$ denoising approaches, this two-class model can be viewed as a "Linear-TV" approach. In fact, it was motivated in [CEN06] as a way to formulate $L^1 - \mathrm{TV}$ *geometry denoising*, i.e. denoising of indicator functions.

Formulation (1.24) has several major advantages:

- The functional is region-based, does not require an explicit parametrization of the boundary, and therefore allows to deal with partitions of arbitrary topology.
- Non-quadratic terms can be trivially included in the data term. In fact, the data term is always linear in u, independent of its original – possibly complicated, statistically derived – form.
- The functional itself is *convex* in u, and therefore does not suffer from local minima.

The last point is particularly important: if one allows u to also assume *fractional* values from the interval $[0, 1]$ instead of only the *integral* values $\{0, 1\}$, one obtains a completely convex problem.

The general form of (1.24) is

$$\min_{u \in \Omega \to [0,1]} f(u) \;=\; \int_\Omega u(x)\, s(x)\, dx + \int_\Omega \|Du\|_2, \tag{1.25}$$

which is equivalent to (1.24) if one sets $s(x) = (c_2 - I)^2 - (c_1 - I)^2$. 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 considered in a maximal-flow setting in [Str83], see also [AT06]. It can be shown that solutions of the *relaxed* problem, with $u(x) \in [0, 1]$, can be thresholded at almost any threshold, and the resulting *integral* u provides a solution of the original problem [CEN06].

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 [BVZ01], however they are inherently formulated on graphs, i.e. they are formulated on the *discretized* problem, which invariably introduces an anisotropy, and prohibits true rotational invariance.

Similar to what can be observed when applying graph-cut techniques, the transition from the two-class problem (1.24) to the general *multiclass* problem (1.11) is a major step. In particular, it is not clear how to represent the labels using u, how to relax the combinatorial constraint set, and how to formulate useful regularizers. Nevertheless, one can hope that by solving these issues at least partially, powerful models for multiclass labeling problems can be derived.

1.3 Contribution

In this work, we investigate an approach for formulating convex relaxations of the multiclass labeling problem (1.11) that combine the strengths of the various models discussed above:

- Our approach is *region-based*, and therefore allows for a statistical interpretation and solutions with arbitrary topology.
- It is formulated in a *spatially continuous* framework, avoiding discretization artifacts caused by an early discretization.
- It is also *convex*, such that the associated optimization problem can be solved globally optimal, and modeling and optimization issues can be clearly separated.

The remainder of this work is structured as follows: In **Chapter 2** we establish a convex extension of (1.11) to the multi-class case. We generally focus on length-based regularizers that allow a certain weighting in form of an interaction potential that depends on the labels on either sides of a boundary. We study the possible interaction potentials and show existence of minimizers for the spatially continuous problem.

The main difficulty when constructing convex extensions for prescribed interaction potentials lies in formulating suitable regularizers that are also computationally tractable. We propose three approaches with different tightness and computational cost, and discuss relations to other regularizers that have been proposed in various contexts. Although we motivate the modified regularizers in the labeling framework, they could as well be interesting for ROF- and $L^1 - \mathrm{TV}$ type problems applied to vector-valued data.

We also take a closer look at the aforementioned thresholding property for the two-class problem, which hinges on the coarea formula from geometric measure theory, and point out why straightforward extensions to the multi-class case fail.

Chapter 3 is devoted to discretizations of the spatially continuous formulation. We relate our approach to "discretize-first" approaches such as Markov Random Fields and graph-cut techniques. We use a result from [CCP08] to show that the problem can be discretized using finite differences such that the discretized functionals Γ-converge to the true (possibly) isotropic functional for vanishing grid spacing.

We show a range of experiments to demonstrate that our approach generally introduces less artifacts than graph-based discretizations. In fact, we conclude that this is not only an effect of the particular discretization, but hinges on the decision of whether one looks for fractional or integral solutions of the discretized problem.

In order to practically solve the discretized problem, in **Chapter 4** we consider several numerical methods. We focus on first-order methods, as these have recently been very successful in image processing when dealing with non-smooth large-scale models. In particular, we consider a controlled smoothing approach suggested by Nesterov and an operator splitting technique based on Douglas-Rachford splitting.

We empirically compare the methods to several other approaches for solving the nonsmooth problem. While the good bounds available for the Nesterov methods turn out to be mainly theoretical, the Douglas-Rachford approach is robust, relatively fast for moderate-accuracy solutions as typically required in image processing, and allows to handle tight regularizers by a suitable introduction of auxiliary variables.

In **Chapter 5** we revisit the question of how integral solutions can be retrieved from solutions of the convex relaxed problem. We show that the two-class case allows an interpretation in a probabilistic rounding framework via the coarea formula.

In order to transfer these results to the multiclass case, we show how an approximate variant of the coarea formula can be obtained from a probabilistic rounding method originally proposed in a finite-dimensional LP relaxation framework [KT99]. We conclude with some empirical comparisons to deterministic rounding techniques.

Finally, in **Chapter 6** we discuss an extension that allows to incorporate higher-level knowledge about the shape of the objects contained in the image. The approach is based on a nonlinear extension of the Sparse Representation problem, and allows to explicitly model shape knowledge using a dictionary. Experiments show that it deals well with heavily occluded objects and also has many other interesting potential applications such as shape decomposition.

The required fundamentals regarding functions of bounded variation, Γ-convergence, and operator splitting methods are collected in the appendix. Together, we hope to provide a motivation for using such specially-designed variational problems in order to solve image labeling problems. While they are numerically more sophisticated to deal with than conventional combinatorial approaches, we found that they provide visually superior results, and are backed by an intriguing theoretical foundation.

1.4 Notation

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

$$A \;=\; (a^1|...|a^l) = (A_{ij}), \quad A_{ij} = (a^j)_i = a^j_i, \quad 1 \leqslant i \leqslant d, 1 \leqslant j \leqslant l.$$

Superscript parentheses $v^{(i)}$ indicate an element of a sequence $(v^{(i)})$. We denote the natural numbers by $\mathbb{N} = \{1, 2, ...\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and the extended real line by $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. We will frequently make use of the Kronecker product [Gra81]

$$\otimes : \mathbb{R}^{n_1 \times m_1} \times \mathbb{R}^{n_2 \times m_2} \to \mathbb{R}^{(n_1 n_2) \times (m_1 m_2)} \tag{1.26}$$

in order to formulate all results for arbitrary-dimensional domains. The *standard simplex* in $\mathbb{R}^l$ is denoted by

$$\Delta_l \;:=\; \{x \in \mathbb{R}^l | x \geqslant 0, e^\top x = 1\}, \tag{1.27}$$

where $e := (1, ..., 1)^\top \in \mathbb{R}^l$. I_n is the identity matrix in $\mathbb{R}^n$ and $\|\cdot\|_2$ the usual Euclidean norm for vectors or the Frobenius norm for matrices. Similarly, the standard inner product $\langle \cdot, \cdot \rangle$ extends to pairs of matrices as the sum over their elementwise product. $\mathcal{B}_r(x)$ denotes the ball of radius r at x, and S^{d-1} the set of $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$. The *characteristic function* $\chi_\mathcal{S}$ and the *indicator function* $\delta_\mathcal{S}$ of a set $\mathcal{S}$ are defined as

$$\chi_\mathcal{S}(x) := \begin{cases} 1, & x \in \mathcal{S}, \\ 0, & x \notin \mathcal{S}, \end{cases} \quad \text{and} \quad \delta_\mathcal{S}(x) := \begin{cases} 0, & x \in \mathcal{S}, \\ +\infty, & x \notin \mathcal{S}. \end{cases} \tag{1.28}$$

For a convex set $\mathcal{C}$,

$$\sigma_\mathcal{C}(u) \;:=\; \sup_{v \in \mathcal{C}} \langle u, v \rangle \tag{1.29}$$

is the support function from convex analysis. ∇u denotes the classical Jacobian for differentiable u. $C_c^k(\Omega)$ is the space of k-times continuously differentiable functions on Ω with compact support, and $C_0(\Omega)$ the completion of $C_c^0(\Omega)$ under the supremum norm. As usual, $\mathcal{L}^d$ denotes the d-dimensional Lebesgue measure, while $\mathcal{H}^k$ denotes the k-dimensional Hausdorff measure. A list of symbols introduced throughout the text can be found in the front matter.

Chapter 2

A Variational Convex Formulation for Multi-Class Labeling

2.1 Introduction and Overview

In this chapter, we consider a method for formulating convex relaxations of the spatially continuous *multiclass* labeling problem (Sect. 1.1). As mentioned, the difficulty lies mainly in the combinatorial nature of the constraint set.

In the following, we will relax this set. This allows to solve the problem in a globally optimal way using convex optimization methods. On the downside, we cannot expect the relaxation to be exact for any problem instance, i.e. *fractional* (non-integral), or integral but suboptimal, artificial solutions may occur. In contrast to existing methods such as LP relaxation, we treat the problem in the fully spatially continuous setting, without resorting to an early discretization.

There are several choices for the relaxation method, of which in our opinion the following is the most transparent (Fig. 2.1): We first identify label i from the label set $\mathcal{I} := \{1, ..., l\}$ with the i-th unit vector $e^i \in \mathbb{R}^l$, set $\mathcal{E} := \{e^1, ..., e^l\}$, and solve

$$\inf_{u \in \mathcal{C}_\mathcal{E}} f(u), \quad f(u) := \langle u, s \rangle + J(u) = \int_\Omega \langle u(x), s(x) \rangle d\, x + J(u)\,, \tag{2.1}$$

$$\mathcal{C}_\mathcal{E} := \mathrm{BV}(\Omega, \mathcal{E}) = \{u \in \mathrm{BV}(\Omega)^l | u(x) \in \mathcal{E} \text{ for a.e. } x \in \Omega\}. \tag{2.2}$$

The labels are thus embedded into a higher-dimensional space. The space of functions of bounded variation $\mathrm{BV}(\Omega, \mathcal{E}) \subset (L^1)^l$ guarantees a minimal regularity of the discontinuities of u, see Appendix A.1.1 for the basic definitions. The data term becomes linear in u and is fully described by the vector

$$s(x) := (s_1(x), ..., s_l(x))^\top := (s(x, 1), ..., s(x, l))^\top \in \mathbb{R}^l. \tag{2.3}$$

Due to the linearization, the local costs s may be arbitrarily complicated, possibly derived from a probabilistic model, without affecting the overall problem class. We generally assume $s \geqslant 0$, 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, we *relax* the label set by allowing u to assume intermediate (fractional) values in the convex hull $\mathrm{conv}\,\mathcal{E}$ of the original label set. This is just the unit simplex Δ_l,

$$\Delta_l \;:=\; \mathrm{conv}\{e^1, ..., e^l\} = \{a \in \mathbb{R}^l | a \geqslant 0, \sum_{i=1}^{l} a_i = 1\}. \tag{2.4}$$

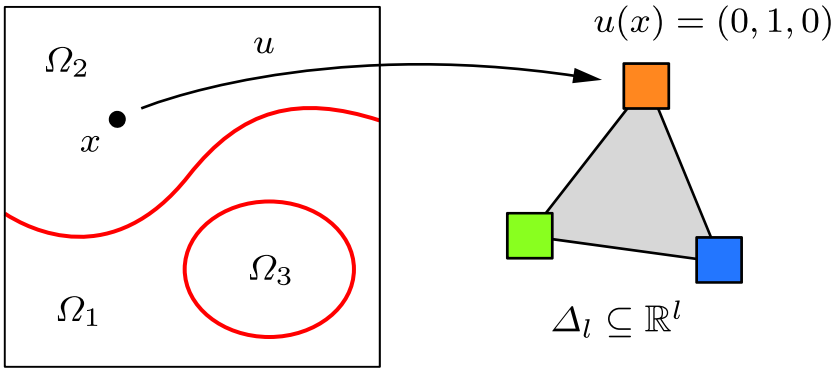

Figure 2.1. Convex relaxation of the multiclass labeling problem. The assignment of one label to each point in the image domain Ω is represented by a vector-valued function $u\colon \Omega \to \mathbb{R}^l$. Ideally, u partitions the image into l sets by assuming one of the unit vectors $\{e^1, \ldots, e^l\}$ everywhere. By relaxing this set to the unit simplex Δ_l, the originally combinatorial problem can be treated in a convex framework.

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

$$\mathcal{C} := \mathrm{BV}(\Omega, \Delta_l) = \{u \in \mathrm{BV}(\Omega)^l | u(x) \in \Delta_l \text{ for a.e. } x \in \Omega\}. \tag{2.5}$$

Assuming we can extend the regularizer J from $\mathcal{C}_\mathcal{E}$ to the whole relaxed set $\mathcal{C}$, we obtain the relaxed problem

$$\inf_{u \in \mathcal{C}} f(u), \quad f(u) := \int_\Omega \langle u(x), s(x) \rangle d\,x + J(u). \tag{2.6}$$

If additionally J can be made convex, the overall problem is convex as well, and it may likely be solvable globally optimal. In addition, J should ideally have a closed-form expression, or at least lead to a computationally tractable problem.

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

$$\mathrm{TV}(u) = \int_\Omega d\,|Du|, \tag{2.7}$$

which generalizes the integral over the Frobenius norm of the gradient, $\|\nabla u\|_2$. Note that ideally u is piecewise constant and thus discontinuous, so the gradient Du has to be understood in a distributional sense (Appendix A.1.1). Using this definition, the relaxation corresponds to the two-class case (1.25) as follows: Setting $l = 2$ and $J = \mathrm{TV}$ in the relaxed problem (2.6), we see that the second component of u is given by the first via $u_2 = 1 - u_1$. We may therefore substitute $u_1 := u'$ and $u_2 = 1 - u'$ for a suitable $u' \in \mathrm{BV}(\Omega, [0, 1])$, and pose the relaxed problem (2.6) in the form

$$\min_{u' \in \mathrm{BV}(\Omega, [0,1])} \int_\Omega u'(x)(s_1(x) - s_2(x)) d\,x + \sqrt{2}\,\mathrm{TV}(u'), \tag{2.8}$$

where $u'(x)$ is a scalar. This is exactly the two-class *continuous cut* introduced in Sect. 1.2, for which globally optimal solutions can be recovered from *any* solution of the relaxed problem by thresholding. In this case the *combinatorial* problem therefore reduces to a *convex* problem. While there are reasons to believe that this procedure cannot be extended to the multi-class case (see Sect. 2.6 below), it is still a strong motivation to consider formulation (2.6) for multiple labels.

Considering again the two-class case (2.8), we see that for integral u' the regularizer penalizes exactly the length of the interface between the two regions where $u' = 0$ and $u' = 1$, respectively. In this chapter, we consider ways to construct *multiclass* regularizers which penalize interfaces between two adjacent regions with labels $i \neq j$ according to the *perimeter* (i.e. length or area) of the interface weighted by $d(i, j)$, for some *interaction potential* $d\colon \{1, ..., l\}^2 \to \mathbb{R}$ depending on the labels (in a slight abuse of notation the interaction potential is also denoted by d, since there is rarely any ambiguity with respect to the ambient space dimension). This will be formalized in the following section, see Fig. 2.2 for an illustration. As a basic prototypical example, consider the *uniform* metric

$$d_u(i, j) \;:=\; \chi_{\{i \neq j\}}(i, j). \tag{2.9}$$

For $d = d_u$, the regularizer should thus penalize the total interface length, as seen above for the total variation. By choosing a different d, one obtains non-uniform regularization as visualized in Fig. 2.3.

For most of this chapter, the regularizer will be of the form

$$J(u) := \int_\Omega d\,\Psi(Du), \tag{2.10}$$

where $\Psi\colon \mathbb{R}^{d\times l} \to \mathbb{R}_{\geqslant 0}$. Note that $\Psi(Du)$ is again a measure, see Appendix A.1.3 for the precise definitions. We generally assume that Ψ is proper, continuous, positively homogeneous and convex. This implies that Ψ is the *support function* of some nonempty closed *dual* set $\mathcal{D}_{\text{loc}} \subseteq \mathbb{R}^{d\times l}$ [RW04, Thm. 8.24],

$$\Psi(z) = \sigma_{\mathcal{D}_{\text{loc}}}(z) = \sup_{v\in\mathcal{D}_{\text{loc}}} \langle z, v\rangle. \tag{2.11}$$

The expression $\Psi(Du)$ in (2.10) should be seen as a transformation of the *measure* Du according to Ψ (cf. (A.42), noting that Ψ is positively homogeneous and therefore coincides with its recession function Ψ_∞): $\Psi(Du) = \Psi(Du/|Du|)\,|Du|$, and

$$J(u) \;=\; \int_\Omega \Psi\!\left(\frac{Du}{|Du|}\right) d\,|Du|. \tag{2.12}$$

Also, we have an equivalent dual formulation in analogy to the definition of the total variation (A.2),

$$J(u) = \sup\{-\int_\Omega \langle u, \operatorname{Div} v\rangle\, dx \,|\, v \in C_c^\infty(\Omega)^{d\times l},\ v(x) \in \mathcal{D}_{\text{loc}} \forall x \in \Omega\}. \tag{2.13}$$

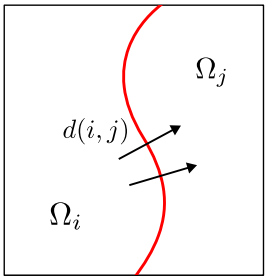

Figure 2.2. Interaction potential d used to define the regularizer. An interface between regions with labels i and j is penalized by its length, weighted by $d(i,j)$.

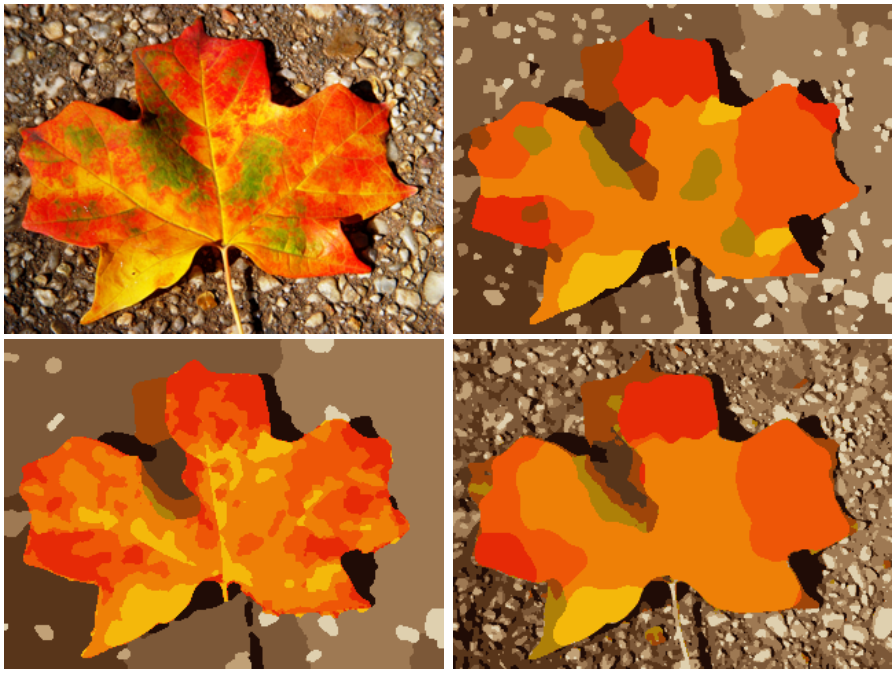

Figure 2.3. Effect of choosing different interaction potentials. **Top row:** The original image (left) is segmented into 12 regions corresponding to prototypical colors vectors. The uniform metric interaction potential penalizes the interface length independently of the labels (right), which leads to a uniformly smooth segmentation. **Bottom row:** By modifying the interaction potential, the regularization strength is selectively adjusted to suppress background (left) or foreground (right) structures while allowing for fine details in the other regions.

We also generally assume that Ψ is rotationally invariant, i.e. $\Psi(R\,z) = \Psi(z)$ for any rotation matrix $R \in \mathrm{SO}(d)$. Equivalently, $R\mathcal{D}_{\mathrm{loc}} = \mathcal{D}_{\mathrm{loc}}$ for any such R, i.e.

$$v = (v^1, \ldots, v^l) \in \mathcal{D}_{\mathrm{loc}} \;\Leftrightarrow\; (R\,v^1, \ldots, R\,v^l) \in \mathcal{D}_{\mathrm{loc}}. \tag{2.14}$$

Under these assumptions, the regularizer J is *isotropic* and *homogeneous*, in the sense that it is invariant under rotation and translation of the coordinates. We will consider in Sect. 2.7 some regularizers which depart from this assumption.

Organization. The remainder of this chapter is organized as follows:

- We formulate natural requirements on the regularizer J and show their implications on the choice of the interaction potential d (Sect. 2.3). In particular, d must necessarily be a metric under these requirements (Prop. 2.1).
- We show under which circumstances a minimizer exists in $\mathrm{BV}(\Omega)^l$ (Sect. 2.4), and how the regularizer is connected to the interaction potential (Sect. 2.5).
- We propose three different regularizers specifically constructed for relaxations of labeling problems with prescribed interaction potentials:
  - The "envelope" approach, which is a generalization of the method recently suggested by Chambolle et al. (Sect. 2.5.1). While there is no simple closed form expression, we show that it can be used to construct an exact regularizer for *any* metric d (Prop. 2.5).
  - The "Euclidean distance" approach (Sect. 2.5.2), which yields exact extensions for Euclidean metrics d only, but has a closed-form expression. We review some methods for the approximation of non-Euclidean metrics.
  - The "emphasized uncertainty" method (Sect. 2.5.3), which is in some sense opposite to the envelope approach, as it strongly tends to non-integral solutions in regions of uncertainty.
- We discuss the connection to the two-class problem with particular emphasis on different ways to generalize the coarea formula to the multi-class case (Sect. 2.6). These considerations motivate the derivation of the optimality bounds in Chap. 5.
- Finally, we give an unified overview, within our framework, over related regularizers that have been proposed in various contexts, and point out connections to other recently proposed techniques (Sect. 2.7).

The fundamental results regarding functions of bounded variation and the coarea formula are summarized in Appendix A.1.1.

2.2 Related Work

For an overview of existing variational approaches we refer to Chap. 1. In contrast to *graph-based, finite-dimensional* methods we generally work in the spatially continuous setting, which prevents early introduction of anisotropy by the discretization. In comparison to existing continuous approaches, we provide a unified framework for arbitrary, non-uniform metric interaction potentials.

When the first approach [LKY+08] was published, two other authors independently published similar ideas: In [ZGFN08], Zach et al. essentially consider the relaxed multi-class problem (2.6) in an informal way, with specific focus on the "decoupling" regularizer $\Psi(z) = \sum_{i=1}^{l} \|z^i\|_2$. This allows to solve the overall problem using several parallel instances of ROF-type problems (cf. Sect. 1.1.). In a sense, [LLT06] can be seen as a predecessor of this approach: In this work, the authors consider the piecewise-constant Mumford-Shah model. They represent the label assignment using a piecewise-constant real-valued function, but parametrize this function using a set of l polynomial basis functions, which can be seen as the individual components of u.

A more systematic treatment was published by Chambolle et al. in [CCP08], based on [PSG+08]. They too motivate their work by the piecewise-constant Mumford-Shah model; however they do not rely on a linear ordering of the labels as e.g. required in [Ish03, BT09b]. Their approach differs from (2.6) in the sense that instead of embedding into $\mathbb{R}^l$ via $\mathcal{E} = \{e^1, ..., e^l\}$, they embed the labels into the space $\mathcal{E}' := \{e'^1, ..., e'^l\} \subseteq \mathbb{R}^{l-1}$, where $(e'^i)_j = \chi_{\{j<i\}}$, i.e.

$$\begin{aligned} e'^1 &= (0,0,0,...,0)^\top, \\ e'^2 &= (1,0,0,...,0)^\top, \\ e'^3 &= (1,1,0,...,0)^\top, \\ &\vdots \\ e'^l &= (1,1,1,...,1)^\top. \end{aligned} \tag{2.15}$$

The relaxation then again consists of taking the convex hull $\Delta'_l := \text{conv}\,\{e'^1, ..., e'^l\}$. Specifically,

$$\Delta'_l = \{a' \in \mathbb{R}^{l-1} | 1 \geqslant a'_1 \geqslant a'_2 \geqslant ... \geqslant a'_l \geqslant 0\}. \tag{2.16}$$

This can be seen as a reparametrization of the unit simplex using $l-1$ coordinates. The non-increasingness property of $a' \in \Delta'_l$ corresponds to the nonnegativity of $a \in \Delta_l$. A constraint similar to the sum condition in Δ_l is then enforced by special boundary conditions. While this parametrization has an intuitive interpretation when the labels represent a quantized range of values (Sect. 2.7.3), we feel that formulation (2.6) provides a clearer view on the problem and simplifies the theoretical treatment, cf. Chap. 5. Moreover, the analysis in [CCP08] is restricted to potentials of the form $\gamma(|i-j|)$ for nondecreasing, positive, concave functions γ, while we consider arbitrary metrics.

2.3 Properties of the Interaction Potential

We begin by formalizing the requirements on the regularizer of the relaxed problem as mentioned in the introduction. Let us assume we are given a general interaction potential $d\colon \mathcal{I}^2 \to \mathbb{R}$. Intuitively, $d(i, j)$ assigns a weight to switching between label i and label j. We require

$$d(i,j) > 0, \quad i \neq j, \tag{2.17}$$

but no other metric properties (i.e. symmetry or triangle inequality) for now. Within this work, we postulate that the regularizer should satisfy:

(P1). J is convex and positively homogeneous on $\text{BV}(\Omega)^l$.

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

(P3). For any partition $(S, \Omega \setminus S)$ of Ω into two sets with finite perimeter $\text{Per}(S) < \infty$, and any $i, j \in \{1, ..., l\}$,

$$J(e^i \chi_S + e^j \chi_{\Omega \setminus S}) = d(i,j)\text{Per}(S), \tag{2.18}$$

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

The requirement (P1) is sensible in order to render global optimization tractable. Indeed, if J is convex, the overall objective function (2.6) will be convex as well due to the linearization of the data term. Positive homogeneity is included as it allows J to be represented in terms of a support function (i.e. $J = \sigma_{\mathcal{D}}$ for some closed convex set $\mathcal{D}$), which will be exploited by our optimization methods.

Requirements (P3) and (P2) formalize the principle that the multilabeling problem should reduce to the classical continuous cut (2.8) when restricted to two classes. Together, these requirements pose a natural restriction on the interaction potential d:

Proposition 2.1. *Let $(J,\ d)$ satisfy (P1) – (P3) as well as the general assumption (2.17). Then d must necessarily be a metric, i.e. for all $i,j,k \in \{1,\dots,l\}$,*

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

Proof. 1. follows from (P2) and (P3) by choosing $i = j$ and S with $\mathrm{Per}(S) > 0$. Symmetry in 2. is obtained from (P3) by replacing S with $\Omega \setminus S$, since $\mathrm{Per}(S) = \mathrm{Per}(\Omega \setminus S)$; the definiteness $d(i,j) > 0$ follows from the assumption (2.17). To show 3., first note that $J(u) = 2J(u/2 + c/2)$ for any constant $c \in \mathbb{R}^l$ and all $u \in \mathrm{BV}(\Omega)^l$, since

$$\begin{aligned} J(u) &= 2\,J((u+c)/2 - c/2) && (2.19)\\ &\leqslant J(u+c) + J(-c) = 2\,J(u/2 + c/2) && (2.20)\\ &\leqslant J(u) + J(c) = J(u) && (2.21) \end{aligned}$$

Fix any set S with perimeter

$$c := \mathrm{Per}(S) > 0. \tag{2.22}$$

Then, using the above mentioned fact and the positive homogeneity of J,

$$\begin{aligned} cd(i,k) &= J(e^i\chi_S + e^k\chi_{\Omega\setminus S}) && (2.23)\\ &= 2J\left(\frac{1}{2}(e^i\chi_S + e^k\chi_{\Omega\setminus S}) + \frac{1}{2}e^j\right) && (2.24)\\ &= 2J\left(\frac{1}{2}(e^i\chi_S + e^j\chi_{\Omega\setminus S}) + \frac{1}{2}(e^j\chi_S + e^k\chi_{\Omega\setminus S})\right) && (2.25)\\ &\leqslant J(e^i\chi_S + e^j\chi_{\Omega\setminus S}) + J(e^j\chi_S + e^k\chi_{\Omega\setminus S}) && (2.26)\\ &= c(d(i,j) + d(j,k)). && (2.27) \end{aligned}$$

□

Note that the general assumption (2.17) is only required for the definiteness. If the positivity requirement (2.17) is dropped, d must still be a semi-metric and it is easy to show that if $d(i,j) = 0$ for some $i \neq j$, then $d(i,k) = d(j,k)$ for *any* k. In this case the classes i and j can be collapsed into a single class as far as the regularizer is concerned. The decision between label i and j is then completely local, i.e. it depends only on the data term and can be postponed to a post-processing step (assuming that a minimizer exists) by modifying the data term to

$$s'_i(x) := s'_j(x) \ := \ \min\{s_i(x), s_j(x)\}. \tag{2.28}$$

Thus the positivity condition (2.17) is not a real limitation and can be always assured. As a side note, it can be shown that, under some assumptions and in the space of piecewise constant functions, the subadditivity of d also follows if J is required to be lower semicontinuous [Bra02, p.88].

It is worth remarking that for graph-based, finite-dimensional models, non-metric potentials are widely in use. In the continuous setting this does not make sense: Consider, for example, a potential d with $d(i,j) < 0$ for some $i \neq j$, and assume that the data term vanishes in some region. Then the objective favors longer interfaces within this region. Due to the continuity of the image domain Ω, the interface can be made arbitrarily long. Therefore the objective is not bounded from below, and the problem is not well-posed. The difference to the finite-dimensional case can be attributed to the fact that in some sense, the finite-dimensional case imposes the additional constraint that regions should have a minimal size.

Proposition 2.1 implies that for non-metric d we generally cannot expect to find a regularizer satisfying (P1)–(P3). Note also that J is not required to be of any particular (e.g. integral) form. In Sect. 2.5.1 we will show that on the other hand, if d is metric as in Proposition 2.1, a suitable regularizer can always be constructed. This implies that the interaction potentials allowing for a regularizer that satisfies (P1)–(P3) are exactly the metric potentials.

2.4 Existence of Minimizers

The complete problem considered here is of the form (cf. (2.6) and (2.12))

$$\inf_{u \in \mathcal{C}} f(u), \quad f(u) := \int_\Omega \langle u, s \rangle d\,x + J(u) \tag{2.29}$$

where $J(u) = \int_\Omega d\Psi(Du)$ as in (2.10), and $\mathcal{C} = \mathrm{BV}(\Omega, \Delta_l)$. Note that f is convex, as it is the pointwise supremum of affine functions (2.13). For simplicity we generally assume $\Omega = (0,1)^d$. Then we have the following

Proposition 2.2. *Let Ψ be positively homogeneous, continuous and convex such that $0 \leqslant \Psi \leqslant \rho_u \|\cdot\|_2$ for some $\rho_u > 0$. Moreover, let $s \in L^\infty(\Omega)^l$, and*

$$f(u) = \int_\Omega \langle u, s \rangle d\,x + \int_\Omega d\Psi(Du). \tag{2.30}$$

Then f is lower semicontinuous in $\mathrm{BV}(\Omega)^l$ with respect to L^1 convergence.

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

$$\tilde{f}(u) := \int_\Omega \langle u, s \rangle d\,x + \int_\Omega \Psi(\nabla u(x)) d\,x, \tag{2.31}$$

on $\mathrm{BV}(\Omega)^l$ with respect to L^1_{loc} (and therefore L^1) convergence and thus lower semicontinuous in $\mathrm{BV}(\Omega)^l$. To apply the theorem, we have to show that Ψ is quasiconvex in the sense of [AFP00, Def. 5.25], which holds as it is convex. The other precondition is (at most) linear growth of Ψ, which holds due to the assumption $0 \leqslant \Psi(x) \leqslant \rho_u \|x\|_2$. □

Proposition 2.3. *Let f, Ψ, s as in Prop. 2.2 and additionally assume that*

$$\Psi(z) \;\geqslant\; \rho_l \|z\|_2 \quad \forall z = (z^1|...|z^l) \in \mathbb{R}^{d\times l} \text{ s.t. } \sum_{i=1}^{l} z^i = 0. \tag{2.32}$$

Then the problem

$$\min_{u\in\mathcal{C}} f(u), \quad \mathcal{C} := \mathrm{BV}(\Omega, \Delta_l) \tag{2.33}$$

has at least one minimizer.

Proof. The constraint $u \in \mathcal{C}$ implies that the distributive derivative $Du = (Du_1|...|Du_l)$ satisfies $Du_1 + ... + Du_l = 0$, since the mapping $u \mapsto Du$ is linear and $D(e^\top u) = 0$ due to the constraint. Therefore the density function $(Du/|Du|)$ satisfies $(Du/|Du|)\, e = 0$ in a $|Du|$-a.e. sense (cf. [AFP00, Cor. 1.29]), which allows to apply the assumption (2.32) to obtain $\rho_l \leqslant \Psi(Du/|Du|) \leqslant \rho_u$ on $\mathcal{C}$ (again in a $|Du|$-a.e. sense). Consequently, we conclude from (A.42) that

$$0 \leqslant \rho_l \mathrm{TV}(u) \leqslant J(u) \leqslant \rho_u \mathrm{TV}(u). \tag{2.34}$$

From

$$\int_\Omega \langle u, s\rangle dx \geqslant -\int_\Omega \|u(x)\|_1 \|s(x)\|_\infty dx, \tag{2.35}$$

the fact that $s \in L^\infty(\Omega)^l$ and $u \in L^1(\Omega)^l$, it follows that the data term is bounded from below.

We now show coercivity of f with respect to the BV norm: Let $(u^{(k)}) \subset \mathcal{C}$ with $\|u^{(k)}\|_1 + \mathrm{TV}(u^{(k)}) \to \infty$. Then, since u and therefore $\|u^{(k)}\|_1$ is bounded, it follows that $\mathrm{TV}(u^{(k)}) \to \infty$. Then $f(u^{(k)}) \to +\infty$ as well, since the data term $\langle u, s\rangle$ is bounded from below and $J(u^{(k)}) \geqslant \rho_l \mathrm{TV}(u^{(k)})$. Thus f is coercive.

Equations (2.34) and (2.35) also show that f is bounded from below, thus there must be a minimizing sequence $(u^{(k)})$. Due to the coercivity, the sequence $\|u^{(k)}\|_1 + \mathrm{TV}(u^{(k)})$ must then be bounded from above, i.e. the sequence $(u^{(k)})$ is bounded in the BV norm. From this and [AFP00, Thm. 3.23] we conclude that there is a subsequence of $(u^{(k)})$ weakly*- (and thus L^1-) converging to some $u \in \mathrm{BV}(\Omega)^l$. With the lower semicontinuity from Prop. 2.2 and closedness of $\mathcal{C}$ with respect to L^1 convergence, existence of a minimizer follows. □

2.5 Regularizers for Specific Interaction Potentials

The following proposition provides the connection between the integrand Ψ and the interaction potential d in view of (P3).

Proposition 2.4. *Let Ψ be as in Prop. 2.2, and additionally isotropic:*

$$\Psi(Rz) \;=\; \Psi(z) \quad \forall R \in \mathrm{SO}(d). \tag{2.36}$$

For some $u' \in \mathrm{BV}(\Omega)$ and vectors $a, b \in \Delta_l$, define $u(x) = (1-u'(x))a + u'(x)b$. Then, for any normal $y \in S^{d-1}$,

$$J(u) \;=\; \Psi(y(b-a)^\top)\mathrm{TV}(u') = \left(\sup_{v \in \mathcal{D}_{\mathrm{loc}}} \|v(b-a)\|_2\right)\mathrm{TV}(u'). \tag{2.37}$$

In particular, if for all $i, j \in \mathcal{I}$ there exists $y \in S^{d-1}$ such that $\Psi(y(e^i - e^j)^\top) = d(i,j)$, then J fulfills (P3).

Proof. To show the first equality, we conclude from (A.42) that

$$\begin{aligned} J(u) \;&=\; \int_\Omega \Psi\left(\frac{Du}{|Du|}\right) d\,|Du| && (2.38)\\ &=\; \int_\Omega \Psi\left(\frac{D(a+u'(b-a))}{|D(a+u'(b-a))|}\right) d\,|D(a+u'(b-a))| && (2.39)\\ &=\; \int_\Omega \Psi\left(\frac{(Du')(b-a)^\top}{|(Du')(b-a)^\top|}\right) d\,|(Du')(b-a)^\top|. && (2.40) \end{aligned}$$

We now use the property $|(Du')(b-a)^\top| = |Du'|\|b-a\|_2$, which is a direct consequence of the definition of the total variation measure and the fact that $\|w(b-a)^\top\|_2 = \|w\|_2\|b-a\|_2$ for any vector $w \in \mathbb{R}^d$ (note that $a, b \in \mathbb{R}^l$ are also vectors). Therefore

$$J(u) \;=\; \int_\Omega \Psi\left(\frac{(Du')(b-a)^\top}{|Du'|\|b-a\|_2}\right) d\,|Du'|\|b-a\|_2, \tag{2.41}$$

which by positive homogeneity of Ψ implies

$$J(u) \;=\; \int_\Omega \Psi\left(\frac{Du'}{|Du'|}(b-a)^\top\right) d\,|Du'|. \tag{2.42}$$

Since the density function $Du'/|Du'|$ assumes only values in S^{d-1}, there exists, for any $y \in S^{d-1}$ and $|Du'|$-a.e. $x \in \Omega$, a rotation matrix mapping $(Du'/|Du'|)(x)$ to y. Together with the rotational invariance of Ψ from (2.14) this implies

$$J(u) \;=\; \int_\Omega \Psi(y(b-a)^\top) d\,|Du'| = \Psi(y(b-a)^\top)\mathrm{TV}(u'), \tag{2.43}$$

which proves the first equality in (2.37). The second equality can be seen as follows:

$$\begin{aligned} r \;&:=\; \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \|v(b-a)\|_2 && (2.44)\\ &=\; \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \sup_{z \in \mathbb{R}^d, \|z\|_2 \leqslant 1} \langle z, v(b-a)\rangle && (2.45)\\ &=\; \sup_{z \in \mathbb{R}^d, \|z\|_2 \leqslant 1} \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \langle z, v(b-a)\rangle && (2.46)\\ &=\; \sup_{z \in \mathbb{R}^d, \|z\|_2 \leqslant 1} \sup_{v \in \mathcal{D}_{\mathrm{loc}}} \langle z\,(b-a)^\top, v\rangle && (2.47)\\ &=\; \sup_{z \in \mathbb{R}^d, \|z\|_2 \leqslant 1} \Psi(z\,(b-a)^\top). && (2.48) \end{aligned}$$

Denote by R_z a rotation matrix mapping z to y, i.e. $R_z z = y$, then due to the rotational invariance

$$\begin{aligned} r &= \sup_{z\in\mathbb{R}^d, \|z\|_2\leqslant 1} \Psi(R_z{}^\top R_z\, z\,(b-a)^\top) && (2.49)\\ &= \sup_{z\in\mathbb{R}^d, \|z\|_2\leqslant 1} \Psi(R_z\, z\,(b-a)^\top) && (2.50)\\ &= \sup_{z\in\mathbb{R}^d, \|z\|_2\leqslant 1} \Psi(y\,(b-a)^\top) && (2.51)\\ &= \Psi(y\,(b-a)^\top). && (2.52) \end{aligned}$$

The second part of the assertion follows directly by setting $u' = \chi_S$, $a = e^i$ and $b = e^j$. □

As a consequence, if the relaxed multiclass formulation is restricted to two classes by parametrizing $u = (1-u')a + u'b$ for $u'(x) \in [0,1]$, it essentially reduces to the scalar continuous cut problem (2.8): solving

$$\min_{u'\in \mathrm{BV}(\Omega,[0,1])} \int_\Omega \langle (1-u')a + u'b, s\rangle d\,x + J(u) \tag{2.53}$$

is equivalent to solving

$$\min_{u'\in \mathrm{BV}(\Omega,[0,1])} \int_\Omega u'(b-a) d\,x + \Psi(y(b-a)^\top)\mathrm{TV}(u'), \tag{2.54}$$

which is just the classical two-class continuous cut approach with data $(b-a)$ and regularizer weight $\Psi(y(b-a)^\top)$, where $y \in \mathbb{R}^d$ is some arbitrary unit vector. For the multiclass case, assume that

$$u = u_P \;=\; e^1\chi_{\Omega_1} + \ldots + e^l\chi_{\Omega_l} \tag{2.55}$$

for some partition $\Omega_1 \cup \ldots \cup \Omega_l = \Omega$ with $\mathrm{Per}(\Omega_i) < \infty$, $i = 1, \ldots, l$. Then the absolutely continuous and Cantor parts of Du_P vanish [AFP00, Thm. 3.59, Thm. 3.84, Rem. 4.22], and only the jump part remains:

$$J(u_P) \;=\; \int_{S_{u_P}} \Psi(\nu_{u_P}(u_P^+ - u_P^-)^\top) d\mathcal{H}^{d-1}, \tag{2.56}$$

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

$$J(u_P) \;=\; \int_{S_{u_P}} \Psi\big(\nu_{u_P}\big(e^{i(x)} - e^{j(x)}\big)^\top\big) d\mathcal{H}^{d-1} = \int_{S_{u_P}} d(i(x), j(x))\, d\mathcal{H}^{d-1}. \tag{2.57}$$

Thus the regularizer locally penalizes jumps between labels i and j along an interface with the interface length, multiplied by the factor $d(i,j)$ depending on the labels of the adjacent regions.

The question is now how to choose Ψ such that $\Psi(y(e^i - e^j)^\top) = d(i,j)$ for a prescribed interaction potential d. We consider three approaches which differ with respect to expressiveness and simplicity of use: In the *local envelope* approach (Sect. 2.5.1), $\mathcal{D}_{\text{loc}}$ is chosen as large as possible. In turn, J is as large as possible in the integral formulation (2.10). This prevents introducing artificial minima generated by the relaxation, and potentially keeps minimizers of the relaxed problem close to minimizers of the original problem. However, Ψ is only implicitly defined, which complicates optimization. In contrast, in the *embedding* approach (Sect. 2.5.2), $\mathcal{D}_{\text{loc}}$ is simpler at the cost of being able to represent only a subset of all metric potentials. In the third approach, the regularizer is deliberately constructed in order to emphasize uncertainty (Sect. 2.5.3). For an illustration of the three approaches, see Fig. 2.4 and Fig. 2.5.

In order to be able to classify the regularizers derived in the following sections, we will briefly state some terminology regarding properties of the regularizer:

Isotropy. As indicated in the introduction, regularizers are considered *isotropic* if they are invariant under coordinate transformations by rotation matrices:

$$\Psi(R\,z) \;=\; \Psi(z) \quad \forall R \in \mathrm{SO}(d) \tag{2.58}$$

This is obviously the case if and only if $\Psi(\nu x^\top) = \Psi(e^1 x^\top)$ for any $\nu \in S^{d-1}$. If Ψ is isotropic with $\Psi \geqslant \rho_l \, \|\cdot\|_2$, according to Prop. 2.4 we may define the corresponding interaction potential

$$d(i,j) \;:=\; \Psi(e^1(e^i - e^j)^\top). \tag{2.59}$$

This is indeed a metric, since $\Psi(e^1(e^i - e^j)^\top) = \Psi(-e^1(e^i - e^j)^\top) = \Psi(e^1(e^j - e^i))$ due to the isotropy, and therefore $d(i,j) = d(j,i)$. From convexity and positive homogeneity of Ψ it follows that d must be subadditive, and from the lower bound ρ_l we get the positivity $d(i,j) = 0 \Leftrightarrow i = j$.

Permutation Invariance. We call some regularizer *permutation invariant*, if it is invariant with respect to permutations of the elements of u, i.e. of the label set. In terms of Ψ, invariance is given if $\Psi(z) = \Psi(z\,P)$ for any permutation matrix $P \in \mathbb{R}^{l \times l}$.

Separability. If Ψ can be written as a sum of terms that depend only on individual components or directional derivatives of u, it is called *separable* in the components of u or in space, respectively. Separability usually simplifies optimization, as it reduces the coupling between variables.

Homogeneity. Instead of (2.10), which is invariant under translation of the coordinates, i.e. *homogeneous*, it is also possible to consider regularizers with an additional dependency on x,

$$J(u) \;=\; \int_\Omega d\Psi_x(D\,u). \tag{2.60}$$

This often occurs in denoising applications where Ψ_x includes an anisotropy which is controlled by local properties of the input or of the current iterate. An even more general approach is to set

$$J(u) \;=\; \int_\Omega dg(x, u, D\,u), \tag{2.61}$$

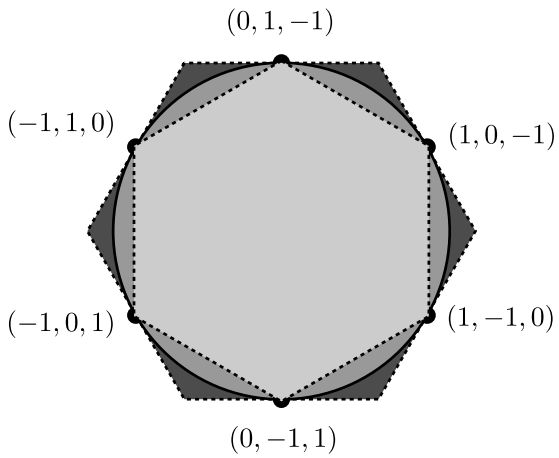

Figure 2.4. Illustration of the set $\mathcal{D}_{\text{loc}}$ used to build the regularizer for the (scaled) uniform metric, for $l = 3$ labels in $d = 1$ – dimensional space. Shown is a cut through the $z^1 + z^2 + z^3 = 0$ plane; the labels correspond to the points $e^i - e^j$ with $i \neq j$. The local envelope method leads to a larger set $\mathcal{D}_{\text{loc}}$ (dashed, outer) than the embedding method (solid). This improves the tightness of the relaxation, but requires more expensive optimization. The "uncertainty" method (dashed, inner) results in the smallest dual set, and thus the least tight relaxation (Fig. 2.5).

Figure 2.5. Tightness of different regularizers for the task of simultaneous inpainting and denoising of a three-color image. The data term was set to zero in the square around the center. **Left to right:** Input image; solutions of the relaxed problem for the "envelope" regularizer (Sect. 2.5.1), "embedding" regularizer (Sect. 2.5.2), and "uncertain" regularizer (Sect. 2.5.3). As the tightness of the relaxation decreases from left to right, the number of fractional labels in the solution increases.

however this considerably complicates the conditions for the existence of solutions [AFP00, Chap. 5].

2.5.1 Relaxation Based on the Local Envelope

In this section, we propose a formulation based on [CCP08], where the authors consider an approach for potentials d of the form $d(i, j) = \gamma(|i - j|)$ for a positive concave function γ. The approach is derived by specifying the value of J on the set of integral u only, and constructing an approximation of the convex envelope by a local approach, i.e. by computing the convex envelope of the *integrand* Ψ. This approach potentially generates tight relaxations and thus one may hope that the convexification process does not generate too many artificial non-integral solutions.

We propose to extend this approach to *arbitrary metric* d by setting

$$\Psi_d := \sigma_{\mathcal{D}^d_{\text{loc}}},\ \mathcal{D}^d_{\text{loc}} \ := \ \bigcap_{i\neq j} \{v=(v^1,...,v^l)\in\mathbb{R}^{d\times l} | \|v^i-v^j\|_2 \leqslant d(i,j), \sum_k v^k = 0\} \tag{2.62}$$

for some given interface potential $d(i,j)$ (cf. Fig. 2.4). This formulation can be derived as follows. We consider only the integrand Ψ, and postulate that for any normal $y\in S^{d-1}$, and any two label indices $i,j\in\mathcal{I}$,

$$\Psi(y\,(e^i-e^j)^\top) \ \stackrel{!}{=}\ d(i,j). \tag{2.63}$$

In the light of Prop. 2.4, this means that J fulfills (P3). We additionally enforce positive homogeneity by requiring $\Psi(y\,(e^i-e^j)^\top) = \|y\|_2\, d(i,j)$ for any $y\in\mathbb{R}^d$, and define

$$\Psi'_d(w) := \begin{cases} \|y\|_2\, d(i,j), & \exists i,j\in\mathcal{I}, y\in\mathbb{R}^d\colon y=w^i=-w^j\,,\, w^k=0\ \forall k\notin\{i,j\}\,, \\ +\infty, & \text{otherwise.} \end{cases}$$

The convex envelope of Ψ'_d can be found by computing the Legendre-Fenchel biconjugate. For the first conjugate, we obtain

$$\begin{aligned} \Psi'^*_d(v) \ &= \ \sup_{w\in\mathbb{R}^{d\times l}} (\langle v,w\rangle - \Psi'_d(w)) && (2.64)\\ &= \ \sup_{w\in\mathbb{R}^{d\times l}} \begin{cases} \langle v^i-v^j,y\rangle - \|y\|_2\, d(i,j), & \exists i,j\in\mathcal{I}, y\in\mathbb{R}^d\colon y=w^i=-w^j, \\ & \quad w^k=0\ \forall k\notin\{i,j\} \\ -\infty, & \text{otherwise,} \end{cases} \\ &= \ \sup_{y\in\mathbb{R}^d, i,j\in\mathcal{I}} \{\langle v^i-v^j,y\rangle - \|y\|_2\, d(i,j)\}. && (2.65) \end{aligned}$$

If $\|v^i-v^j\|_2 > d(i,j)$ for *any* $i,j\in\mathcal{I}$, the supremum is $+\infty$ (choose $y=c\,(v^i-v^j)$ and let $c\to\infty$). On the other hand, if $\|v^i-v^j\|_2\leqslant d(i,j)$ holds for *all* $i,j\in\mathcal{I}$, then necessarily $\langle v^i-v^j,y\rangle \leqslant |\langle v^i-v^j,y\rangle| \leqslant \|y\|_2\,\|v^i-v^j\|_2 \leqslant \|y\|_2\, d(i,j)$, therefore the inner expression in (2.65) is nonpositive, and the supremum is zero (choose $y=0$). We conclude that Ψ'^*_d is the indicator function

$$\Psi'^*_d(v) = \delta_{\|v^i-v^j\|_2\leqslant d(i,j)\ \forall i,j\in\mathcal{I}}(v). \tag{2.66}$$

The desired convex envelope Ψ_d of Ψ'_d is therefore

$$\Psi_d(w) \ = \ (\Psi'^*_d)^*(w) = \sup_{\|v^i-v^j\|_2\leqslant d(i,j)\ \forall i,j\in\mathcal{I}} \langle w,v\rangle. \tag{2.67}$$

Since Ψ_d is only applied to the density $Du/|Du|$, which satisfies $(Du/|Du|)\,e=0$ due to the simplex constraint $u\in\mathcal{C}=\text{BV}(\Omega,\Delta_l)$, we may include the additional constraint $v^1+...+v^l=0$, and obtain

$$\Psi_d(w) \ = \ \sigma_{\mathcal{D}^d_{\text{loc}}}(w). \tag{2.68}$$

Such Ψ_d satisfies the lower and upper boundedness required for the existence of a minimizer in Prop. 2.3, however evaluating Ψ_d is nontrivial. Nevertheless, since its conjugate Ψ^*_d is known, optimization is still possible using primal-dual techniques (Chap. 4).

It remains to show that Ψ_d has the desired property (2.63), i.e. it coincides with Ψ'_d where the latter is finite. The inequality $\Psi_d\leqslant\Psi'_d$ always holds since Ψ_d is the convex envelope of Ψ'_d, however the converse has to be shown. We first show that any metric d can be reconstructed from $\mathcal{D}^d_{\text{loc}}$:

Proposition 2.5. *Let $d\colon \{1,\dots,l\}^2 \to \mathbb{R}_{\geqslant 0}$ be a metric. Then for any i, j,*

$$\sup_{v \in \mathcal{D}^d_{\mathrm{loc}}} \left((v^i)_1 - (v^j)_1\right) \;=\; d(i,j). \tag{2.69}$$

Proof. "$\leqslant$" follows from the definition (2.62). "$\geqslant$" can be shown using a network flow argument:

$$\begin{aligned}
&\sup_{v \in \mathcal{D}^d_{\mathrm{loc}}} \left((v^i)_1 - (v^j)_1\right) && (2.70)\\
\geqslant\;& \sup\left\{p_i - p_j \mid p \in \mathbb{R}^l \colon e^\top p = 0, \forall i', j' \colon p_{i'} - p_{j'} \leqslant d(i',j')\right\} && (2.71)\\
\overset{(*)}{=}\;& \sup\left\{p_i - p_j \mid p \in \mathbb{R}^l \colon \forall i', j' \colon p_{i'} - p_{j'} \leqslant d(i',j')\right\} && (2.72)\\
\overset{(**)}{=}\;& d(i,j). && (2.73)
\end{aligned}$$

Equality $(*)$ holds since each p in the set in (2.72) can be associated with the vector $\tilde{p} := p - \frac{1}{l}\sum_k p_k$, which is contained in the set in (2.71) and satisfies $p_i - p_j = \tilde{p}_i - \tilde{p}_j$. The last equality $(**)$ follows from [Mur03, 5.1] with the notation $\gamma = d$ (and $\bar{\gamma} = d$, since d is a metric and therefore the triangle inequality implies that the length of the shortest path from i to j is always $d(i,j)$). □

The main result of this section is the following:

Proposition 2.6. *Let $d\colon \mathbb{R}^{l\times l} \to \mathbb{R}_{\geqslant 0}$ be a metric. Define Ψ_d as in (2.62), and*

$$J_d(u) \;:=\; \int_\Omega d\Psi_d(D\,u). \tag{2.74}$$

Then J_d satisfies (P1)–(P3).

Proof. (P1) and (P2) are clear from the definition of J_d. (P3) follows directly from Prop. 2.5 and Prop. 2.4 with $y = e^1$. □

Defining $\mathcal{D}^d_{\mathrm{loc}}$ as in (2.62) provides us with a way to extend the desired regularizer for *any* metric d to non-integral $u \in \mathcal{C}$ via (2.12). The price to pay is that there is no simple closed expression for Ψ_d and thus for J_d, which potentially complicates optimization.

By construction, Ψ_d is isotropic. Permutation invariance only occurs if d is a scaled uniform metric, $d(i,j) = c\,\chi_{\{i\neq j\}}$ for some $c > 0$. Note that in order to define Ψ_d, d does not have to be a metric. However Prop. 2.5 then cannot be applied and (2.63) does not hold, so J is not a true extension of the desired regularizer, although it still provides a lower bound.

2.5.2 Relaxation Based on Embeddings

In this section, we consider a regularizer which is less powerful but more efficient to evaluate. Recall the classical total variation for vector-valued u as defined in (A.2). Using $\Psi = \|\cdot\|_2$, it can be written as

$$\mathrm{TV}(u) = \int_\Omega d\|Du\|_2. \tag{2.75}$$

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

$$J_A(u) \;:=\; \mathrm{TV}(A\,u). \tag{2.76}$$

This corresponds to substituting the Frobenius matrix norm on the distributive gradient with a linearly weighted variant. In the framework of (2.12), it amounts to setting $\mathcal{D}_{\mathrm{loc}} = \mathcal{D}_{\mathrm{loc}}^A$ (cf. Fig. 2.4) with

$$\mathcal{D}_{\mathrm{loc}}^A \;:=\; \{v'A \,|\, v' \in \mathbb{R}^{d\times k}, \|v'\|_2 \leqslant 1\} = \mathcal{B}_1(0)A. \tag{2.77}$$

Clearly $0 \in \mathcal{D}_{\mathrm{loc}}^A$ and

$$\begin{aligned} \Psi_A(z) \;&=\; \sigma_{\mathcal{D}_{\mathrm{loc}}^A}(z) = \sup_{v'\in\mathcal{B}_1(0)A} \langle z, v'\rangle = \sup_{v\in\mathcal{B}_1(0)} \langle z, v\,A\rangle && (2.78)\\ &=\; \sup_{v\in\mathcal{B}_1(0)} \langle z\,A^\top, v\rangle = \|z\,A^\top\|_2. && (2.79) \end{aligned}$$

In particular, we have

$$\Psi_A(D\,u) \;=\; \|(D\,u)A^\top\|_2 = \|D(A\,u)\|_2, \tag{2.80}$$

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

$$\mathrm{TV}_A(u) \;=\; \int_\Omega d\sqrt{\|D_1 u\|_A^2 + \ldots + \|D_d u\|_A^2}, \tag{2.81}$$

with $\|w\|_A := (w^\top A^\top A\, w)^{1/2}$. Therefore the approach can be understood as replacing the Euclidean norm by a linearly weighted, Mahalanobis-type variant.

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

Proposition 2.7. *Let d be an* Euclidean metric*, i.e. there exist $k \in \mathbb{N}$, $a^1, \ldots, a^l \in \mathbb{R}^k$ such that $d(i,j) = \|a^i - a^j\|_2$, and define $A := (a^1|\ldots|a^l)$. Then the regularizer $J_A := \mathrm{TV}_A$ satisfies (P1)–(P3).*

Proof. (P1) and (P2) are clearly satisfied. In order to show (P3) we apply Prop. 2.4 and assume $\|y\|_2 = 1$ to obtain

$$\Psi_A(y(e^i - e^j)^\top) \overset{(2.79)}{=} \|y(e^i - e^j)^\top A^\top\|_2 = \|y(a^i - a^j)^\top\|_2 \overset{\|y\|_2=1}{=} \|a^i - a^j\|_2. \tag{2.82}$$

□

The class of Euclidean metrics comprises some important special cases:

- The *uniform*, *discrete* or *Potts* metric as also considered in [ZGFN08, LKY+09] and as a special case in [KT99, KT07]. Here $d(i,j) = 0$ iff $i = j$ and $d(i,j) = 1$ in any other case, which corresponds to $A = (1/\sqrt{2})I$.

- The *linear* (label) metric, $d(i,j) = c|i-j|$, with $A = (c, 2c, ..., lc)$. This regularizer is suitable to problems where the labels can be naturally ordered, e.g. depth from stereo or grayscale image denoising.

- More generally, if label i corresponds to a prototypical vector z^i in k-dimensional feature space and the Euclidean norm is an appropriate metric on the features, it is natural to set $d(i,j) = \|z^i - z^j\|_2$, which is Euclidean by construction. This corresponds to a regularization in feature space, rather than in "label space".

Note that the lower boundedness condition (2.32) involving $\Psi \geqslant \rho_l \, \|\cdot\|_2$ required for the existence proof (Prop. 2.3) is only fulfilled if the kernel of A is sufficiently small, i.e. $\ker A \subseteq \{t\ e | t \in \mathbb{R}\}$, with $e = (1, ..., 1)^\top \in \mathbb{R}^l$: otherwise, the constraint set $\mathcal{D}^A_{\mathrm{loc}} = (\mathcal{B}_1(0)A) \cap \{(v^1, ..., v^l) | \sum_i v^i = 0\}$ is contained in a subspace of at most dimension $(l-2)d$, and (2.32) cannot be satisfied for any $\rho_l > 0$. Thus if d is a degenerate Euclidean metric which can be represented by an embedding into a lower-dimensional space, as is the case with the linear metric, it has to be regularized for the existence result in Prop. 2.3 to hold. This can for example be achieved by choosing an orthogonal basis $(b^1, ..., b^j)$ of $\ker A$, where $j = \dim \ker A$, and substituting A with the matrix $A' := (A^\top, \varepsilon b^1, ..., \varepsilon b^j)^\top$, enlarging k as required. However these observations are mostly of theoretical interest, since for the discretized problem, the existence of minimizers follows already from compactness of the (finite-dimensional) discretized constraint set.

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

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

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

$$\min_{S \in \mathcal{P}_{l-1}} \|B(S) - D\|_M. \tag{2.83}$$

Problem (2.83) can be solved using available numerical solvers. Then $E = B(S) \in \mathcal{Q}_l$, and A can be extracted by factorizing $-\frac{1}{2} C\ E\ C$. Since both E and D are explicitly known, the quantity

$$\varepsilon_E \ := \ \max_{i,j} |(E_{ij})^{1/2} - (D_{ij})^{1/2}| \tag{2.84}$$

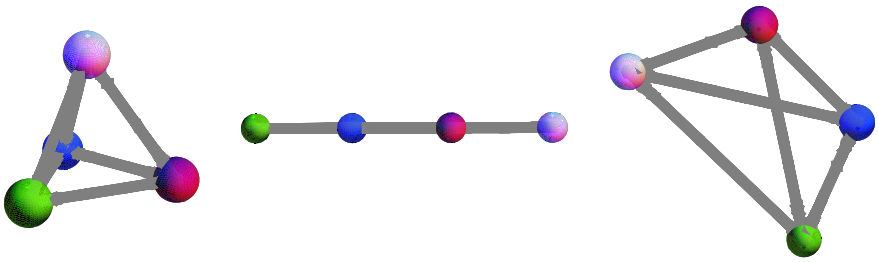

Figure 2.6. Euclidean embeddings into $\mathbb{R}^3$ for several interaction potentials with four classes. **Left to right:** Uniform metric; linear metric; non-Euclidean truncated linear metric. The vertices correspond to the columns $a^1, \ldots, a^l$ of the embedding matrix A. For the truncated linear metric an optimal approximate embedding was computed as outlined in Sect. 2.5.2 with the matrix norm $\|X\|_M := \max_{i,j} |X_{ij}|$.

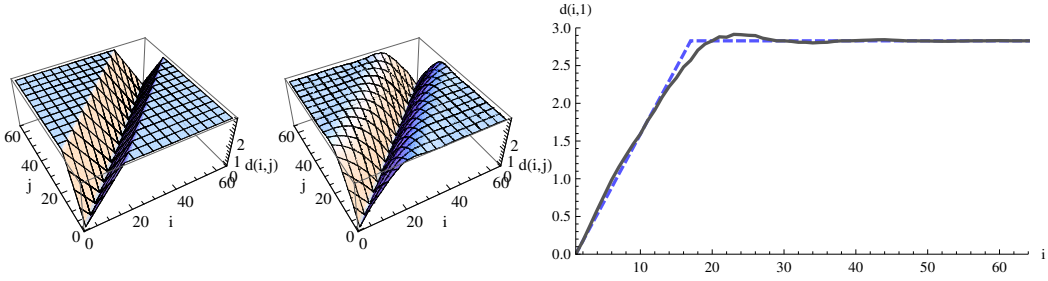

Figure 2.7. Euclidean approximation of the Non-Euclidean truncated linear metric with interaction potential $d(i,j) = (\sqrt{2}/8) \min\{|i-j|, 16\}$. **Left to right:** Original potential for 64 classes; potential after Euclidean approximation; cross section of the original (dashed) and approximated (solid) metric at $i = 1$. The approximation was computed using semidefinite programming as outlined in Sect. 2.5.2. It represents the closest Euclidean metric with respect to the matrix norm $\|X - Y\|_M := \sum_{i,j} |X_{ij} - Y_{ij}|$. The maximal elementwise error with respect to the original potential is $\varepsilon_E = 0.2720$.

can be computed and provides an *a posteriori* bound on the distortion due to the embedding. Fig. 2.6 shows a visualization of some embeddings for a four-class problem. In many cases, in particular when the number of labels is large, the Euclidean embedding provides a good approximation for non-Euclidean metrics (Fig. 2.7). Therefore, the Euclidean embedding approach can be used to solve approximations of the labeling problem with arbitrary metric interaction potentials. Compared to the envelope approach in Sect. 2.5.1 the relaxation is less tight, but the regularizer has a much simpler structure and can be evaluated in closed form.

The approach can also be generalized to embeddings into non-Euclidean spaces, such as ℓ^1. In fact, any regularizer Ψ can be modified by introducing an embedding matrix $A = (a^1|\ldots|a^l) \in \mathbb{R}^{k \times l}$ for some $k \leqslant l$, and defining

$$\Psi_A(D\,u) \;:=\; \Psi(D(A\,u)) = \Psi((D\,u)A^\top). \tag{2.85}$$

The approach preserves isotropy of the underlying norm, but neither permutation invariance nor separability. Applied to a jump from label i to label j, this results in the modified potential

$$\Psi_A(\nu(e^i - e^j)^\top) \;=:\; d_A(i,j), \tag{2.86}$$

and amounts to transforming the dual set $\mathcal{D}_{\text{loc}}$ of Ψ to $\mathcal{D}_{\text{loc}}\, A^\top$, similarly to (2.77). Alternatively, it is often easier to formally merge A into the linear gradient operator $D\,u$, which preserves the structure of the dual set and requires only few modifications to the optimization method (Chap. 4).

2.5.3 Relaxation with Emphasized Uncertainty

For illustration, we will consider also an extreme case where the regularizer is deliberately constructed such that it emphasizes uncertainty, i.e. it provokes non-integral solutions in areas where the data term does not dominate.

Consider the restriction of the dual set for the envelope method (2.62) to the one-dimensional case, i.e. $d=1$, for the uniform metric:

$$\begin{aligned}\mathcal{D}^d_{\text{loc}} &:= \{v=(v^1,...,v^l)\in\mathbb{R}^l \,|\, |v^i-v^j|\leqslant 1 \forall i\neq j\,,\, \sum_k v^k=0\} &(2.87)\\ &= \{v\in\mathbb{R}^l \,|\, (e^i-e^j)^\top v^\top \leqslant 1 \forall i\neq j\,,\, \sum_k v^k=0\}. &(2.88)\end{aligned}$$

In this restricted setting, it is obvious that $\mathcal{D}^d_{\text{loc}}$ is constructed by an intersection of affine half-spaces defined by hyperplanes with normals (e^i-e^j) through the points $(e^i-e^j)/2$. We ask the question what happens if $\mathcal{D}_{\text{loc}}$ is instead constructed as the *convex hull* of these points, see also Fig. 2.4. In a sense, this will create the *smallest* regularizer that still satisfies (P3).

For simplicity, we only consider the case where d is the uniform metric and $l=3$. In one dimension, we arrive at

$$\mathcal{D}^e_{\text{loc}} := \operatorname{conv} V\,, \tag{2.89}$$

$$V := \frac{1}{2}\left\{ y^1=\begin{pmatrix}-1\\1\\0\end{pmatrix}, y^2=\begin{pmatrix}-1\\0\\1\end{pmatrix}, y^3=\begin{pmatrix}0\\-1\\1\end{pmatrix}, -y^1, -y^2, -y^3 \right\}. \tag{2.90}$$

The points $v=(v^1,v^2,v^3)\in\mathcal{D}^e_{\text{loc}}$ are characterized by the inequalities

$$|2\,v^1-v^2-v^3|\leqslant\frac{3}{2},\;\; |-v^1+2\,v^2-v^3|\leqslant\frac{3}{2},\;\; |-v^1-v^2+2\,v^3|\leqslant\frac{3}{2}. \tag{2.91}$$

For $d\geqslant 2$ and $l=3$ we extend this definition in an isotropic way, similar to Ψ_d:

$$\begin{aligned}\Psi_e := \sigma_{\mathcal{D}^e_{\text{loc}}},\; \mathcal{D}^e_{\text{loc}} := \{v=(v^1,...,v^l)\in\mathbb{R}^{d\times l} \,|\, \sum_k v^k=0,\\ \|2\,v^1-v^2-v^3\|_2\leqslant\frac{3}{2}, \|-v^1+2\,v^2-v^3\|_2\leqslant\frac{3}{2}, \|-v^1-v^2+2\,v^3\|_2\leqslant\frac{3}{2}\}.\end{aligned} \tag{2.92}$$

This definition generates the rightmost result shown in Fig. 2.5 above. Interestingly, the regularizer still favors integral labels in regions where the data term – although noisy – contains a sufficient amount of information. However, in regions where the regularizer is largely the only source of information, it does not advocate a specific solution. This property could e.g. be useful in medical or biological applications where wrong classifications in some region due to corrupt or missing measurements potentially have a worse effect than no labeling at all. In such situations, an uncertainty-emphasizing regularizer provides an integrated way to detect these regions.

2.6 Optimality

As noted in the introduction above, [CEN06] showed that for the two-class continuous cut (2.8), thresholding u' at almost any threshold results in an *integral* optimal solution. A natural question is whether a similar relation exists for the multiclass case. In this section we will provide a unified view on the problem in the form of a generalized coarea formula, and show why some natural extensions of the scalar case fail.

2.6.1 The Two-Class Case

In order to motivate the following sections, we will give a short summary of the main theorem of [CEN06] adopted to our notation, together with some implications. Note that, as in the original work, the theorem is formulated for *scalar-valued* u, i.e. in the sense of two-class continuous cuts (2.8).

Theorem 2.8. *(Thresholding Theorem for the Two-Class Case) [CEN06, Thm. 2] Let* $s \in L^\infty(\Omega)$*, and* $f\colon \mathrm{BV}(\Omega) \to \mathbb{R}$ *defined by*

$$f(u) \;:=\; \int_\Omega u(x)\, s(x)\, dx + \int_\Omega d\, |Du|. \tag{2.93}$$

Assume that u^* *is a minimizer of* f *over the relaxed set* $\mathcal{C} := \mathrm{BV}(\Omega, [0,1])$*. Then, for almost every* $\alpha \in [0,1]$*, the* thresholded *function* $\bar{u}^*_\alpha := \chi_{\{u>\alpha\}}$ *is a minimizer of* f *over the set* $\mathcal{C}_{\{0,1\}} := \mathrm{BV}(\Omega, \{0,1\})$*.*

Remark 2.9. In [Ber09] it was shown, using lower-semicontinuity of f, that the optimality of $\bar{u}^*_\alpha$ actually holds for *every* $\alpha \in [0,1]$.

Proof. The proof [CEN06] of Thm. 2.8 centers around the "coarea-like" property,

$$\int_0^1 f(\bar{u}^*_\alpha)\, d\alpha \;=\; f(u^*). \tag{2.94}$$

The property can be shown separately for each of the terms in u. For the linear part it follows directly from Fubini's theorem or [AFP00, Prop. 1.78]. The corresponding relation for the regularizer,

$$\int_\Omega d|Du| \;=\; \int_0^1 \int_\Omega d\, |D\chi_{\{u>\alpha\}}|\, d\alpha = \int_0^1 \int_\Omega d\, |D\, \bar{u}^*_\alpha|\, d\alpha, \tag{2.95}$$

is exactly the coarea formula for BV functions (Thm. A.32).

We define the set of α violating the assertion, $\mathcal{S} := \{\alpha \in [0,1] | f(\bar{u}^*_\alpha) \neq f(u^*)\}$. Since $\bar{u}^*_\alpha \in \mathcal{C}_{\{0,1\}}$ and $\mathcal{C}_{\{0,1\}} \subseteq \mathcal{C}$, we have for any minimizer $u^*_{\{0,1\}}$ of f over $\mathcal{C}_{\{0,1\}}$,

$$f(u^*) \leqslant f(u^*_{\{0,1\}}) \leqslant f(\bar{u}^*_\alpha), \tag{2.96}$$

thus $\mathcal{S} = \{\alpha \in [0,1] | f(u^*) < f(\bar{u}^*_\alpha)\}$. Moreover, if $\alpha \notin \mathcal{S}$, then $f(u^*) = f(u^*_{\{0,1\}}) = f(\bar{u}^*_\alpha)$ by (2.96). Therefore, in order to show the theorem it suffices to show that $\mathcal{S}$ is an $\mathcal{L}^1$-zero set.

Assume the contrary holds, i.e. $\mathcal{L}^1(\mathcal{S}) > 0$. Then there must be $\varepsilon > 0$ such that

$$\mathcal{S}_\varepsilon \;:=\; \{\alpha \in [0,1] | \, f(u^*) \leqslant f(u^*_\alpha) - \varepsilon\} \tag{2.97}$$

has also nonzero measure, since otherwise $\mathcal{S}$ would be the countable union of zero measure sets, $\mathcal{S} = \bigcup_{i\in\mathbb{N}} \mathcal{S}_{1/n}$, and would consequently have zero measure as well. Then

$$\begin{aligned}
f(u^*) \quad &= \quad \int_{[0,1]\setminus\mathcal{S}_\varepsilon} f(u^*) d\alpha + \int_{\mathcal{S}_\varepsilon} f(u^*) d\alpha && (2.98)\\
&\overset{u^* \text{ optimal}}{\leqslant} \quad \int_{[0,1]\setminus\mathcal{S}_\varepsilon} f(\bar{u}^*_\alpha) d\alpha + \int_{\mathcal{S}_\varepsilon} f(u^*) d\alpha && (2.99)\\
&\overset{\text{definition of } \mathcal{S}_\varepsilon}{\leqslant} \quad \int_{[0,1]\setminus\mathcal{S}_\varepsilon} f(\bar{u}^*_\alpha) d\alpha + \int_{\mathcal{S}_\varepsilon} (f(\bar{u}^*_\alpha) - \varepsilon) d\alpha && (2.100)\\
&\overset{\text{linearity}}{=} \quad \int_0^1 f(\bar{u}^*_\alpha) d\alpha - \varepsilon \, \mathcal{L}^1(\mathcal{S}_\varepsilon). && (2.101)
\end{aligned}$$

But we assumed $\mathcal{L}^1(\mathcal{S}_\varepsilon) > 0$, therefore

$$f(u^*) \;<\; \int_0^1 f(\bar{u}^*_\alpha) d\alpha. \tag{2.102}$$

This is a contradiction to (2.94), therefore $\mathcal{L}^1(\mathcal{S}) = 0$ and the assertion follows. □

At the heart of the proof is the coarea-like property (2.94). It has the following intuitive interpretation:

1. The function u may be written in the form of a "generalized convex combination" of (an infinite number of) extreme points $E_u := \{\bar{u}_\alpha | \alpha \in [0,1]\}$ of the constraint set, i.e. the unit ball in $\mathrm{BV}(\Omega)$. As shown in [Fle57] based on a result by Choquet [Cho56], and noted in [Str83, p.127], extreme points of this constraint set are (multiples of, but in this case equal to) indicator functions.

2. The extreme points $(\bar{u}_\alpha)$ and coefficients in this convex combination can be explicitly found. In fact, the coefficients are all equal to $1/|[0,1]| = 1$, i.e. u is the barycenter of the points in E_u.

3. For any convex f, the inequality

$$\int_0^1 f(\bar{u}_\alpha) \, d\alpha \;\geqslant\; f(u) \tag{2.103}$$

always holds. The coarea formula (2.94) is therefore equivalent to the reverse inequality.

In fact, the original proof of the coarea formula [FR60] relies on showing (2.103) and using the fact that (2.94) holds for piecewise linear u [FR60, (1.5c)]. Approximating an arbitrary $u \in \mathrm{BV}(\Omega)$ by a sequence of piecewise linear functions, this result is then transported to the general case.

2.6.2 Generalized Coarea Formulas

Generalizing Thm. 2.8 to more than two labels hinges on a property similar to (2.94) that holds for *vector-valued* u. In a general setting, the question is whether there exist

- a probability space (Γ, μ), and
- a *parametrized rounding method*, i.e. for μ-a.e. $\gamma \in \Gamma$:

$$R_\gamma\colon\ \mathcal{C} \to \mathcal{C}_\mathcal{E}, \tag{2.104}$$

$$u \in \mathcal{C} \mapsto \bar{u}_\gamma := R_\gamma(u) \in \mathcal{C}_\mathcal{E} \tag{2.105}$$

  satisfying $R_\gamma(u') = u'$ for all $u' \in \mathcal{C}_\mathcal{E}$,

such that a *multiclass coarea-like property* (or *generalized coarea formula*)

$$f(u) \;=\; \int_\Gamma f(\bar{u}_\gamma)\, d\mu(\gamma) \tag{2.106}$$

holds. If this could be achieved, all arguments in the proof of Thm. 2.8 would apply equally to the multiclass formulation. Therefore the integral problem

$$\arg\min_{u \in \mathcal{C}_\mathcal{E}} \; f(u) \tag{2.107}$$

could be solved by computing a solution $u^* \in \mathcal{C}$ of the *relaxed* problem and then thresholding using R_γ for μ-a.e. $\gamma \in \Gamma$ to obtain an *integral* solution $\bar{u}^*_\mathcal{E} = R_\gamma(u^*) \in \mathcal{C}_\mathcal{E}$. For our particular problem (2.1), condition (2.106) is fulfilled if

$$\langle u, s\rangle = \int_\Gamma \langle \bar{u}_\gamma, s\rangle\, d\mu(\gamma) \quad \text{and} \quad J(u) = \int_\Gamma J(\bar{u}_\gamma)\, d\mu(\gamma). \tag{2.108}$$

Additionally, minimization of $f(u) := \langle u, s\rangle + J(u)$ should be feasible over the relaxed set $\mathcal{C}$. In the following sections, we consider two straightforward approaches to construct such R_γ, and see why they fail.

2.6.3 Higher Codimensions

One possible approach is to try to apply higher-dimensional variants of the coarea formula, as provided by Thm A.29. For some $m \geqslant k$ (with adopted notation),

$$\int_E \boldsymbol{C}_k d^E u'_x \, d\mathcal{H}^d(x) \;=\; \int_{\mathbb{R}^k} \mathcal{H}^{d-k}\big(E \cap u'^{-1}(\gamma)\big)\, d\gamma. \tag{2.109}$$

for a (Lipschitz) function $u'\colon \mathbb{R}^m \to \mathbb{R}^k$ and a countably $\mathcal{H}^d$-rectifiable set $E \subseteq \mathbb{R}^m$. The scalar case corresponds to $m = d$, $E = \Omega$ and $k = 1$, therefore $\boldsymbol{C}_k d^E u'_x = \|\nabla u'\|_2$.

In order to apply (2.109) to vector-valued $u'\colon \Omega \to \mathbb{R}^l$, one would again choose $E = \Omega$. Unfortunately, since $k = l$, (2.109) can then only be applied to problems where $d \geqslant l$, i.e. the number of spatial dimensions is coupled with the number of labels. Moreover, the left-hand side becomes an integral over the absolute value of the product of the singular values of $\nabla u'$ (cf. (A.47)),

$$\int_\Omega \left(\det\left((\nabla u')^\top \nabla u'\right)\right)^{1/2} dx. \tag{2.110}$$

Compared to the integral over the norm of the gradient in the scalar case, this does not seem to have a useful interpretation when applied to images or labeling functions, nor is it convex in general.

Finally, for multiple classes the codimension for the Hausdorff measures on the right-hand side of (2.109) becomes larger than one. In the scalar case, the Hausdorff measures can be represented in terms of the distributional gradient of the rounded functions via $\mathrm{TV}(\chi_{\{u'>\alpha\}})$. In the vector-valued case, there is no obvious extension of this principle; in fact it is not even clear how to define a rounding operation analogous to $\chi_{\{u'>\alpha\}}$ for vector-valued functions u'.

2.6.4 Separable Regularizer

Another straightforward approach is to choose J separable in the components of u, e.g.

$$J(u) \;=\; \int_\Omega d\,|Du_1| + \ldots + \int_\Omega d|Du_l|. \tag{2.111}$$

The same argument as in the proof of Thm. 2.8 then shows that

$$\begin{aligned} f(u) \;&=\; \sum_{i=1}^{l} \left(\int_\Omega u_i(x)\, s(x)\, dx + \int_\Omega d|Du_i| \right) && (2.112)\\ &=\; \sum_{i=1}^{l} \int_0^1 \left(\int_\Omega \left(\bar{u}_i\right)_{\gamma_i} s(x)\, dx\, d\alpha + \int_\Omega d|D\left(\bar{u}_i\right)_{\gamma_i}| \right) d\gamma_i && (2.113)\\ &=\; \int_{\gamma\in[0,1]^l} f(\tilde{u}_\gamma)\, d\gamma, && (2.114) \end{aligned}$$

where the *vector* of "rounding parameters" $\gamma \in \mathbb{R}^l$ replaces the scalar α, and

$$\tilde{u}_\gamma \;:=\; \begin{pmatrix} \chi_{\{u_1>\gamma_1\}} \\ \vdots \\ \chi_{\{u_l>\gamma_l\}} \end{pmatrix}. \tag{2.115}$$

This provides a coarea-like property as in (2.94) by the method of applying the coarea formula to each component separately. However, it has the severe drawback that $\tilde{u}_\gamma$ is generally not in $\mathcal{C}_\mathcal{E}$, or even $\mathcal{C}$, for $l \geqslant 3$: If $u \in \mathcal{C}$ is not integral from the beginning, it has at least two nonzero components, without loss of generality u_1 and u_2. Then, for sufficiently small ε and $\gamma = (\varepsilon, \varepsilon, \ldots)$, $(\tilde{u}_\gamma)_1 = (\tilde{u}_\gamma)_2 = 1$ and therefore $\tilde{u}_\gamma \notin \mathcal{C}_\mathcal{E}$.

Thus the approach provides a coarea-like property, but unfortunately does not induce a method for obtaining integral solutions, i.e. a representation in terms of integral functions. This invalidates the remainder of the proof of Thm. 2.8, since (2.96) does not hold anymore.

In summary, the existence of such an *exact* relation for the discretized problem seems unlikely. This is also supported by the fact that for the uniform metric, the problem is equivalent to a multiterminal cut (see e.g. [BVZ01]), which is a known NP-hard problem for more than 2 labels [DJPS94, Thm. 2a]. A multiclass coarea formula would allow to solve such problems using convex optimization, which can generally be achieved in polynomial time (at least to some finite precision). However, in Chap. 5 we derive an approximate variant, which – while it does not allow to recover *exact* solutions of the original problem – permits to obtain approximate solutions with an *a priori* optimality bound.

2.7 Relation to other Approaches and Extensions

In this section, we give an overview over several related regularizers that have been proposed in various contexts, and how they can be interpreted in the context of labeling problems. We also point out several recent extensions to the general relaxed multiclass labeling formulation (2.6).

2.7.1 Isotropic Regularizers

Frobenius Norm. The classical choice

$$\Psi_F(Du) := \|Du\|_2 = \left(\sum_{i,j} (D_i u_j)^2 \right)^{\frac{1}{2}}. \tag{2.116}$$

is the basis for large parts of geometric measure theory and the theory of functions of bounded variation [AFP00], and is sometimes referred to as MTV in the context of denoising of vector-valued data [SR96, CS05, DAV08, YYZW08]; see also [CEPY05] for an overview of TV-based research and applications. It is isotropic and permutation invariant, however it is neither separable in the components of u nor in space. The associated potential is

$$\left(1/\sqrt{2}\right)\Psi(\nu(e^i - e^j)^\top) \;=\; \chi_{\{i \neq j\}} \;=\; d_u(i,j), \tag{2.117}$$

with the uniform metric d_u.

Channel-By-Channel. The regularizer considered in Sect. 2.6.4 can be formalized as

$$\Psi_1(Du) \;:=\; \|Du_1\|_2 + \ldots + \|Du_l\|_2. \tag{2.118}$$

In this formulation, the objective is separable in the components of u, which potentially simplifies numerical optimization [Blo98, ZGFN08]. Similar to the Frobenius norm, Ψ_1 implements the uniform metric,

$$(1/2)\Psi_1(\nu(e^i - e^j)^\top) \;=\; d_u(i,j), \tag{2.119}$$

and is isotropic, with $\mathcal{D}^1_{\text{loc}} = \{(v^1|\ldots|v^l) \,|\, \|v^i\|_2 \leqslant 1 \;\forall i \in \{1,\ldots,l\}\}$. As in the case of the Frobenius norm, linearly transformed variants of the form

$$\Psi_{1,A}(Du) \;:=\; \Psi_1(D(Au)) \tag{2.120}$$

could be used (Sect. 2.5.2). A straightforward transformation shows that the modified integrand satisfies $\Psi_{1,A} = \|a^i - a^j\|_1$. Therefore this approach covers metrics that can be represented using a linear embedding into a space that is now endowed with the norm given by Ψ_1 instead of the Euclidean norm.

In the context of color denoising, [Blo98] observed that the Frobenius norm prefers transitions with similar magnitude in all channels, which leads to a color smearing effect at edges and a color shift towards the grayscale image: the one-step transition $(0,0) \to (1,1)$ is assigned a much lower penalty than the two consecutive transitions $(0,0) \to (1,0) \to (0,1)$. This phenomenon does not occur with the Channel-by-Channel regularizer Ψ_1. A similar effect was observed in [CCP08] for multiclass segmentation, where the preference towards similar gradients leads to minimizers that assume non-integral values more frequently than is the case for e.g. Ψ_d.

Eigenvalue-Based Norms. In the anisotropic diffusion community it is common practice ([SR96], see also [WS01] and the references therein) to employ weighted norms based on the eigenvalues $\lambda_1 \geqslant ... \geqslant \lambda_d \geqslant 0$ of the structure tensor

$$G(Du) \;:=\; (Du)(Du)^\top. \tag{2.121}$$

We denote $\lambda^2(Du) := (\lambda_1^2, ..., \lambda_d^2)$, i.e. the λ_i represent the magnitudes of the singular values of Du. While originally rooted in a diffusion framework, the approach can also be used to construct TV-like regularizers. It includes the Frobenius norm, since

$$\Psi_F(Du) \;=\; \sqrt{e^\top \lambda^2(Du)}. \tag{2.122}$$

In addition, for some rotation matrix $R \in \mathbb{R}^{d\times d}$ and permutation matrix $P \in \mathbb{R}^{l\times l}$,

$$G(R(Du)P) = R(Du)PP^\top(Du)^\top R^\top = RG(Du)R^\top, \tag{2.123}$$

therefore $\lambda^2(Du) = \lambda^2(R(Du)P)$, i.e. all norms derived from these singular values are isotropic and permutation invariant. In [GC10a] it was proposed to employ

$$\Psi_2(Du) \;:=\; \sqrt{\lambda_1^2(Du)}, \tag{2.124}$$

which amounts to the standard ℓ^2 operator norm on Du. The corresponding dual set can be represented as

$$\mathcal{D}^2_{\text{loc}} \;=\; \{\nu x^\top | \nu \in \mathbb{R}^d, x \in \mathbb{R}^l, \|\nu\|_2 \leqslant 1, \|x\|_2 \leqslant 1\}. \tag{2.125}$$

While Ψ_2 is not trivial to handle numerically, it can be dealt with reasonably well using primal-dual methods, and experimentally reduces color smearing and channel coupling in denoising, deblurring and superresolution applications [GC10a]. Applied to labeling approaches, one obtains

$$\big(1/\sqrt{2}\big)\Psi_2(\nu(e^i - e^j)^\top) \;=\; d_u(i,j), \tag{2.126}$$

i.e. Ψ_2 again represents the uniform metric. However, since $\Psi_2 \leqslant \Psi_F$, for multiclass labeling the energy when using Ψ_2 potentially generates more artificial minima than the standard choice Ψ_F.

2.7.2 Anisotropic and Inhomogeneous Regularizers

In this section, we will consider approaches that are not rotation invariant. Note that most of these have been developed for scalar-valued total variation, however they extend to the vector-valued case in a straightforward manner and could be coupled with any of the above approaches for implementing different metrics.

Wulff Shapes. For scalar-valued u, the use of anisotropic variants of the total variation has been studied in [EO04] for the Rudin-Osher-Fatemi model, where the authors characterize minimizers of such functionals. They base their analysis on the "Wulff shape" associated with Ψ, which is the equivalent of $\mathcal{D}_{\text{loc}}$ for the scalar-valued case, i.e. it defines the norm Ψ via the unit ball of its dual norm. As an example, consider the choice $\mathcal{D}^b_{\text{loc}} := [0,1]^d$. Then, for scalar $u \in \text{BV}(\Omega)$,

$$\Psi_b(z) = \sigma_{\mathcal{D}^b_{\text{loc}}}(z) = \|z\|_1, \;\; z \in \mathbb{R}^{d\times 1}. \tag{2.127}$$

It can be shown that the structure of $\mathcal{D}_{\text{loc}}$ is reflected in the *geometry* of the minimizer of the ROF functional (1.9) in the sense that geometric structures in the shape of $\mathcal{D}_{\text{loc}}$ itself are retained: for ν small enough, the minimizer u^* of

$$f(u) \;=\; \frac{1}{2}\int_\Omega (u-\chi_{\mathcal{D}_{\text{loc}}})^2 dx + \nu \int_\Omega d\Psi(Du) \tag{2.128}$$

is just a multiple of the input, i.e. $u^* = c\chi_{\mathcal{D}_{\text{loc}}}$ [EO04, Thm. 4.1]. Applied to the above definition for Ψ_b, this implies that the unit box may occur as the minimizer of the anisotropic ROF model, which cannot happen for the standard ROF model [Mey01].

Based on these ideas, it was shown in [ZNF09] that the thresholding property for two-class isotropic continuous cuts can be transferred to their anisotropic counterparts, i.e. it is still possible to recover integral solutions of the anisotropic continuous cut problems by thresholding. When combined with an adaptive, edge-driven version of the Wulff shape, the authors observed improved visual quality when applied to the reconstruction of depth maps and 3D structure. Similar results have independently been derived in [OBOK09].

The "Wulff shape" anisotropies could also be extended to the vector-valued case, for example by setting

$$\Psi(Du) \;=\; \left(\sum_i \Psi_b(Du_i)\right)^2, \tag{2.129}$$

or by replacing $\|\cdot\|_2$ by Ψ_b in the definition of Ψ_d (2.62). However, as in the isotropic case, this invalidates the thresholding property.

Anisotropy from Discretization. A large class of anisotropies that occurs in practice are actually induced by the discretization for approximating the total variation on grids. A very common scheme is to add the total variation of the individual components,

$$\Psi_{a,\|\cdot\|}(Du) \;:=\; \|D_1 u\|_2 + \ldots + \|D_1 u\|_2, \tag{2.130}$$

where $\|\cdot\|$ refers to some norm on $\mathbb{R}^l$. Notable cases include $\|\cdot\|_2$ and the completely separable case

$$\Psi_{a,1}(Du) \;:=\; \Psi_{a,\|\cdot\|_1}(Du) \;:=\; \|D_1 u\|_1 + \ldots + \|D_d u\|_1 = \sum_{i=1}^{d}\sum_{j=1}^{l} |D_i u_j|, \tag{2.131}$$

which is the natural limit of the usual 4-neighborhood discretization (Chap. 3). This discretization often occurs in LP relaxations and is tremendously popular as it is convex, contains only pairwise terms (i.e. terms depending on only two different variables) and is therefore easy to implement and analyze. Moreover, for two-class problems (or scalar-valued u), the energy satisfies a discrete variant of the coarea formula, which allows to carry over thresholding properties to the discretized problem [BVZ01]. The main drawback is that edges parallel to the coordinate axes are preferred to diagonal edges, which often leads to "zig-zag" artifacts on diagonal structures.

To some extent, the discretization-induced anisotropy can be reduced by increasing the neighborhood, i.e. by increasing the number of pairwise terms and adding proper weighting factors. In [Boy03, KB05] it was shown that anisotropies formulated as a certain class of metrics can asymptotically be approximated arbitrarily well using pairwise terms. However this requires the grid spacing and the neighborhood size to approach zero and infinity, respectively: for discretizations with a fixed number of neighbors, true isotropy cannot be guaranteed even for an arbitrarily fine grid. In practice, increasing the neighborhood size generally reduces artifacts but considerably increases the runtime of graph-based solvers. This effect is even more pronounced in higher-dimensional data [KSK+08]. We refer to Chap. 3 for a detailed discussion.

Inhomogeneous Regularizers. The usefulness of *homogeneous* anisotropic regularizers as mentioned above is quite limited. Therefore, anisotropic regularizers have traditionally been formulated in an inhomogeneous way, i.e.

$$J(u) \;=\; \int_\Omega d\Psi(x, Du). \tag{2.132}$$

Often the regularizer is of the multiplicative form $\Psi(x, Du) = g(x)\,\Psi'(Du)$ for some local weighting function g. In the context of two-class segmentation, such approaches have for instance been used for multiview reconstruction [KKBC09] and tracking [UMPB09].

2.7.3 Linearly Ordered Label Set

Lifting Approach. An interesting special case is when the labels in the label set $\mathcal{I}$ correspond to quantized values, i.e. they can be naturally ordered. This occurs in particular in so-called *lifting* approaches for finding *scalar-valued* minimizers of variational problems, e.g.

$$\min_{u' \in \mathcal{C}'} f'(u'), \quad f'(u') := \int_\Omega h(x, u'(x), \nabla u'(x)) d\,x \tag{2.133}$$

for some function h which is convex in $\nabla u'$, and a constraint set $\mathcal{C}' \subseteq \{u' : \Omega \to \mathbb{R}\}$. Often the functional f' is non-convex in u', as e.g. in the case of depth reconstruction from calibrated pairs of stereo images: here the scalar $u'(x)$ represents the local disparity, which corresponds to depth, and the – generally highly nonconvex – data term describes how well the corresponding parts of the images agree, given a specific disparity $u'(x)$.

In this case we can exploit the fact that the disparity is bounded, and thus u has a bounded range. This is also true in many other applications, such as when u' represents a grayscale image with intensities in [0, 1]. Then a possible solution to remove the nonconvexity is to quantize the range of u' into l values $\{c_1, ..., c_l\}$, and identify these with the l labels $\{1, ..., l\}$. The original nonconvex variational problem is then turned into the combinatorial problem of assigning, to each point x in the image domain Ω, the label $\ell(x) \in \{1, ..., l\}$ indicating $u'(x) = c_{\ell(x)}$.

Assuming that the data term is fully local, this often results in combinatorial problems which admit a convex relaxation. In a finite-dimensional setting, such approaches have been considered in [Ish03, BT09b]. At the same time, the "calibration" idea was developed in the continuous setting [ABDM03], which was used by [CCP08, PCBC10] to show that for $\mathcal{C} \subseteq W^{1,1}(\Omega, \mathbb{R})$, f' can be expressed in terms of the $\{0,1\}$-characteristic function $\chi_{u'}$ of the hypograph of u',

$$\begin{aligned} \operatorname{hyp} u' &= \{(x,t) \in \mathbb{R}^d \times \mathbb{R} | u'(x) > t\}, && (2.134) \\ \chi_{u'}(x,t) &:= \chi_{\operatorname{hyp} u'}(x,t) = \begin{cases} 0, & t \geqslant u'(x), \\ 1, & t < u'(x). \end{cases} && (2.135) \end{aligned}$$

Specifically,

$$f'(u') = \int_{\Omega \times \mathbb{R}} d\Psi(x,t,D\chi_{u'}), \tag{2.136}$$

where $\Psi(x,t,z) := \sigma_{\mathcal{D}^{x,t}_{\text{loc}}}(z)$ is defined implicitly via the Legendre-Fenchel conjugate of h:

$$\mathcal{D}^{x,t}_{\text{loc}} := \{(v,w) \in \mathbb{R}^d \times \mathbb{R} | w \geqslant h^*(x,t,v)\}. \tag{2.137}$$

Here h^* denotes the conjugate of h with respect to the last argument. Essentially, this transforms the problem of finding the optimal *function* u' into the problem of finding the *set of points* below its graph, which can be seen as a *two-class* segmentation problem in $\mathbb{R}^{d+1}$ with an anisotropic, inhomogeneous regularizer. This effectively linearizes the nonconvexity of h with respect to $u'(x)$. On the other hand, depending on the integrand h the dual constraint set $\mathcal{D}$ may be very complicated.

The problem is then relaxed in the usual manner via

$$\min_{u \in \mathrm{BV}(\Omega,[0,1])} f(u), \quad f(u) := \int_{\Omega \times \mathbb{R}} d\Psi(x,t,Du), \tag{2.138}$$

with the additional constraints that $u(x,\cdot)$ is nondecreasing, and that $u(x,t) \to 0/1$ for $t \to +\infty/-\infty$. After discretization, one obtains the alternative parametrization of the unit simplex (2.16). In this functional lifting setting, the connection between this method and our approach (2.6) is that the former represents u' in terms of its *superlevelsets* $\chi_{\{u'>t\}}$ with the above relaxation, while the latter represents it in terms of a family of *Dirac measures* $\{\delta_{u'(x)} | x \in \Omega\}$ corresponding to the values of u', with the relaxation to a family of *probability measures*.

Mumford-Shah. We consider again the Mumford-Shah problem (Sect. 1.2) in the weak formulation (1.21), with the normalization $\alpha = \lambda/\mu$, $\beta = \nu/\mu$:

$$\inf_{u' \in \mathrm{SBV}(\Omega)} f_{\mathrm{MS}}(u') := \alpha \int_\Omega (u' - I)^2 \, dx + \int_\Omega d|D^a u'|(\Omega) + \beta \mathcal{H}^{d-1}(S_{u'}). \tag{2.139}$$

Due to the nonconvex Hausdorff measure, f_{MS} does not directly admit a representation in the form (2.133). However, as shown in [ABDM03, (3.5)], for $\tau_1 \leqslant I \leqslant \tau_2$ the problem can be rewritten in the following way:

Proposition 2.10. *Define*

$$f(u) := \sup_{v\in\mathcal{D}} \int_{\Omega\times\mathbb{R}} \langle v, Du\rangle, \quad u\in \mathrm{SBV}(\Omega\times\mathbb{R}), \tag{2.140}$$

$$\mathcal{D} := C_c^1(\Omega\times\mathbb{R}, \mathbb{R}^{d+1}) \cap \mathcal{R} \cap \mathcal{S}, \tag{2.141}$$

$$\mathcal{R} := \left\{ (v^x,\, v^t) | v^t(x',\, t') + \alpha(t' - f(x'))^2 \geqslant \frac{\|v^x(x',t')\|_2^2}{4} \ \forall x'\in\Omega,\, \forall t'\in[\tau_1,\, \tau_2] \right\}, \tag{2.142}$$

$$\mathcal{S} := \bigcap_{\tau_1\leqslant p<q\leqslant\tau_2} \mathcal{S}_{p,q}, \quad \mathcal{S}_{p,q} := \left\{ (v^x, v^t) | \left\| \int_p^q v^x(x',t')dt' \right\|_2 \leqslant \beta \forall x'\in\Omega \right\}. \tag{2.143}$$

Then

$$f_{\mathrm{MS}}(u') = f(\chi_{u'}). \tag{2.144}$$

for any $u'\in\mathrm{SBV}(\Omega)$.

This provides an augmented (or *lifted*) form of the original objective on a higher-dimensional domain. The definition of $\mathcal{R}$ in (2.142) corresponds to the linearization of the smooth, convex part of (2.139), i.e. the data term and the squared norm regularizer, similar to the convex case (2.137), while $\mathcal{S}$ encodes the additional nonconvex length term $\beta\,\mathcal{H}^{d-1}(S_{u'})$. However, in contrast to the approach in the last paragraph for the functional f' (2.133), the lifting technique is applied to the *nonconvex* functional f_{MS}.

Note that in (2.140), the supremum is generally not attained. In order to obtain a maximum, one has to drop the compact support and continuity of v. Specifically, if there exists, for some fixed u, a sufficiently regular (in the sense that its divergence exists in a distributional sense, cf. [ABDM03, Def. 2.1]) mapping $v=(v^x,v^t)\colon\Omega\times\mathbb{R}\to\mathbb{R}^{d+1}$ that is divergence free, satisfies $v\in\mathcal{R}\cap\mathcal{S}$, $\langle v^x, \nu_{\partial\Omega}\rangle = 0$ $\mathcal{H}^{d-1}$-a.e. on $\partial\Omega\times\mathbb{R}$, and maximizes the supremum in the objective (2.140), i.e.

$$f(u) = \int_{\Omega\times\mathbb{R}} \langle v, Du\rangle, \tag{2.145}$$

then v is called an (absolute) *calibration* for u. Such a calibration acts as a certificate for optimality: If some function $\chi_{u'}$ admits an absolute calibration, u' is necessarily a global minimizer of f_{MS}, and v is a calibration for any other global minimizer [ABDM03, Thm. 3.4]. This is a strong result, as it provides sufficient conditions for the *global* optimality of minimizers for the *nonconvex* energy f_{MS}.

On the other hand, the definition in (2.140) can be used to solve the Mumford-Shah problem approximately using the same relaxation technique as in the convex case, i.e. replacing $\chi_{u'}$ in (2.144) by some $u\in\mathrm{SBV}(\Omega,[0,1])$, as suggested in [PCBC09].

Solving this relaxed problem does not necessarily provide an integral u, i.e. a minimizer of f_{MS}. However, it is still useful for computing approximate minimizers. Fig. 2.8 shows an exemplary result obtained using the above technique.

Levelable Functions. A particularly well-behaved special case of (2.133) is

$$f'(u') = \int_\Omega g(x, f(x))\, dx + J(u'), \tag{2.146}$$

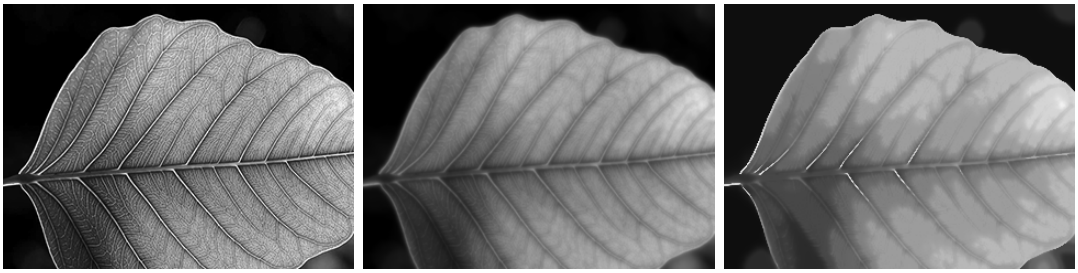

Figure 2.8. Approximation of the solution of the nonconvex Mumford-Shah problem using the "lifting" technique. **Left:** Input image. **Center:** Result of classical $L^2 - L^2$ denoising. Details are removed uniformly, which leads to loss of detail and smoothing at hard edges. **Right:** Result of variational denoising using the Mumford-Shah functional with 8 levels. Noise or fine details can be removed without blurring sharp edges. The lifting approach allows to approximately minimize the full Mumford-Shah functional by solving an associated labeling problem in a higher-dimensional space.

where g is convex in the second argument and J *has a coarea-like property*,

$$J(u') \;=\; \int_0^1 J(\chi_{\{u'>\alpha\}})\,d\alpha. \tag{2.147}$$

Such special J are also called *levelable functions* [DS06a, DS06b] or *discrete total variations* [CD09] and have been studied in [SKO09]. The difference to two-class continuous cuts lies in g, which is generally not linear and not levelable. As a consequence, minimizers are not necessarily integral. The characteristic property for such functionals it that the problem of finding $u = \chi_{\{u'>t\}}$ as in (2.135) *decouples in* t, i.e. it can be solved for all t *independently*, and the obtained characteristic function is in fact a hypograph, i.e. it is decreasing in t.

Under a suitable discretization, these problems can then be solved using a sequence of two-class cuts for finding the individual sublevelsets, and subsequent reconstruction of u from the obtained level sets. In particular this is the case for the ROF [CD09] and $L^1 - \mathrm{TV}$ [CEN06] models (Sect. 1.1). The latter has an extensive theory concerning the geometric structure of its sublevelsets [DAG09].

A related question is whether it is possible to specify J only on indicator functions,

$$J(\chi_S) \;:=\; g(S), \tag{2.148}$$

for some monotone function $g\colon 2^\Omega \to \mathbb{R}$, and *define* J such that is satisfies a generalized coarea formula:

$$J(u') \;:=\; \int_0^1 g(\chi_{\{u'>\alpha\}})\,d\alpha. \tag{2.149}$$

This definition is also known as the *Choquet integral* [Cho54] (also *Lovász extension* [Mur03]). The corresponding g are known as *capacities* (*set functions*), and it can be shown that J is monotone and positively homogeneous. Moreover, if g is *2-alternating* (*submodular*), J is convex. Therefore, any such g implicitly defines a convex regularizer via the integral (2.149).

2.7.4 Other Extensions

Partially Separable Norms. For linearizations of labeling problems that involve a large number of labels at each point, optimization can be made more efficient by exploiting separability in the regularizer. This occurs for example in optical flow estimation, where the two-dimensional flow vectors $u = (u_1, u_2)$ at each point are quantized using M^2 labels, which requires a prohibitively large amount of memory and computation time for fine quantizations.

If the regularizer decomposes with respect to u_1 and u_2, i.e. $J(u) = J_1(u_1) + J_2(u_2)$, it is possible to apply the relaxation technique in [GC10b], which reduces the memory requirements to the order of $O(2M)$ as opposed to $O(M^2)$. A related approach has also been suggested in [GBO09b].

Segmentation on Manifolds. The presented multiclass labeling techniques have recently been extended to segmentation on manifolds [DFPH09, WZDT11]. The main difficulty lies in a proper discretization of the gradient operator for a mesh-based representation of the manifold.

2.8 Summary and Further Work

In this chapter we have introduced the general framework for multiclass labeling on continuous domains and shown the existence of minimizers. We characterized the possible potentials and showed different ways to construct regularizers with prescribed interaction potentials.

Concerning the construction of regularizers, there are several paths for future work: The "envelope" relaxation in Sect. 2.5.1 seems to be related to the Wasserstein distance which occurs in transportation problems, for the finite-dimensional setting see [CKNZ01, KKMR06]. Investigating this connection and deriving results for the anisotropic, inhomogeneous case seem to be promising directions for further work.

Also, it is an open question whether the envelope relaxation could be tightened even more by constructing *nonlocal* regularizers. A related question concerns the thresholding theorem for nonlocal functionals, i.e. what conditions are required for nonlocal functionals to fulfill a generalized coarea formula.

Regarding the "embedding" approach in Sect. 2.5.2, an open question is if there are any *a priori* bounds on the quality of the embedding. For the class of *tree metrics*, such bounds exist, and the general case is the subject of ongoing research [CDG+09].

Chapter 3

Discretization of Functionals with Length-Based Regularizers

3.1 Introduction and Overview

In this chapter, 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 $\ell^*\colon \Omega \to \mathcal{I} = \{1, ..., l\}$ based on input data I modeled as $s\colon \Omega \to \mathbb{R}^l$, and again assume that there is some function space $\mathcal{U}$ and an objective function f depending on s, whose minimizer

$$u^* \;=\; \arg\min_{u \in \mathcal{U}} f(u) \tag{3.1}$$

provides some information about ℓ^*. In our setting, $\mathcal{U} = \mathrm{BV}(\Omega, \mathcal{E})$ corresponds to the combinatorial problem (2.1), and $\mathcal{U} = \mathrm{BV}(\Omega, \Delta_l)$ to the relaxed problem (2.6).

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,

$$u^{h,*} \;=\; \arg\min_{u^h \in \mathcal{U}^h} f^h(u^h). \tag{3.2}$$

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_l?
- 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 in some sense 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?

Figure 3.1. Segmentation of an image into 12 classes using a graph-based pairwise discretization. **Top left:** Input image, **Top right:** Result obtained by solving a graph-based *combinatorial* discretized problem with 4-neighborhood. The bottom row shows detailed views of the marked parts of the image. The minimizer of the combinatorial problem exhibits blocky artifacts caused by the choice of discretization.

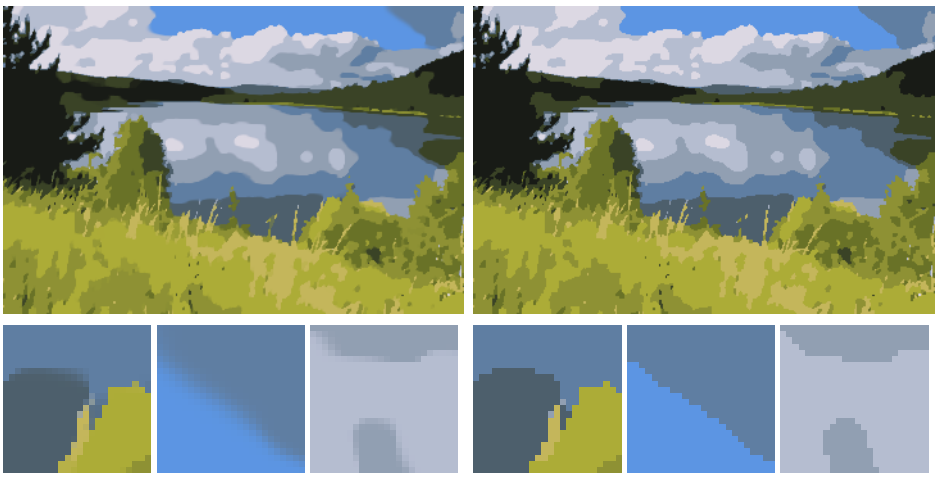

Figure 3.2. Segmentation obtained by solving a finite-differences discretization of the *relaxed* problem. **Left:** Non-integral solution obtained as a minimizer of the discretized relaxed problem. **Right:** Integral labeling obtained by rounding the fractional labels in the solution of the relaxed problem to the nearest integral label. The rounded result contains almost no visible artifacts.

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

As a motivation, consider the color segmentation problem in Fig. 3.1. The task is to segment the image into 12 classes, each class corresponding to a prototypical color, with the uniform metric regularizer. As can be seen, the classical approach of (approximately) solving a graph-based *combinatorial* discretized problem with $\mathcal{U}^h = (\mathcal{E})^n$ using a standard 4-neighborhood generates artifacts. On the other hand, discretizing and

solving the *relaxed* problem using finite differences and the constraint set $\mathcal{U}^h = (\Delta_l)^n$ – and subsequent thresholding to integral values if required – leads to visually far more convincing results (Fig. 3.2).

Organization. In the following sections we investigate the theoretical and practical advantages and disadvantages of the various discretization approaches for the multiclass labeling problem:

- We review several classical approaches including graph-based, pairwise, and LP formulations (Sect. 3.2).
- We consider in more detail a finite-differences approximation of the *relaxed* problem (Sect. 3.3) as proposed in [CCP08]. For infinitesimal scale, these functionals Γ-converge to the original relaxed functional (2.6). In particular, this implies that minimizers of the discretized functionals approximate minimizers of the original functional.
- We thoroughly compare the different choices in an experimental evaluation (Sect. 3.4). In particular, we try to answer the question whether the good results observed for the finite-differences method are an effect of the particular discretization, or whether they are related to the more fundamental question of whether one should minimize a combinatorial objective or obtain a solution by thresholding a solution of the relaxed problem (Sect. 3.5).

For an overview of the terminology regarding Γ-convergence, see Appendix A.2.

3.2 Related Work

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 multiindices, 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 $\mathcal{J} = \{0, ..., k-1\}^d$, where k is the grid size and $h = 1/k$ is the scale/grid spacing. We consider the regular grid

$$\Omega^h \;=\; \{x^{\bar{i}} = (\bar{i} + e/2)\, h \,|\, \bar{i} \in \mathcal{J}\}. \tag{3.3}$$

With any $\ell\colon \Omega \to \mathcal{I}$, we associate its discretization $\ell^h\colon \Omega^h \to \mathcal{I}$, $\ell^h_{\bar{i}} := \ell(x^{\bar{i}})$. We could alternatively consider $u^h\colon \Omega \to \mathcal{E}$ in order to be more consistent with (3.2), however we prefer to stick to the notation ℓ^h within this section, since it makes clear that there is no relaxation step involved, and it is more compatible with the traditional label assignment vector notation used in the related literature. Classically, a discretized *combinatorial* energy $f^h\colon \mathcal{I}^n \to \mathbb{R}$ is then constructed in a way such that ideally $f^h(\ell^h)$ approximates the energy $f(\ell)$. Minimizing f^h, one obtains

$$\ell^{h,*} \;=\; \arg \min_{\ell^h : \Omega^h \to \mathcal{I}} f^h(\ell^h). \tag{3.4}$$

The resulting problem is combinatorial in nature and thus 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. However, 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 $\ell^h_{\bar{i}} = \ell^h\big(x^{\bar{i}}\big)$ the labels. In the MRF approach, f^h is (non-uniquely) written as

$$f^h(\ell^h) \;=\; \sum_{C \in \mathrm{cl}(G)} \psi_C(\ell^h_C), \tag{3.5}$$

where the sum is taken over all sets $\mathrm{cl}(G)$ of *cliques* of G, $\ell^h_C \in \mathcal{I}^{|C|}$ denotes the *restriction* of ℓ to the vertices in the clique, and $\psi_C : \mathcal{I}^{|C|} \to \mathbb{R}$ are the individual *factors* of f. The "factor" terminology originates in the MRF setting, where one considers (among others) the problem of maximizing the *a posteriori* probability (MAP) of the labeling variables conditioned on the measurements/input data I, i.e.

$$\ell^{h,*} \;=\; \arg\max_{\ell^h : \Omega^h \to \mathcal{I}} \mathbb{P}(\ell^h | I). \tag{3.6}$$

Under the assumption that the edges in G represent the conditional dependence of the random variables $\ell^h\big(x^{\bar{i}}\big)$ in a particular sense [Lau96, KF09], the conditional probability $\mathbb{P}(\cdot|I)$ can be factorized over the cliques,

$$\mathbb{P}(\ell^h | I) \;=\; \prod_{C \in \mathrm{cl}(G)} \phi_C\big(\ell^h_C\big), \tag{3.7}$$

where the functions ϕ_C represent the *factors* of the joint probability $\mathbb{P}(\cdot|I)$. Maximizing $\mathbb{P}(\ell^h|I)$ is equivalent to minimizing its negative logarithm, thus

$$\arg\max_{\ell^h : \Omega^h \to \mathcal{I}} \mathbb{P}(\ell^h | I) \;=\; \arg\min_{\ell^h : \Omega^h \to \mathcal{I}} -\log \prod_{C \in \mathrm{cl}(G)} \phi_C\big(\ell^h_C\big) \tag{3.8}$$

$$= \arg\min_{\ell^h : \Omega^h \to \mathcal{I}} \sum_{C \in \mathrm{cl}(G)} -\phi_C\big(\ell^h_C\big). \tag{3.9}$$

The substitution $\psi_C\big(\ell^h_C\big) = -\phi_C\big(\ell^h_C\big)$ provides the connection to (3.5).

An important special case is 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,

$$f^h(\ell^h) \;=\; \sum_{x \in V} \psi_x(\ell^h(x)) + \sum_{(x,y) \in E} \psi_{x,y}(\ell^h(x), \ell^h(y)). \tag{3.10}$$

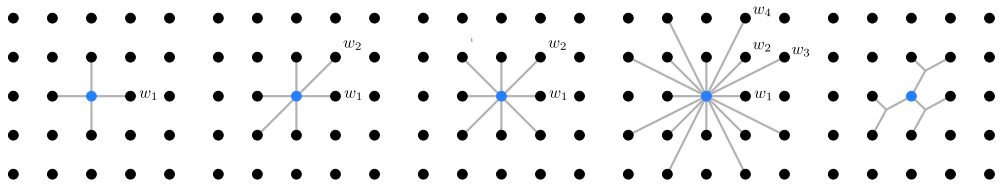

Figure 3.3. Graph-based discretization on a grid. **Left to right:** Pairwise terms with 4-, 6-, 8- and 16-neighborhood; higher-order discretization with ternary terms. The dots correspond to the vertices of the graph, the lines indicate factors in the representation (3.5) and weights in the pairwise representation (3.14).

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, a classical choice is to set

$$\psi_x(\ell^h(x)) = s(x, \ell^h(x)) = \mathbb{P}(I(x)|\ell^h(x)), \quad \psi_{x,y}(p,q) = \begin{cases} w_{x,y}, & p \neq q, \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

for some $w_{x,y} \geqslant 0$, 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. 3.3).

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

$$\psi_{x,y}(1,1) = \psi_{x,y}(2,2) = 0, \tag{3.12}$$
$$\psi_{x,y}(1,2) = \psi_{x,y}(2,1) = w_{x,y}, \tag{3.13}$$

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

$$f^h(\ell^h) \;=\; \sum_{x \in V} \psi_x(\ell^h(x)) + \sum_{(x,y) \in E} w_{x,y}\, \chi_{\{\ell^h(x) \neq \ell^h(y)\}}. \tag{3.14}$$

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 (3.10) 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 special "source" and "sink" nodes.

For nonnegative $w_{x,y}$ such problems with pairwise terms can be solved in polynomial time by min cut/max flow algorithms [Ber98]. Therefore an important question concerns whether it is possible to discretize some given *spatially continuous* functional f in this way, such that the discretized energy $f^h(\ell^h)$ approximates the continuous energy $f(\ell)$ if one sets $\ell^h_i = \ell(x^i)$.

In [Boy03], 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 [KB05] for a generalization to a larger class of metrics). Specifically, let $C\colon [0, |C|] \to \mathbb{R}^2$ denote some curve parametrized by arc length with tangent τ_C, and define its anisotropic length $|C|_{\mathcal{R}}$ by

$$|C|_{\mathcal{R}} \;:=\; \int_0^{|C|} \|A(C(s))\,\tau_C(s)\|_2\,ds, \tag{3.15}$$

where $A\colon \mathbb{R}^2 \to \mathrm{GL}_2(\mathbb{R})$ 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 (1.11) this corresponds, for some suitable $A'\colon \mathbb{R}^2 \to \mathrm{GL}_2(\mathbb{R})$, to the length-based regularizer

$$J(\ell) \;=\; \int_{\mathbb{R}^d} \left\| A' \frac{D\chi_{P^1}}{|D\chi_{P^1}|} \right\|_2 d\,|D\chi_{P^1}|. \tag{3.16}$$

For $A = A' = I$ we obtain the classical isotropic length $J(\ell) = \mathrm{TV}(\chi_{\Omega_1}) = \mathrm{Per}(\Omega_1)$.

We denote by $\mathcal{N} = \mathcal{N}(x) = \{y^1, ..., y^{|\mathcal{N}|}\}$ 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. 3.3. Then, for some fixed x, the vectors

$$g^m \;:=\; y^m - x, \quad m = 1, ..., |\mathcal{N}| \tag{3.17}$$

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. [Boy03] construct a regularizer J^h with pairwise potentials as in (3.14) by choosing the weights according to

$$\psi_{x,y^m}(p,q) \;=\; w_{x,y^m}\,\chi_{\{p\neq q\}}, \quad w_{x,y^m} = \frac{h^2\,\|g^m\|_2^2\,(\alpha_{m+1} - \alpha_m)\det(A(x))}{2\,\|A(x)\,g^m\|_2^3}. \tag{3.18}$$

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

$$h \to 0, \quad \sup_m |g^m| \to 0 \quad \text{and} \quad \sup_m |\alpha_m| \to 0. \tag{3.19}$$

This establishes a consistency result for the representation of metrics using pairwise terms, however it has several 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 (3.19) 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 in the discretization for $h \to 0$. 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.

LP Relaxation and Pairwise Functionals. One distinct advantage of the pairwise energy (3.14) is that it has a very well-behaved natural linear programming (LP) relaxation in the two-class case: Associate ℓ^h with $u^h\colon \Omega^h \to \{0, 1\}$ in the sense that $u^h(x) = 0 \Leftrightarrow \ell^h(x) = 1$, and consider the relaxed problem

$$\min_{u^h \in [0,1]^n} f^h_{\mathrm{LP}}(u^h), \tag{3.20}$$

$$f^h_{\mathrm{LP}}(u^h) := \sum_{x \in V} \left(\psi_x(1) - \psi_x(2)\right) u^h(x) + \sum_{(x,y) \in E} w_{x,y}\, |u^h(x) - u^h(y)|. \tag{3.21}$$

The problem can clearly be solved as a linear program. Moreover, since for $w_{x,y} \geqslant 0$ it satisfies the generalized coarea formula (cf. Sect. 2.6)

$$f^h_{\mathrm{LP}}(u^h) \;=\; \int_0^1 f^h_{\mathrm{LP}}\big(\chi_{\{u^h > \alpha\}}\big)\, d\alpha, \tag{3.22}$$

integral solutions of the pairwise energy f^h from (3.14) can be found by minimizing the LP relaxation f^h_{LP} and thresholding (cf. [CD09] and Thm. 2.8).

By setting $u^h(x) := u(x)$, the LP energy (3.20) can be extended to the set of relaxed functions $u\colon \Omega \to [0, 1]$ as follows:

$$f_{\mathrm{LP}}(u) := \sum_{x \in V} \left(\psi_x(1) - \psi_x(2)\right) u(x) + \sum_{(x,y) \in E} w_{x,y}\, |u(x) - u(y)|. \tag{3.23}$$

Although formulated on functions defined on continuous domains, energies such as (3.23) 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 [GM01], where the authors consider energies of the form

$$J_{G,h}(u) \;=\; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \eta(g)\, \varphi_{h\|g\|_2}\!\left(\frac{|u(x + h\,g) - u(x)|}{h\,\|g\|_2}\right) dg\,dx, \tag{3.24}$$

where $u \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$, $\eta \in L^1(\mathbb{R}^d)$ with $0 \not\equiv \eta \geqslant 0$, and for each h, $\varphi_h\colon \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ 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 $h \to 0$ of $h\,\varphi_h(z/h)$ and $\varphi_h(z)$, respectively, and define

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

Then, under some technical assumptions, the functionals $J_{G,h}$ Γ-converge to J_G in L^1_{loc} as $h \to 0$ [GM01, Thm. 4.3]. As a consequence, one obtains pointwise convergence of the functionals, as well as convergence of their minimizers. We refer to Appendix A.2 for the precise definitions and properties of Γ-converging sequences.

This convergence result can be seen as an extended variant of the result from [Boy03] mentioned previously, formulated for nonlocal functionals in terms of (possibly non-integral) functions u. However, in formulation (3.24) the finite sum of pairwise terms (3.10) 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 (3.10) has been considered in [Cha99], in order to approximate the nonconvex part of the Mumford-Shah energy,

$$J_{\mathrm{MS}}(u) \;=\; \lambda \int_{\mathbb{R}^d} \|\nabla u\|_2^2 \, dx + \mu \, \mathcal{H}^{d-1}(S_u). \tag{3.25}$$

The considered nonlocal energies are of the form

$$J_N^h(u^h) \;=\; \sum_{\bar{i} \in \mathcal{J}} h^d \sum_{\bar{j} \in \mathbb{Z}^d} \frac{1}{h} \, \varphi_h\!\left(\frac{\big(u^h\big(x^{\bar{i}+\bar{j}}\big) - u^h\big(x^{\bar{i}}\big)\big)^2}{h} \right) \eta(\bar{j}), \tag{3.26}$$

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

$$J_N^h(u^h) \;=\; \sum_{\bar{i} \in \mathcal{J}} h^d \sum_{\bar{j} \in \mathbb{Z}^d} \frac{1}{h} \, \varphi_h\!\left(\frac{\big(u\big(x^{\bar{i}} + h\,\bar{j}\big) - u\big(x^{\bar{i}}\big)\big)^2}{h} \right) \eta(\bar{j}). \tag{3.27}$$

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

$$J_N(u) = \int_\Omega \sum_{\bar{j} \in \mathbb{Z}^d} \eta(\bar{j}) \, \alpha_{\bar{j}} \, |\langle \nabla u, \bar{j} \rangle|^2 \, dx + \int_{S_u} \sum_{\bar{j} \in \mathbb{Z}^d} \eta(\bar{j}) \, \beta_{\bar{j}} \, |\langle \nu_u, \bar{j} \rangle| \, d\mathcal{H}^{d-1} \tag{3.28}$$

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

As a special case assume that u is integral, i.e. $u\colon \Omega \to \{0, 1\}$. 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 (3.28). 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 a sum of (separate) 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 [KT99, KT07]

$$f_{\mathrm{MLP}}^h(u^h) := \sum_{x \in V} \sum_{j=1}^{l} \psi_x(j) \, \big(u^h(x)\big)_j + \frac{1}{2} \sum_{(x,y) \in E} w_{x,y} \, \|u^h(x) - u^h(y)\|_1, \tag{3.29}$$

where f^h_{MLP} is minimized over all $u^h\colon \Omega^h \to \Delta_l$. While the energy itself still satisfies the coarea-like property (3.22), integral solutions cannot trivially be obtained using thresholding due to the constraints, see also Sect. 2.6.4.

Finite Elements. Another possibility is to use finite-element approximations. However, for fixed triangulations this again invariably introduces an anisotropy, as explicitly computed in [Neg99] for the Mumford-Shah functional. In [CDM99, BC00] is was shown that the Mumford-Shah functional can be exactly approximated in terms of Γ-convergence using finite elements, however this requires an involved adaptive triangulation.

3.3 Convergent Finite-Differences Approximation of the Relaxed Problem

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 [CCP08] which has a fixed neighborhood but still provides isotropy in the limit.

Since this discretization will be used in the optimization part, we will consider it in more detail. Again we represent multidimensional functions $u\colon V \to \mathbb{R}^l$ on Ω by a matrix $u^h = (u^{h,1}|\ldots|u^{h,l}) \in \mathbb{R}^{n\times l}$. The l-dimensional vector associated with $\bar{i}$ or $x^{\bar{i}}$ is denoted by $u^h_{\bar{i}} = u^h\big(x^{\bar{i}}\big) \in \mathbb{R}^l$. The standard forward-differences approximation for $\nabla u\big(x^{\bar{i}}\big)$ is

$$\nabla_{\bar{i}} u^h \;=\; \frac{1}{h}\begin{pmatrix} \big(u^h_{\bar{i}+e^1} - u^h_{\bar{i}}\big)^\top \\ \vdots \\ \big(u^h_{\bar{i}+e^d} - u^h_{\bar{i}}\big)^\top \end{pmatrix}, \tag{3.30}$$

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

$$s^h_{\bar{i}} := s^h\big(x^{\bar{i}}\big) := \frac{1}{h^d}\int_{C^h_{\bar{i}}} s(x)\,dx, \tag{3.31}$$

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

$$C^h_{\bar{i}} \;:=\; x^{\bar{i}} + \frac{1}{2}(-h,h)^d = (i_1 h,(i_1+1)h) \times \cdots \times (i_d h,(i_d+1)h). \tag{3.32}$$

Then, for the relaxed multiclass labeling functional (cf. (2.6) and (2.10))

$$f(u) \;=\; \int_\Omega \langle u, s\rangle\,dx + \int_\Omega d\Psi(Du), \tag{3.33}$$

we define the discretization

$$f^h(u^h) \;:=\; \sum_{\bar{i}\in\mathcal{J}} h^d \left\langle u^h_{\bar{i}}, s^h_{\bar{i}} \right\rangle + \sum_{\bar{i}\in\mathcal{J}} h^d\, \Psi(\nabla_{\bar{i}} u^h). \tag{3.34}$$

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 $u^h \in \mathcal{U}^h := \{u^h \colon \Omega^h \to \Delta_l\} = (\Delta_l)^n$ with the piecewise constant function

$$\tilde{u}^h \in \mathrm{BV}(\Omega)^l, \quad \tilde{u}^h(x) = u^h\big(x^{\bar{i}}\big) \in \mathbb{R}^l, \quad \text{for } \mathcal{L}^d\text{-a.e. } x \in C^h_{\bar{i}}. \tag{3.35}$$

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

$$\tilde{\mathcal{U}}^h \;=\; \{u \in \mathrm{BV}(\Omega)^l \,|\, \exists\, u^h \in \mathcal{U}^h \colon u = \tilde{u}^h\}. \tag{3.36}$$

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

$$\tilde{f}^h \colon \mathrm{BV}(\Omega)^l \to \bar{\mathbb{R}} \tag{3.37}$$

$$\tilde{f}^h(u') := \begin{cases} f^h(u^h), & \text{if there exists } u^h \in \mathcal{U}^h \text{ s.t. } u' = \tilde{u}^h, \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.38}$$

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

$$f_{\mathcal{C}} \colon \mathrm{BV}(\Omega) \to \mathbb{R}, \tag{3.39}$$

$$f_{\mathcal{C}}(u) := \int_\Omega \langle u, s\rangle \, dx + \int_\Omega d\Psi(Du) + \delta_{\mathcal{C}}(u). \tag{3.40}$$

Then, from Prop. A.40 we conclude that minimizers of the discretized functionals converge to those of the original functional. The involved part is showing Γ-convergence of the regularizer, which we treat first.

In order to not obscure the notation by having to deal with a fractional number of pixels, we formally denote, for some sequence $(z^{(k)})$, $z^h := z^{1/k} := z^{(k)}$, and

$$\lim_{h\to 0} z^h \;:=\; \lim_{k\to\infty} z^{1/k}, \tag{3.41}$$

with similar notations for $\liminf$, $\limsup$, and Γ-convergence. Using these conventions, the number of points in Ω^h is always $n = k^d$. Note that the proofs in the following do not require Ψ to be isotropic.

Proposition 3.1. *Let $\Psi \colon \mathbb{R}^{d\times l} \to \mathbb{R}$ be continuous, convex and positively homogeneous with $\rho_l \|z\|_2 \leqslant \Psi(z) \leqslant \rho_u \|z\|_2 \,\forall z \in \mathbb{R}^{d\times l}$ (but not necessarily isotropic), and*

$$J_{\mathcal{C}}(u) \;:=\; \int_\Omega d\Psi(Du) + \delta_{\mathcal{C}}(u). \tag{3.42}$$

Denote

$$\tilde{J}^h(u') := \begin{cases} \sum_{\bar{i}\in\mathcal{J}} h^d\,\Psi(\nabla_{\bar{i}}u^h), & \text{if there exists } u^h\in\mathcal{U}^h \text{ s.t. } u'=\tilde{u}^h, \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.43}$$

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

Proof. The proof largely follows [CCP08, Prop. 3.1]. The second part additionally uses an argument derived from [Cha99].

"lim inf" inequality: The proof of the "lim inf" inequality (cf. Def. A.34)

$$\liminf_{h\to 0} \tilde{J}^h(\tilde{u}^h) \;\geqslant\; J_{\mathcal{C}}(u), \quad \forall u\in \mathrm{BV}(\Omega)^l \tag{3.44}$$

is basically identical to the one in [CCP08, Prop. 3.1]: Let $\tilde{u}^h\to u$ in L^1 (or in the strong topology of BV). If there exists h_0 such that $\tilde{J}^h(\tilde{u}^h)=+\infty$ for *all* $h<h_0$, it follows that $\liminf_{h\to 0}\tilde{J}^h(\tilde{u}^h)=+\infty$, and the first in equality in Def. A.34 holds trivially. Otherwise, we may restrict the following considerations to the subsequence satisfying $\tilde{J}^h(\tilde{u}^h)<+\infty$, i.e. without loss of generality we may assume that $\tilde{u}^h\in\tilde{\mathcal{U}}^h$ for all h, with associated vector representation $u^h\in\mathcal{U}^h$. It therefore remains to show that

$$\liminf_{h\to 0} \tilde{J}^h(\tilde{u}^h) \geqslant J(u), \tag{3.45}$$

$$J(u) = \sup\{-\int_\Omega \langle u, \mathrm{Div}\, v\rangle\, dx \,|\, v\in C_c^\infty(\Omega)^{d\times l}, v(x)\in\mathcal{D}_{\mathrm{loc}} \forall x\in\Omega\}. \tag{3.46}$$

This is the case if we can show that for any fixed $v\in C_c^\infty(\Omega)^{d\times l}, v(x)\in\mathcal{D}_{\mathrm{loc}}\ \forall x\in\Omega$,

$$\liminf_{h\to 0} \tilde{J}^h(\tilde{u}^h) \;\geqslant\; -\int_\Omega \langle u, \mathrm{Div}\, v\rangle\, dx. \tag{3.47}$$

To show this, we use the fact that $\|\mathrm{Div}\, v\|_\infty<\infty$ due to the smoothness and compact support of v. Therefore, the fact that $\tilde{u}^h\to u$ with respect to L^1-convergence implies

$$L(u) := -\int_\Omega \langle u, \mathrm{Div}\, v\rangle\, dx \;=\; \lim_{h\to 0} -\int_\Omega \langle \tilde{u}^h, \mathrm{Div}\, v\rangle\, dx. \tag{3.48}$$

Since $\tilde{u}^h$ is constant on the boxes $C_{\bar{i}}^h$ defining the grid, we may represent $L(u)$ as

$$L(u) \;=\; \lim_{h\to 0} -\sum_{\bar{i}\in\mathcal{J}} \int_{C_{\bar{i}}^h} \langle u_{\bar{i}}^h, \mathrm{Div}\, v\rangle\, dx = \lim_{h\to 0} -\sum_{\bar{i}\in\mathcal{J}} \left\langle u_{\bar{i}}^h, \int_{C_{\bar{i}}^h} \mathrm{Div}\, v dx \right\rangle. \tag{3.49}$$

By partial integration [AFP00, Thm. 3.36], we get

$$L(u) \;=\; \lim_{h\to 0} \sum_{\bar{i}\in\mathcal{J}} \left\langle u_{\bar{i}}^h, \int_{\Omega\cap\partial C_{\bar{i}}^h} v^\top \nu_{C_{\bar{i}}^h}\, ds \right\rangle = \lim_{h\to 0} \sum_{\bar{i}\in\mathcal{J}} \int_{\Omega\cap\partial C_{\bar{i}}^h} \langle u_{\bar{i}}^h, v^\top \nu_{C_{\bar{i}}^h}\rangle\, d\mathcal{H}^{d-1}, \tag{3.50}$$

where $\nu_{C_{\bar{i}}^h}$ denotes the inner unit normal on the boundary of the box $C_{\bar{i}}^h$. Rearranging the expression in terms of the line segments $\overline{C_{\bar{i}+e^j}^h} \cap \overline{C_{\bar{i}}^h}$ bounding the $C_{\bar{i}}^h$ yields

$$L(u) \;=\; \lim_{h\to 0} \sum_{\bar{i}\in\mathcal{J}} \sum_{j=1}^{d} \int_{\Omega\cap\overline{C_{\bar{i}+e^j}^h}\cap\overline{C_{\bar{i}}^h}} \langle u_{\bar{i}+e^j}^h - u_{\bar{i}}^h, v^j\rangle \, d\mathcal{H}^{d-1}, \tag{3.51}$$

here $v^j(x)\in\mathbb{R}^l$ is the j-th *row* of $v(x)$ (recall that $v(x)\in\mathbb{R}^{d\times l}$). Then

$$\begin{aligned}
L(u) \;&=\; \lim_{h\to 0} \sum_{\bar{i}\in\mathcal{J}} \sum_{j=1}^{d} \int_{\Omega\cap\overline{C_{\bar{i}+e^j}^h}\cap\overline{C_{\bar{i}}^h}} \langle u_{\bar{i}+e^j}^h - u_{\bar{i}}^h, v^j(x^{\bar{i}}) + (v^j - v^j(x^{\bar{i}}))\rangle \, d\mathcal{H}^{d-1} && (3.52)\\
&=\; \lim_{h\to 0} \sum_{\bar{i}\in\mathcal{J}} \Bigg(h^{d-1} \left\langle h\,\nabla_{\bar{i}} u^h, v(x^{\bar{i}})\right\rangle \\
&\qquad + \sum_{j=1}^{d} \int_{\Omega\cap\overline{C_{\bar{i}+e^j}^h}\cap\overline{C_{\bar{i}}^h}} \left\langle h\,\nabla_{\bar{i}} u^h, v^j - v^j(x^{\bar{i}})\right\rangle d\mathcal{H}^{d-1} \Bigg) && (3.53)\\
&\leqslant\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{J}} \Bigg(h^{d-1} \langle h\,\nabla_{\bar{i}} u^h, v(x^{\bar{i}})\rangle \\
&\qquad + \sum_{j=1}^{d} \int_{\Omega\cap\overline{C_{\bar{i}+e^j}^h}\cap\overline{C_{\bar{i}}^h}} \|h\,\nabla_{\bar{i}} u^h\|_2 \, \|v^j - v^j(x^{\bar{i}})\|_2 \, d\mathcal{H}^{d-1} \Bigg). && (3.54)
\end{aligned}$$

Denote $C_v := \max_j \|\nabla v^j\|_\infty$. Then $C_v < \infty$ due to the smoothness and compact support of v, and $\|v^j(x) - v^j(x^{\bar{i}})\|_2 \leqslant h\sqrt{d}\, C_v$ for all $x\in C_{\bar{i}}^h$. Therefore

$$\begin{aligned}
L(u) \;&\leqslant\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{I}} \{h^{d-1} \langle h\,\nabla_{\bar{i}} u^h, v(x^{\bar{i}})\rangle \\
&\qquad + \sum_{j=1}^{d} \int_{\Omega\cap\overline{C_{\bar{i}+e^j}^h}\cap\overline{C_{\bar{i}}^h}} \|h\,\nabla_{\bar{i}} u^h\|_2 \, h\sqrt{d}\, C_v \, d\mathcal{H}^{d-1}\} && (3.55)\\
&=\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{I}} \left(h^{d-1} \langle h\,\nabla_{\bar{i}} u^h, v(x^{\bar{i}})\rangle + h^{d-1} \|h\,\nabla_{\bar{i}} u^h\|_2 \, h\, d^{3/2} C_v\right) && (3.56)\\
&=\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{I}} \left(h^{d} \langle \nabla_{\bar{i}} u^h, v(x^{\bar{i}})\rangle + h^{d+1} \|\nabla_{\bar{i}} u^h\|_2 \, d^{3/2}\, C_v\right). && (3.57)
\end{aligned}$$

Since $\Psi(z) = \sup_{v\in\mathcal{D}_{\mathrm{loc}}} \langle z, v\rangle \geqslant \langle z, v\rangle$ for all $v\in\mathcal{D}_{\mathrm{loc}}$ by definition, this can be bounded via

$$L(u) \;\leqslant\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{I}} \left(h^{d} \,\Psi(\nabla_{\bar{i}} u^h) + h^{d+1} \|\nabla_{\bar{i}} u^h\|_2 \, d^{3/2}\, C_v\right). \tag{3.58}$$

Using $\rho_l \|z\|_2 \leqslant \Psi(z)$, we arrive at

$$\begin{aligned}
L(u) \;&\leqslant\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{I}} \left(h^{d} \,\Psi(\nabla_{\bar{i}} u^h) + h^{d+1} \frac{1}{\rho_l} \Psi(\nabla_{\bar{i}} u^h)\, d^{3/2}\, C_v\right) && (3.59)\\
&=\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{I}} \left(h^{d} \,\Psi(\nabla_{\bar{i}} u^h) \left(1 + \frac{h}{\rho_l} d^{3/2}\, C_v\right)\right) && (3.60)\\
&=\; \liminf_{h\to 0} \sum_{\bar{i}\in\mathcal{I}} h^{d} \,\Psi(\nabla_{\bar{i}} u^h) && (3.61)\\
&=\; \liminf_{h\to 0} J^h(\tilde{u}^h). && (3.62)
\end{aligned}$$

Starting from (3.48), we conclude

$$-\int_\Omega \langle u, \operatorname{Div} v\rangle \, dx \;\leqslant\; \liminf_{h\to 0} \tilde{J}^h(\tilde{u}^h). \tag{3.63}$$

Since $v \in \mathcal{D}$ was arbitrary, this proves (3.47) and finally (3.45), which shows the "lim inf" inequality required for Γ-convergence.

"lim sup" inequality: Showing the "lim sup" inequality (Def. A.34) amounts to finding, for arbitrary but fixed $u \in \mathrm{BV}(\Omega)^l$, a sequence $(\tilde{u}^h)$ converging to u in L^1 s.t.

$$\limsup_{h\to 0} \tilde{J}^h(\tilde{u}^h) \;\leqslant\; J_{\mathcal{C}}(u). \tag{3.64}$$

First, choose a sequence of $u^{(j)} \subseteq \{u\colon \Omega \to \Delta_l \,|\, u \in C^\infty(\Omega)^l\}$ such that $u^{(j)} \overset{j\to\infty}{\to} u$ in terms of L^1-convergence and $\mathrm{TV}(u^{(j)}) \to \mathrm{TV}(u)$. The unconstrained case was shown in [AFP00, Thm. 3.9] for $u \in \mathrm{BV}(\Omega)^l$; the constraint $u^{(j)}(x) \in \Delta_l$ follows from the same proof since the sequence $u^{(j)}$ is constructed by mollification of spatial restrictions of u and Δ_l is convex. Then, by [AFP00, Thm. 3.15] and the continuity of Ψ,

$$\begin{aligned} J_{\mathcal{C}}(u) &= \int_\Omega \Psi\Big(\frac{Du}{|Du|}\Big) d\,|Du| && (3.65)\\ &\overset{\text{[AFP00, Thm. 3.15]}}{=} \lim_{j\to\infty} \int_\Omega \Psi\Big(\frac{Du^{(j)}}{|Du^{(j)}|}\Big) d|Du^{(j)}| && (3.66)\\ &= \lim_{j\to\infty} J_{\mathcal{C}}(u^{(j)}). && (3.67) \end{aligned}$$

For *fixed* j, consider the discretized functions $u_y^{(j),h}$ for $y \in (-h,h)^d$, where

$$u_{y,\bar{i}}^{(j),h} := u_y^{(j),h}\big(x^{\bar{i}}\big) := u^{(j)}\big(x^{\bar{i}} + y\big) = u^{(j)}(\bar{i}\,h + (h/2)\,e + y). \tag{3.68}$$

The associated piecewise constant functions $\tilde{u}_y^{(j),h}$ assume the value $u^{(j)}\big(x^{\bar{i}} + y\big)$ on $C_{\bar{i}}^h$. In order to handle the boundary condition, we follow the convention that any finite differences terms involving at least one point outside of Ω should be treated as zero. Instead of the limit of $\tilde{J}^h$ we consider the limit of its mean over all *shifts* $z = y/h$:

$$\begin{aligned} &\lim_{h\to 0} \int_{z\in\frac{1}{2}(-1,1)^d} \tilde{J}^h(\tilde{u}_{hz}^{(j),h})\, dz && (3.69)\\ =\;& \lim_{h\to 0} h^d \int_{y\in\frac{1}{2}(-h,h)^d} \tilde{J}^h(\tilde{u}_y^{(j),h})\, dy && (3.70)\\ =\;& \lim_{h\to 0} \int_{y\in\frac{1}{2}(-h,h)^d} \sum_{\bar{i}\in\mathcal{J}} \Psi\left(\frac{1}{h}\begin{pmatrix} (u^{(j)}\big(x^{\bar{i}} + y + h\,e^1\big) - u^{(j)}\big(x^{\bar{i}} + y\big))^\top \\ \vdots \\ (u^{(j)}\big(x^{\bar{i}} + y + h\,e^d\big) - u^{(j)}\big(x^{\bar{i}} + y\big))^\top \end{pmatrix}\right) dy && (3.71)\\ =\;& \lim_{h\to 0} \int_\Omega \Psi\left(\frac{1}{h}\begin{pmatrix} \big(u^{(j)}(x + h\,e^1) - u^{(j)}(x)\big)^\top \\ \vdots \\ \big(u^{(j)}(x + h\,e^d) - u^{(j)}(x)\big)^\top \end{pmatrix}\right) dx. && (3.72) \end{aligned}$$

In order to swap the limit and the integral we show the conditions for the dominated convergence theorem. The integrand in (3.72) can be absolutely bounded via

$$\Psi\left(\frac{1}{h}\begin{pmatrix} (u^{(j)}(x+h\,e^1)-u^{(j)}(x)^\top \\ \vdots \\ \big(u^{(j)}(x+h\,e^d)-u^{(j)}(x)\big)^\top \end{pmatrix}\right) \tag{3.73}$$

$$\leqslant \frac{\rho_u}{h}\left\|\begin{pmatrix} \big(u^{(j)}(x+h\,e^1)-u^{(j)}(x)\big)^\top \\ \vdots \\ (u^{(j)}(x+h\,e^d)-u^{(j)}(x))^\top \end{pmatrix}\right\|_2 \tag{3.74}$$

$$\leqslant \sum_{k=1}^{d}\frac{\rho_u}{h}\left\|u^{(j)}(x+h\,e^k)-u^{(j)}(x)\right\|_1 \tag{3.75}$$

$$\leqslant \sum_{k=1}^{d}\frac{\rho_u}{h}\int_0^h \left\|(e^k)^\top\nabla u^{(j)}(x+t\,e^k)\right\|_1 dt \tag{3.76}$$

$$\leqslant \sum_{k=1}^{d}\frac{C}{h}\int_0^h \left\|(e^k)^\top\nabla u^{(j)}(x+t\,e^k)\right\|_2 dt =: p(x) \tag{3.77}$$

for some $C>0$ independent of h, again with the convention that $\nabla u^{(j)}(x+t\,e^d)=0$ for $x+t\,e^d\notin\Omega$. Integrating this upper bound over Ω shows

$$\int_\Omega p(x)\,dx = \sum_{k=1}^{d}\frac{C}{h}\int_\Omega\int_0^h \left\|(e^k)^\top\nabla u^{(j)}(x+t\,e^k)\right\|_2 dt\,dx \tag{3.78}$$

$$\overset{(*)}{=} \sum_{k=1}^{d}\frac{C}{h}\int_0^h\int_\Omega \left\|(e^k)^\top\nabla u^{(j)}(x+t\,e^k)\right\|_2 dx\,dt \tag{3.79}$$

$$\leqslant \sum_{k=1}^{d}\frac{C}{h}\int_0^h\int_\Omega \left\|(e^k)^\top\nabla u^{(j)}(x)\right\|_2 dx\,dt \tag{3.80}$$

$$= \sum_{k=1}^{d} C\int_\Omega \left\|(e^k)^\top\nabla u^{(j)}(x)\right\|_2 dx \tag{3.81}$$

$$\leqslant \sum_{k=1}^{d} C\int_\Omega \left\|\nabla u^{(j)}(x)\right\|_2 dx \tag{3.82}$$

$$\leqslant (C\,d)\,\mathrm{TV}\big(u^{(j)}\big) \quad < \quad \infty. \tag{3.83}$$

The bound in the last equation justifies the application of Fubini's theorem at $(*)$. We conclude that one may apply the dominated convergence theorem to (3.72) to obtain

$$\lim_{h\to 0}\int_{z\in\frac{1}{2}(-1,1)^d}\tilde{J}^h(\tilde{u}_{hz}^{(j),h})\,dz \tag{3.84}$$

$$\overset{\text{dom. conv.}}{=} \int_{x\in\Omega}\lim_{h\to 0}\Psi\left(\frac{1}{h}\begin{pmatrix} (u^{(j)}(x+h\,e^1)-u^{(j)}(x))^\top \\ \vdots \\ (u^{(j)}(x+h\,e^d)-u^{(j)}(x))^\top \end{pmatrix}\right)dx \tag{3.85}$$

$$\overset{\Psi\text{ contin.}}{=} \int_{x\in\Omega}\Psi(\nabla u^{(j)}(x))\,dz \tag{3.86}$$

$$\overset{u^{(j)}\,\text{smooth}}{=} \int_\Omega \Psi(Du^{(j)}/|Du^{(j)}|)\,d\,|Du^{(j)}| \tag{3.87}$$

$$= J\big(u^{(j)}\big), \tag{3.88}$$

and we conclude that

$$\lim_{h\to 0}\int_{z\in\frac{1}{2}(-1,1)^d} \tilde{J}^h(\tilde{u}^{(j),h}_{hz})\,dz \;=\; J\big(u^{(j)}\big). \tag{3.89}$$

For each h, choose $z^h\in\frac{1}{2}\,(-1,1)^d$ such that

$$\tilde{J}^h(\tilde{u}^{(j),h}_{hz^h}) \;\leqslant\; \int_{z\in\frac{1}{2}(-1,1)^d} \tilde{J}^h(\tilde{u}^{(j),h}_{hz})\,dz, \tag{3.90}$$

which is possible since the integral in (3.90) is a mean. Denote $\tilde{u}^{(j),h}:=\tilde{u}^{(j),h}_{hz^h}$, then

$$\limsup_{h\to 0}\tilde{J}(\tilde{u}^{(j),h})\leqslant\lim_{h\to 0}\int_{z\in\frac{1}{2}(-1,1)^d}\tilde{J}^h(\tilde{u}^{(j),h}_{hz})\,dz\overset{(3.89)}{=}J\big(u^{(j)}\big). \tag{3.91}$$

Since $u^{(j)}(x)\in\Delta_l$ and therefore $\tilde{u}^{(j),h}(x)\in\Delta_l$ for all $x\in\Omega$, $\|u^{(j)}(x)-\tilde{u}^{(j),h}(x)\|_2$ is bounded from above, and from boundedness of Ω and Fatou's Lemma we obtain

$$\limsup_{h\to 0}\int_\Omega \|u^{(j)}(x)-\tilde{u}^{(j),h}(x)\|_2\,dx \tag{3.92}$$

$$\leqslant \int_\Omega \limsup_{h\to 0}\|u^{(j)}(x)-\tilde{u}^{(j),h}(x)\|_2\,dx \tag{3.93}$$

$$= \int_\Omega \limsup_{h\to 0}\|u^{(j)}(x)-u^{(j)}(h\,(\lfloor x/h\rfloor+e/2+z^h))\|_2\,dx. \tag{3.94}$$

The integrand in (3.94) is zero for all x, since $\big\|x-h\,(\lfloor x/h\rfloor+e/2+z^{(h)})\big\|_\infty\leqslant h$ and $u^{(j)}$ is continuous. Therefore

$$\tilde{u}^{(j),h}\;\to\;u^{(j)}\quad\text{in } L^1(\Omega). \tag{3.95}$$

From (3.91) and (3.95) we see that for each j we may choose $h'(j)>0$ such that

$$\max\big\{\|\tilde{u}^{(j),h}-u^{(j)}\|_{L^1},\tilde{J}^h(\tilde{u}^{(j),h})-J(u^{(j)})\big\}\;\leqslant\;\frac{1}{j}\quad\forall h\leqslant h'(j). \tag{3.96}$$

We set $h(1)=\min\{1,h'(1)\}$, $h'(j+1)=\min\Big\{\frac{1}{j+1},h(j),h'(j+1)\Big\}$. Then the sequence $(h(j))_{j\in\mathbb{N}}$ is nonincreasing with $0<h(j)\leqslant 1/j$, and

$$\max\big\{\|\tilde{u}^{(j),h}-u^{(j)}\|_{L^1},\tilde{J}^h(\tilde{u}^{(j),h})-J(u^{(j)})\big\}\;\leqslant\;\frac{1}{j}\quad\forall h\leqslant h(j). \tag{3.97}$$

For sufficiently small h, $j(h):=\max\{j\,|h\leqslant h(j)\}$ exists and is finite. Moreover we have $j(h)\geqslant 1/h\to+\infty$ and $h\leqslant h(j(h))$. Then, due to $h\leqslant h(j(h))$ and (3.97),

$$\limsup_{h\to 0}\,(\tilde{J}^h(\tilde{u}^{(j(h)),h})-J_{\mathcal{C}}(u)) \tag{3.98}$$

$$\leqslant \limsup_{h\to 0}\,(\tilde{J}^h\big(\tilde{u}^{(j(h)),h}\big)-J_{\mathcal{C}}(u^{(j(h))}))+\limsup_{h\to 0}\,(J_{\mathcal{C}}\big(u^{(j(h))}\big)-J_{\mathcal{C}}(u)) \tag{3.99}$$

$$\overset{(3.97)}{\leqslant} \limsup_{h\to 0}\frac{1}{j(h)}+\limsup_{h\to 0}\,(J_{\mathcal{C}}\big(u^{(j(h))}\big)-J_{\mathcal{C}}(u)) \tag{3.100}$$

$$= 0+0. \tag{3.101}$$

The limit in the last equation follows from $j(h) \to \infty$ and (3.67). In the same manner it can be shown that

$$\tilde{u}^{(j(h)),h} \overset{h\to 0}{\to} u \tag{3.102}$$

in L^1. This finally proves (3.64) for the sequence $(\tilde{u}^h)$ with $\tilde{u}^h := \tilde{u}^{(j(h)),h}$. □

Theorem 3.2. *Let f^h, $f_{\mathcal{C}}$ as defined in (3.34), (3.39) for $s \in L^\infty(\Omega), s \geqslant 0$. Then $\tilde{f}^h$ as defined in (3.37) Γ-converges with respect to L^1-convergence to the constrained functional $f_{\mathcal{C}}$ for $h \to 0$ and is equicoercive with respect to the L^1-topology.*

Proof. Denote the data term by

$$g^h(u^h) \;:=\; \sum_{\bar{i}\in\mathcal{J}} h^d \,\langle u^h_{\bar{i}}, s^h_{\bar{i}}\rangle, \tag{3.103}$$

then $f^h = g^h + J^h$. For any $\tilde{u}^h \in \tilde{\mathcal{U}}^h$, we have

$$\begin{aligned}
\tilde{g}^h(\tilde{u}^h) &= \sum_{\bar{i}\in\mathcal{J}} h^d \,\langle u^h_{\bar{i}}, \frac{1}{h^d}\int_{C^h_{\bar{i}}} s\,dx\rangle && (3.104)\\
&= \sum_{\bar{i}\in\mathcal{J}} \int_{C^h_{\bar{i}}} \langle u^h_{\bar{i}}, s\rangle\,dx && (3.105)\\
&= \int_\Omega \langle u^h_{\bar{i}}, s\rangle\,dx && (3.106)\\
&= \int_\Omega \langle \tilde{u}^h, s\rangle\,dx. && (3.107)
\end{aligned}$$

Therefore, since $\tilde{J}^h(u) = +\infty$ for all $u \notin \tilde{\mathcal{U}}^h$, $\tilde{f}^h$ can be represented as

$$\tilde{f}^h(u) = \int_\Omega \langle u, s\rangle\,dx + \tilde{J}^h(u) \tag{3.108}$$

for any $u \in \mathrm{BV}(\Omega)^l$, 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 [Bra02, Rem. 1.7], and $\tilde{J}^h$ Γ-converges to $J_{\mathcal{C}}$ due to Prop. 3.1, $\tilde{f}^h$ Γ-converges to $f_{\mathcal{C}}$.

In order to show equicoercivity, by Prop. A.39 it suffices to provide a lower semicontinuous, coercive function f' with $\tilde{f}^h \geqslant f'$ uniformly for all h. We define the spatially separable $\Psi' \colon \mathbb{R}^{d\times l} \to \mathbb{R}$ by

$$\Psi'\left(\begin{pmatrix} (z^1)^\top \\ \vdots \\ (z^d)^\top \end{pmatrix}\right) \;:=\; \frac{\rho_l}{\sqrt{d}} \sum_{j=1}^d \|z^j\|_2 \tag{3.109}$$

and denote by $f'_{\mathcal{C}}$ and $\tilde{f}'^h$ the corresponding functional and discretization. Then

$$\Psi(z) \geqslant \rho_l \,\|z\|_2 = \rho_l \left(\sum_{j=1}^d \|z^j\|^2\right)^{1/2} \geqslant \frac{\rho_l}{\sqrt{d}} \sum_{j=1}^d \|z^j\|_2 = \Psi'(z), \tag{3.110}$$

therefore

$$\tilde{f}^h \;\geqslant\; \tilde{f}'^h. \tag{3.111}$$

For some $u \in \mathrm{BV}(\Omega)^l$ and $h > 0$, if $u \notin \tilde{\mathcal{U}}^h$ then $\tilde{f}'^h(u) = +\infty \geqslant f'_{\mathcal{C}}(u)$. If $u \in \tilde{\mathcal{U}}^h$ then u is piecewise constant on the grid. In this case, since Ψ' is separable in the directional derivatives, it can be seen that $\tilde{f}'^h(u) = f'_{\mathcal{C}}(u)$. Therefore we conclude $\tilde{f}'^h \geqslant f'_{\mathcal{C}}$ for any $u \in \mathrm{BV}(\Omega)^l$. Substituting this result into (3.111), we obtain the required uniform bound

$$\tilde{f}^h \geqslant f'_{\mathcal{C}}. \tag{3.112}$$

By Prop. 2.2, $f'_{\mathcal{C}}$ is sequentially lower semicontinuous and therefore lower semicontinuous. Moreover, the proof in Prop. 2.3 shows that $f'_{\mathcal{C}}$ is sequentially coercive, and therefore coercive (Prop. A.38). Using these properties, Prop. A.39 provides equicoercivity of the sequence $(\tilde{f}^h)$. □

Remark 3.3. In view of Prop. A.40, Thm. 3.2 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. 3.2 can also be in part applied to pairwise energies for multiclass problems. In particular, consider the multiclass LP relaxation (3.29) with a simple 4-neighborhood, setting $\psi_x(j) = h^d \left(s^h\left(x^{\bar{i}}\right)\right)_j$ and $w_{x,y} = h^{d-1}$. This energy coincides with f^h as in (3.34) for the integrand

$$\Psi(z) \;=\; \sum_{i=1}^{d} \sum_{j=1}^{l} |(z^j)_i|. \tag{3.113}$$

Therefore Thm. 3.2 shows that the (properly weighted) LP relaxation objective actually Γ-converges to the *anisotropic* objective

$$f_1(u) \;=\; \int_\Omega \langle u, s \rangle \, dx + \sum_{i=1}^{d} \int_\Omega d \|D_i u\|_1 + \delta_{\mathcal{C}}(u), \tag{3.114}$$

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

Remark 3.4. Finite-differences discretizations generally do *not* fulfill a generalized coarea formula 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 is often not the optimal approach.

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 (Sect. 2.6), the original continuous problem then admits an integral minimizer. Consequently, 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.

Anisotropy of the Discretization. In order to get a quantitative impression on the anisotropy induced by the various methods, we evaluated several graph-based and finite-differences energies on a labeling u^h rotated by different angles. We compared the following energies:

- Classical pairwise energies with 4-, 8-, and 16- neighborhood (3.14) as depicted in Fig. 3.3. The weights were chosen according to (3.18), see Table 3.1. For non-integral labelings, the pairwise LP relaxation (3.23) was employed.

- The "isometric" finite-differences scheme as outlined in Sect. 3.3, and a variant that also involves backward differences in order to make it more symmetric.

The rotated source labelings were generated in a resolution of 512×512 pixels and downscaled to 128×128 pixels in order to reduce artifacts (Fig. 3.4).

We first evaluated the functionals on *integral* labelings (Fig. 3.5). Since the images were artificially generated, the true expected length can be computed as $\pi + 2 = 5.14$ 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 energies do not seem to work very well: The energy is overestimated, and they show larger oscillations than in the case of the 8-neighborhood.

When the edges of the shape are slightly blurred, the picture changes completely (Fig. 3.6): For a light 4-pixel Gaussian blur, the range of the finite-differences energies over all rotations is already close to one pixel width (an energy difference of 0.024 in

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 finite-differences	$\frac{\sqrt{2}}{2}$	$\frac{2-\sqrt{2}}{2}$		
8-nb finite-differences	$\frac{\sqrt{2}}{2}$	$\frac{2-\sqrt{2}}{4}$		

Table 3.1. Weights used for the graph-based pairwise discretizations (cf. Fig. 3.3).

Figure 3.4. Artificial labelings u^h used for the tests in Fig. 3.5 and Fig. 3.6. **First row:** The integral labelings were downscaled from a larger source image of 512 × 512 pixels, rotated by a number of 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.

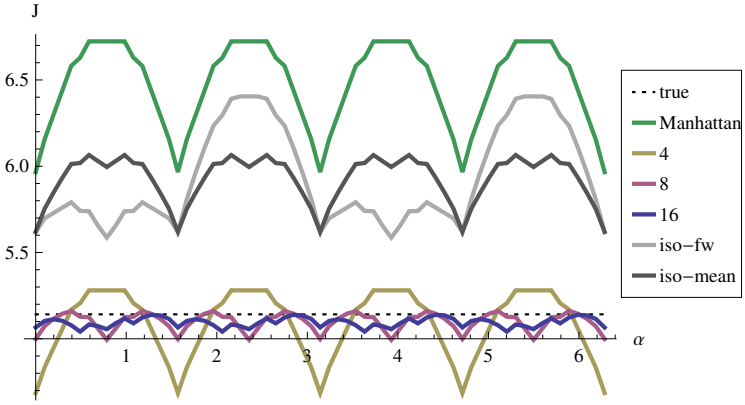

Figure 3.5. Energy of the rotated integral labelings in Fig. 3.4, first row, vs. rotation angle. The pairwise discretizations with 4-, 8-, and 16-neighborhoods exhibit a distinct anisotropy, which decreases as the neighborhood size increases. On such integral labelings the isotropic finite-differences energies overestimate the true length, and are close to the length defined by the "Manhattan" (ℓ^1-) distance.

this scale). For a 10-pixel blur, the anisotropy is barely noticeable. Thus, by allowing a moderate amount of fractional labels, the isotropy of the finite-differences energy can be greatly increased.

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 (16-neighborhood), 6 (8-neighborhood) and 25 (4-neighborhood) pixels, clearly preferring some directions over others.

Consistency of the Discretization. As noted in Sect. 3.2, 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 in fact a problem in practice, we first computed a large half-disc shaped template labeling with a size of 2048 × 2048 pixels. From this template we generated a range of downscaled copies with resolutions down to 32 ×32 (Fig. 3.7). For each resolution, we computed the energies using the above-mentioned regularizers.

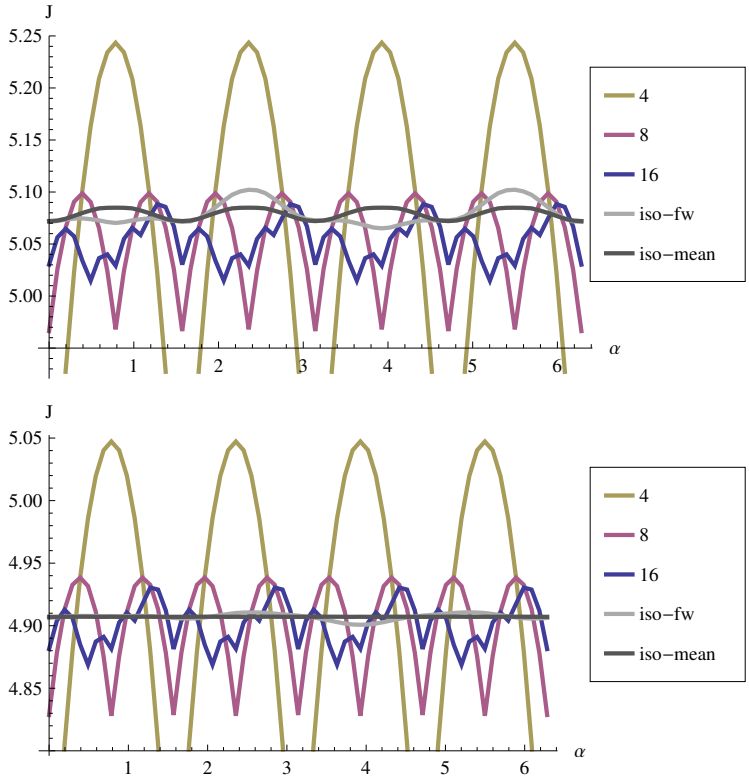

Figure 3.6. Energies for the fractional labelings in Fig. 3.4, second and third row. By allowing a certain amount of factional labels, the isotropy of the finite-differences energies is greatly increased. Note that the true total variation is not available since it is slightly reduced by the smoothing process. This also explains the overall lower energies compared to Fig. 3.5.

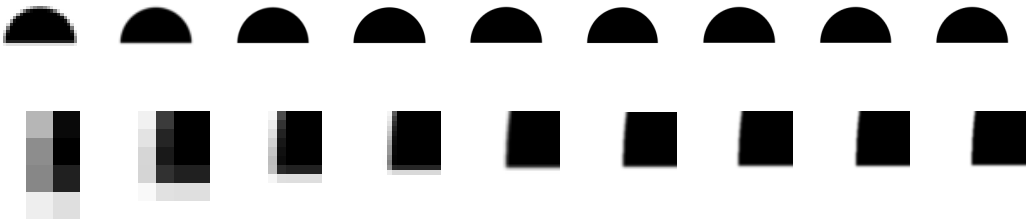

Figure 3.7. Labelings used for comparing the isotropy of the regularizer at different resolutions (Fig. 3.8). **Top row:** Discretization of labeling functions for grid sizes between 32×32 and 2048×2048. **Bottom row:** Detail (lower left corner).

It becomes apparent that the graph cut-, respective LP relaxation-based, energies exhibit the same anisotropy over all scales if the neighborhood size is kept the same, and systematically underestimate the true length (Fig. 3.8). This is in accordance with the observation at the end of Sect. 3.2: 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. 3.1.

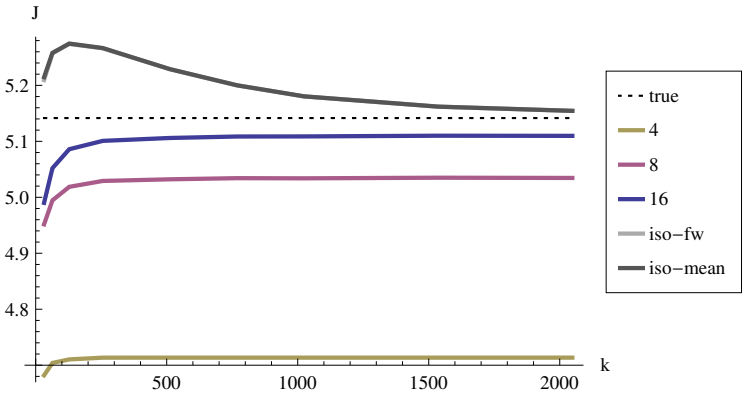

Figure 3.8. Energy comparison for different resolutions. Shown are the energies for the templates in Fig. 3.7 vs. the horizontal grid size $k \in \{32, \ldots, 2048\}$. For a fixed neighborhood, the graph-based energies exhibit a systematic anisotropy, while the finite-differences energies (iso-fw and iso-mean, coinciding in this case) approximate the true energy better as the resolution increases.

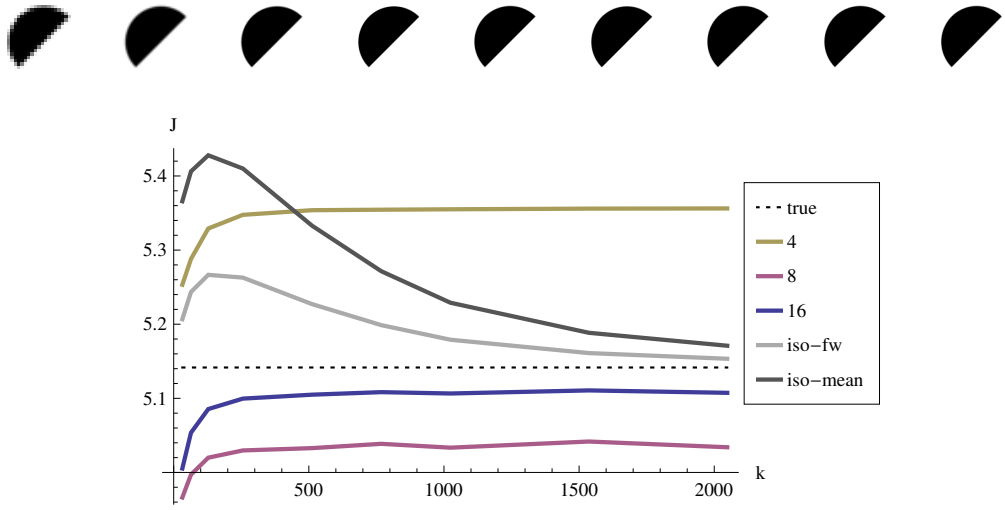

Figure 3.9. Variant of the experiment in Fig. 3.8 for a rotated template. Again, the graph-based discretizations exhibit a systematic error. Due to the rotation, the true length is now overestimated when using a 4-neighborhood. The finite-differences energy converges to the true isotropic length.

Another example can be seen in Fig. 3.9. Here the original template was rotated by 45 degrees. Consequently, the 4-neighborhood energy increases beyond the ground truth, while the finite-differences energies again converge to the true length.

Integral and Fractional Minimizers. In order to see what effect the choice of the discretization has on the minimizer, consider the problem in Fig. 3.10. The input data consists of an image with two circular segments, with several instances generated by rotating the original input. The gray regions are uncertain, and have thus to be filled in by the regularizer. We also added subtle Gaussian noise in order to render the minimizer unique, and therefore integral.

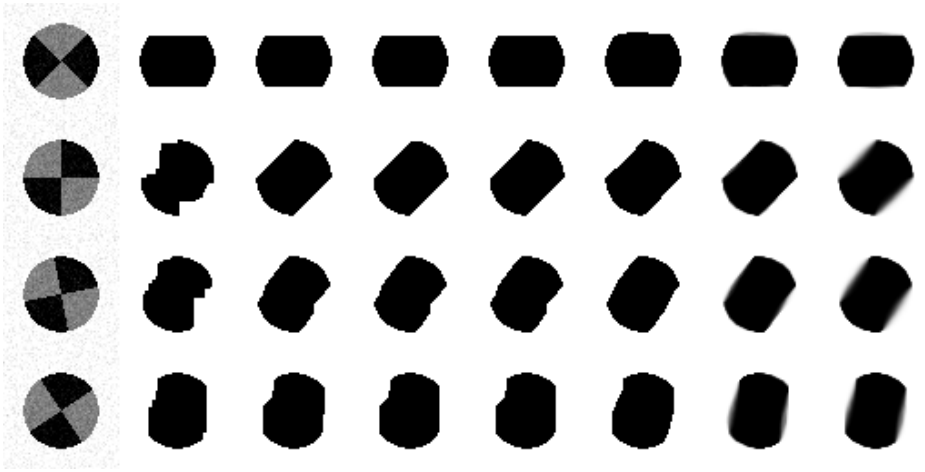

Figure 3.10. Segmentation results for different discretizations. The graph-based discretizations generate artifacts in the solution, even for the large 16-neighborhood. The finite-differences energies (iso-fw, iso-mean) show considerable less artifacts and generate fractional labelings at the boundaries.

The minimization problems were solved to a relative gap of 10^{-8} (cf. Chap. 4) in order to minimize effects caused by suboptimal solutions. The results of the graph-based energies are integral and thus global minimizers of their energies over the set of *integral* labelings.

We compared the results for the graph-based 4-, 8- and 16-neighborhood discretizations and the two finite-differences schemes outlined above. In addition, we included two regularizers proposed in [KSK+08]. 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 [KSK+08] 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, corresponding to a 6- and 8-neighborhood, respectively (Fig. 3.3, Table 3.1). 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 distinct artifacts due to the anisotropy, for at least one of the rotated inputs (Fig. 3.11). Switching to larger neighborhoods reduces the artifacts, however they cannot be completely avoided. In contrast, the finite-differences formulation results in 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. 3.11). 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.

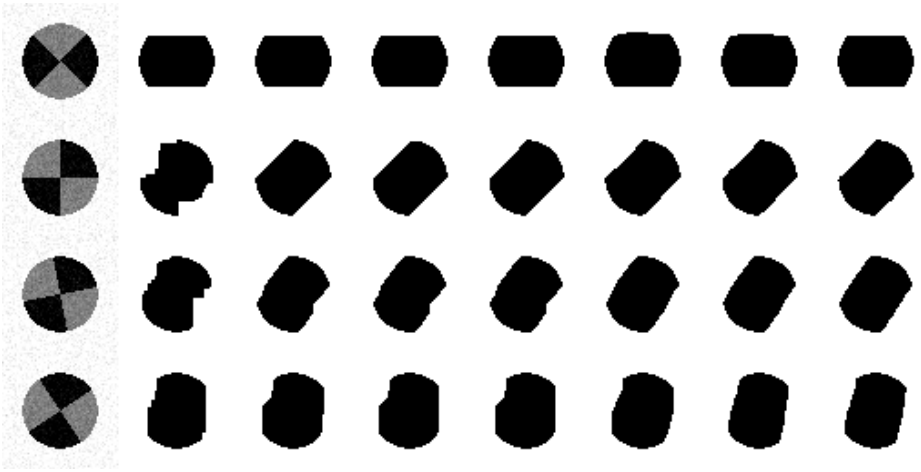

Figure 3.11. Thresholded results from Fig. 3.10. The *thresholded fractional* solutions for the finite-differences energies (iso-fw, iso-mean) result in a better approximation of the desired shape than the *integral* solutions of the graph-based pairwise energies.

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. 3.12 it can be seen that this is not a recommendable approach: integral minimizers of the finite-differences energy are visually clearly inferior to those obtained by thresholding a fractional solution. In fact, the latter are *not* integral minimizers, as can be seen from the energies (Table. 3.2). This indicates that minimizing the finite-differences energy 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 these results were computed on the two-class problem. In the multi-class setting, additional uncertainties are potentially introduced by the relaxation of the continuous problem. Therefore, fractional solutions cannot be unambiguously classified as caused by either the discretization or relaxation the. However, in principle these considerations also apply – with less theoretical justification – to the multi-class case, see also Fig. 3.2.

3.5 Discussion: Integral and Fractional Models

The experiments in the previous chapter 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.

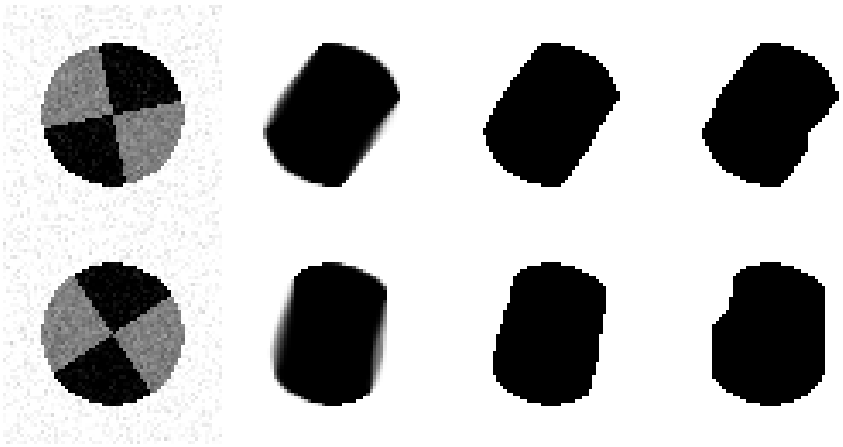

Figure 3.12. Minimizing energies over fractional vs. integral labelings. Finding a minimizer (second from left) of the finite-differences discretization and subsequent thresholding (second from right) results in less artifacts than solving the combinatorial problem of minimizing *exactly the same* energy over the set of integral labelings (right).

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 logical 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 $\ell^h(x^{\bar{i}})$ denotes *in which of the class regions* $\Omega_1, ..., \Omega_l$ *the point* $x^{\bar{i}}$ *is contained*. This is clearly a combinatorial decision, and does not allow any intermediate values. 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 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.

problem	fractional	thresholded	combinatorial
1st row in Fig. 3.12	-2207.13	-2164.41	-2166.86
2nd row in Fig. 3.12	-2213.08	-2177.46	-2181.04

Table 3.2. Energies for the solutions in Fig. 3.12. The thresholded fractional solution is *not* a combinatorial minimizer of the energy, but it is visually clearly superior. Therefore it is not reasonable to minimize the energy under an integrality constraint.

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. 3.2 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 much more better 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 $u^*\colon \Omega \to \{0,1\}$ 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 $u^h\big(x^{\bar{i}}\big)$ 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.

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 $\bar{u}^h\colon \mathbb{R}^n \to \{0,1\}$ *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,

$$\bar{u}^h \;=\; \arg \min_{u'^h\colon \Omega^h \to \{0,1\}} \|u^* - \tilde{u}'^h\|_{L^1}, \tag{3.115}$$

or a similar formulation with a different norm. For the usual L^1 distance, this corresponds to setting $\bar{u}^h\big(x^{\bar{i}}\big) = 1$ 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 $x^{\bar{i}}$.

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.

$$\bar{u}^{h,*} = \arg \min_{u'^h:\Omega^h\to\{0,1\}} \|u^{h,*} - \tilde{u}'^h\|_{L^1}, \quad u^{h,*} = \arg \min_{u^h:\Omega^h\to[0,1]} f^h(u^h). \tag{3.116}$$

This amounts exactly to *thresholding* the fractional values of $u^{h,*}$ to integral values, and again is a purely local operation. The thresholding 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

$$\bar{u}^h = \arg \min_{u^h:\Omega^h\to\{0,1\}} f^h(u^h). \tag{3.117}$$

For an illustration, see Fig. 3.13. 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) threshold.

2. Analogously, it is not reasonable to compare the energy of a thresholded fractional solution to the energy of some other solution obtained via a combinatorial optimization method of the same energy. In particular, the thresholding 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

$$\bar{u}_c^{h,*} = \arg \min_{u^h:\Omega^h\to\{0,1\}} f_c^h(u^h). \tag{3.118}$$

The graph cut approach can be seen as an implementation of this idea (3.118), while the relaxed approach considered in this work conforms to (3.116). We argue that, if one actually requires integral labelings, it is much easier to construct *relaxed* energies f^h such that the thresholded minimizer $\bar{u}^h$ from (3.116) 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. (3.28).

With this in mind, (3.116) 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.

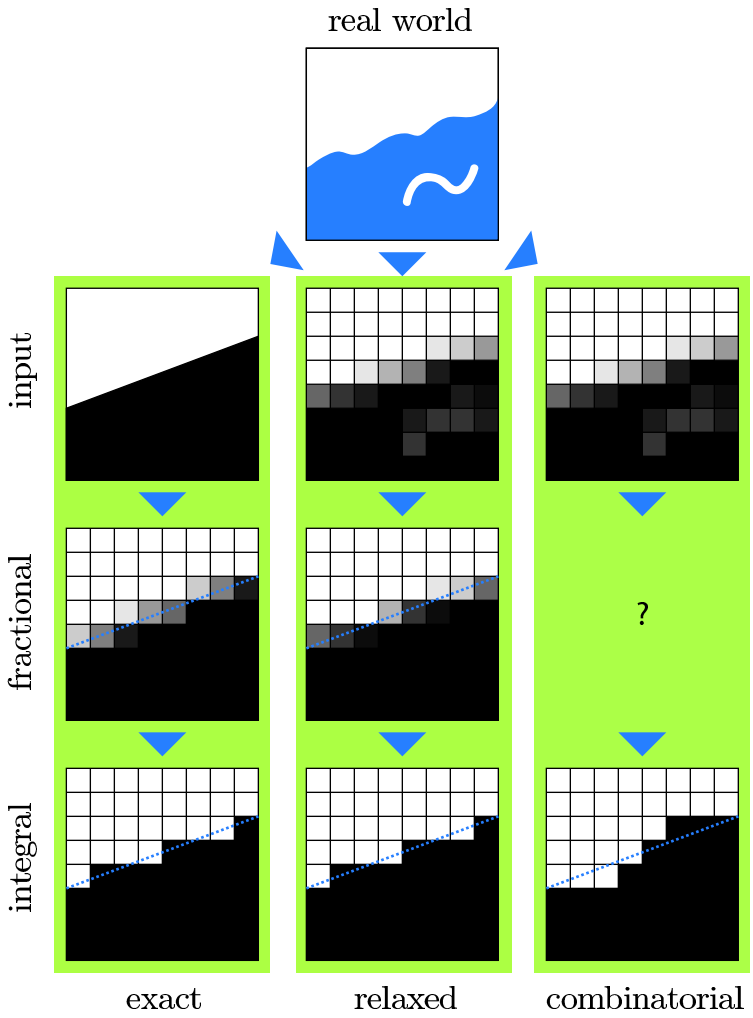

Figure 3.13. Approximation of the optimal segmentation (left) by solving the discretized relaxed problem and thresholding (center) vs. minimizing the same energy as a combinatorial problem. The relaxed approach respects the origin of the data from a continuous world, and returns an approximation to the *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).

Multi-Class Case. Note that the thresholding 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 (Sect. 2.5.1). 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.

3.6 Summary and Further Work

We outlined several approaches to discretize labeling problems formulated on continuous domains. In particular, these include Markov Random Field formulations, functionals with pairwise terms both in the combinatorial and in the relaxed setting, LP relaxation techniques and finite-differences schemes.

The finite-differences functionals converge in the sense of Γ-convergence to the original, spatially continuous functional, and their solutions therefore approximate the true continuous solution as the resolution increases. In contrast to graph-based methods, this can be achieved using local terms, without resorting to an infinitely large neighborhood.

Experiments showed that, if a certain amount of fractional labels is allowed, the finite-differences energies exhibit much less anisotropy than the LP- or graph-based formulations. We also observed that thresholded minimizers of the finite-differences energy closer resemble the optimal segmentation than integral minimizers of graph-based pairwise energies with similar or even larger neighborhoods. In the discussion we investigated this behavior and concluded that:

- Labels in discretizations of a spatially continuous labeling can be associated to *points* or to *pixels/regions*. This decision is important to keep in mind when designing energies.
- In the region-based interpretation, *fractional values occur naturally* even without an explicit relaxation, and are a desired effect in order to better approximate the true spatially continuous solution.
- In order to obtain the optimal *integral approximation* for the true spatially continuous solution, it may be much easier to minimize a relaxed energy and threshold, than to formulate and minimize a combinatorial energy, since the former can convey much more information about the continuous problem.

In particular the last point is important for the following chapter: the approach (3.116) considered here should not be seen as a technique for minimizing a simple combinatorial energy according to (3.117), but rather to approximate the minimizer of the optimal combinatorial energy (3.118), which would otherwise be too difficult to represent.

Concerning further work, an idea to achieve solutions with similar quality using the graph cut approach could be to iteratively increase the neighborhood size around the edges of the segmentation in order to render the energy more isotropic. Moreover, the concept of solving complicated combinatorial problems on originally continuous domains by exploiting the degrees of freedom in formulating a *relaxed* energy and subsequent thresholding seems appealing, and justifies further investigation.

Chapter 4
Nonsmooth Optimization

4.1 Introduction and Overview

When solving the discretized relaxed problem (2.6),

$$\inf_{u\in\mathcal{C}} f(u),\ f(u) := \int_\Omega \langle u(x), s(x)\rangle d\,x + \int_\Omega d\Psi(Du), \tag{4.1}$$

two main issues occur:

- The problem is *large-scale*, since there is at least one variable per label per pixel.
- The problem is also *nonsmooth* due to the regularizer.

However, in view of the dual representation of the regularizer (2.13), the original problem can be reformulated as

$$\inf_{u\in\mathcal{C}} \sup_{v\in\mathcal{D}} g(u,v), \quad g(u,v) := \int_\Omega \langle u(x), s(x)\rangle\, dx - \int_\Omega \langle u(x), \operatorname{Div} v(x)\rangle\, dx, \tag{4.2}$$

$$\mathcal{D} := \{v\in C_c^\infty(\Omega)^{d\times l},\ v(x)\in\mathcal{D}_{\mathrm{loc}}\forall x\in\Omega\}, \tag{4.3}$$

for suitable $\mathcal{D}_{\mathrm{loc}}$, i.e. $\Psi = \sigma_{\mathcal{D}_{\mathrm{loc}}}$. This effectively removes the nonsmoothness at the cost of introducing the dual variables $v(x)$.

We now apply the finite-differences scheme from Chap. 3, i.e. $\Omega^h = \{x^{\bar{i}}\in\mathbb{R}^d | \bar{i}\in\mathcal{J}\}$, $u^h = (u^{h,1}|...|u^{h,l}) \in \mathbb{R}^{n\times l}$. We denote by grad: $=(\mathrm{grad}_1^\top|...|\mathrm{grad}_d^\top)^\top \in \mathbb{R}^{(nd)\times n}$ the standard d-dimensional forward differences gradient operator for Neumann boundary conditions associated with the grid Ω^h, cf. (3.30). Accordingly, div: $=-\mathrm{grad}^\top$ is the backward differences divergence operator for Dirichlet boundary conditions.

Identifying $u^h\in\mathbb{R}^{n\times l}$ with the vector in $\mathbb{R}^{nl}$ obtained by concatenating the columns, these operators extend to $\mathbb{R}^{n\times l}$ via Grad: $=(I_l\otimes\mathrm{grad})$, Div: $=(I_l\otimes\mathrm{div})$. Using these definitions, the discretization of (4.2) can be posed as the finite-dimensional problem

$$\min_{u^h\in\mathcal{C}^h}\max_{v^h\in\mathcal{D}^h} g^h(u^h,v^h), \quad g^h(u^h,v^h) := \langle u^h, s^h\rangle - \langle u^h, \operatorname{Div} v^h\rangle \tag{4.4}$$

$$= \langle u^h, s^h\rangle + \langle \operatorname{Grad} u^h, v^h\rangle, \tag{4.5}$$

$$\mathcal{C}^h := \{u^h\in\mathbb{R}^{n\times l} | u_{\bar{i}}^h\in\Delta_l, \bar{i}\in\mathcal{J}\}, \tag{4.6}$$

$$\mathcal{D}^h := \prod_{\bar{i}\in\mathcal{J}} \mathcal{D}_{loc}\subseteq\mathbb{R}^{n\times d\times l}. \tag{4.7}$$

In the remainder of this chapter we will consider exclusively the finite-dimensional case. Therefore, in a slight abuse of notation, we drop the superscript h, i.e. we denote $u = u^h$, $\mathcal{C} = \mathcal{C}^h$, etc. Motivated by (4.4), we consider the general problem class

$$\min_{u \in \mathcal{C}} \max_{v \in \mathcal{D}} g(u, v), \quad g(u, v) := \langle u, s \rangle + \langle L\, u, v \rangle - \langle b, v \rangle, \tag{4.8}$$

where $\mathcal{C} \subseteq \mathbb{R}^{n \times l}$ and $\mathcal{D} \subseteq \mathbb{R}^{n \times d \times k}$ for some $k \geqslant 1$ are bounded closed convex sets, $L \in \mathbb{R}^{ndk \times nl}$, $s \in \mathbb{R}^{n \times l}$ and $b \in \mathbb{R}^{n \times d \times k}$. The formulation covers (4.4) by setting $k = l$, $L = \mathrm{Grad}$ and $b = 0$, but also more general – even non-uniform, non-isotropic or non-local – regularizers, as considered in Sect. 2.7.

The primal and dual objectives associated with the saddle-point problem (4.8) are

$$f(u) := \max_{v \in \mathcal{D}} g(u, v) \quad \text{and} \quad f_D(v) := \min_{u \in \mathcal{C}} g(u, v), \tag{4.9}$$

respectively. The associated primal and dual problems are

$$\min_{u \in \mathcal{C}} f(u) = \min_{u \in \mathcal{C}} \max_{v \in \mathcal{D}} g(u, v) \quad \text{and} \quad \max_{v \in \mathcal{D}} f_D(v) = \max_{v \in \mathcal{D}} \min_{u \in \mathcal{C}} g(u, v). \tag{4.10}$$

As $\mathcal{C}$ and $\mathcal{D}$ are assumed to be bounded, it follows from [Roc70, Cor. 37.6.2] that a saddle point (u^*, v^*) of g exists. With [Roc70, Lemma 36.2] this implies strong duality, i.e.

$$\min_{u \in \mathcal{C}} f(u) = f(u^*) = g(u^*, v^*) = f_D(v^*) = \max_{v \in \mathcal{D}} f_D(v). \tag{4.11}$$

For multiclass labeling, the sets $\mathcal{C}$ and $\mathcal{D}$ exhibit a product structure, which allows to efficiently apply first-order methods that rely on projections on $\mathcal{C}$ and $\mathcal{D}$, as we demonstrate in the following sections.

Moreover, for $\mathcal{C} = (\Delta_l)^n$, the minimum in the dual objective f_D decouples spatially. Therefore, since $\min_{y \in \Delta_l} \langle y, z \rangle = \mathrm{vecmin}(z) := \min_i z_i$ for all $z \in \mathbb{R}^l$, the dual objective can always be evaluated by summing, over all points $x \in \Omega$, the per-pixel minima over the components of the corresponding entries $(L^\top v + s)_{\bar{i}}$ of $L^\top v + s$,

$$\begin{aligned} f_D(v) &= -\langle b, v \rangle + \min_{u \in (\Delta_l)^n} \langle u, L^\top v + s \rangle & (4.12) \\ &= -\langle b, v \rangle + \sum_{\bar{i} \in \mathcal{J}} \mathrm{vecmin}((L^\top v + s)_{\bar{i}}). & (4.13) \end{aligned}$$

In contrast, the evaluation of the primal objective f can be more difficult, depending on the definition of $\mathcal{D}_{\mathrm{loc}}$ resp. $\mathcal{D}$:

- For simple regularizers, a closed-form expression may be available, such as

$$f(u) = \langle u, s \rangle + \sum_{\bar{i} \in \mathcal{J}} \|\nabla_{\bar{i}}\, u\|_2 \tag{4.14}$$

  in the case of the classical vector-valued total variation $\Psi = \|\cdot\|_2$.

- For the Euclidean embedding approach (Sect. 2.5.2) with some embedding matrix $A \in \mathbb{R}^{k \times l}$, the straightforward approach $L = \mathrm{Grad}$, $b = 0$, $\mathcal{D}_{\mathrm{loc}} = \mathcal{D}^A_{\mathrm{loc}} = \mathcal{B}_1(0)\, A \subseteq \mathbb{R}^{d \times l}$ still permits a closed-form computation,

$$f(u) = \langle u, s \rangle + \sum_{\bar{i} \in \mathcal{J}} \|(\nabla_{\bar{i}}\, u)\, A^\top\|_2, \tag{4.15}$$

however the dual sets $\mathcal{D}_{\text{loc}}$ are more complicated. Alternatively, we may equivalently merge all linear transformations into the linearity L, i.e. set

$$L := (\text{Grad})(A \otimes I_n) \in \mathbb{R}^{ndk \times nl}, \tag{4.16}$$
$$\mathcal{D}_{\text{loc}} := \mathcal{D}_{\text{loc}}^I = \mathcal{B}_1(0) \subseteq \mathbb{R}^{d \times k}. \tag{4.17}$$

This sufficiently simplifies the structure of $\mathcal{D}_{\text{loc}}$ so that the projection-based methods as discussed below can be applied in a straightforward way. This relation was the original motivation for considering the embedding techniques, since it allows to cover a large class of non-standard regularizers at almost no additional cost compared to the standard total variation.

- For the local envelope approach (Sect. 2.5.1) with given metric $d\colon \mathcal{I}^2 \to \mathbb{R}$, the regularizer is only available in implicit form, defined by $L = \text{Grad}$, $b = 0$ and

$$\mathcal{D}_{\text{loc}}^d = \{v = (v^1, ..., v^l) \in \mathbb{R}^{d \times l} | e^\top v = 0\} \cap \bigcap_{i \neq j} \{v | \|v^i - v^j\|_2 \leqslant d(i,j)\}. \tag{4.18}$$

Due to the generality of the regularizer (Prop. 2.6), the primal objective can only be computed approximately (for a special case with three labels there is a derivation in [CCP08]). The same holds for projections on $\mathcal{D}_{\text{loc}}^d$, which have to be approximated iteratively if required. Similar difficulties occur when other advanced regularizers are applied (Sect. 2.5.3, Sect. 2.7).

In the following sections we investigate numerical methods for solving the bilinear saddle-point problem (4.8). Due to the considerations in Chap. 3, we will not consider combinatorial optimization approaches, but focus on the nonsmooth, constrained, convex problem (4.8).

Organization. In this chapter, we are concerned with the numerical optimization of the relaxed problem (4.8). To this end:

- We provide a brief overview over related combinatorial and convex optimization methods (Sect. 4.2).
- We provide and analyze two different methods that are capable of minimizing the specific class of saddle point problems (Sect. 4.3):
  - A specialization of a method for nonsmooth optimization as suggested by Nesterov (Sect. 4.3.1). The method relies on a controlled smoothing technique, is virtually parameter-free and provides explicit a priori and a posteriori optimality bounds for the relaxed problem.
  - A Douglas-Rachford splitting approach (Sect. 4.3.2). We show that the approach allows to compute a sequence of dual iterates that provide an optimality bound and stopping criterion in form of the primal-dual gap, and provide two extensions that allow to deal with complicated constraint sets, as in the case of the local envelope regularizer.

Both methods are highly parallelizable and are shown to converge. For reference, we also summarize the primal-dual technique from [PCBC09], and show how the substeps of the proposed methods can be computed for the multi-class labeling problem (Sect. 4.4). Note that due to the generality of the saddle-point problem (4.8) all methods in this chapter apply equally to pairwise graph-based formulations and general MRF energies with higher-order terms, as long as they are jointly convex in the unit vectors representing the labels.

- We evaluate and compare the performance of the performance of the proposed methods under varying conditions and demonstrate their applicability on real-world problems (Sect. 4.5).

For convenience, a brief summary of the terminology for operator splitting approaches can be found in Appendix A.3.

4.2 Related Work

Combinatorial Optimization. As already indicated in Chap. 3, a large focus in current literature lies on solving *combinatorial* problem formulations, cf. Sect. 3.2. For the reasons discussed in Sect. 3.5, we will not investigate such methods in detail, however we will point out the basic strategies. For a general overview we refer to [BW05].

The two-class graph-based pairwise energy (3.14) is a *minimal cut* problem, and usually considered in its dual form, which is a *maximum flow* problem through the associated graph, i.e. the problem of pushing as much flow as possible through the graph under capacity constraints on the edges [FF62]. Any maximal flow induces a minimal cut of the graph along its saturated edges. These problems occur in many application areas, and have been thoroughly researched [Ber98].

In the field of computer vision, such methods have first been considered in [GPS89] for the restoration of binary images, a special case of the two-class labeling problem. Earlier methods relied on simulated annealing [GG84] or greedy strategies such as Iterated Conditional Modes [Bes86].

Algorithms for solving the maximum flow problem can be categorized into two classes: *Augmenting-path* approaches maintain a feasible flow, i.e. a set of feasible dual variables, and iteratively find paths through the graph along which additional flow can be pushed, which can be achieved by solving a sequence of shortest-path problems [FF56, EK72]. *Preflow-push* (also called *push-relabel*) methods start with an infeasible flow, and gradually decrease infeasibility [GT88]. For general graphs, the latter generally exhibit better overall performance, however for special graphs such as grid- or planar graphs, variants of the augmenting-path method have also proven to be very successful [BK04, STC09]. We also refer to [ABKM10] for a recent improvement of the preflow-push method for imaging applications. All these algorithms are inherently sequential, and therefore difficult to parallelize [GSS82, DKP05], although lately some progress has been made [HVD07, VN08]. However, they allow to solve the two-class pairwise energy problem in polynomial time.

In contrast, the multi-class problem is generally NP-hard [DJPS94, BVZ01, CN04]. For the special case of a submodular objective function (cf. Sect. 2.7.3), several polynomial-time methods exist [Mur03]. The classical approach to approximate a solution of

the *general* multi-class problem is to iteratively solve a sequence of two-class graph cut problems [BVZ01]; such methods are also known as *move-making* methods. Classical approaches include α-*expansion*, where in each step one label competes against all others, and α-β-*swap*, where in each step a pair of labels is selected to compete against each other. These approaches can be improved for linearly ordered label sets by allowing special moves, as in the case of the truncated-linear metric [KVT11], or by greedy heuristics [ZHW10]. In [KT07, KTP08, Kom10], several move-making methods were analyzed and improved from an LP-relaxation perspective, cf. (3.20). Such methods can also be extended in part to a spatially continuous formulation; we refer to Chap. 5 below for a more detailed discussion. In contrast, we directly solve the relaxed problem.

Another classical approach is to relax the combinatorial problem as a *semidefinite program* [WSV00, KSSC03], however this greatly increases the problem size and therefore is only applicable to very small problems. Similar restrictions apply to generic mixed-integer approaches, see e.g. [MPR98]. Besides these specialized methods, the combinatorial labeling problem can also be approached from the more general *graphical model* viewpoint, where an abundance of methods is available; see [Sud06] for an overview. Successful general-purpose methods include tree-reweighted belief propagation [WJW05, Kol06], and more general dual decomposition approaches [KPT07], as well as *pseudo-boolean* optimization [BH02].

However, all these methods share the drawback of being inherently formulated on a discrete, finite feasible set. Due to the reasons outlined in Sect. 3.5, we will instead focus on methods for solving the *relaxed, convex* problem formulation.

Convex Optimization. A large amount of numerical solvers has been proposed for functionals involving total variation terms. Much focus has been laid on the $L^2-\mathrm{TV}$ (ROF) model with quadratic data term, therefore we will consider several approaches and relate them to formulation (4.8). Generally, the approaches can be classified into *primal*, *dual*, and *primal-dual* methods.

Primal methods tackle the primal objective directly. For the unconstrained ROF functional (1.9), this amounts to solving the associated Euler-Lagrange equation

$$-\lambda \mathrm{div}\left(\frac{\nabla u}{\|\nabla u\|_2}\right) + u - I \;=\; 0. \tag{4.19}$$

Eq. (4.19) is then solved by gradient descent [ROF92] or general fixpoint iterations. This approach could in principle also be applied to the primal formulation of (4.8). However, it requires to introduce a smoothing parameter to avoid the case where $\nabla u = 0$, e.g. solve

$$-\lambda \mathrm{div}\left(\frac{\nabla u}{\sqrt{\|\nabla u\|_2^2 + \beta^2}}\right) + u - I \;=\; 0 \tag{4.20}$$

for some $\beta > 0$. Large β lead to smeared edges, and for small β the individual steps become ill-conditioned, and convergence speed, especially for Newton methods, decreases rapidly [VO96]. A similar approach was used in [YYZW08, WABF09], where the Euclidean norm was replaced by a Huber term, with the same drawbacks. As an additional difficulty, and unlike in the case of the ROF problem, solutions of the labeling problem (4.8) are ideally piecewise constant, i.e. satisfy $\nabla u = 0$ almost everywhere. Therefore smoothing techniques should only be applied in a very controlled way.

Due to these issues, *dual* methods have been very successful for solving the ROF problem. It can be seen that for the ROF model, the dual problem is to solve

$$v^* \;=\; \arg \min_{\|v(x)\|_2 \leqslant \lambda \forall x \in \Omega} \frac{1}{2} \|\operatorname{div} v - I\|_2^2, \tag{4.21}$$

which removes the nonsmoothness from the objective, at the cost of introducing pointwise constraints. Formulation (4.21) can then be approached by interior-point [Car01] or particular simple gradient-projection methods [Cha04, Cha05], see also [ZWC10] for further references. A primal solution is then obtained from the dual via $u^* = I - \operatorname{div} v$. Unfortunately, in the case of the labeling problem (4.8), the dual objective (4.13) is nonsmooth as well, i.e. there is no inherent advantage in considering the dual problem. Moreover, reconstructing a primal solution from a dual solution is not trivial, and not necessarily unique. However, very recently it has been shown that by smoothing the dual problem using a log-sum-exp approximation, the dual problem can be solved using projected gradient descent, and under certain conditions a primal solution can still be recovered [BT09a].

We will be primarily concerned with *primal-dual* methods, i.e. methods that track both the primal and dual variables. A first related idea can be found in [CGM99] for the ROF problem, where the authors explicitly introduce a dual variable; however in order to apply a Newton method they again apply a smoothing similar to (4.20). A more recent primal-dual approach for the ROF problem can be found in [HS06], however again the dual problem is regularized, which amounts to replacing the total variation by a Huber term.

If the dual set $\mathcal{D}_{\text{loc}}$ can be formulated in terms of Euclidean norms, problem (4.8) can be posed as a *second-order cone program* (SOCP). Such problems can be solved using primal-dual interior-point approaches, which combine Newton updates with a barrier term that enforces the constraints, and is controlled in a way to keep the Newton method inside its region of superlinear convergence [NN93]. We also refer to [Boy04, Chap. 11] and [BTN01, Ren01] for an overview of the subject.

While interior-point methods have excellent asymptotical convergence properties, their implementation is involved, and exploiting sparsity of the operator L in order to speed up the Newton steps and reduce memory requirements is nontrivial. Moreover, they are not particularly suited well for massively parallel computation, such as on the upcoming GPU platforms.

Therefore we focus on *first-order* primal-dual methods that use only simple operations involving L, and projections on the primal and dual constraint sets $\mathcal{C}$ and $\mathcal{D}$. Such methods have recently been tremendously popular in connection with the minimization of TV-related imaging problems (see [EZC10] for an overview), since they

- allow to exploit sparse problem structure – as is the case if L discretizes a gradient operator – in a particularly straightforward way,
- are relatively simple to implement and analyze,
- can also be formulated in general Hilbert spaces [CP10b],
- involve only basic operations that can be easily parallelized due to their local nature, such as evaluations of L and $L^\top$ and projections on $\mathcal{C}$ and $\mathcal{D}$, and are therefore amenable to the massive parallelization available on the upcoming GPU platforms [ZGFN08].

First-order methods are generally surpassed by higher-order approaches – such as interior-point methods – in terms of the theoretical convergence rate. However, as will be seen in Sect. 4.5, for the low- to medium-accuracy results usually required for imaging applications they may outperform even commercial interior-point solvers.

4.3 First-Order Schemes for Multiclass Labeling

One of the most straightforward first-order approaches for optimizing (4.8) is to fix small primal and dual step sizes τ_P and τ_D, and alternatingly apply projected gradient descent/ascent on the primal/dual variables. This *Arrow-Hurwicz* approach [AHU64] was proposed in a PDE framework for solving the two-class labeling problem in [AT06] and recently used in [CCP08]. An application to denoising problems can be found in [ZC08]. However it seems nontrivial to derive sufficient conditions for convergence. Therefore in [PCBC09, CP10a] the authors propose the *Fast Primal-Dual* (FPD) method, a variant of Popov's saddle point method [Pop80] with provable convergence. The algorithm is summarized in Alg. 4.1.

Due to the explicit steps involved, there is an upper bound condition on the step size to assure convergence, which can be shown to be $\tau_P \tau_D < 1/\|L\|^2$ [PCBC09]. In [CP10a] it was noted that a generalization of the method can be sped up by adapting the step sizes; however this requires at least a part of the objective to be uniformly (and therefore strictly) convex, which is not fulfilled by (4.8). In the experimental section we compare Alg. 4.1 to the two methods proposed below.

Other successful applications of first-order methods include the FISTA algorithm [BT09c, BT10] used for sparse reconstruction, however this requires to compute proximal steps for $J(u)$, i.e. to solve problems of the form

$$\min_{u \in \mathcal{C}} \ \frac{1}{2}\|u - u'\|_2^2 + J(u). \tag{4.22}$$

For sparse reconstruction applications, one usually has $J(u) = \|u\|_1$ and $\mathcal{C} = \mathbb{R}^n$, therefore (4.22) can be solved in closed form using a "shrinkage" operation. In the general case (4.8) however, due to the additional linearity L, (4.22) is only marginally easier than the original problem – it is strictly convex and therefore has a unique solution, in contrast to (4.8).

In the following sections we propose two alternative first-order approaches for solving the saddle-point problem (4.8). The first method is due to Nesterov [Nes04b] and relies on a controlled smoothing combined with a smooth optimization method. The second method relies on the Douglas-Rachford splitting scheme [DR56] and is directly applied to the nonsmooth formulation. Although both approaches are originally formulated on the primal objective, from Thm. 4.1 and Thm. 4.4 it can be seen that they are essentially primal-dual approaches, since they track the dual variables as well.

A common advantage shared by primal-dual methods is that they provide a convenient stopping criterion in form of the numerical *primal-dual* gap $f(u) - f_D(v)$. If both objectives can be evaluated, it provides, for any feasible primal-dual pair $u \in \mathcal{C}$, $v \in \mathcal{D}$, an optimality bound on the primal objective:

$$0 \leqslant f(u) - f(u^*) \leqslant f(u) - f_D(v). \tag{4.23}$$

Algorithm 4.1. Fast Primal-Dual Method (FPD)

Choose $\bar{u}^{(0)} \in \mathbb{R}^{n\times l}$, $v^{(0)} \in \mathbb{R}^{n\times d\times l}$.
Choose $\tau_P > 0, \tau_D > 0$.
$k \leftarrow 0$.
while (not converged)
    $v^{(k+1)} \leftarrow \Pi_{\mathcal{D}}\big(v^{(k)} + \tau_D\big(L\bar{u}^{(k)} - b\big)\big)$.
    $u^{(k+1)} \leftarrow \Pi_{\mathcal{C}}\big(u^{(k)} - \tau_P\big(L^\top v^{(k+1)} + s\big)\big)$.
    $\bar{u}^{(k+1)} \leftarrow 2u^{(k+1)} - u^{(k)}$.
    $k \leftarrow k+1$.
end while

In order to improve scale invariance, the gap is often normalized to the *relative numerical gap* $(f(u) - f_D(v))/f_D(v)$, with the bound

$$\frac{f(u) - f(u^*)}{f(u^*)} \leqslant \frac{f(u) - f_D(v)}{f_D(v)}. \tag{4.24}$$

4.3.1 Nesterov Approach

We first demonstrate how to apply Nesterov's method [Nes04b] to the saddle point problem (4.8). The algorithm has a theoretical worst-case complexity of $O(1/\varepsilon)$ for finding an ε-*optimal* solution, i.e. for finding $u^{(k)} \in \mathcal{C}$ satisfying $f(u^{(k)}) - f(u^*) \leqslant \varepsilon$. It has been shown to give accurate results for denoising [Auj08] and general ℓ_1-norm based problems [WABF09, BBC09]. Besides the desired accuracy, no further parameters have to be provided.

The bound of $O(1/\varepsilon)$ improves on the bound of $O(1/\varepsilon^2)$ for general subgradient methods, which has been shown to be optimal for oracle-based problem formulations [Nes04a]. This is possible since some additional structure is required; specifically, the problem must be of the form

$$\min_{u\in\mathcal{C}} f(u), \qquad f(u) = \hat{f}(u) + \max_{v\in\mathcal{D}} \big(\langle L\,u, v\rangle - \hat{\phi}(v)\big), \tag{4.25}$$

where $\mathcal{C}$ and $\mathcal{D}$ are closed, bounded, convex sets, $\hat{f}$ is differentiable and convex with Lipschitz-continuous gradient with constant $M > 0$, and $\hat{\phi}$ is continuous and convex. It can be seen that this formulation applies to the saddle-point problem (4.8) with $\hat{f}(u) = \langle u, s\rangle$ and $\hat{\phi}(v) = \langle b, v\rangle$.

The inherent nonsmoothness is taken care of by first formally applying a smoothing step and then using a smooth constrained optimization method, however the amount of smoothing is balanced in such a way that the overall number of iterations to produce a solution with a specific accuracy is minimized.

Note that all the considerations below can be extended to general finite-dimensional real vector spaces, i.e. to spaces equipped with other than the Euclidean norm, however for simplicity we only consider the Euclidean case.

Controlled Smooth Approximation. In order to obtain a smooth approximation of f, choose two strongly convex *prox-functions* $p_i \colon \mathcal{D} \to \mathbb{R}_{\geqslant 0}$ with parameter $\sigma_i > 0$, i.e.

$$p_i(v) \;\geqslant\; \frac{1}{2}\sigma_i \,\|v - c^i\|_2^2, \quad \forall v \in \mathcal{D}, i \in \{1,2\} \tag{4.26}$$

for some $c^1 \in \mathcal{C}$, $c^2 \in \mathcal{D}$. Furthermore, set $D_1 := \max_{u \in \mathcal{C}} d_1(u)$ and $D_2 := \max_{v \in \mathcal{D}} d_2(v)$. Then, for $\mu > 0$,

$$f_\mu(u) \;:=\; \hat{f}(u) + \max_{v \in \mathcal{D}} \big(\langle L\,u, v\rangle - \hat{\phi}(v) - \mu\, p_2(v) \big) \tag{4.27}$$

is a *differentiable* function, and its gradient is Lipschitz continuous with constant

$$L_\mu \;:=\; M + \frac{1}{\mu\,\sigma_2}\,\|L\|^2, \tag{4.28}$$

where $\|L\|$ denotes the operator norm, i.e. the spectral norm in our case [Nes04b, Thm. 1]. Moreover, f_μ approximates f via

$$f_\mu(u) \leqslant f(u) \leqslant f_\mu(u) + \mu\, D_2, \tag{4.29}$$

therefore an ε-optimal solution u_μ of f_μ is an $(\varepsilon + \mu\ D_2)$-optimal solution of f. A concise representation of the smoothing process can be obtained by noting that f can be expressed as $f(u) = \hat{f}(u) + \hat{\phi}^*(L\,u)$ and

$$f_\mu(u) \;=\; \hat{f}(u) + (\hat{\phi} + \mu\, p_2)^*(L\,u). \tag{4.30}$$

For $p_2 = \|\cdot\|_2$, the right summand is also known as the *Moreau-envelope* of $\hat{\phi}$ [RW04, Def. 1.22]. Accordingly, evaluating the gradient $\nabla f_\mu(u)$ amounts to solving an optimization problem:

$$\nabla f_\mu(u) \;=\; \nabla \hat{f}(u) + L^\top \arg\max_{v \in \mathcal{D}} \big(\langle L\,u, v\rangle - \hat{\phi}(v) - \mu\, p_2(v) \big). \tag{4.31}$$

Smooth Minimization of Nonsmooth Functions. After obtaining a smooth approximation to f a smooth optimization method can be applied. For a sequence of step sizes $(\alpha^{(k)})$, the method generates a sequence of "helper" points $(x^{(k)})$ and current iterates $(u^{(k)})$ as in defined by Alg. 4.2, such that the invariant

$$f_\mu(u^{(k)}) \;\leqslant\; \frac{1}{A^{(k)}}\frac{L_\mu}{\sigma_1}\,p_1(u_\mu^*) + \sum_{i=0}^{k} \frac{\alpha^{(i)}}{A^{(k)}} \big(f_\mu(x^{(i)}) + \langle \nabla f_\mu(x^{(i)}), u_\mu^* - x^{(i)} \rangle \big) \tag{4.32}$$

is maintained, where $A^{(k)} := \sum_{i=0}^{k} \alpha^{(k)}$, and u_μ^* is the minimizer of f_μ over $\mathcal{C}$. Then, from convexity of f_μ, one obtains

$$f_\mu(u_\mu^*) \leqslant f_\mu(u^{(k)}) \leqslant \frac{1}{A^{(k)}}\frac{L_\mu}{\sigma_1}\,p_1(u_\mu^*) + \sum_{i=0}^{k} \frac{\alpha^{(i)}}{A^{(k)}}\, f_\mu(u_\mu^*) \leqslant \frac{1}{A^{(k)}}\,L_\mu\,\frac{D_1}{\sigma_1} + f_\mu(u_\mu^*). \tag{4.33}$$

Therefore, if (4.32) can be maintained, it follows from (4.29) and (4.33) that

$$f(u^{(k)}) - f(u^*) \leqslant f_\mu(u^{(k)}) - f_\mu(u_\mu^*) + \mu\, D_2 \leqslant \frac{1}{A^{(k)}}\,L_\mu\,\frac{D_1}{\sigma_1} + \mu\, D_2. \tag{4.34}$$

Algorithm 4.2. Basic Nesterov Method

Choose $x^{(0)} \in \mathcal{C}$.
$u^{(0)} \leftarrow \arg\min_{u \in \mathcal{C}} \{ f_\mu(x^{(0)}) + \langle \nabla f_\mu(x^{(0)}), u - x^{(0)} \rangle + \frac{1}{2} L_\mu \, \|u - x^{(0)}\|_2^2 \}$.
for $k = 0, 1, 2, \ldots$ **do**
  $\tau^{(k)} \leftarrow \frac{\alpha^{(k+1)}}{A^{(k+1)}}$.
  $x^{(k+1)} \leftarrow \tau^{(k)} \, z^{(k)} + (1 - \tau^{(k)}) \, u^{(k)}$.
  $u^{(k+1)} \leftarrow \arg\min_{u \in \mathcal{C}} \{ \langle \nabla f_\mu(x^{(k+1)}), u \rangle + \frac{1}{2} L_\mu \, \|u - x^{(k+1)}\|_2^2 \}$.
  $z^{(k+1)} \leftarrow \arg\min_{z \in \mathcal{C}} \{ (L_\mu / \sigma_1) \, p_1(z) + \sum_{i=0}^{k+1} \alpha^{(i)} \, \langle \nabla f_\mu(x^{(i)}), z \rangle \}$.
end for

Setting $\alpha^{(k)} = \frac{k+1}{2}$ as in the original publication, we obtain $A^{(k)} = \frac{(k+1)\,(k+2)}{4}$. For this sequence, and for fixed μ, the algorithm therefore generates ε-optimal solutions of f_μ in $O(1/\sqrt{\varepsilon})$. Substituting (4.28) into (4.34), we obtain

$$f(u^{(k)}) - f(u^*) \;\leqslant\; \frac{1}{A^{(k)}} \left(M + \frac{1}{\mu \, \sigma_2} \, \|L\|^2 \right) \frac{D_1}{\sigma_1} + \mu \, D_2 \tag{4.35}$$

$$\leqslant\; \frac{4}{(k+1)^2} \left(M + \frac{1}{\mu \, \sigma_2} \, \|L\|^2 \right) \frac{D_1}{\sigma_1} + \mu \, D_2. \tag{4.36}$$

Minimizing (4.36) with respect to μ, for some fixed $k = N$, yields the optimal smoothing

$$\mu_N \;:=\; 2 \, \frac{\|L\|}{N+1} \sqrt{\frac{D_1}{\sigma_1 \, \sigma_2 \, D_2}}. \tag{4.37}$$

Using this choice, by substitution of $\mu = \mu_N$ and $k = N$ into (4.36) one obtains the bound

$$f(u^{(N)}) - f(u^*) \;\leqslant\; \frac{4}{(N+1)^2} \, M \, \frac{D_1}{\sigma_1} + \frac{4}{(N+1)} \, \|L\| \sqrt{\frac{D_1 \, D_2}{\sigma_1 \, \sigma_2}}. \tag{4.38}$$

In particular, in our case $\hat{f}(u) = \langle u, s \rangle$, therefore $\nabla \hat{f}$ has Lipschitz constant $M = 0$, and from the resulting

$$f(u^{(N)}) - f(u^*) \;\leqslant\; \frac{4}{(N+1)} \, \|L\| \sqrt{\frac{D_1 \, D_2}{\sigma_1 \, \sigma_2}} \tag{4.39}$$

we conclude that the method can be used to obtain an ε-optimal minimizer of f on $\mathcal{C}$ in $O(1/\varepsilon)$, which improves on the optimal complexity class for general subgradient-based methods, $O(1/\varepsilon^2)$. However, note that (4.39) only holds for the *final* iterate $u^{(N)}$. Intermediate estimates for $k < N$ can be obtained by substituting $\mu = \mu_N$ into (4.36).

Nesterov Method for Multiclass Labeling. The prox-functions p_1 and p_2 can be modified to suit the application and improve the bound. In [Nes04b] the author proposes to use an entropy prox-function for problems involving simplex constraints. However, we found that this only provides good bounds if the constraints consist of few high-dimensional simplices, rather than a large number of low-dimensional simplices as in multiclass labeling problem.

Therefore we chose the Euclidean distance for the prox-functions, $p_1 = p_2 = \frac{1}{2}\|\cdot\|_2^2$. Under this choice, $\sigma_1 = \sigma_2 = 1$, and the inner optimization steps in Alg. 4.2 and (4.31) reduce to the projections $\Pi_{\mathcal{C}}, \Pi_{\mathcal{D}}$ on the sets $\mathcal{C}$ and $\mathcal{D}$ (Appendix A.3). For reference, the complete method applied to the saddle-point problem (4.8) is outlined in Alg. 4.3. While the method is originally derived solely from the primal objective, it is actually a primal-dual method:

Proposition 4.1. *In Alg. 4.3, the iterates $u^{(k)}$ and $v^{(k)}$ are primal and dual feasible, i.e. $u^{(k)} \in \mathcal{C}$, $v^{(k)} \in \mathcal{D}$. Moreover, for any solution u^* of the relaxed problem (4.8), the* a priori *bound*

$$f(u^{(N)}) - f(u^*) \leqslant f(u^{(N)}) - f_D(v^{(N)}) \leqslant \frac{2\, r_1\, r_2\, C}{(N+1)} \tag{4.40}$$

holds for the final iterates $u^{(N)}$, $v^{(N)}$.

Proof. The left inequality in (4.39) is always satisfied. The right inequality follows from [Nes04b, Thm. 3] using the notation $\hat{f}(u) = \langle u,\, s\rangle$, $A = L$, $\hat{\phi}(v) = \langle b,\, v\rangle$, $d_1(u) := \frac{1}{2}\|u - c_1\|^2$, $d_2(v) := \frac{1}{2}\|v - c_2\|^2$, $D_1 = \frac{1}{2}r_1^2$, $D_2 = \frac{1}{2}r_2^2$, $\sigma_1 = \sigma_2 = 1$, $M = 0$. □

Corollary 4.2. *For given $\varepsilon > 0$, applying Alg. 4.3 with*

$$N \;=\; \lceil 2 r_1 r_2\, C\, \varepsilon^{-1} - 1 \rceil \tag{4.41}$$

yields an ε-optimal solution of (4.8), i.e. $f(u^{(N)}) - f(u^*) \leqslant \varepsilon$.

For the finite-differences discretization in Chap. 3, we may choose $c_1 = \frac{1}{l}\, e$ and $r_1 = \sqrt{n(l-1)/l}$, which leads to the following complexity bounds for $u^{(N)}$ with respect to the optimality ε:

Local Envelope Method. (Sect. 2.5.1) For $L = \mathrm{Grad}$, we have $C := 2\sqrt{d} \geqslant \|L\|$ and $c_2 = 0$. Then, we claim that for $v = (v^1, ..., v^l) \in \mathcal{D}_{\mathrm{loc}}^d$,

$$\|v\|_2 \leqslant \min_{i \in \mathcal{I}} \left(\sum_{j \in \mathcal{I}} d(i,j)^2 \right)^{\frac{1}{2}} \tag{4.42}$$

holds. In fact, from the constraint $\sum_{j\in\mathcal{I}} v^j = 0$ in (2.62) we deduce, for arbitrary but fixed label $i \in \mathcal{I}$,

$$\sum_{j \in \mathcal{I}} \|v^j\|_2^2 \;\leqslant\; \left(\sum_{j \in \mathcal{I}} \|v^j\|_2^2 \right) + l\,\|v^i\|_2^2 = \left(\sum_{j \in \mathcal{I}} \|v^j\|_2^2 \right) - 2\langle v^i, \sum_{j \in \mathcal{I}} v^j \rangle + l\,\|v^i\|_2^2 \tag{4.43}$$

$$= \sum_{j \in \mathcal{I}} (\|v^j\|_2^2 - 2\langle v^i, v^j\rangle + \|v^i\|_2^2) = \sum_{j \in \mathcal{I}} \|v^j - v^i\|_2^2 \leqslant \sum_{j \in \mathcal{I}} d(i,j)^2. \tag{4.44}$$

Since i was arbitrary this proves (4.42). Therefore $\mathcal{D}_{\mathrm{loc}}^d \subseteq \mathcal{B}_{\rho_d}(0)$ with

$$\rho_d \;:=\; \min_{i \in \mathcal{I}} \left(\sum_{j \in \mathcal{I}} d(i,j)^2 \right)^{1/2}, \tag{4.45}$$

Algorithm 4.3. Nesterov Multi-Class Labeling

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

and we may set $r_2 = \rho_d \sqrt{n}$. Substituting C, r_1 and r_2 into (4.41) yields the total complexity in terms of the number of iterations

$$O(\varepsilon^{-1} n \sqrt{d}\, \rho_d). \tag{4.46}$$

Embedding method. (Sect. 2.5.2) Here we may set $C = 2\sqrt{d}\|A\|$, $c_2 = 0$ and $r_2 = \sqrt{n}$ for a total complexity of

$$O(\varepsilon^{-1} n \sqrt{d} \|A\|). \tag{4.47}$$

In summary, we arrive at a parameter-free algorithm, with the exception of the desired suboptimality bound. The sequence $(u^{(k)}, v^{(k)})$ allows to compute the current primal-dual gap at each iteration. In addition, as a non-standard feature, the number of required iterations can be determined a priori and independently of the variables in the data term, which could be an advantage in real-time applications, where a fixed response time is required. However, it should be noted that, for fixed μ, the method does *not* necessarily converge to a minimizer of f, but rather to a minimizer of the smoothed function f_μ. An approach to iteratively adapt the smoothing is currently in the process of publication [SKSS11].

4.3.2 Douglas-Rachford Splitting

We now demonstrate how to apply the Douglas-Rachford splitting approach [DR56] to our problem. The Douglas-Rachford approach was thoroughly investigated in [Eck89]. While it was found to be inferior to the specialized polynomial-time methods on classical maximal-flow problems (see also the references in the related work section), it has the distinct advantage that – in contrast to the latter – it can also be applied to the relaxed problem (4.8) or its dual, without assuming a pairwise, graph-based discretization. Therefore it is a good candidate to optimize (4.8) using the finite-differences scheme.

Douglas-Rachford methods have been used successfully for denoising with non-Gaussian noise [CP07], image inpainting, matrix denoising and Poisson noise removal [Set09b, DFN09]. A strong point of the method is that it does not require any part of the objective to be smooth or finite, which allows to introduce constraints as required. Additionally, it has a comparably simple implementation, and is globally convergent.

Under a special splitting, the basic Douglas-Rachford iteration applied to the dual problem can be shown to be equivalent to the Alternating Direction Method of Multipliers [Gab83] and the recently proposed Alternating Split Bregman method [GBO09a, GO09, Set09a, Set09b], hence our results equally apply in these formulations. Very recently, a generalization of the FPD method [PCBC09] has also been proposed [CP10a], that is equivalent to a "preconditioned" Alternating Direction of Multipliers Method and Douglas-Rachford splitting under certain circumstances.

The Douglas-Rachford approach is formulated in the *operator splitting* framework (Appendix A.3) as follows: Assume that the subdifferential $T := \partial f$ can be decomposed into two "simple" operators, $T = A + B$, of which forward and backward steps can practically be computed. In view of Prop. A.45, this is given if $f = f_1 + f_2$ for proper, convex, lsc f_i such that

$$\operatorname{ri}(\operatorname{dom} f_1) \cap \operatorname{ri}(\operatorname{dom} f_2) \neq \emptyset. \tag{4.48}$$

If this is the case, then $T = A + B$ with $A = \partial f_1$, $B = \partial f_2$. Denoting by $J_{\tau S} := (I + \tau S)^{-1}$ the *resolvent* of an operator S (Appendix A.3), the two most basic splitting techniques are the forward-forward and backward-backward fixpoint iterations [Eck89],

$$(\mathrm{FW} - \mathrm{FW}) \quad \bar{u}^{(k+1)} \leftarrow \left(I - \tau^{(k)} B\right)\left(I - \tau^{(k)} A\right)\bar{u}^{(k)}, \tag{4.49}$$

$$(\mathrm{BW} - \mathrm{BW}) \quad \bar{u}^{(k+1)} \leftarrow J_{\tau^{(k)} B}\, J_{\tau^{(k)} A}\, \bar{u}^{(k)}, \tag{4.50}$$

The former corresponds to alternating subgradient descent with the sequence of step sizes $(\tau^{(k)})$, which is generally problematic due to non-uniqueness of the subgradient if the problem is not strictly convex. The latter, while convergent under certain restrictions on the sequence $(\tau^{(k)})$, converges only in the mean, which provokes numerical difficulties [Eck89, Thm. 3.11].

A very common scheme is the forward-backward iteration,

$$(\mathrm{FW} - \mathrm{BW}) \quad \bar{u}^{(k+1)} \leftarrow J_{\tau^{(k)} B}\left(I - \tau^{(k)} A\right)\bar{u}^{(k)}. \tag{4.51}$$

For the special case $f(u) = \delta_{\mathcal{C}}(u) + h(u)$ with Lipschitz-continuous gradient ∇h, the forward-backward method corresponds to projected gradient descent and can be shown to converge with an upper bound on $\tau^{(k)}$ [LP66][Eck89, Thm. 3.12], however convergence is not clear in general. Here we consider the (tight) *Douglas-Rachford-Splitting* iteration [DR56, LM79] with the fixpoint iteration

$$\bar{u}^{(k+1)} \;=\; (J_{\tau A}(2J_{\tau B} - I) + (I - J_{\tau B}))(\bar{u}^{(k)})\,. \tag{4.52}$$

Under the very general constraint that A and B are maximal monotone and $A + B$ has at least one zero, the sequence $(\bar{u}^{(k)})$ is uniquely defined and converges to a point $\bar{u}$ for *any* step size τ, with the additional property that $u := J_{\tau B}(\bar{u})$ is a zero of T and thus a minimizer of f [Eck89, Thm. 3.15, Prop. 3.20, Prop. 3.19]. Maximal monotonicity follows directly if f_1 and f_2 are proper, convex, lower semicontinuous functions.

In addition to the exact convergence result for the Douglas-Rachford approach, there also exists a finite precision convergence result (again, we restrict ourselves to $X := \mathbb{R}^n$):

Algorithm 4.4. Basic Douglas-Rachford Method

1: Choose $\bar{u}^{(0)} \in X$, $\tau > 0$.
2: Choose $(\theta^{(k)}) \subset (0,2)$ s.t. $\sum_{n\in\mathbb{N}} \theta^{(k)}\,(2-\theta^{(k)}) = +\infty$.
3: $k \leftarrow 0$.
4: **while** (not converged)
5: $\quad u^{(k)} \leftarrow P_{\tau f_2}(\bar{u}^{(k)}) = \arg\min_{u\in X} \left\{ \frac{1}{2}\|u - \bar{u}^{(k)}\|^2 + \tau f_2(u) \right\}$.
6: $\quad u'^{(k)} \leftarrow P_{\tau f_1}(2\,u^{(k)} - \bar{u}^{(k)}) = \arg\min_{u\in X} \left\{ \frac{1}{2}\|u - (2\,u^{(k)} - \bar{u}^{(k)})\|^2 + \tau f_1(x) \right\}$.
7: $\quad \bar{u}^{(k+1)} \leftarrow \bar{u}^{(k)} + u'^{(k)} - u^{(k)}$.
8: $\quad k \leftarrow k+1$.
9: **end while**

Proposition 4.3. *[Com04, Cor. 5.2] Let $A, B\colon X \rightrightarrows X$ be maximal monotone with $0 \in A + B$. Let $\tau > 0$, $(\theta^{(k)}) \subseteq (0,2)$, $(a^{(k)}), (b^{(k)}) \subseteq X$ such that*

$$\sum_{k\in\mathbb{N}} \theta^{(k)}\,(2-\theta^{(k)}) = +\infty \quad \text{and} \quad \sum_{k\in\mathbb{N}} \theta^{(k)} \left(\|a^{(k)}\|_2 + \|b^{(k)}\|_2 \right) < +\infty. \tag{4.53}$$

Then the sequence $(\bar{u}^{(k)})$ generated by the iteration

$$u^{(k)} \leftarrow J_{\tau B}\bar{u}^{(k)} + b^{(k)}, \tag{4.54}$$
$$u'^{(k)} \leftarrow J_{\tau A}(2\,u^{(k)} - \bar{u}^{(k)}) + a^{(k)},$$
$$\bar{u}^{(k+1)} \leftarrow \bar{u}^{(k)} + \theta^{(k)} \left(u'^{(k)} - u^{(k)} \right). \tag{4.55}$$

converges to some $u \in X$ with

$$0 \;\in\; (A+B)(J_{\tau B}\,u). \tag{4.56}$$

Alg. 4.4 shows the complete algorithm including the proximal steps for computing a minimizer of the problem

$$\min_{u\in X} \{ f_1(u) + f_2(u) \}. \tag{4.57}$$

In the following, we will generally set the overrelaxation parameter $\theta^{(k)}$ to 1.

There is generally no unique way of how to choose the splitting $f = f_1 + f_2$. If the splitting is not chosen carefully, evaluating the resolvents (steps 5 and 6 in Alg. 4.4) becomes a nontrivial problem, which then has to be solved iteratively. This is undesirable, as it would require to solve the inner problems with increasing accuracy to ensure convergence, see (4.53). In the following, we consider several splitting techniques with an increasing number of auxiliary variables, leading to a decreasing difficulty of the inner problems.

4.3.2.1 Primal Constraint Splitting

The most straightforward approach is to split (4.8) according to

$$\min_u \{ \underbrace{\delta_{\mathcal{C}}(u)}_{f_1(u)} + \underbrace{\max_{v\in\mathcal{D}} g(u,v)}_{f_2(u)} \}, \tag{4.58}$$

which results in Alg. 4.5 as published in [LKY+09]. Convergence follows from Prop. A.45 and Prop. 4.3 if $\mathrm{ri}(\mathcal{C}) \neq 0$, since $\mathcal{D}$ is bounded and thus f_2 has full domain.

Step 5 can be easily solved by computing

$$u'^{(k)} \leftarrow \Pi_{\mathcal{C}}(2u^{(k)} - \bar{u}^{(k)}). \tag{4.59}$$

Solving the inner problem in step 4 is more difficult. While it is an unconstrained problem, it involves the full primal objective f_2 together with a quadratic data term. This problem is similar to the Rudin-Osher-Fatemi problem (cf. Sect. 1.1), and can be solved using related methods, such as forward-backward [Cha05, DAV08] or half-quadratic methods [YYZW08]; see [LKY+09] for a comparison.

However, we observed that the inner problem in step 4 requires considerable parameter tuning and is a major obstacle to quickly obtaining accurate solutions. In fact, it is not much simpler than the original problem itself. Therefore we do not evaluate this approach in detail and refer to [LKY+09] instead.

A notable special case occurs when Ψ – and therefore the dual constraint set $\mathcal{D}$ – is separable in the labels, for instance

$$\Psi(z = (z^1 | \ldots | z^l)) \;=\; \|z^1\|_2 + \ldots + \|z^l\|_2. \tag{4.60}$$

In this case, step 4 additionally decouples into l separate ROF-type problems, and the method can be seen as a parallel analogon to the α-expansion method [BVZ01]: in the i-th of the l separate ROF-problems, label i ($u_i(x) = 1$) competes against all other labels ($u_i(x) = 0$). However, compared to α-expansion, the individual problems are solved in parallel, rather than sequentially, and the obtained (fractional) solutions are merged afterwards in step 5. A similar approach has very recently been proposed for the dual problem [YBTB10], with the same issue of requiring to iteratively solve the difficult inner problems.

4.3.2.2 Auxiliary Variables

Due to the above-mentioned shortcomings, we propose an alternative approach, following the procedure in [EB92, Set09b] of adding auxiliary variables before splitting the objective in order to simplify the individual steps of the algorithm; see [LS10] for the corresponding technical report. We introduce $w = L\,u$ and split according to

$$\min_{u \in \mathcal{C}} \max_{v \in \mathcal{D}} \{\langle u, s\rangle + \langle L\,u, v\rangle - \langle b, v\rangle\} \tag{4.61}$$

$$= \min_{u} \{\langle u, s\rangle + \sigma_{\mathcal{D}}(L\,u - b) + \delta_{\mathcal{C}}(u)\} \tag{4.62}$$

$$= \min_{u,w} h(u,w), \quad h(u,w) := \underbrace{\delta_{Lu=w}(u,w)}_{h_1(u,w)} + \underbrace{\langle u, s\rangle + \delta_{\mathcal{C}}(u) + \sigma_{\mathcal{D}}(w - b)}_{h_2(u,w)}. \tag{4.63}$$

We apply the tight Douglas-Rachford iteration (Alg. 4.4) to this formulation using $A = \partial h_1$ and $B = \partial h_2$: Denote

$$(u^{(k)}, w^{(k)}) \;:=\; J_{\tau B}(\bar{u}^{(k)}, \bar{w}^{(k)}), \tag{4.64}$$

$$(u'^{(k)}, w'^{(k)}) \;:=\; J_{\tau A}(2J_{\tau B} - I)(\bar{u}^{(k)}, \bar{w}^{(k)}) \tag{4.65}$$

$$=\; J_{\tau A}(2u^{(k)} - \bar{u}^{(k)}, 2w^{(k)} - \bar{w}^{(k)}). \tag{4.66}$$

Algorithm 4.5. Basic Douglas-Rachford Multi-Class Labeling

1: Choose $u^{(0)} \in \mathbb{R}^{n\times l}$, $\tau > 0$.
2: $k \leftarrow 0$.
3: **while** (not converged)
4: $\quad u^{(k)} \leftarrow \arg\min_u \Big\{ \frac{1}{2}\|u - \bar{u}^{(k)}\|_2^2 + \tau\,(\langle u, s\rangle + \sigma_{\mathcal{D}}(L\,u - b)) \Big\}$.
5: $\quad u'^{(k)} \leftarrow \arg\min_u \{ \frac{1}{2}\|u - (2u^{(k)} - \bar{u}^{(k)})\|_2^2 + \delta_{\mathcal{C}}(u) \}$.
6: $\quad \bar{u}^{(k+1)} \leftarrow \bar{u}^{(k)} + \theta^{(k)}\,(u'^{(k)} - u^{(k)})$.
7: $\quad k \leftarrow k+1$.
8: **end while**

Then, according to (4.52), $(\bar{u}^{(k+1)}, \bar{w}^{(k+1)}) = (\bar{u}^{(k)} + u'^{(k)} - u^{(k)}, \bar{w}^{(k)} + w'^{(k)} - w^{(k)})$. Evaluating the resolvent $J_{\tau B}$ is equivalent to a proximal step on h_2; moreover due to the introduction of the auxiliary variables the computation decouples:

$$\begin{aligned} u^{(k)} &= \arg\min_u \{\frac{1}{2}\|u - \bar{u}^{(k)} + \tau s\|_2^2 + \delta_{\mathcal{C}}(u)\} && (4.67)\\ &= \Pi_{\mathcal{C}}\big(\bar{u}^{(k)} - \tau s\big), && (4.68)\\ w^{(k)} &= \arg\min_w \{\frac{1}{2\tau}\|w - \bar{w}^{(k)}\|_2^2 + \sigma_{\mathcal{D}}(w - b)\} && (4.69)\\ &= \bar{w}^{(k)} - \tau\Pi_{\mathcal{D}}\Big(\frac{1}{\tau}(\bar{w}^{(k)} - b)\Big). && (4.70) \end{aligned}$$

In a similar manner, $J_{\tau A}$ resp. the proximal step on h_1 amounts to the least-squares minimization problem

$$\begin{aligned} (u'^{(k)}, w'^{(k)}) &= \arg\min_{u',w'} \{\delta_{Lu'=w'} + \\ &\qquad \frac{1}{2\tau}\big(\|u' - (2\,u^{(k)} - \bar{u}^{(k)})\|_2^2 + \|w' - \big(2\,w^{(k)} - \bar{w}^{(k)}\big)\|_2^2\big). \end{aligned} \tag{4.71}$$

Substituting the constraint $w'^{(k)} = L\,u'^{(k)}$ yields

$$\begin{aligned} u'^{(k)} &= \arg\min_{u'} \{\|u' - (2\,u^{(k)} - \bar{u}^{(k)})\|_2^2 + \|L\,u' - \big(2\,w^{(k)} - \bar{w}^{(k)}\big)\|_2^2\} && (4.72)\\ &= (I + L^\top L)^{-1}\big((2\,u^{(k)} - \bar{u}^{(k)}) + L^\top\big(2\,w^{(k)} - \bar{w}^{(k)}\big)\big). && (4.73) \end{aligned}$$

By the substitution $w''^{(k)} := \Pi_{\mathcal{D}}\big(\tau^{-1}(\bar{w}^{(k)} - b)\big) = \frac{1}{\tau}(\bar{w}^{(k)} - w^{(k)})$, one obtains Alg. 4.6.

Compared to Alg. 4.5, the individual steps in Alg. 4.6 are much simpler, involving mainly projections on $\mathcal{C}$ and $\mathcal{D}$, and evaluations of L and $L^\top$. For the finite-differences discretization with Neumann boundary conditions, solving the linear equation system (4.73) can be greatly accelerated by exploiting the fact that $\mathrm{grad}^\top\mathrm{grad}$ diagonalizes under the discrete cosine transform (DCT-2) [Str99, LKY+09]:

$$\mathrm{grad}^\top\mathrm{grad} = B^{-1}\mathrm{diag}(c)B \tag{4.74}$$

where B is the orthogonal transformation matrix of the DCT and c is the vector of eigenvalues of the discrete Laplacian. More generally, assume that L is of the form $L = A \otimes \mathrm{grad}$ for some (possibly identity) matrix $A \in \mathbb{R}^{k\times l}$, $k \leqslant l$. This also covers the embedding approach (Sect. 2.5.2).

Algorithm 4.6. Douglas-Rachford with Auxiliary Variables (DR)

Choose $\tau > 0$, $\bar{u}^{(0)} \in \mathbb{R}^{n \times l}$, $\bar{w}^{(0)} \in \mathbb{R}^{n \times d \times l}$.
$k \leftarrow 0$.
while (not converged)
  $u^{(k)} \leftarrow \Pi_{\mathcal{C}}\big(\bar{u}^{(k)} - \tau s\big)$.
  $w''^{(k)} \leftarrow \Pi_{\mathcal{D}}\Big(\frac{1}{\tau}(\bar{w}^{(k)} - b)\Big)$.
  $u'^{(k)} \leftarrow (I + L^\top L)^{-1}\Big((2u^{(k)} - \bar{u}^{(k)}) + L^\top(\bar{w}^{(k)} - 2\tau w''^{(k)})\Big)$.
  $w'^{(k)} \leftarrow L\, u'^{(k)}$.
  $\bar{u}^{(k+1)} \leftarrow \bar{u}^{(k)} + u'^{(k)} - u^{(k)}$.
  $\bar{w}^{(k+1)} \leftarrow w'^{(k)} + \tau w''^{(k)}$.
  $k \leftarrow k + 1$.
end while

We first compute the decomposition $A^\top A = V^{-1}\mathrm{diag}(a)V$ with $a \in \mathbb{R}^l$ and an orthogonal matrix $V \in \mathbb{R}^{l \times l}$, $V^{-1} = V^\top$. We claim that

$$(I + L^\top L)^{-1} \;=\; (V^\top \otimes I_n)(I_l \otimes B^{-1})\,(I + \mathrm{diag}(a) \otimes \mathrm{diag}(c))^{-1}(I_l \otimes B)(V \otimes I_n). \tag{4.75}$$

To see (4.75), note that $(P \otimes Q)(R \otimes S) = (PR) \otimes (Q\,S)$ for matrices P, Q, R, S with compatible dimensions, and therefore

$$\begin{aligned}
(I + L^\top L)^{-1} &= (I + (A \otimes \mathrm{grad})^\top (A \otimes \mathrm{grad}))^{-1} && (4.76)\\
&= (I + (A^\top A) \otimes (\mathrm{grad}^\top \mathrm{grad}))^{-1} && (4.77)\\
&= (I + (V^{-1}\mathrm{diag}(a)V) \otimes (B^{-1}\mathrm{diag}(c)B))^{-1} && (4.78)\\
&= (I + (V^{-1} \otimes B^{-1})(\mathrm{diag}(a) \otimes \mathrm{diag}(c))(V \otimes B))^{-1} && (4.79)\\
&= ((V^{-1} \otimes B^{-1})(I + \mathrm{diag}(a) \otimes \mathrm{diag}(c))(V \otimes B))^{-1} && (4.80)\\
&= (V^{-1} \otimes B^{-1})(I + \mathrm{diag}(a) \otimes \mathrm{diag}(c))^{-1}(V \otimes B). && (4.81)
\end{aligned}$$

Using $V^{-1} = V^\top$, (4.75) follows. Thus step 6 in Alg. 4.6 can be achieved fast and accurately through matrix multiplications with V, discrete cosine transforms, and one $O(n\,l)$ product for inverting the inner diagonal matrix.

Similar to the Nesterov approach, it can be shown that Alg. 4.6 is a primal-dual method, with the sequence $(u^{(k)}, w''^{(k)})$ of primal-dual feasible pairs (see also [Eck89, Prop. 3.42] for a similar result):

Proposition 4.4. *Let $\mathcal{C}$, $\mathcal{D}$ closed convex bounded sets with $\mathrm{ri}(\mathcal{C}) \neq \emptyset$ and $\mathrm{ri}(\mathcal{D}) \neq \emptyset$. Then Alg. 4.6 generates a sequence of primal/dual feasible pairs $(u^{(k)}, w''^{(k)}) \in \mathcal{C} \times \mathcal{D}$ converging to a saddle point (u^*, v^*) of the relaxed problem (4.8).*

Proof. The primal-dual feasibility is clear from the definition of the algorithm, since $u^{(k)}$ and $w''^{(k)}$ are obtained by projections on $\mathcal{C}$ and $\mathcal{D}$, respectively. Convergence follows as a special case from the more general Prop. 4.5 and Prop. 4.6 below applied to the dual problem (4.10), i.e. substitute $v \leftrightarrow u$, $\mathcal{C} \leftrightarrow \mathcal{D}$, $b \leftrightarrow s$, $L \leftrightarrow -L^\top$, and set $r = 1$, $\mathcal{D}_1 = \mathcal{D}$. □

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

$$f(u^{(k)}) - f_D(w''^{(k)}) \tag{4.82}$$

as a stopping criterion.

4.3.2.3 Multiple-Constraint Dual Variables

The above approach is still restricted in that it requires projections on the constraint sets. This generally does not pose a problem with respect to the primal constraint set $\mathcal{C}$, however the tight relaxation proposed in Sect. 2.5.1 and the lifting methods in 2.7 both result in an intricate dual constraint set $\mathcal{D}$.

This prohibits a projection on $\mathcal{D}$ in closed form (Alg. 4.6 step 5) and requires an inexact iterative projection instead, causing a number of issues: From a theoretical viewpoint, convergence of the outer algorithm usually requires the inner problem to be solved with an increasing accuracy at each step (Prop. 4.3), which is impractical. Thus in practice convergence is no longer guaranteed. In addition, the projections become very slow and raise many questions on how to choose suitable and matching stopping criteria, possibly introducing accuracy and convergence issues.

However, note that in both the envelope-regularized labeling problems and the lifting problems one faces discretized problems of the form

$$\min_{u \in \mathcal{C}} \max_{v \in \mathcal{D}_1 \cap \ldots \cap \mathcal{D}_r} \{\langle u, s\rangle + \langle L\,u, v\rangle - \langle b, v\rangle\}, \tag{4.83}$$

where $\mathcal{D} = \mathcal{D}_1 \cap \ldots \cap \mathcal{D}_r$ for some $r \in \mathbb{N}$, and projections on $\mathcal{D}_i$ can be computed in closed form, cf. (2.62) and (2.141)–(2.143). We now show how to add auxiliary variables before splitting the objective in order to exploit this structure, avoiding the iterative projections and the associated accuracy and convergence issues [LBS10]. Instead of solving (4.83) directly, we solve the *dual* problem and additionally introduce auxiliary variables z and $v_1, \ldots, v_r$, leading to the equivalent problem

$$\min_{v_i} \{\underbrace{\delta_{-L^\top\left(\frac{1}{r}\sum_i v_i\right) = z, v_1 = \ldots = v_r}}_{f_1} + \underbrace{\textstyle\sum_i \delta_{\mathcal{D}_i}(v_i) + \langle \frac{1}{r}\sum_i v_i, b\rangle + \max_{u \in \mathcal{C}} \langle u, z - s\rangle}_{f_2}\}. \tag{4.84}$$

The extra constraints are represented as characteristic functions δ assuming the values $\{0, +\infty\}$. Applying the Douglas-Rachford method to the above splitting formulation leads to the complete *Dual Multiple-Constraint Douglas-Rachford* (DMDR) algorithm as outlined in Alg. 4.7.

Due to the auxiliary variables, the backward step for f_2 requires only separate projections on the $\mathcal{D}_i$ instead of the complete set $\mathcal{D}$. The backward step for f_1 can be accelerated as in the previous section. Convergence of Alg. 4.7 follows directly from a mild condition on the relative interiors ri of the domains:

Proposition 4.5. *Assume that $\mathcal{D} = \mathcal{D}_1 \cap \ldots \cap \mathcal{D}_r$ and $\mathcal{C}$ be closed convex bounded sets such that all $\mathcal{D}_i$ are closed, convex and bounded,* $\mathrm{ri}(\mathcal{D}_1) \cap \ldots \cap \mathrm{ri}(\mathcal{D}_r) \neq \emptyset$ *and* $\mathrm{ri}(\mathcal{C}) \neq \emptyset$. *Then the sequence $(v_1^{(k)}, \ldots, v_r^{(k)}, z''^{(k)})$ in Alg. 4.7 converges to a primal-dual feasible point $(v_1, \ldots, v_r, z'') \in \mathcal{D}_1 \times \ldots \times \mathcal{D}_r \times \mathcal{C}$.*

Algorithm 4.7. Dual Multiple-Constraint Douglas-Rachford (DMDR)

Choose $\tau > 0$, $\bar{v}_i^{(0)} \in \mathbb{R}^{n \times d \times l}$, $\bar{z}^{(0)} \in \mathbb{R}^{n \times d}$.
$k \leftarrow 0$.
while (not converged)
  $v_i^{(k)} \leftarrow \Pi_{\mathcal{D}_i}\Big(\bar{v}_i^{(k)} - \frac{\tau}{r} b \Big)$.
  $z''^{(k)} \leftarrow \Pi_{\mathcal{C}}\Big(\frac{1}{\tau}\big(\bar{z}^{(k)} - s \big) \Big)$.
  $v'^{(k)} \leftarrow (r\, I + L\, L^\top)^{-1}\Big(\sum_i \Big(2 v_i^{(k)} - \bar{v}_i^{(k)} \Big) - L\Big(\bar{z}^{(k)} - 2\tau z''^{(k)} \Big) \Big)$.
  $v_1'^{(k)} = ... = v_r'^{(k)} \leftarrow v'^{(k)}$.
  $z'^{(k)} \leftarrow (-L^\top) v'^{(k)}$.
  $\bar{v}_i^{(k+1)} \leftarrow \bar{v}_i^{(k)} + v_i'^{(k)} - v_i^{(k)}$.
  $\bar{z}^{(k+1)} \leftarrow z'^{(k)} + \tau z''^{(k)}$.
  $k \leftarrow k + 1$.
end while

Proof. Since $\mathcal{C}$ is closed we have

$$\begin{aligned} \operatorname{ri}(\operatorname{dom} f_2) \cap \operatorname{ri}(\operatorname{dom} f_1) &= \operatorname{ri}(\operatorname{dom} f_2) \cap \{ v_1 = ... = v_r, -A^\top v_i = z \} && (4.85) \\ &= \{ (v, ..., v, -A^\top v)^\top | v \in \operatorname{ri}(\mathcal{D}_1) \cap ... \cap \operatorname{ri}(\mathcal{D}_r) \} && (4.86) \end{aligned}$$

This set is nonempty by the assertion, which with Prop. A.45, Prop. 4.3 and continuity of the projections implies convergence. □

In particular, the convergence of the sequence $(v_i^{(k)})$ guarantees that from some point on the constraints hold *exactly*. Then $v_1^{(k)} = ... = v_r^{(k)} =: v^{(k)}$, and $v^{(k)}$ converges to a solution v of the dual problem (4.10). Unfortunately, it is nontrivial to generate a primal solution u from a single dual solution, as both the dual and the primal problem are usually not strictly convex. However, it turns out that the above algorithm additionally returns a primal solution:

Proposition 4.6. *Let $(v := v_1 = ... = v_r, z'')$ be a limit of Alg. 4.7. Then (z'', v) is a saddle point of the problem (4.8), i.e. $u := z''$ is a primal solution.*

Proof. We show the saddle point property

$$g(z'', \tilde{v}) \leqslant g(z'', v) \leqslant g(\tilde{u}, v) \quad \forall \tilde{u} \in \mathcal{C}, \tilde{v} \in \mathcal{D}. \tag{4.87}$$

In the following, all variables without index denote the limits of their corresponding sequences, i.e. $v^{(k)} \to v$, $z''^{(k)} \to z''$ etc. As all operations in the algorithm are continuous, the relations between the iterates transfer to their limits. We further define $z := \bar{z} - \tau z''$. From the definition of the algorithm we obtain

$$\begin{aligned} v_i &= v_i', && (4.88) \\ -L^\top v = -L^\top v_i = -L^\top v_i' &= \bar{z} - \tau z'' = z. && (4.89) \end{aligned}$$

From the Douglas-Rachford convergence theorem [Eck89, Prop. 3.19], it follows that

$$\tau^{-1}(\bar{v}_1 - v_1, ..., \bar{v}_r - v_r, \bar{z} - z)^\top \in \partial f_2(v_1, ..., v_r, z). \tag{4.90}$$

As f_2 is separable in v_i and z, the subdifferential decomposes into a direct product and thus $r+1$ separate equations:

$$\tau^{-1}(\bar{v}_i - v_i) \in N_{\mathcal{D}_i}(v_i) + r^{-1} b, \quad u \in \{1, ..., r\}. \tag{4.91}$$

$$\tau^{-1}(\bar{z}_i - z_i) \in \arg\max_{u \in \mathcal{C}} \langle u, z - s \rangle. \tag{4.92}$$

We now use the fact that

$$N_{\mathcal{D}_1}(v_1) + ... + N_{\mathcal{D}_r}(v_r) = N_{\mathcal{D}_1}(v) + ... + N_{\mathcal{D}_r}(v) = N_{\mathcal{D}_1 \cap ... \cap \mathcal{D}_r}(v) = N_{\mathcal{D}}(v), \tag{4.93}$$

which follows from the assumption $v = v_i$ and [RW04, Cor. 10.9]; here the convexity, closedness and nonseparability of the $\mathcal{D}_i$ is required. Summing up (4.91) and using (4.93), we arrive at

$$\tau^{-1} \sum_i (\bar{v}_i - v_i) \in \sum_i N_{\mathcal{D}_i}(v_i) + b = N_{\mathcal{D}}(v) + b. \tag{4.94}$$

From the definition of the algorithm we also obtain $\tau^{-1} \sum_i (\bar{v}_i - u_i) = L\, z''$, therefore

$$L\, z'' \in N_{\mathcal{D}}(v) + b, \tag{4.95}$$

which shows that $v \in \arg\max_{v \in \mathcal{D}} \langle L\, z'' - b, v \rangle$, i.e. the left inequality in (4.87). To show the right inequality, we use (4.89) and (4.92) to obtain

$$z'' \in \arg\max_{u \in \mathcal{C}} \langle u, z - s \rangle = \arg\max_{u \in \mathcal{C}} \langle u, -L^\top v - s \rangle = \arg\min_{u \in \mathcal{C}} (\langle L\, u, v \rangle + \langle u, s \rangle). \tag{4.96}$$

Together, (4.95) and (4.96) show the saddle-point property of (z'', v). Therefore z'' must be a primal solution. □

By duality, the same scheme can be applied to solve problems where the *primal* constraint set is more complicated, i.e. $\mathcal{C} = \mathcal{C}_1 \cap ... \cap \mathcal{C}_r$. Also note that for $r = 1$, the algorithm reduces to the Douglas-Rachford method from Sect. 4.3.2.2, applied to the dual problem (Prop. 4.4).

4.4 Implementation Details

Projection on the Primal Constraints. Projections on the set $\mathcal{C}$ are highly separable and can be computed exactly in a finite number of steps [Mic86]. Alg. 4.8 summarizes the pointwise operation.

Projection on the Dual Constraints. As mentioned in the introduction, for the Euclidean embedding approach, projecting onto the unit ball $\mathcal{D}^I_{\text{loc}}$ is trivial:

$$\Pi_{\mathcal{D}^I_{\text{loc}}}(v) = \begin{cases} v, & \|v\|_2 \leqslant 1, \\ \dfrac{v}{\|v\|_2}, & \text{otherwise}. \end{cases} \tag{4.97}$$

Algorithm 4.8. Projection of $y \in \mathbb{R}^l$ onto the standard simplex Δ_l

$y^{(0)} \leftarrow y$.
$Z^{(0)} \leftarrow \emptyset$.
repeat
  $\tilde{y}_i^{(k+1)} \leftarrow \begin{cases} 0, & i \in Z^{(k)}, \\ y_i^{(k)} - \frac{e^\top y^{(k)} - 1}{l - |Z^{(k)}|}, & \text{otherwise}, \end{cases} \quad (i \in \mathcal{I})$.
  $Z^{(k+1)} \leftarrow Z^{(k)} \cup \{i \in \mathcal{I} \,|\, \tilde{y}_i^{(k+1)} < 0\}$.
  $y_i^{(k+1)} \leftarrow \max\{\tilde{y}_i^{(k+1)}, 0\} \quad (i \in \mathcal{I})$.
  $k \leftarrow k + 1$.
until $\tilde{y}^{(k+1)} \geqslant 0$.

For the envelope method, $\mathcal{D}_{\text{loc}} = \mathcal{D}_{\text{loc}}^d$ is more complicated. We represent $\mathcal{D}_{\text{loc}}^d$ as the intersection of convex sets,

$$\mathcal{D}_{\text{loc}}^d = \mathcal{R} \cap \mathcal{S}, \; \mathcal{R} := \{v \in \mathbb{R}^{d \times l} | \sum_i v^i = 0\}, \tag{4.98}$$

$$\mathcal{S} := \bigcap_{i<j} \mathcal{S}^{i,j}, \; \mathcal{S}^{i,j} := \{v \in \mathbb{R}^{d \times l} | \|v^i - v^j\|_2 \leqslant d(i,j)\}. \tag{4.99}$$

Enumerating the $\mathcal{S}^{i,j}$ as $\mathcal{S}^{i_1,j_1}, ..., \mathcal{S}^{i_r,j_r}$, this provides the decomposition of $\mathcal{D}$,

$$\mathcal{D} = \mathcal{D}_1 \cap ... \cap \mathcal{D}_r, \quad \mathcal{D}_t := \mathcal{R} \cap \mathcal{S}^{i_t,j_t}, \tag{4.100}$$

as required for the DMDR method (Alg. 4.7). Since $\Pi_{\mathcal{R}}$ is amounts to a translation along the vector $e = (1, ..., 1)$, and the $\mathcal{S}^{i,j}$ are translation-invariant in the direction of e, the projection can be decomposed:

$$\Pi_{\mathcal{D}_t}(v) = \Pi_{\mathcal{R}}(\Pi_{\mathcal{S}^{i_t,j_t}}(v)). \tag{4.101}$$

The individual projections $\Pi_{\mathcal{S}^{i,j}}$ can be computed in closed form, as shown by the following proposition:

Proposition 4.7. *Let $D \geqslant 0$ and $c \in \mathbb{R}^l$, $c \neq 0$. For $v = (v^1|...|v^l) \in \mathbb{R}^{d \times l}$, denote $C(v) := \sum_i c_i v^i$ and*

$$\mathcal{K} := \{v \in \mathbb{R}^{d \times l} \,|\, \|C(v)\|_2 \leqslant D\}. \tag{4.102}$$

Then

$$\Pi_{\mathcal{K}}(v) \;=\; \begin{cases} v, & \|C(v)\|_2 \leqslant D, \\ (w^1|...|w^l), \; w^i = v^i - c_i \frac{\|C(v)\|_2 - D}{\|c\|_2^2} \frac{C(v)}{\|C(v)\|_2}, & \|C(v)\|_2 > D. \end{cases} \tag{4.103}$$

Proof. Computing $w = \Pi_{\mathcal{K}}(v)$ amounts to solving

$$\arg \min_{w \in \mathbb{R}^{d \times l}, \|c_1 w^1 + ... + c_l w^l\|_2 \leqslant D} \sum_i \frac{1}{2} \|w^i - v^i\|_2^2. \tag{4.104}$$

The Lagrangian is

$$L(w,\mu) \;=\; \sum_i \frac{1}{2}\|w^i - v^i\|_2^2 + \mu\,(\|C(w)\|_2^2 - D^2), \tag{4.105}$$

with KKT conditions

$$0 \overset{!}{=} \frac{\partial}{\partial w^i} L(w,\mu) = w^i - v^i + 2\,\mu\,c_i\,C(w), \quad i=1,...,l, \tag{4.106}$$

$$0 \leqslant \mu \perp (\|C(w)\|_2^2 - D^2) \leqslant 0. \tag{4.107}$$

If $\|C(v)\|_2 \leqslant D$, we may set $w=v$ and $\mu=0$. If $\|C(v)\|_2 > D$, set

$$\mu \;=\; \frac{\|C(v)\|_2 - D}{2\,D\,\|c\|_2^2} > 0 \tag{4.108}$$

and w as in (4.103). Summing up the equalities in (4.106) weighted by the c_i leads to

$$C(w) \;=\; C(v) - \frac{\|C(v)\|_2 - D}{\|C(v)\|_2}\,C(v) \tag{4.109}$$

$$\;=\; D\,\frac{C(v)}{\|C(v)\|_2}, \tag{4.110}$$

from which the second set of KKT conditions (4.107) follows. Moreover,

$$w^i \overset{(4.103)}{=} v^i - c_i\,\frac{\|C(v)\|_2 - D}{\|c\|_2^2}\,\frac{C(v)}{\|C(v)\|_2} \tag{4.111}$$

$$\overset{(4.110)}{=} v^i - c_i\,\frac{\|C(v)\|_2 - D}{D\,\|c\|_2^2}\,C(w) \tag{4.112}$$

$$\overset{(4.108)}{=} v^i - 2\,\mu\,c_i\,C(w), \tag{4.113}$$

which shows that also the first set of KKT conditions (4.106) holds. □

Corollary 4.8. *(Specialization to $\mathcal{S}^{i,j}$)*

$$\Pi_{\mathcal{S}^{i,j}}(v) \;=\; \begin{cases} v, & \|v^i - v^j\|_2 \leqslant d(i,j), \\ (w^1,...,w^l), & \textit{otherwise}, \end{cases} \tag{4.114}$$

where

$$w^t \;=\; \begin{cases} v^t, & t \notin \{i,j\}, \\ v^i - \dfrac{\|v^i - v^j\|_2 - d(i,j)}{2}\cdot\dfrac{v^i - v^j}{\|v^i - v^j\|_2}, & t=i, \\ v^j + \dfrac{\|v^i - v^j\|_2 - d(i,j)}{2}\cdot\dfrac{v^i - v^j}{\|v^i - v^j\|_2}, & t=j. \end{cases} \tag{4.115}$$

Proof. Set $c_i=1$, $c_j=-1$ and $c_k=0$ for $k \notin \{i,j\}$, and apply Prop. 4.7. □

Remark 4.9. The special case of Cor. 4.8 corresponds to the method outlined in [CCP08], where a different linearization and restricted set of metrics $d(i,j)$ was used.

Algorithm 4.9. Dykstra's Method for Projecting onto the Intersection of Convex Sets

$x \leftarrow v \in \mathbb{R}^{d \times l}$.
$y^1, ..., y^k \leftarrow 0 \in \mathbb{R}^{d \times l}$
while (x not converged)
    for $t = 1, ..., k$ **do**
        $x' \leftarrow \Pi_{\mathcal{S}^{(i_t, j_t)}}(x + y^t)$.
        $y^t \leftarrow x + y^t - x'$.
        $x \leftarrow x'$.
    end for
end while

For applications where an approximation of the complete projection $\Pi_{\mathcal{D}}$ is required, we follow the idea of [CCP08] to use Dykstra's method [BD86]. However, any recent multiple-splitting method could be used [CP08, GM09]. One has to take caution when applying large-scale methods such as sequential/parallel "projection on convex sets" (*POCS*) and row-action methods [You78, Com96, SY98] who are shown to converge to a point in $\mathcal{C}$, which is not necessarily the Euclidean projection on $\mathcal{C}$. While suitable for finding a *feasible point* in $\mathcal{C}$, they cannot by applied to our setting, since we require the actual (Euclidean) projection on $\mathcal{C}$.

The complete method for projecting a vector v onto $\mathcal{S} = \mathcal{S}^{(i_1, j_1)} \cap ... \cap \mathcal{S}^{(i_k, j_k)}$ is outlined in Alg. 4.9. While the sequence of y may be unbounded, x converges to $\Pi_{\mathcal{S}}(v)$ (cf. [GM89, p.40][Xu00]).

4.5 Experimental Comparison

In this section we present some observations regarding the practical performance of the proposed optimization methods. Due to the reasons outlined in Chap. 3, we restrict ourselves to methods for solving the convex relaxed problem, rather than combinatorial approaches. We first quantitatively compare the algorithms in terms of runtime and the number of inner iterations for problems where the inner problems can be solved exactly, as in the case of the embedding method, and then provide some results on the envelope regularization. We do not claim to provide a comprehensive benchmark, but rather illustrate and highlight the characteristic properties of the individual methods on selected examples.

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

4.5.1 General Observations

To compare the convergence speed of the different approaches, we computed the primal-dual gap at each iteration. As it bounds the optimality of the current iterate (see Sect. 4.3), it constitutes a reliable and convenient criterion for performance comparison.

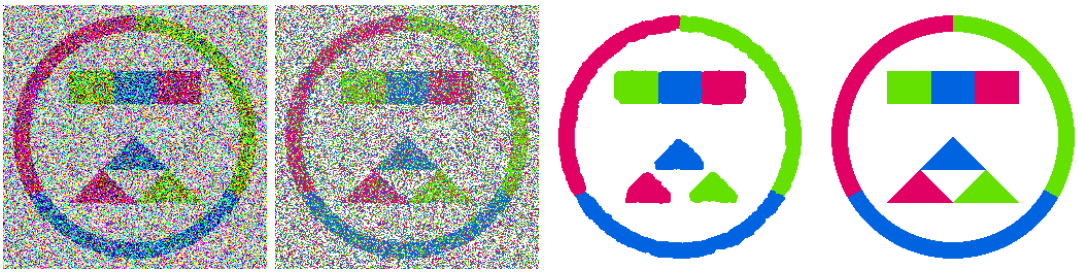

Figure 4.1. Synthetical "four colors" image for the performance tests. **Left to right:** Input image with Gaussian noise, $\sigma = 1$; local labeling without regularizer; result with uniform metric regularizer and Douglas-Rachford optimization; ground truth.

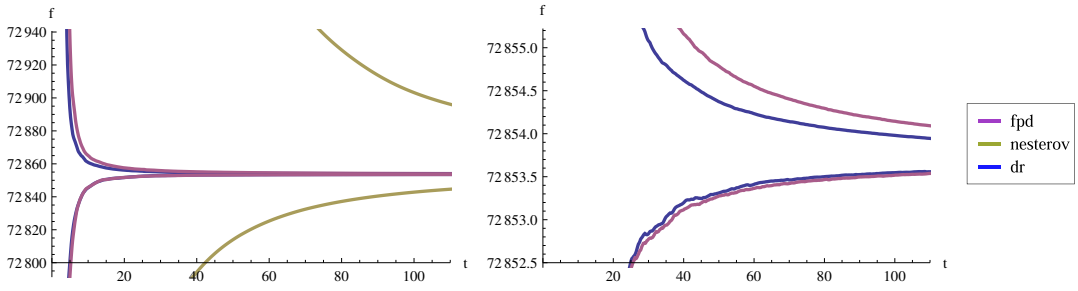

Figure 4.2. Convergence speed for the "four colors" image in Fig. 4.1. **Left:** Primal (upper) and dual (lower) objectives vs. computation time for the (from top to bottom) Nesterov, Fast Primal-Dual (FPD) and Douglas-Rachford (DR) methods. **Right:** Detailed view of the FPD and DR methods. The primal and dual objectives provide upper and lower bounds for the objective of the true optimum. Douglas-Rachford and FPD perform similarly, while the Nesterov method falls behind by a large margin.

Unfortunately the gap is not available for the envelope method, as it requires the primal objective to be evaluatable. Using a numerical approximation is not a reliable option, as this would only provide a *lower* bound for the objective, i.e. an underestimation of the gap, which is critical as one is interested in the behavior when the gap is very close to zero. Therefore we restricted the gap-based performance tests to the embedding relaxation (Sect. 2.5.2). In this setting the projections can be computed exactly, and thus the Douglas-Rachford (DR) and Dual Multiple-Splitting Douglas-Rachford (DMDR) methods coincide, therefore all results for DR apply as well to DMDR. In order to make a fair comparison we generally analyzed the progression of the gap with respect to computation time, rather than the number of iterations.

For the first tests we used the synthetical 256×256 "four colors" input image (Fig. 4.1). It represents a typical segmentation problem with several objects featuring sharp corners and round structures above a uniform background. The label set consists of three classes for the foreground and one class for the background. The image was overlaid with i.i.d. Gaussian noise with standard deviation $\sigma = 1$ and truncated to the interval $[0, 1]$ on all RGB channels. We used a simple ℓ^1 data term, $s_i(x) = \|I(x) - c^i\|_1$, where $I(x) \in [0,\ 1]^3$ are the RGB color values of the input image in x, and c^i is a prototypical color vector for the i-th class.

The runtime analysis shows that FPD and DR perform similar, while the Nesterov method falls behind with respect to both the primal and the dual objective (Fig. 4.2).

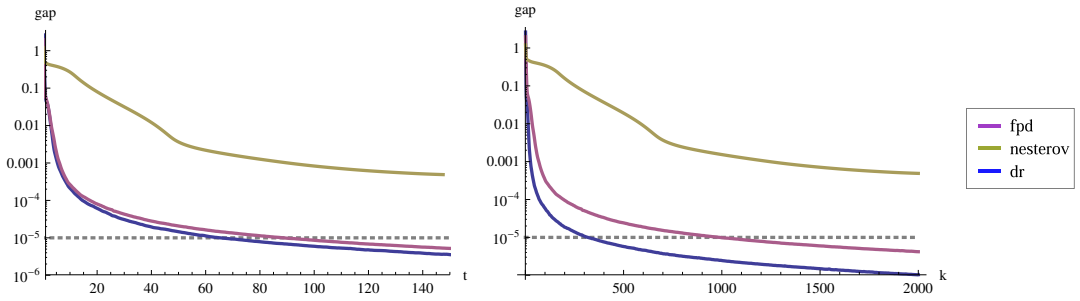

Figure 4.3. Relative primal-dual gap for Fig. 4.2 with respect to time and number of iterations. **Left:** Relative gap vs. time and number of iterations. The Nesterov method (top) again falls behind, while FPD (center) and Douglas-Rachford (bottom) are equally fast. **Right:** Primal-Dual gap vs. number of iterations. The Douglas-Rachford method requires only one third of the FPD iterations, which makes it suitable for problems with expensive inner steps.

The picture changes when considering the gap with respect to the number of iterations rather than time, eliminating influences of the implementation and system architecture. To achieve the same accuracy, DR requires only one third of the iterations compared to FPD (Fig. 4.3). This advantage does not fully translate to the time-based analysis as the DCT steps increase the per-step computational cost significantly. However in this example the projections on the sets $\mathcal{C}$ and $\mathcal{D}$ were relatively cheap compared to the DCT. In situations where the projections dominate the time per step, the reduced iteration count can be expected to lead to an approximately proportional reduction in computation time.

One could ask if these relations are typical to the synthetical data used. However, we found them confirmed on a large majority of the problems tested. As a real-world example, consider the "leaf" image from Chap. 2 (Fig. 2.3). We computed a segmentation into 12 classes with uniform metric regularizer, based on ℓ^1 distances for the data term, with very similar relative performance as for the "four colors" image (Fig. 4.4).

4.5.2 Varying Problem Size and Parameters

Problem Size and Interior-Point Methods. To examine how the methods scale with an increasing problem size, we evaluated the "four colors" problem at various scales ranging from 16×16 to 512×512. Note that if the grid spacing is held constant, the regularizer weights must be scaled proportionally to the image width and height in order to obtain structurally comparable results, and in order to not mix up effects of the problem size and of the regularization strength.

In addition to the FPD, DR and Nesterov algorithms, we also implemented the problem in an SOCP formulation, and compared the first-order methods to two interior-point solvers: the non-commercial MATLAB-based SDPT3 and the commercial MOSEK package. The latter consistently achieves the fastest runtimes in the independent SOCP benchmark [Mit03, Mit11].

In order to compensate for the increasing number of variables, the stopping criterion was based on the relative gap (4.24), i.e. the algorithms were terminated as soon as the relative gap fell below a threshold. All tests were repeated 10 times with different noise.

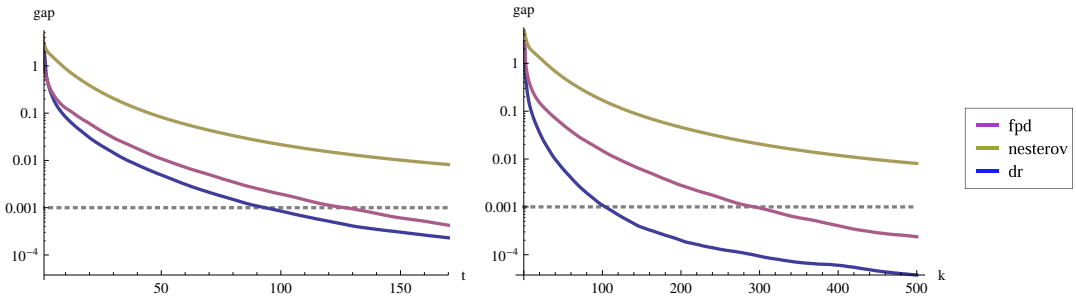

Figure 4.4. Relative primal-dual gap for the real-world leaf image in Fig. 2.3 with 12 classes and uniform metric regularizer. **Left:** Relative gap vs. number of iterations. As in the synthetic examples, the Nesterov method (top) falls behind the FPD (center) and Douglas-Rachford (bottom) methods. **Right:** Relative gap vs. number of iterations. As with the synthetical four-colors image (Fig. 4.3), the Douglas-Rachford approach reduces the number of required iterations by approximately a factor of three.

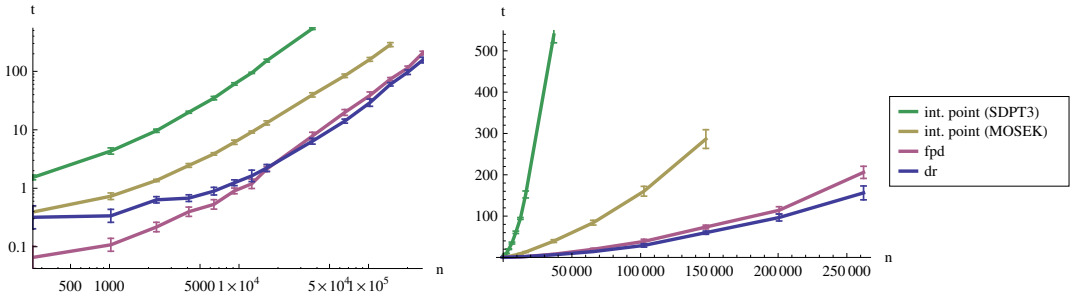

Figure 4.5. Performance of first-order and interior-point methods for increasing problem size (number of pixels) n. Shown is the mean time in seconds over 10 problems with different noise, with error bars at 2σ. The Nesterov method never converged to the required relative gap of 10^{-4} within the limit of 2000 iterations, despite an optimal selection of the smoothing parameter, and is therefore not shown. For this relatively low accuracy, FPD and DR outperform both interior-point solvers, which additionally exceeded the available memory at image sizes of 192×192 and 384×384, respectively. The computational effort scales slightly superlinearly with the number of pixels in the image.

As can be seen in Fig. 4.5, for the moderate accuracy of 10^{-4}, the FPD and DR outperform the interior-point solvers by a large margin. The Nesterov method consistently produced relative gaps in the 10^{-3} range and never achieved the threshold within the limit of 2000 iterations, despite the built-in optimal selection of the smoothing parameter. Compared to FPD, the Douglas-Rachford approach seems to have a slight advantage on larger images. Both approaches seem to scale only slightly superlinearly with the problem size, which is quite encouraging given the comparatively simple first-order methods (Fig. 4.5).

The interior-point solvers exceeded the available memory at resolutions of 192×192, corresponding to 500.000 variables and 180.000 constraints (SDPT3), and 384×384, corresponding to 2 million variables and 730.000 constraints (MOSEK).

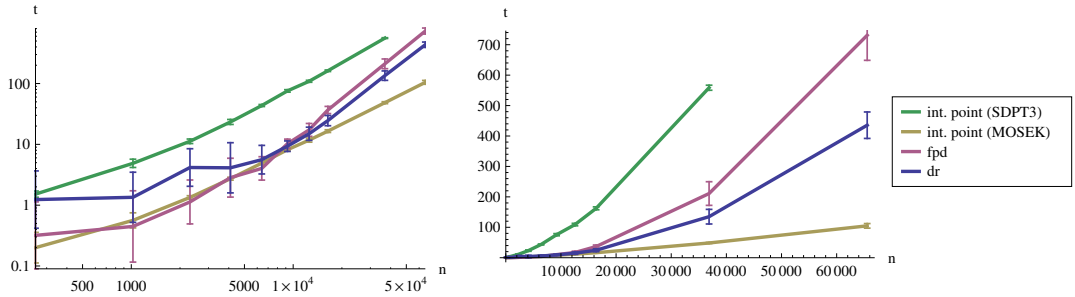

Figure 4.6. Performance of first-order and interior-point methods for a relative gap of 10^{-6}, cf. Fig. 4.5 (note the different scales). For higher requested accuracy, the better asymptotical convergence rate of the interior-point methods becomes advantageous. However, the first-order methods still outperform the SDPT3 solver.

For applications where a high accuracy is required, interior-point approaches are better suited due to their better asymptotical convergence rate. This can be seen in Fig. 4.6, where the stopping criterion was set to 10^{-6}. However, such accuracy is often not required in imaging applications, and the first-order methods are very amenable to parallelization, which can lead to an additional speedup of $30-100$ [ZGFN08]. We also observed that the interior-point methods generated infeasible solutions that fully utilize the permitted infeasibility, while the solutions returned by FPD and DR were feasible up to numerical precision.

Regularization Strength. While we deliberately excluded influences of the regularizer in the previous experiment, it is also interesting to examine how algorithms cope with varying regularization strength. We fixed a resolution of 256×256 and evaluated the performance of the FPD and DR algorithms, scaling the regularization term by an increasing sequence of λ in the $[0.1, 5]$ range (Fig. 4.7).

For low regularization, where much of the noise remains in the solution, FPD clearly takes the lead, while for scenarios with large structural changes, DR performs better. Again, the Nesterov method never achieved the required accuracy. We observed two peaks in the runtime plot which we cannot completely explain. However we found that at the first peak, structures in the image did not disappear in parallel during the course of the optimization process, but rather one after the other, i.e. neither the regularizer nor the data term dominate the other.

4.5.3 Breaking Points

We have no clear explanation why the Nesterov method clearly performs worst in most settings. However it is possible to compare its behavior with the theoretical bound from Prop. 4.1. It can be seen that exactly one half of the final bound comes from the smoothing step, while the other half is caused by the finite number of iterations:

$$\varepsilon_{\text{total}} = \varepsilon_{\text{smooth}} + \varepsilon_{\text{iter}}, \quad \text{where} \quad \varepsilon_{\text{smooth}} = \varepsilon_{\text{iter}}. \tag{4.116}$$

Moreover, $\varepsilon_{\text{iter}}$ decreases with $1/(k+1)^2$ according to (4.36), which gives a per-iteration optimality bound of

$$\varepsilon_{\text{total}}^{(k)} = \varepsilon_{\text{smooth}} + \left(\frac{N+1}{k+1}\right)^2 \varepsilon_{\text{iter}}. \tag{4.117}$$

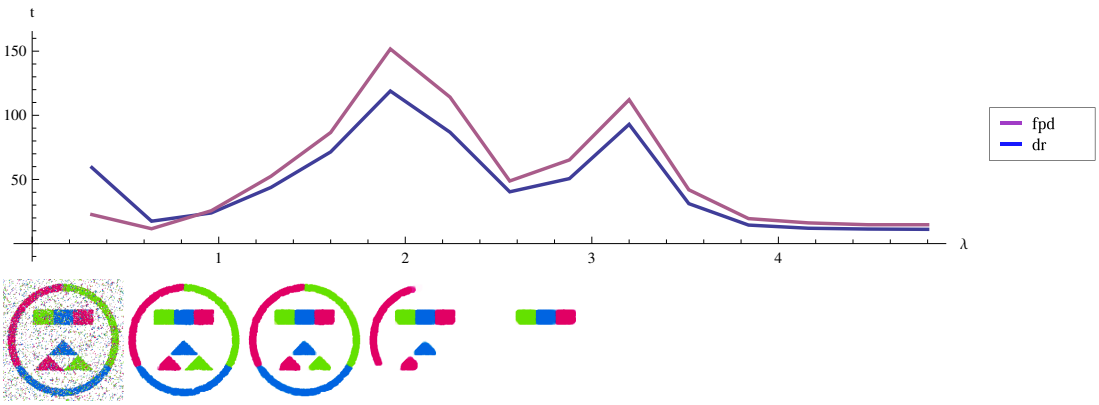

Figure 4.7. Computation time for varying regularization strength λ for the Douglas-Rachford (dark) and FPD (light) methods. The images at the bottom show the final result for the corresponding λ above. FPD is strong on low regularization problems, while Douglas-Rachford is better suited for problems with large structural changes. The Nesterov method never achieved the relative gap of 10^{-5} within 2000 iterations.

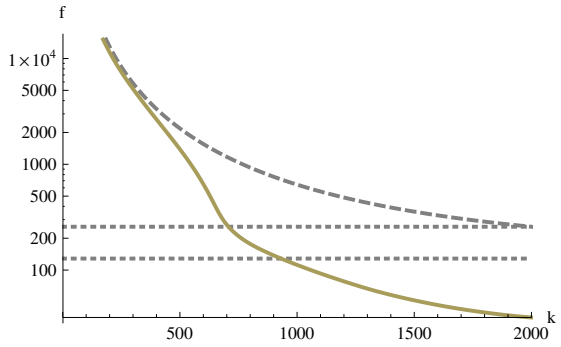

Figure 4.8. Theoretical vs. practical performance of the Nesterov method for the "four-colors" problem in Fig. 4.1. As expected, the primal objective (solid) stays below the theoretical bound $\varepsilon^{(k)}_{\text{total}}$ (dashed). At the final iteration, the worst-case total bound $\varepsilon_{\text{total}}$ (dotted, top) is outperformed by a factor of 7, which implicates that the error introduced by the smoothing is also well below its worst-case bound $\varepsilon_{\text{smooth}}$ (dotted, bottom). During the first iterations the method stays close to the theoretical bound, indicating that the theoretical bound cannot be considerably improved by a better choice of the constants.

On the "four colors" image, the actual gap stays just below $\varepsilon^{(k)}_{\text{total}}$ in the beginning (Fig. 4.8). This implies that the theoretical bound can hardly be improved, e.g. by choosing constants more precisely. Unfortunately, the bound is generally rather large, in this case at $\varepsilon_{\text{total}} = 256.8476$ for 2000 iterations. While the Nesterov method outperforms the theoretical bound $\varepsilon_{\text{total}}$ by a factor of 2 to 10 and even drops well below the worst-case smoothing error δ_{smooth}, it still cannot compete with the other methods, which achieve a gap of 0.3052 (FPD) and 0.0754 (Douglas-Rachford). This indicates that the bounds provided by Prop. 4.1 – although they constitute true *a priori* bounds, and not only in an asymptotical sense – are not very relevant in practice.

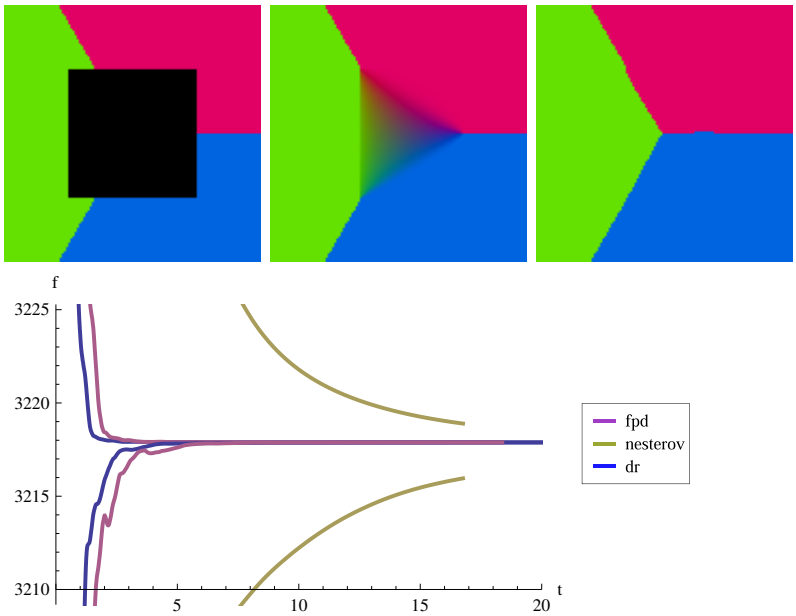

Figure 4.9. Primal and dual objectives for the "triple point" inpainting problem with uniform metric regularizer. **Top row, left to right**: Input image with zeroed-out region around the center; relaxed solution; rounded solution. **Bottom row:** Primal (upper) and dual (lower) energies vs. time. The triple junction in the center is reconstructed solely by the regularizer.

There is an interesting extreme case where the Nesterov method seems to come to full strength. Consider the noise-free "triple point" inpainting problem (Fig. 4.9). The triple junction in the center can only be reconstructed by the regularizer, as the ℓ^1 data term has been zeroed out around the center. By reversing the sign of the data term, one obtains the "inverse triple point" problem, an extreme case that has also been studied in [CCP08] and shown to be an example where the relaxation leads to a strictly non-integral solution (Fig. 4.10).

On the inverse problem, the Nesterov method catches up and even surpasses FPD, in contrast to the regular triple point problem, which is more closely related to real-world data. We conjecture that this sudden strength comes from the inherent averaging over all previous gradients (step 10 in Alg. 4.3): In fact, on the inverse problem, DR and FPD display a pronounced oscillation in the primal and dual objectives, which is accompanied by slow convergence. The Nesterov method consistently shows a monotone and smooth convergence. However, it is still outperformed by the DR approach.

It remains a mystery why the Nesterov method performs so slow, despite its good reported performance for general TV-regularized problems [WABF09]. We suspect that its performance hinges on the strict convexity of the objective, which is fulfilled by the usual ROF models – due to the quadratic data term –, but not by the bilinear saddle-point problem (4.8).

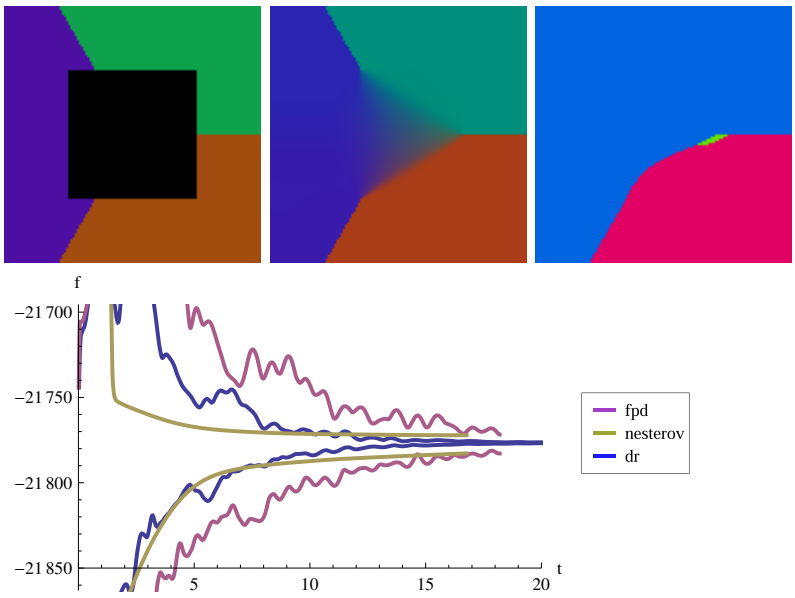

Figure 4.10. Inverse triple point problem (cf. Fig. 4.9). **Top row, left to right:** Input image with zeroed-out region around the center; relaxed solution; rounded solution. The inverse triple point problem exhibits a *strictly* non-integral relaxed solution. The small irregularities are due to the finite accuracy of the solution, which is amplified by the rounding step. **Bottom row:** For the inverse triple point, Douglas-Rachford (bottom) and FPD (center) show an oscillatory behavior with slow convergence. The Nesterov approach (top) does not suffer from oscillation due to the inherent averaging, and surpasses FPD on the inverse problem.

4.5.4 Choice of the Relaxation

High Label Count Using Euclidean Embeddings. As an example for a problem with a large number of labels, we analyzed the "penguin" inpainting problem from [SZS+06]. We chose 64 labels corresponding to 64 equally spaced gray values. The input image contains a region where the image must be inpainted in addition to removing the considerable noise. Again the data term was generated by the ℓ^1 distance, which in this case reduces to the absolute difference of the gray values. In order to remove noise but not overly penalize hard edges such as between the black wing and the white front, we chose a regularizer based on the truncated linear metric (Sect. 2.5.2).

Due to the large number of labels, this problem constitutes an example where the embedding approach is very useful. As the computational effort and memory requirements for the DMDR method grow quadratically in the number of labels, they quickly become prohibitively large for a moderate amount of classes. In contrast, the embedding method requires considerably less computational effort and still approximates the potential function to a reasonable accuracy (cf. Fig. 2.7). In the practical evaluation, the DR method converged in 1000 iterations to a relative gap of $8.3 \cdot 10^{-4}$, and recovered both smooth details near the beak, and hard edges in the inpainting region (Fig. 4.11).

Tightness of the Relaxations. Besides the properties of the optimization methods, it is interesting to study the effect of the relaxation technique on the relaxed and rounded solutions. To get an insight into the tightness of the relaxations, we used the

Figure 4.11. Denoising/inpainting problem with 64 classes and the nontrivial truncated linear metric. **Left:** Noisy input image with inpainting region marked black [SZS+06]; **Right:** Result of the DR method. The truncated linear metric d was approximated by an Euclidean embedding, which allows to handle problems with a large number of labels.

Douglas-Rachford method to repeat the "triple point" inpainting experiment with both the embedding and the envelope regularizer (Fig. 4.12). We did not further compare the "uncertainty emphasizing" regularizer since it seems to be mainly of theoretical interest and is numerically even more involved than the envelope method.

Despite the inaccuracies in the projections, the envelope regularizer generates a nearly integral solution: 97.6% of all pixels were assigned "almost integral" labels with an ℓ^∞ distance of less than 0.05 to one of the unit vectors $\{e^1, ..., e^l\}$. For the embedding approach, this constraint was only satisfied at 88.6% of the pixels. Note that in contrast to the two-class problems considered in Chap. 3, in multiclass problems fractional labels may also be caused by the relaxation, therefore a larger number of integral pixels gives at least an indication that the relaxation may be more tight.

The result for the envelope relaxation is very close to the sharp triple junction one would expect from the continuous formulation, which conforms to the theoretically improved tightness of the envelope relaxation compared to the embedding method. However, after rounding both approaches generate almost identical integral results, and the embedding regularizer is more than four times faster, with 41.1 seconds per 1000 iterations vs. 172.2 seconds for the envelope regularizer.

While the triple point is a problem specifically designed to challenge the regularizer, real-world images usually contain more structure as well as noise, while the data term is available for most or all of the image. However, we observed that the above results are quite representative. As a real-world example, consider the four-class "sailing" color segmentation problem in Fig. 4.13. The improved tightness of the envelope relaxation was also noticeable, with 96.2% vs. 90.6% of "almost integral" labels. However, due to the larger number of labels and the larger image size of 360×240, runtimes increased to 4253 (envelope) vs. 420 (embedding) seconds. Therefore the embedding method seems to provide a fast alternative, which is however slightly less accurate.

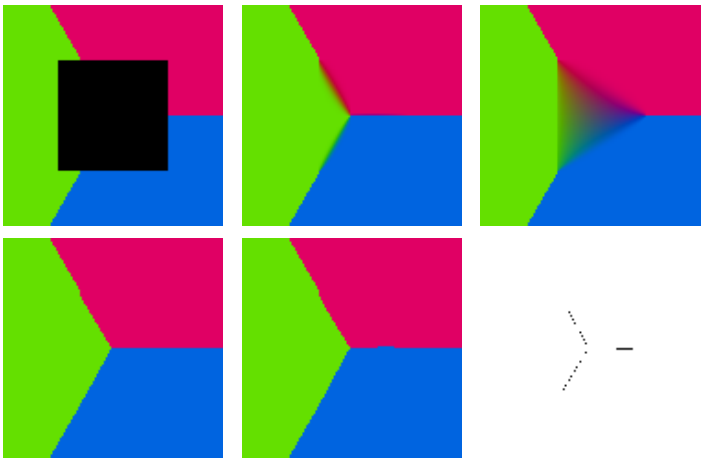

Figure 4.12. Tightness of the relaxation. **Top row:** In the input image (left), the data term was blanked out in a quadratic region. All structure within the region is generated purely by the regularizer with a standard uniform metric interaction potential. The envelope relaxation is tighter and generates a more integral solution (center) than the embedding method (right). **Bottom row:** After rounding of the fractional solutions, the envelope (left) and embedding (center) methods generate essentially the same solution, as can be seen in the difference image (right). The embedding method performed more than four times faster due to the simpler structure of the regularizer.

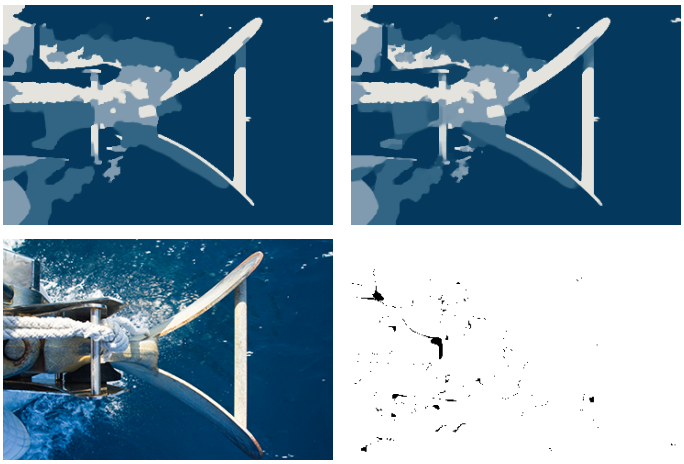

Figure 4.13. Effect of the choice of relaxation method on the real-world "sailing" image (image courtesy of F. Becker). **Top row:** Four-class segmentation using envelope (left) and Euclidean metric (right) methods. Shown are the solutions of the relaxed problem. **Bottom row:** Original image (left); difference image of the rounded solutions (right). While the envelope relaxation leads to substantially more "almost integral" labels in the relaxed solution, it also runs more than 10 times slower and does not provide a suboptimality bound. The generated solutions are visually almost identical.

The relaxed as well as the rounded solutions show some differences but are hard to distinguish visually. It is difficult to pinpoint if these differences are caused by the tighter relaxation or by numerical issues: while the Douglas-Rachford method applied to

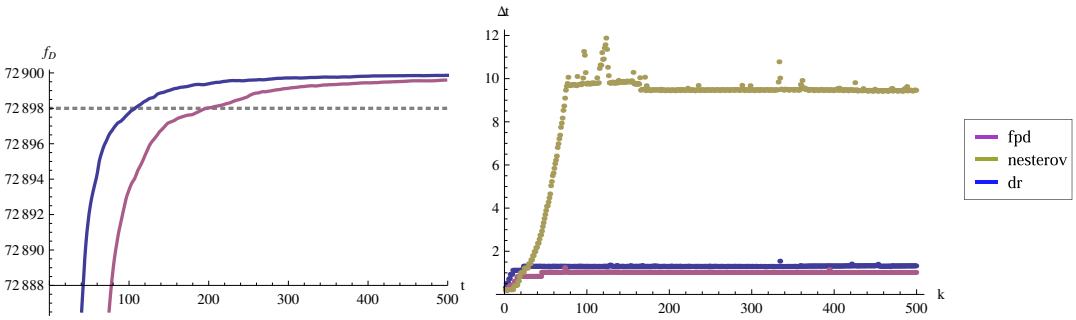

Figure 4.14. Performance on the "four colors" image with uniform metric interaction potential and the envelope regularizer. **Left:** Dual objectives of for DR (top) and FPD (bottom) vs. time. The reduced iteration count of the Douglas-Rachford method becomes more apparent in the time plot as the time per iteration is now dominated by the projection rather than the DCT. **Right:** Time per iteration for Nesterov (top), DR (center) and FPD (bottom). The Nesterov method fails to converge as it accumulates errors from the approximate projections, which in turn leads to slower and more inexact projections.

the embedding relaxation converged to a final relative gap of $1.5 \cdot 10^{-6}$, no such bound is available to estimate the accuracy of the solution for the envelope relaxation, due to the primal objective not being available.

4.5.5 Dual Multiple-Constraint Douglas-Rachford

Inaccuracies Caused by Inexact Projections. As a motivation for the DMDR method in connection with the envelope regularizer, consider the experiment in Fig. 4.14, where we compare DMDR to FPD and DR with iterative, inexact projections. The iterative Dykstra projection (Alg. 4.9) was stopped when the iterates differed by at most $\delta = 10^{-2}$, with an additional limit of 50 iterations. While the gap cannot be computed in this case, the dual objective can still be evaluated and provides an indicator for the convergence speed.

We found that compared to the embedding regularizer from the previous examples, the margin between FPD and DR increases significantly. This is consistent with the remarks in Sect. 4.5.1: the lower iteration count of the DR method becomes more important, as the projections dominate the per-iteration runtime (Fig. 4.14).

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

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

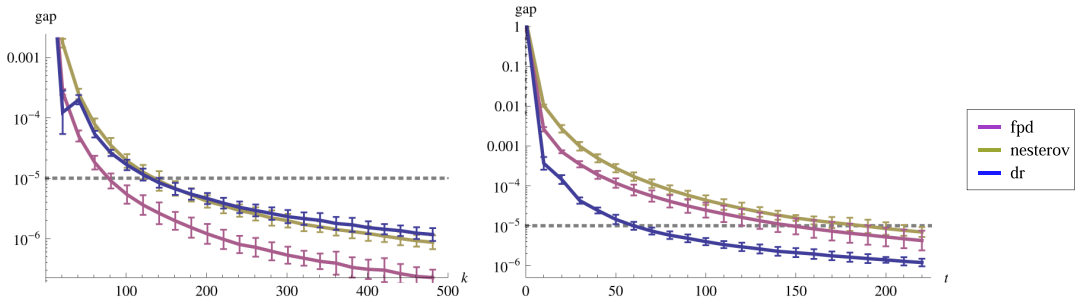

Figure 4.15. Runtime comparison on the "four colors" test set for the envelope regularizer. Shown is the relative gap to the objective of the computed reference solution vs. number of iterations (left) and time (right), averaged over 10 problem with different noise, and error indicators at 2σ. With respect to the number of iterations, the DMDR method performs comparable to FPD, and slightly worse than DR. However, it requires significantly less time per iteration, since it does not require to iteratively solve the inner projection steps, resulting in a total speedup of $2-3$ compared to both methods.

Runtime Comparison for the Envelope Regularizer. In order to demonstrate that the DMDR approach avoids these problems associated with the envelope regularizer, we compared FPD and DR with iterative projections to the DMDR method. Note that in this case the primal objective f cannot be accurately evaluated due to the complexity of $\mathcal{D}$, therefore we must resort to a more elementary stopping criterion such as the difference between two consecutive iterates, $\|z''^{(k)} - z''^{(k-1)}\|_2$. In order to still get an objective measure on convergence speed, we computed a reference solution u^+ using 5000 iterations of the DR method and recorded the gap $f(u^+) - f_D(u^{(k)})$. Again, the experiment was repeated 10 times with varying noise (Fig. 4.15).

In terms of the number of iterations, the DMDR method converges as fast as FPD, and slightly slower than DR. However, as it requires significantly less effort per iteration, it outperforms FPD and DR by a factor of $2-3$ with respect to the overall runtime.

Improved Numerical Robustness. For a larger number of labels, the runtime advantage of DMDR is expected to become more apparent, since the cost per iterative projection increases. As an example, consider the 12-class segmentation of the real-world images in Fig. 4.16, again with ℓ^1 data term and uniform metric regularizer.

For this moderate number of labels, the iterative projections for DR and FPD are already quite slow, so we fixed a maximum of 5 inner iterations per outer step in order to get a reasonable computation time. For the shown problems, the DMDR method is about $6-10$ times faster than DR, and $7-17$ times faster than FPD. Moreover, due to the inexact projections, DR and FPD converge to infeasible dual points, i.e. they generate dual solutions v that do not lie inside the dual constraint set $\mathcal{D}$ (Fig. 4.16). In contrast, using DMDR the infeasibility gradually decreases, and is guaranteed to eventually drop to zero given exact arithmetic.

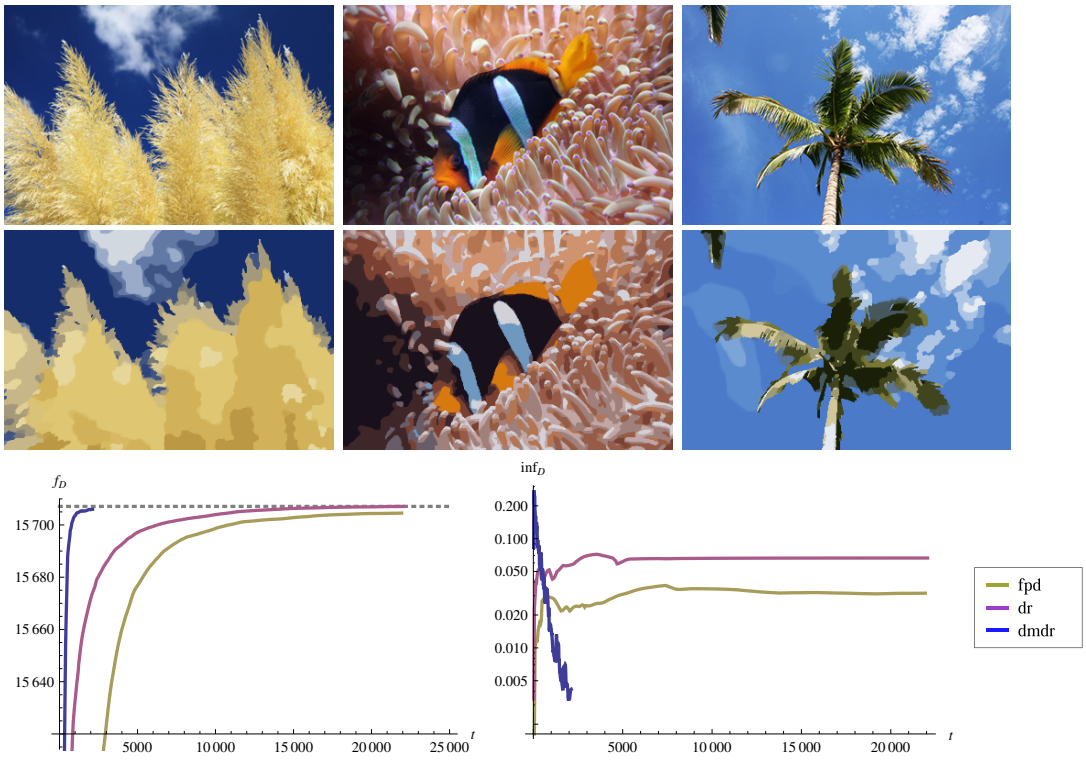

Figure 4.16. Runtime performance on segmentation problems with a higher label count. **Top row:** Input images (top) and segmentation into 12 classes (bottom) computed using the proposed DMDR method. **Bottom left:** Dual objective vs. time for 500 iterations on the "crop" image. DMDR outperforms DR and FPD by a factor of 10 resp. 17. **Bottom right:** Dual infeasibility vs. time. Due to the inexact projections, FPD and DR get stuck and converge to infeasible solutions. In contrast, DMDR gradually decreases the infeasibility to zero in theory and practice.

Histogram-Based Segmentation and Accuracy. Fig. 4.17 shows the application of our method to a real-world histogram-based three-class segmentation where the data term is based on probabilities computed from histograms over regions preselected by the user. In order to preserve more details, we chose a low regularization with $\lambda = 0.025$.

As in the previous section, it can be seen that FPD and DR get stuck at infeasible solutions for the envelope regularizer, while DMDR converges smoothly. Increasing the accuracy of the approximate projections reduces the infeasibility, but leads to a much slower convergence.

It remains to ask how the dual gap relates to actual visual differences. Therefore at each step we evaluated the ℓ^2 distance of the current iterate to a reference solution computed using 5000 DMDR iterations (Fig. 4.17). Again it becomes clear that the inexact projections cause convergence issues for FPD and DR, while DMDR does not suffer from these problems. After 500 iterations, DMDR recovered a solution u^k with $\|u^k - u^*\|_2 \leqslant 10$, which amounts to an average of $1.3 \cdot 10^{-4}$ per pixel, suggesting that only few iterations are required for visually high quality results.

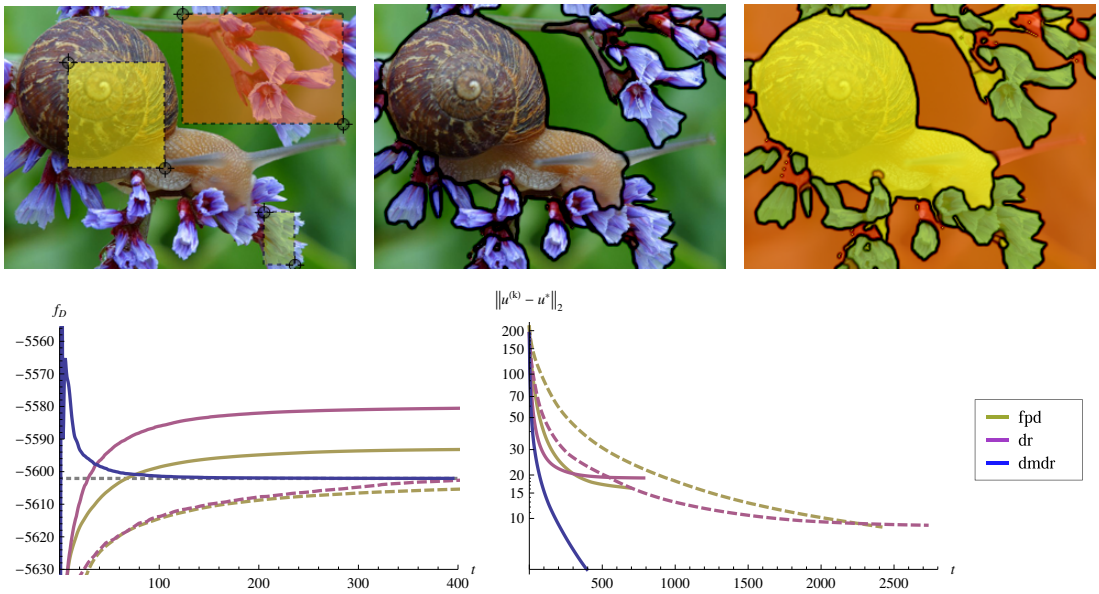

Figure 4.17. Application to histogram-based segmentation. **Top row, left to right:** Input image with seed regions marked by the user; minimizer of the three-class variational segmentation using the proposed approach. **Bottom row:** Dual objectives (left) and ℓ^2 distance to the reference solution (right) vs. time. With low-accuracy approximate projections, FPD and DR get stuck in an infeasible solution (solid). Increasing the projection accuracy reduces the effect but slows down convergence (dashed). The proposed DMDR method avoids these problems and returns high-quality solutions after only a few iterations.

4.6 Summary and Further Work

We presented two approaches for solving the relaxed saddle-point problem (4.8). Both methods only rely on inexpensive first-order operations. While the Nesterov method, despite its theoretically relatively fast convergence in the class of first-order methods, seems to only come to full strength on very special applications, the Douglas-Rachford method successfully competes with more advanced interior-point solvers for low- to medium accuracy, as typically required in imaging applications.

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

In order to avoid the problems caused by inexact iterative projections, we proposed the DMDR extension of the Douglas-Rachford approach involving auxiliary variables, and showed that it copes well with the more difficult "local envelope" relaxation. Experiments indicated that it outperforms the FPD and DR methods by a factor of $4 - 20$, and avoids the inaccuracies and convergence issues of the FPD and DR methods that rely on inexact projections.

A straightforward extension that comes into mind is the implementation of a coarse-to-fine or multigrid strategy. However, first experiments indicated that this generally does not substantially speed up convergence: in fact, basic structures and edges are obtained very fast by the Douglas-Rachford algorithm, and much time is spent in determining the exact values around the edges and converging to integral values within the regions. Therefore better starting values only yield a marginal improvement. These observations are in line with [ZWC10], where the authors also found that the considered first-order methods generally did not provide good warmstarts for higher-order methods. Another direction which deserves thorough consideration is to evaluate to which amount higher-order and interior-point methods can be suitably sped up by exploiting the specific problem formulation.

However, in order to achieve large speed-ups, the most promising direction seems to be investigating strategies how to translate the concepts for polynomial max-flow methods, formulated on pairwise energies, to the finite-differences discretization (or generally to the optimize-first approach). This is by no means trivial, since concepts such as augmenting paths do not have a direct analogon for non-pairwise energies.

In the meantime, as seen above, for many applications the DMDR method provides relatively accurate results in a reasonable time. The splitting formulation of the DR and DMDR approaches also seems to be a good compromise between speed and flexibility, and can easily be adapted to related convex models, such as TV-constrained optimization and "ratio" cut [KB05].

Chapter 5
Optimality and Rounding

5.1 Introduction and Overview

In this chapter, we consider optimality properties of the variational multiclass labeling problem as introduced previously,

$$\begin{aligned} &\inf_{u\in\mathcal{C}_\mathcal{E}} f(u), \quad f(u) := \int_\Omega \langle u(x), s(x)\rangle d\,x + \int_\Omega d\Psi(D\,u), \\ &\mathcal{C}_\mathcal{E} = \mathrm{BV}(\Omega,\mathcal{E}), \quad \Omega = (0,1)^d, \quad \mathcal{E} = \{e^1,...,e^l\}, \end{aligned} \tag{5.1}$$

Throughout this chapter we assume that $s \in L^\infty(\Omega)^l$, $s \geqslant 0$ are the local costs representing the data term, and $\Psi\colon \mathbb{R}^{d\times l} \to \mathbb{R}_{\geqslant 0}$ is positively homogeneous, convex and continuous, and defines the regularizer. Again we consider the convenient relaxation of the combinatorial problem as introduced in Sect. 2.1,

$$\inf_{u\in\mathcal{C}} f(u), \quad \mathcal{C} := \mathrm{BV}(\Omega,\Delta_l), \tag{5.2}$$

where $\Delta_l = \{x \in \mathbb{R}^l | x \geqslant 0,\ \sum_i x_i = 1\}$ is the convex hull of $\mathcal{E} = \{e^1,...,e^l\}$, i.e. the l-dimensional unit simplex.

As noted above, problem (5.2) 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 $\bar{u}^*$. This may may increase the objective, and lead to a suboptimal solution of the original problem (5.1).

Note that this behavior is independent of the effects discussed in the discretization chapter: In Chap. 3, 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 chapter, we consider the *second* source of fractional solutions, which is the *relaxation* of the original combinatorial *multiclass* problem to a convex problem, and rule out discretization effects by working completely in the spatially continuous setting.

In the following, we are concerned with the question whether it is possible to obtain, using the convex relaxation (5.2), *integral* solutions with an upper bound on the objective. Specifically, we concentrate on inequalities of the form

$$f(\bar{u}^*) \ \leqslant\ (1+\varepsilon) f(u^*_\mathcal{E}) \tag{5.3}$$

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 $u^*_\mathcal{E}$ of (5.1). Note that the reverse inequality

$$f(u^*_\mathcal{E}) \ \leqslant\ f(\bar{u}^*) \tag{5.4}$$

always holds since $\bar{u}^* \in \mathcal{C}_\mathcal{E}$ and $u^*_\mathcal{E}$ is an optimal integral solution. The specific form (5.3)

can be attributed to the alternative interpretation

$$\frac{f(\bar{u}^*) - f(u^*_{\mathcal{E}})}{f(u^*_{\mathcal{E}})} \;\leqslant\; \varepsilon, \tag{5.5}$$

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 $\bar{u} \in \mathcal{C}_{\mathcal{E}}$ has been obtained from u by a rounding process. Then the pair $(\bar{u}, v)$ is feasible as well since $\mathcal{C}_{\mathcal{E}} \subseteq \mathcal{C}$, and from the considerations in Sect. 4.3 we obtain an *a posteriori* optimality bound of the form (5.5) with respect to the optimal *integral* solution $u^*_{\mathcal{E}}$:

$$\frac{f(\bar{u}) - f_D(u^*_{\mathcal{E}})}{f_D(u^*_{\mathcal{E}})} \leqslant \frac{f(\bar{u}) - f_D(u^*_{\mathcal{E}})}{f_D(v)} \leqslant \frac{f(\bar{u}) - f_D(v)}{f_D(v)} =: \varepsilon'. \tag{5.6}$$

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. In this chapter, we will analyze several rounding methods in order to derive such bounds. 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, we will again assume an "optimize first" perspective due to the reasons outlined in Chap. 3. Unfortunately, the proofs for the finite-dimensional results often rely on pointwise arguments that cannot directly be transferred to the continuous setting. Deriving similar results for continuous problems therefore requires considerable additional work.

Organization. This chapter is organized as follows:

- We propose an improved *deterministic* local rounding technique that takes into account the specific structure of the regularizer (Sect. 5.3).
- In order to motivate the proposed rounding approach, we point out a connection between *probabilistic* rounding methods and an approximate variant of the coarea formula (Sect. 5.4).
- We provide a probabilistic rounding method and prove that it allows to obtain integral solutions with an *a priori* upper bound on the objective of the form

  $$\mathbb{E} f(\bar{u}^*) \;\leqslant\; (1+\varepsilon) f(u^*_{\mathcal{E}}), \tag{5.7}$$

  similar to (5.3) (Sect. 5.5). The approach is based on the work of Kleinberg and Tardos [KT99], which is set in an LP relaxation framework. However their results are restricted in that they assume a graph-based representation and extensively rely on the finite dimensionality. In contrast, our results hold in the continuous setting without assuming a particular problem discretization.
- We conclude the chapter with an experimental comparison and verification of the obtained *a priori* and *a posteriori* bounds (Sect. 5.6).

5.2 Related Work

As shown in Sect. 2.6.1, in the spatially continuous setting the *two-class* problem admits the trivial thresholding approach with

$$\bar{u}^*_\alpha \;:=\; e^1\,\chi_{\{u_1^*>\alpha\}} + e^2\,\chi_{\{u_1^*\leqslant\alpha\}} \tag{5.8}$$

for almost every $\alpha>0$ due to the coarea formula. In view of the ε-optimality bound (5.3), this amounts to $f(\bar{u}^*)=f(u^*_{\mathcal{E}})$, i.e. $\varepsilon=0$. After discretization, the same property holds if the regularizer satisfies a generalized coarea formula, cf. Sect. 2.6 and Sect. 3.2. In the 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 provide a brief overview in order to motivate our approach.

5.2.1 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. [BVZ01], and is based on the α-expansion method.

The α-expansion method provides a way to reduce the multiclass segmentation problem to a sequence of two-class problems, which can then be solved globally optimal, for instance using graph cuts, for metric d. Denote by $G=(V,E)$ the (undirected) graph representation of the problem, where the energy for a labeling $\ell\colon V\to\mathcal{I}:=\{1,\dots,l\}$ is

$$f(\ell) \;=\; \sum_{x\in V} s_x(\ell(x)) + \sum_{e=(x^1,x^2)\in E} d(\ell(x^1),\ell(x^2)) \tag{5.9}$$

for some nonnegative $s_x\colon\mathcal{I}\to\mathbb{R}_{\geqslant 0}$ and metric $d\colon\mathcal{I}^2\to\mathbb{R}_{\geqslant 0}$.

An early idea for generating integral solutions from the solution of the relaxed problem was provided by [DJPS94] 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 [DJPS94] for the finite-dimensional problem (5.9).

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. 5.1: In each step, one label j is selected, and $\ell_{j,\mathcal{V}'}$ 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 [BVZ01]. The inner problem (5.10) 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.

Algorithm 5.1. Graph-Based α-Expansion

1: Choose $\ell^{(0)}: V \to \mathcal{I}$.
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.

$$\ell' := \ell_{j,V'} \;=\; \operatorname*{arg\,min}_{V' \supseteq \{x \in V \,|\, \ell^{(k)}(x) = j\}} f(\ell_{j,V'}). \qquad (5.10)$$

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** $f\big(\ell^{(k)}\big)$ did not decrease in at least one of the inner iterations.
11: **Output:** $\ell^+ := \ell^{(k)}$.

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 5.1. *[BVZ01, Thm. 6.1] Let ℓ^+ be a local minimum with respect to expansion moves, and ℓ^* be a global minimizer of (5.9). Then*

$$f(\ell^+) \;\leqslant\; 2\,c\,f(\ell^*), \qquad (5.11)$$

where

$$c \;:=\; \frac{\max_{i \neq i'} d(i,i')}{\min_{i \neq i'} d(i,i')} \geqslant 1. \qquad (5.12)$$

Proof. We will sketch the proof in order to highlight the differences to the spatially continuous framework, and motivate the reasoning behind the arguments in Sect. 5.5. The outline is as follows:

1. If ℓ^* is a true minimizer and $j \in \mathcal{I}$, then

$$\ell^{+,j}(x) \;:=\; \begin{cases} j, & \ell^*(x) = j, \\ \ell^+(x), & \text{otherwise} \end{cases} \qquad (5.13)$$

is a valid α-expansion from ℓ^+. Therefore, since ℓ^+ is a local minimum,

$$f(\ell^+) \;\leqslant\; f(\ell^{+,j}). \qquad (5.14)$$

2. Define $V^j := (\ell^*)^{-1}(\{j\})$. We denote the *restriction* of the energy (5.9) to the – unary and pairwise – potentials involving only vertices in V^j by $f|_{\operatorname{int} V^j}$, and the restriction to the – only pairwise – potentials involving *exactly one* vertex in V^j (and one in $V \setminus V^j$) by $f|_{\operatorname{bd} V^j}$.

Since $\ell^{+,j}$ and ℓ^+ coincide in $V \setminus V^j$ by definition, from (5.14) we conclude

$$f|_{\operatorname{int} V^j}(\ell^+) + f|_{\operatorname{bd} V^j}(\ell^+) \leqslant \underbrace{f|_{\operatorname{int} V^j}(\ell^{+,j})}_{=f|_{\operatorname{int} V^j}(\ell^*)} + \underbrace{f|_{\operatorname{bd} V^j}(\ell^{+,j})}_{\leqslant c f|_{\operatorname{bd} V^j}(\ell^*)} \tag{5.15}$$

$$\leqslant c\,(f|_{\operatorname{int} V^j}(\ell^*) + f|_{\operatorname{bd} V^j}(\ell^*)). \tag{5.16}$$

The inequality involving c holds because d is a metric: for each of the terms $d(\ell^*(x^1), \ell^*(x^2))$ that occurs in $f|_{\operatorname{bd} V^j}(\ell^*)$, exactly one of the (x^1, x^2) is contained in V^j, therefore $\ell^*(x^1) \neq \ell^*(x^2)$ and $d(\ell^*(x^1), \ell^*(x^2)) \geqslant \min_{i \neq i'} d(i, i') \geqslant c^{-1}\, d(\ell^{+,j}(x^1), \ell^{+,j}(x^2))$.

3. We now use the fact that each unary term is contained in exactly *one* of the $f|_{\operatorname{int} V^j}$, $j \in \mathcal{I}$. Likewise, each nonzero pairwise term is contained in exactly *one* of the $f|_{\operatorname{int} V^j}$, or in exactly *two* of the $f|_{\operatorname{bd} V^j}$. Using this relation on the summation of (5.16) over all $j \in \mathcal{I}$, we obtain the assertion,

$$f(\ell^+) \leqslant f(\ell^+) + \frac{1}{2} \sum_{j=1}^{l} f|_{\operatorname{bd} V^j}(\ell^+) \tag{5.17}$$

$$\overset{(5.16)}{\leqslant} c \left(f(\ell^*) + \underbrace{\frac{1}{2} \sum_{j=1}^{l} f|_{\operatorname{bd} V^j}(\ell^*)}_{\leqslant f(\ell^*)} \right) \leqslant 2\, c\, f(\ell^*). \tag{5.18}$$

□

From (5.11) we therefore obtain $\varepsilon = 2\,c - 1$ for the α-expansion method. The principle of reducing multi-class problems to a sequence of two-class problems such as (5.10) is also the basis for the α-β-swap technique from the same authors, which can handle the case of semi-metric d, but provides no bound similar to (5.11). A generalization can be found in [LRB07, LRRB10]: The authors view the problem of finding the optimal expansion step (5.10) as the decision between two solutions: identifying V' with its characteristic function $u' := \chi_{V'} \colon V \to \{0, 1\}$, the expansion step becomes

$$u' = \arg \min_{u' : V \to \{0,1\}} f\big((1 - u')\, \ell^{(k)} + u'\, j\big), \tag{5.19}$$

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.

5.2.2 Continuous Binary Fusion

The finite-dimensional approach (5.19) can be generalized to the *spatially continuous* case by essentially replacing V with Ω. This was proposed in [TPCB08] in an informal way, without specifying the actual function spaces and assumptions on the functionals. In [Ols09, OBOK09], the authors argue that Prop. 5.1 similarly holds for functionals of the form (5.1), with the separable (but anisotropic) regularizer

$$\Psi(z) = \sum_{j=1}^{l} \|A\, z^j\|_2, \quad z \in \mathbb{R}^{d \times l}, \tag{5.20}$$

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.
- They use a pointwise argument much as the vertex- and edge-wise argument used in Prop. 5.1. This requires additional justification when dealing with BV functions, which are defined only almost everywhere.
- The authors seem to employ the classical interior and exterior as replacements for the interior and boundary in (5.16). For functionals on BV involving the total variation, this causes the issue that the restriction of f to these sets is not well-defined, since sets with nonzero $\mathcal{H}^{d-1}$ measure can be added or removed from such a set by choosing a different representative of ℓ in the same L^1 equivalence class (Rem. A.14).

Apart from these problems, it is nontrivial to show that the continuous analogon to Alg. 5.1 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.

Properly addressing these issues would require considerable additional work, and still provide a result which is tied to a specific optimization method that requires to solve a *sequence* of convex problems. Instead, we decided to base our optimality results on an approach which more closely resembles (5.2), as motivated in the following section.

5.2.3 LP Relaxation with Derandomization

In [CKR98], 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 (5.9) with (locally weighted) uniform metric regularizer $d(i,j) = \chi_{\{i\neq j\}}$. As seen in the previous chapter, for grid graphs this corresponds to $\Psi = \|\cdot\|_1$.

In order to cope with general metrics, [KT99] adapted a variant of the LP formulation (Sect. 3.2), which raises the bound to $\varepsilon = 1$ for the uniform metric. For the uniform metric, the LP relaxation has the form

$$\min_{y,z} \quad \sum_{x\in V} \sum_{j\in\mathcal{I}} s_x(j)\, y_{x,j} + \sum_{e\in E} w_e\, z_e \tag{5.21}$$

$$s.t. \quad y_{x,\cdot} \in \Delta_l, \quad x \in V, \tag{5.22}$$

$$z_e = \frac{1}{2} \sum_{j\in\mathcal{I}} z_{e,j}, \tag{5.23}$$

$$z_{e,j} \geqslant y_{x^1,j} - y_{x^2,j}, \quad (x^1,x^2) \in E, \tag{5.24}$$

$$z_{e,j} \geqslant y_{x^2,j} - y_{x^1,j}, \quad (x^1,x^2) \in E. \tag{5.25}$$

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

$$\min_{y\in(\Delta_l)^n} \left\{ \sum_{x\in V} \sum_{j\in\mathcal{I}} s_x(j)\, y_{x,j} + \frac{1}{2} \sum_{e=(x^1,x^2)\in E} w_e \sum_{j\in\mathcal{I}} |y_{x^1,j} - y_{x^2,j}| \right\}. \tag{5.26}$$

Assuming $w_e \equiv 1$, this is equivalent to a graph-based discretization of (5.9) and closely resembles (5.2), which motivates to adapt the proof for our spatially continuous setting.

The bound of $\varepsilon = 1$ is proven in [KT99] by first considering a randomized rounding method and then showing that it can be derandomized in polynomial time (see [Sri99, BW05] for an overview of randomized algorithms and derandomization strategies, and [KMN93] for an application to network flow problems). We restrict ourselves to a probabilistic result, since the derandomization techniques crucially depend on fixing single labels, which is not available in a well-defined sense in the continuous setting.

However, much as the proof of Prop. 5.1, the proof in [KT99] strongly relies on the vertex/edge representation, which poses a problem when transferring the results to the infinite-dimensional setting. In particular, several steps directly involve the slack variables z_e and $z_{e,j}$, which do not have a direct analogon in the continuous setting. These difficulties are aggravated by the particular randomized rounding algorithm, which requires a possibly infinite number of stages. We show how to overcome these difficulties in Sect. 5.5.

5.3 Improved Deterministic Schemes

As indicated previously, the two-class approach gives rise to a simple deterministic thresholding approach. We first discuss two similar methods for the multi-class case in order to compare them to the probabilistic approach in the later sections.

5.3.1 First-Max

The simplest deterministic rounding scheme is the *first-max* approach: The label $\ell(x)$ is set to the index of the first maximal component of the relaxed solution $u^*(x)$,

$$\ell(x) \;=\; \min \left\{ \arg \max_{j \in \{1,\dots,l\}} u_j \right\}. \tag{5.27}$$

While this works well for the uniform metric, it may lead to undesired effects for non-standard metrics: Consider the segmentation of a grayscale image with the three labels $1, 2, 3$ corresponding to the gray level intensity and the linear metric $d(i, j) = |i - j|$. Assume there is a region where

$$u^*(x) \;=\; \begin{pmatrix} \frac{1}{3} + \delta(x) \\ \frac{1}{3} \\ \frac{1}{3} - \delta(x) \end{pmatrix} \tag{5.28}$$

for some small $\delta(x) \in \mathbb{R}$, which may occur due to small noise in the image, or due to inexact optimization. The most "natural" choice given the interpretation as grayscale values is the constant labeling $\ell(x) = 2$. However, the first-max approach results in $\ell(x) \in \{1, 3\}$, depending on the sign of $\delta(x)$, which leads to a noisy final segmentation.

5.3.2 Modified First-Max

On closer inspection, the first-max approach amounts to choosing

$$\ell(x) \;=\; \min\left\{\arg\min_{\ell\in\{1,\dots,l\}} \|u(x)-e^\ell\|_2\right\}. \tag{5.29}$$

We propose to extend this to non-uniform distances by setting (for isotropic Ψ)

$$\begin{aligned}\ell(x) \;&=\; \arg\min_{\ell\in\{1,\dots,l\}} \bar\Psi(u(x)-e^\ell), \\ &\quad \bar\Psi\colon \Delta^l\to\mathbb{R},\quad \bar\Psi(y) := \Psi(e^1 y^\top).\end{aligned} \tag{5.30}$$

That is, we select the label corresponding to the nearest unit vector *with respect to* $\bar\Psi$ (note that instead of e^1 we could choose any normalized vector as Ψ is assumed to be rotationally invariant). We thereby introduce knowledge about the structure of Ψ into the rounding process, which potentially improves the result. Note that this requires a numerical approximation in the case where there is no closed form expression for Ψ.

For the linear distance example above we obtain, for the corresponding (exact) Euclidean embedding Ψ_A with $A = \begin{pmatrix} -1 & 0 & 1 \end{pmatrix}$,

$$\bar\Psi(z) \;=\; |-z_1+z_3|. \tag{5.31}$$

Thus

$$\begin{aligned}\bar\Psi(u(x)-e^1) \;&=\; |1-2\delta(x)|, \\ \bar\Psi(u(x)-e^2) \;&=\; |2\delta(x)|, \\ \bar\Psi(u(x)-e^3) \;&=\; |1+2\delta(x)|.\end{aligned} \tag{5.32}$$

In contrast to the "first-max" rounding in the previous section, for small δ we get the stable and semantically correct choice $\ell(x) = 2$. While it is a heuristic, this method proved to work well in practice, and considerably improved both the *a posteriori* bounds as well as the quality of the solution (Sect. 5.6).

5.4 Coarea Formula and Probabilistic Rounding

As a motivation for the following sections, we first provide a probabilistic interpretation of the generalized coarea formula outlined in Sect. 2.6. From Thm. A.32, we know that for $u' \in \mathrm{BV}(\Omega, [0,1])$, the coarea formula states that the total variation of u' can be represented by summing the boundary lengths of its superlevelsets:

$$\mathrm{TV}(u') \;=\; \int_0^1 \mathrm{TV}(\chi_{\{u'>\alpha\}})\,d\alpha. \tag{5.33}$$

The coarea formula provides a connection between problem (5.1) and the relaxation (5.2) in the two-class case, where $\mathcal{E} = \{e^1, e^2\}$, $u \in \mathcal{C}_\mathcal{E}$ and $u_1 = 1 - u_2$: From Prop. 2.7,

$$\mathrm{TV}(u) = \|e^1-e^2\|_2\,\mathrm{TV}(u_1) = \sqrt{2}\,\mathrm{TV}(u_1), \tag{5.34}$$

therefore the coarea formula (5.33) can be rewritten as

$$\begin{aligned} \mathrm{TV}(u) &= \sqrt{2}\int_0^1 \mathrm{TV}(\chi_{\{u_1>\alpha\}})d\alpha && (5.35)\\ &= \int_0^1 \mathrm{TV}(e^1\chi_{\{u_1>\alpha\}} + e^2\chi_{\{u_1\leqslant\alpha\}})d\alpha && (5.36)\\ &= \int_0^1 \mathrm{TV}(\bar{u}_\alpha)d\alpha, \quad \bar{u}_\alpha := e^1\chi_{\{u_1>\alpha\}} + e^2\chi_{\{u_1\leqslant\alpha\}}. && (5.37) \end{aligned}$$

Consequently, the total variation of u can be expressed as the *mean* over the total variations of a set of *integral* labelings $\{\bar{u}_\alpha \in \mathcal{C}_\mathcal{E} | \alpha \in [0,\,1]\}$, obtained by *rounding* u *at different thresholds* α. We now adopt a *probabilistic* view of (5.37): We regard the mapping

$$R\colon (u,\alpha) \in \mathcal{C}\times[0,1] \;\mapsto\; \bar{u}_\alpha \in \mathcal{C}_\mathcal{E} \quad (\text{for a.e. } \alpha\in[0,1]) \tag{5.38}$$

as a *parametrized, deterministic* rounding algorithm that depends on u and on an additional parameter α. From this we obtain a *probabilistic* (randomized) rounding algorithm by assuming α to be a uniformly distributed random variable. Under these assumptions the coarea formula (5.37) can be written as

$$\mathrm{TV}(u) \;=\; \mathbb{E}_\alpha \mathrm{TV}(\bar{u}_\alpha). \tag{5.39}$$

This has the probabilistic interpretation that applying the probabilistic rounding to (arbitrary, but fixed) u does – in a probabilistic sense, i.e. in the mean – not change the objective. It can be shown that this property extends to the full functional f in (5.2). A well-known implication is that if $u = u^*$, i.e. u minimizes (5.2), then almost every $\bar{u}_\alpha = \bar{u}^*_\alpha$ is a minimizer of (5.1) [CEN06].

Unfortunately, property (5.39) is intrinsically restricted to the two-class case with TV regularizer. In the general case, one would hope to obtain a relation

$$f(u) \;=\; \int_\Gamma f(\bar{u}_\gamma)d\mu(\gamma) = \mathbb{E}_\gamma f(\bar{u}_\gamma) \tag{5.40}$$

for some probability space $(\Gamma,\,\mu)$. For $l = 2$ and $\Psi(x) = \|\cdot\|_2$, (5.39) shows that (5.40) holds with $\gamma = \alpha$, $\Gamma = [0,1]$, $\mu = \mathcal{L}^1$, and $R\colon \mathcal{C}\times\Gamma\to\mathcal{C}_\mathcal{E}$ as defined in (5.38).

In the multiclass case, the difficulty lies in providing a suitable combination of a probability space $(\Gamma,\,\mu)$ and a parametrized rounding step $(u,\,\gamma)\mapsto\bar{u}_\gamma$. Unfortunately, obtaining a relation such as (5.39) for the full functional (5.1) is unlikely, as it would mean that solutions to the (after discretization) NP-hard problem (5.1) could be obtained by solving the convex relaxation (5.2) and subsequent rounding.

The main result of this chapter will be a bound of the form

$$(1+\varepsilon)f(u) \;\geqslant\; \int_\Gamma f(\bar{u}_\gamma)d\mu(\gamma) = \mathbb{E}_\gamma f(\bar{u}_\gamma). \tag{5.41}$$

This can be seen as an *approximate* variant of the coarea formula. While (5.41) is not sufficient to provide a bound on $f(\bar{u}_\gamma)$ for *particular* γ, it permits a *probabilistic* bound in the sense of (5.7): for any minimizer u^* of the relaxed problem (5.2),

$$\mathbb{E}_\gamma f(\bar{u}^*_\gamma) \leqslant (1+\varepsilon)f(u^*) \leqslant (1+\varepsilon)f(u^*_\mathcal{E}), \tag{5.42}$$

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

In the following sections we construct a suitable parametrized rounding method and probability space in order to obtain an approximate coarea formula of the form (5.41).

5.5 A Priori Bounds

5.5.1 Probabilistic Rounding for Multiclass Image Partitions

We consider the probabilistic rounding approach based on [KT99] as defined in Alg. 5.2. Where possible without ambiguities, we omit the parentheses for elements of a sequence, i.e. we denote $u^k = u^{(k)}$, in order to avoid notational clutter. The algorithm proceeds in a number of phases. At each iteration, a label and a threshold

$$\gamma^k := (i^k, \alpha^k) \ \in \ \Gamma' := \mathcal{I} \times [0,1] \tag{5.43}$$

are randomly chosen (step 3), and label i^k is assigned to all yet unassigned points x where $u_{i^k}^{k-1}(x) > \alpha^k$ holds (step 5). In contrast to the two-class case considered above, the randomness is provided by a *sequence* (γ^k) of uniformly distributed random variables, i.e. $\Gamma = (\Gamma')^{\mathbb{N}}$.

After iteration k, all points in the set $U^k \subseteq \Omega$ are still *unassigned*, while all points in $\Omega \setminus U^k$ have been assigned an (integral) label in iteration k or in a previous iteration. Iteration $k+1$ potentially modifies points only in the set U^k. The variable c_j^k stores the lowest threshold α chosen for label j up to and including iteration k, and is only required for the proofs.

While the algorithm is defined using pointwise operations, it is well-defined in the sense that for fixed γ, the sequence (u^k), viewed as elements in L^1, does not depend on the specific representative of u from the equivalence class in L^1. The sequences (M^k) and (U^k) depend on the representative, but are unique up to $\mathcal{L}^d$-negligible sets.

In an actual implementation, the algorithm could be terminated as soon as all points in Ω have been assigned a label, i.e. $U^k = \emptyset$. However, in our framework used for analysis the algorithm never terminates explicitly. Instead, for fixed input u we regard the algorithm as a mapping between *sequences* of parameters (or instances of random variables) $\gamma = (\gamma^k) \in \Gamma$ and *sequences* of states (u_γ^k), (U_γ^k) and (c_γ^k). We drop the subscript γ if it does not create ambiguities. The elements of the sequence $(\gamma^{(k)})$ are independently uniformly distributed, and by the Kolmogorov extension theorem [Øks03, Thm. 2.1.5] there exists a probability space and a stochastic process on the set of sequences γ with compatible marginal distributions.

Algorithm 5.2. Continuous Probabilistic Rounding

$u^0 \leftarrow u$, $U^0 \leftarrow \Omega$, $c^0 \leftarrow (1,...,1) \in \mathbb{R}^l$.
for $k = 1, 2, ...$ **do**
  Randomly choose $\gamma^k := (i^k, \alpha^k)$ uniformly from $\mathcal{I} \times [0,1]$.
  $M^k \leftarrow U^{k-1} \cap \{x \in \Omega | u^{k-1}_{i^k}(x) > \alpha^k\}$.
  $u^k \leftarrow e^{i^k} \chi_{M^k} + u^{k-1} \chi_{\Omega \setminus M^k}$.
  $U^k \leftarrow U^{k-1} \setminus M^k$.
  $c^k_j \leftarrow \begin{cases} \min\{c^{k-1}_j, \alpha^k\}, & j = i^k, \\ c^{k-1}_j, & \text{otherwise.} \end{cases}$
end for

In order to define the parametrized rounding step $(u, \gamma) \mapsto \bar{u}_\gamma$, we observe that once $U^{k'}_\gamma = \emptyset$ occurs for some $k' \in \mathbb{N}$, the sequence (u^k_γ) becomes stationary at $u^{k'}_\gamma$. In this case the algorithm may be terminated, with output $\bar{u}_\gamma := u^{k'}_\gamma$:

Definition 5.2. *Let $u \in \mathrm{BV}(\Omega)^l$ and $f\colon \mathrm{BV}(\Omega)^l \to \mathbb{R}$. For some $\gamma \in \Gamma$, if $U^{k'}_\gamma = \emptyset$ in Alg. 5.2 for some $k' \in \mathbb{N}$, we denote $\bar{u}_\gamma := u^{k'}_\gamma$. We define*

$$\begin{aligned} f(\bar{u}_{(.)})\colon\ & \Gamma \to \mathbb{R} \cup \{+\infty\} \\ \gamma \in \Gamma \quad & \mapsto f(\bar{u}_\gamma) := \begin{cases} f(u^{k'}_\gamma), & \text{if there ex. } k' \in \mathbb{N}\colon U^{k'}_\gamma = \emptyset \wedge u^{k'}_\gamma \in \mathrm{BV}(\Omega)^l, \\ +\infty, & \text{otherwise.} \end{cases} \end{aligned} \tag{5.44}$$

We denote by $f(\bar{u})$ the corresponding random variable induced by assuming γ to be uniformly distributed on Γ.

As indicated above, $f(\bar{u}_\gamma)$ is well-defined: if $U^{k'}_\gamma = \emptyset$ for some (γ, k') then $u^{k'}_\gamma = u^{k''}_\gamma$ for all $k'' \geqslant k'$. Instead of focusing on local properties of the random sequence (u^k_γ) as in the proofs for the finite-dimensional case (Sect. 5.2.1 and 5.2.3), we will derive our results directly for the sequence $(f(u^k_\gamma))$.

In particular, we will show that the expectation of $f(\bar{u})$ over all sequences γ can be bounded according to

$$\mathbb{E} f(\bar{u}) = \mathbb{E}_\gamma f(\bar{u}_\gamma) \ \leqslant\ (1+\varepsilon) f(\bar{u}) \tag{5.45}$$

for some $\varepsilon \geqslant 0$, cf. (5.41). Consequently, the rounding process may only increase the average objective in a controlled way.

5.5.2 Termination Properties

Theoretically, the algorithm may produce a sequence (u^k_γ) that does *not* become stationary, or does become stationary with a solution that is not an element of $\mathrm{BV}(\Omega)^l$. In Thm. 5.5 we show that this happens only with zero probability, i.e. almost surely Alg. 5.2 generates (in a finite number of iterations) an *integral* labeling function $\bar{u}_\gamma \in \mathcal{C}_\mathcal{E}$. The following two propositions are required for the proof.

Proposition 5.3. *For the sequence (c^k) produced by Alg. 5.2,*

$$\mathbb{P}(e^\top c^k < 1) \;\geqslant\; \sum_{p\in\{0,1\}^l} (-1)^{e^\top p} \left(\sum_{j=1}^{l} \frac{1}{l} \left(\left(1 - \frac{1}{l} \right)^{p_j} \right) \right)^k \tag{5.46}$$

holds. In particular,

$$\mathbb{P}(e^\top c^k < 1) \;\overset{k\to\infty}{\to}\; 1. \tag{5.47}$$

Proof. Denote by $n_j^k \in \mathbb{N}_0$ the number of $k' \in \{1,...,k\}$ such that $i^{k'} = j$, i.e. the number of times label j was selected up to and including the k-th step. Then

$$(n_1^k, ..., n_l^k) \;\sim\; \text{Multinomial}\Big(k; \frac{1}{l}, ..., \frac{1}{l} \Big), \tag{5.48}$$

i.e. the probability of a specific instance is

$$\mathbb{P}((n_1^k, ..., n_l^k)) \;=\; \begin{cases} \frac{k!}{n_1^k! \cdot ... \cdot n_l^k!} \left(\frac{1}{l} \right)^k, & \sum_j n_j^k = k, \\ 0, & \text{otherwise.} \end{cases} \tag{5.49}$$

Therefore,

$$\mathbb{P}(e^\top c^k < 1) \;=\; \sum_{n_1^k,...,n_l^k} \mathbb{P}(e^\top c^k < 1 | (n_1^k, ..., n_l^k)) \, \mathbb{P}((n_1^k, ..., n_l^k)) \tag{5.50}$$

$$= \sum_{n_1^k+...+n_l^k=k} \frac{k!}{n_1^k! \cdot ... \cdot n_l^k!} \left(\frac{1}{l} \right)^k \mathbb{P}(e^\top c^k < 1 | (n_1^k, ..., n_l^k)). \tag{5.51}$$

Since $c_1^k, ..., c_l^k < \frac{1}{l}$ is a sufficient condition for $e^\top c < 1$, we may bound the probability according to

$$\mathbb{P}(e^\top c < 1) \;\geqslant\; \sum_{n_1^k+...+n_l^k=k} \frac{k!}{n_1^k! \cdot ... \cdot n_l^k!} \left(\frac{1}{l} \right)^k \mathbb{P}\Big(c_j^k < \frac{1}{l} \, \forall j \in \mathcal{I} | (n_1^k, ..., n_l^k) \Big). \tag{5.52}$$

We now consider the distributions of the components c_j^k of c^k conditioned on the vector $(n_1^k, ..., n_l^k)$. Given n_j^k, the probability of $\{c_j^k \geqslant t\}$ is the probability that in each of the n_j^k steps where label j was selected the threshold α was randomly chosen to be *at least as large as* t. For $0 < t < 1$, we conclude

$$\mathbb{P}(c_j^k < t | (n_1^k, ..., n_l^k)) \;=\; \mathbb{P}(c_j^k < t | n_j^k) \tag{5.53}$$

$$= 1 - \mathbb{P}(c_j^k \geqslant t | n_j^k) \tag{5.54}$$

$$\overset{0<t<1}{=} 1 - (1-t)^{n_j^k}. \tag{5.55}$$

The above formulation also covers the case $n_j^k = 0$ (note that we assumed $0 < t < 1$). For fixed k the distributions of the c_j^k are independent when conditioned on $(n_1^k, ..., n_l^k)$. Therefore we obtain from (5.52) and (5.55)

$$\mathbb{P}(e^\top c < 1) \overset{(5.52)}{\geqslant} \sum_{n_1^k+...+n_l^k=k} \frac{k!}{n_1^k! \cdot ... \cdot n_l^k!} \left(\frac{1}{l} \right)^k \prod_{j=1}^{l} \mathbb{P}\Big(c_j^k < \frac{1}{l} | (n_1^k, ..., n_l^k) \Big) \tag{5.56}$$

$$\overset{(5.55)}{=} \sum_{n_1^k+...+n_l^k=k} \frac{k!}{n_1^k! \cdot ... \cdot n_l^k!} \left(\frac{1}{l} \right)^k \prod_{j=1}^{l} \left(1 - \left(1 - \frac{1}{l} \right)^{n_j^k} \right). \tag{5.57}$$

Expanding the product and swapping the summation order, we derive

$$\begin{aligned}
\mathbb{P}(e^\top c^k < 1) &\geqslant \sum_{n_1^k+\ldots+n_l^k=k} \frac{k!}{n_1^k!\cdot\ldots\cdot n_l^k!}\left(\frac{1}{l}\right)^k \sum_{p\in\{0,1\}^l} \prod_{j=1}^{l}\left(-\left(1-\frac{1}{l}\right)^{n_j^k}\right)^{p_j} && (5.58)\\
&= \sum_{p\in\{0,1\}^l} \sum_{n_1^k+\ldots+n_l^k=k} \frac{k!}{n_1^k!\cdot\ldots\cdot n_l^k!}\left(\frac{1}{l}\right)^k \prod_{j=1}^{l}\left(-\left(1-\frac{1}{l}\right)^{n_j^k}\right)^{p_j} && (5.59)\\
&= \sum_{p\in\{0,1\}^l} \sum_{n_1^k+\ldots+n_l^k=k} \frac{k!}{n_1^k!\cdot\ldots\cdot n_l^k!}\cdot \\
&\qquad (-1)^{e^\top p}\left(\frac{1}{l}\right)^k \prod_{j=1}^{l}\left(\left(1-\frac{1}{l}\right)^{p_j}\right)^{n_j^k} && (5.60)\\
&= \sum_{p\in\{0,1\}^l} (-1)^{e^\top p} \sum_{n_1^k+\ldots+n_l^k=k} \frac{k!}{n_1^k!\cdot\ldots\cdot n_l^k!} \prod_{j=1}^{l}\left(\frac{1}{l}\left(1-\frac{1}{l}\right)^{p_j}\right)^{n_j^k} && (5.61)\\
&\overset{(*)}{=} \sum_{p\in\{0,1\}^l} (-1)^{e^\top p}\left(\underbrace{\sum_{j=1}^{l}\frac{1}{l}\left(1-\frac{1}{l}\right)^{p_j}}_{=:q_p}\right)^k, && (5.62)
\end{aligned}$$

which proves (5.46). At $(*)$ the multinomial summation formula was invoked. Note that in (5.62) the n_j^k do not occur explicitly anymore. To show the second assertion (5.47), we use the fact that $0<q_p<1$ for any $p\neq(0,\ldots,0)$. Therefore

$$\begin{aligned}
\mathbb{P}(e^\top c^k < 1) &\geqslant q_0 + \sum_{p\in\{0,1\}^l, p\neq 0} (-1)^{e^\top p}(q_p)^k && (5.63)\\
&= 1 + \sum_{p\in\{0,1\}^l, p\neq 0} (-1)^{e^\top p}\underbrace{(q_p)^k}_{\overset{k\to\infty}{\to}0} && (5.64)\\
&\overset{k\to\infty}{\to} 1, && (5.65)
\end{aligned}$$

which proves (5.47). □

We now show that Alg. 5.2 generates a sequence in $\mathrm{BV}(\Omega)^l$ almost surely.

Proposition 5.4. *For the sequences (u^k), (U^k) generated by Alg. 5.2, define*

$$A := \bigcap_{k=1}^{\infty}\{\gamma\in\Gamma \,|\, \mathrm{Per}(U_\gamma^k)<\infty\}. \tag{5.66}$$

Then

$$\mathbb{P}(A) = 1. \tag{5.67}$$

If $\mathrm{Per}(U_\gamma^k)<\infty$ for all k, then $u_\gamma^k\in\mathrm{BV}(\Omega)^l$ for all k as well. Moreover,

$$\mathbb{P}(u^k\in\mathrm{BV}(\Omega)^l \wedge \mathrm{Per}(U^k)<\infty \forall k\in\mathbb{N}) = 1, \tag{5.68}$$

i.e. the algorithm almost surely generates a sequence of BV *functions (u^k) and a sequence of sets of finite perimeter (U^k).*

Proof. We first show that if $\mathrm{Per}(U^{k'}) < \infty$ for all $k' \leqslant k$, then $u^k \in \mathrm{BV}(\Omega)^l$ for all $k' \leqslant k$ as well. For $k = 0$, the assertion holds, since $u^0 = u \in \mathrm{BV}(\Omega)^l$ by assumption. For $k \geqslant 1$,

$$u^k \;=\; e^{i^k}\,\chi_{M^k} + u^{k-1}\,\chi_{\Omega\setminus M^k}. \tag{5.69}$$

Since $M^k = U^{k-1} \cap (\Omega \setminus U^k)$, and U^k, U^{k-1} are assumed to have finite perimeter, M^k also has finite perimeter. Applying [AFP00, Thm. 3.84] together with the boundedness of u^{k-1} and $u^{k-1} \in \mathrm{BV}(\Omega)^l$ by induction then provides $u^k \in \mathrm{BV}(\Omega)^l$.

We now denote

$$I^k \;:=\; \{\gamma \in \Gamma \,|\, \mathrm{Per}(U^k_\gamma) = \infty\}, \tag{5.70}$$

and the event that the *first* set with non-finite perimeter is encountered at step $k \in \mathbb{N}_0$,

$$B^k \;:=\; I^k \cap (\Gamma \setminus I^{k-1}) \cap \ldots \cap (\Gamma \setminus I^0). \tag{5.71}$$

Then

$$\mathbb{P}(A) \;=\; 1 - \mathbb{P}\left(\bigcup_{k=0}^{\infty} B^k \right). \tag{5.72}$$

As the sets B^k are pairwise disjoint, and due to the countable additivity of the probability measure, we have

$$\mathbb{P}(A) \;=\; 1 - \sum_{k=0}^{\infty} \mathbb{P}(B^k). \tag{5.73}$$

Now $U^0 = \Omega$, therefore $\mathrm{Per}(U^0) = \mathrm{TV}(\chi_{U^0}) = 0 < \infty$ and $\mathbb{P}(B^0) = 0$. For $k \geqslant 1$, we have

$$\begin{aligned}
\mathbb{P}(B^k) \;&\leqslant\; \mathbb{P}\big(\mathrm{Per}(U^k) = \infty \wedge \mathrm{Per}(U^{k'}) < \infty \;\forall k' < k\big) && (5.74)\\
&\leqslant\; \mathbb{P}\big(\mathrm{Per}(U^k) = \infty \,|\, \mathrm{Per}(U^{k'}) < \infty \;\forall k' < k\big) && (5.75)\\
&=\; \mathbb{P}\big(\mathrm{Per}(U^{k-1} \cap \{x \in \Omega \,|\, u^{k-1}_{i^k}(x) \leqslant \alpha^k\}) = \infty \,|\, \mathrm{Per}(U^{k'}) < \infty \;\forall k' < k\big). && (5.76)
\end{aligned}$$

By the argument from the beginning of the proof, we know that $u^{k-1} \in \mathrm{BV}(\Omega)^l$ under the condition on $\mathrm{Per}(U^{k'})$, therefore from [AFP00, Thm. 3.40] we conclude that the perimeter $\mathrm{Per}(\{x \in \Omega \,|\, u^{k-1}_{i^k}(x) \leqslant \alpha^k\})$ is finite for $\mathcal{L}^1$-a.e. α^k and all i^k. As the sets of finite perimeter are closed under finite intersection, and the α^k are drawn from an uniform distribution, this implies that

$$\mathbb{P}(\mathrm{Per}(U^k) < \infty \,|\, \mathrm{Per}(U^{k-1}) < \infty) \;=\; 1. \tag{5.77}$$

Together with (5.76) we arrive at

$$\mathbb{P}(B^k) \;=\; 0. \tag{5.78}$$

Substituting this result into (5.73) leads to the assertion,

$$\mathbb{P}(A) \;=\; 1. \tag{5.79}$$

Equation (5.68) follows immediately. □

Using these propositions, we now formulate the main result of this section: Alg. 5.2 almost surely generates an integral labeling that is of bounded variation.

Theorem 5.5. *Let $u \in \mathrm{BV}(\Omega)^l$ and $f(\bar{u})$ as in Def. 5.2. Then*

$$\mathbb{P}(f(\bar{u}) < \infty) \;=\; 1. \tag{5.80}$$

Proof. The first part is to show that (u^k) becomes stationary almost surely, i.e.

$$\mathbb{P}(\exists k \in \mathbb{N} \colon U^k = \emptyset) \;=\; 1. \tag{5.81}$$

Assume there exists k such that $e^\top c^k < 1$, and assume further that $U^k \neq \emptyset$, i.e. there exists some $x \in U^k$. Then $u_j(x) \leqslant c_j^k$ for all labels j. But then $e^\top u(x) \leqslant e^\top c^k < 1$, which is a contradiction to $u(x) \in \Delta_l$. Therefore U^k must be empty. From this observation and Prop. 5.3 we conclude, for all $k' \in \mathbb{N}$,

$$1 \geqslant \mathbb{P}(\exists k \in \mathbb{N} \colon U^k = \emptyset) \geqslant \; \mathbb{P}(e^\top c^{k'} < 1) \overset{k' \to \infty}{\to} 1, \tag{5.82}$$

which proves (5.81).

In order to show that $f(\bar{u}_\gamma) < \infty$ with probability 1, it remains to show that the result is almost surely in $\mathrm{BV}(\Omega)^l$. A sufficient condition is that almost surely *all* iterates u^k are elements of $\mathrm{BV}(\Omega)^l$, i.e.

$$\mathbb{P}(u^k \in \mathrm{BV}(\Omega)^l \, \forall k \in \mathbb{N}) \;=\; 1. \tag{5.83}$$

This is shown by Prop. 5.4. Then

$$\begin{aligned}
\mathbb{P}(f(\bar{u}) < \infty) \quad &\geqslant \quad \mathbb{P}(\{\exists k \in \mathbb{N} \colon U^k = \emptyset\} \wedge \{u^k \in \mathrm{BV}(\Omega)^l \, \forall k \in \mathbb{N}\}) && (5.84)\\
&= \quad \mathbb{P}(\{u^k \in \mathrm{BV}(\Omega)^l \, \forall k \in \mathbb{N}\}) \\
&\quad\; -\mathbb{P}(\{\forall k \in \mathbb{N} \colon U^k \neq \emptyset\} \wedge \{u^k \in \mathrm{BV}(\Omega)^l \, \forall k \in \mathbb{N}\}) && (5.85)\\
&\overset{(5.83)}{=} \quad \mathbb{P}(\{u^k \in \mathrm{BV}(\Omega)^l \, \forall k \in \mathbb{N}\}) - 0 && (5.86)\\
&= \quad 1. && (5.87)
\end{aligned}$$

Thus $\mathbb{P}(f(\bar{u}) < \infty) = 1$, which proves the assertion. □

5.5.3 Intermediate Results

In order to show the bound (5.45), we first need several technical propositions regarding the composition of two BV functions along a set of finite perimeter. We denote by $(E)^1$ and $(E)^0$ the measure-theoretic interior and exterior, and refer to Appendix A.1 for the precise definitions.

It turns out that for deriving the bounds, it is more suitable to replace the upper and lower boundedness of Ψ, $\rho_l \|\cdot\|_2 \leqslant \Psi \leqslant \rho_u \|\cdot\|_2$, by the assumption that there exist a lower bound $\lambda_l > 0$ such that

$$\Psi(z = (z^1, \ldots, z^l)) \;\geqslant\; \lambda_l \frac{1}{2} \sum_{i=1}^{l} \|z^i\|_2 \quad \forall z \in \mathbb{R}^{d \times l}, \; \sum_{i=1}^{l} z^i = 0, \tag{5.88}$$

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

$$\Psi(y\,(e^i - e^j)^\top) \;\leqslant\; \lambda_u \quad \forall i, j \in \{1, \ldots, l\}, y \in \mathbb{R}^d, \|y\|_2 = 1. \tag{5.89}$$

Note that $\lambda_u \geqslant \lambda_l$ in case both are defined, since (5.88) implies, for any y with $\|y\|_2 = 1$,

$$\lambda_u \geqslant \Psi(y\,(e^i - e^j)^\top) \geqslant \frac{\lambda_l}{2}\,(\|y\|_2 + \|y\|_2) = \lambda_l. \tag{5.90}$$

Also, the upper boundedness by λ_u implies a similar bound in the 2-norm:

Proposition 5.6. *Let $\Psi: \mathbb{R}^{d\times l} \to \mathbb{R}_{\geqslant 0}$ be positively homogeneous and convex, and satisfy the upper-boundedness condition (5.89). Then*

$$\Psi(y\,(z^1 - z^2)^\top) \;\leqslant\; \lambda_u \quad \forall z^1, z^2 \in \Delta_l,\; y \in \mathbb{R}^d, \|y\|_2 = 1. \tag{5.91}$$

Moreover, there exists a constant $C < \infty$ such that

$$\Psi(w) \;\leqslant\; C\,\|w\|_2 \quad \forall w \in W := \{w = (w^1|\ldots|w^l) \in \mathbb{R}^{d\times l} | \sum_{i=1}^{l} w^i = 0\}. \tag{5.92}$$

Proof. In order to prove the first assertion (5.91), note that the mapping $w \mapsto \Psi(y\,w^\top)$ is convex, therefore it must assume its maximum on the polytope

$$\Delta_l - \Delta_l := \{z^1 - z^2 | z^1, z^2 \in \Delta_l\}. \tag{5.93}$$

Since the polytope $\Delta_l - \Delta_l$ is the difference of two polytopes, its vertex set is at most the difference of their vertex sets, $V := \{e^i - e^j | i, j \in \{1, \ldots, l\}\}$. On this set, $\Psi(y\,w^\top) \leqslant \lambda_u$ holds for $w \in V$ due to the upper-boundedness condition (5.89), which proves (5.91).

The second equality (5.92) follows from the fact that the set

$$G \;:=\; \{b^{ik} := e^k\,(e^i - e^{i+1})^\top | k \in \{1, \ldots, d\}, i \in \{1, \ldots, l-1\}\} \tag{5.94}$$

is a basis of the linear subspace W satisfying $\Psi(b^{ik}) \leqslant \lambda_u$, and Ψ is positively homogeneous and convex, and thus subadditive. Specifically, there exists a linear transform T: $W \to \mathbb{R}^{d\times(l-1)}$ such that $w = \sum_{i,k} b^{ik}\,\alpha_{ik}$ for $\alpha = T(w)$. Then

$$\Psi(w) = \Psi(\sum_{i,k} b^{ik}\,\alpha_{ik}) = \Psi(\sum_{i,k} |\alpha_{ik}|\,\mathrm{sgn}(\alpha_{ik})\,b^{ik}) \leqslant \sum_{i,k} |\alpha_{ik}|\,\Psi(\mathrm{sgn}(\alpha_{ik})\,b^{ik}) \tag{5.95}$$

Since (5.89) provides $\Psi(\pm b^{ik}) \leqslant \lambda_u$, we obtain

$$\Psi(w) \leqslant \lambda_u \sum_{ik} |\alpha_{ik}| \leqslant \lambda_u\,\|T\|\,\|w\|_2 \tag{5.96}$$

for some suitable operator norm $\|\cdot\|$ and any $w \in W$. □

Proposition 5.7. *Let $E, F \subseteq \Omega^d$ be $\mathcal{L}^d$-measurable sets. Then*

$$(E \cap F)^1 \;=\; (E)^1 \cap (F)^1. \tag{5.97}$$

Proof. We prove mutual inclusion:

- $''\subseteq''$: From the definition of the measure-theoretic interior,

$$x \in (E \cap F)^1 \;\Rightarrow\; \lim_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap E \cap F|}{|\mathcal{B}_\delta(x)|} = 1. \tag{5.98}$$

Since $|\mathcal{B}_\delta(x)| \geqslant |\mathcal{B}_\delta(x) \cap E| \geqslant |\mathcal{B}_\delta(x) \cap E \cap F|$ (and vice versa for $|\mathcal{B}_\delta(x) \cap F|$), it follows by the "sandwich" criterion that both $\lim_{\delta \searrow 0} |\mathcal{B}_\delta(x) \cap E|/|\mathcal{B}_\delta(x)|$ and $\lim_{\delta \searrow 0} |\mathcal{B}_\delta(x) \cap F|/|\mathcal{B}_\delta(x)|$ exist and are equal to 1, which shows $x \in E^1 \cap F^1$.

- $''\supseteq''$: Assume that $x \in E^1 \cap F^1$. Then

$$
\begin{aligned}
1 &\geqslant \limsup_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap E \cap F|}{|\mathcal{B}_\delta(x)|} && (5.99)\\
&\geqslant \liminf_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap E \cap F|}{|\mathcal{B}_\delta(x)|} && (5.100)\\
&= \liminf_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap E| + |\mathcal{B}_\delta \cap F| - |\mathcal{B}_\delta \cap (E \cup F)|}{|\mathcal{B}_\delta(x)|} && (5.101)\\
&\geqslant \liminf_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap E|}{|\mathcal{B}_\delta(x)|} + \liminf_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap F|}{|\mathcal{B}_\delta(x)|} + \liminf_{\delta \searrow 0} \left(-\frac{|\mathcal{B}_\delta \cap (E \cup F)|}{|\mathcal{B}_\delta(x)|} \right)\\
&= 2 - \underbrace{\limsup_{\delta \searrow 0} \frac{|\mathcal{B}_\delta \cap (E \cup F)|}{|\mathcal{B}_\delta(x)|}}_{\leqslant 1} \geqslant 1. && (5.102)
\end{aligned}
$$

  Therefore

$$
\limsup_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap E \cap F|}{|\mathcal{B}_\delta(x)|} = \liminf_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap E \cap F|}{|\mathcal{B}_\delta(x)|} = 1, \tag{5.103}
$$

  i.e. $x \in (E \cap F)^1$. □

Proposition 5.8. *Let $u, v \in \mathrm{BV}(\Omega, \Delta_l)$ and $E \subseteq \Omega$ such that $\mathrm{Per}(E) < \infty$. Define*

$$
w \;:=\; u\,\chi_E + v\,\chi_{\Omega \setminus E}. \tag{5.104}
$$

Then $w \in \mathrm{BV}(\Omega)^l$, and

$$
Dw \;=\; Du\llcorner(E)^1 + Dv\llcorner(E)^0 + \nu_E\,(u^+_{\mathcal{F}E} - v^-_{\mathcal{F}E})^\top\,\mathcal{H}^{d-1}\llcorner(\mathcal{F}E \cap \Omega), \tag{5.105}
$$

where $u^+_{\mathcal{F}E}$ and $v^-_{\mathcal{F}E}$ denote the one-sided approximate limits of u and v on $\mathcal{F}E$, and ν_E is the generalized inner normal of E (Def. A.12). Moreover, for continuous, convex and positively homogeneous Ψ satisfying the upper-boundedness condition (5.89) and some Borel set $A \subseteq \Omega$,

$$
\int_A d\Psi(Dw) \;\leqslant\; \int_{A \cap (E)^1} d\Psi(Du) + \int_{A \cap (E)^0} d\Psi(Dv) + \lambda_u\,\mathrm{Per}(E). \tag{5.106}
$$

Proof. First note that

$$
\begin{aligned}
& \int_{\mathcal{F}E \cap \Omega} \|w^+_{\mathcal{F}E} - w^-_{\mathcal{F}E}\|_2\, d\mathcal{H}^{d-1} && (5.107)\\
\leqslant\;& \sup\{\, \|w^+_{\mathcal{F}E}(x) - w^-_{\mathcal{F}E}(x)\|_2 \,|\, x \in \mathcal{F}E \cap \Omega\} \cdot \mathcal{H}^{d-1}(\mathcal{F}E \cap \Omega) && (5.108)\\
\overset{(*)}{\leqslant}\;& \sup\{\, \|w(x) - w(y)\|_2 \,|\, x, y \in \Omega\} \cdot \mathrm{TV}(\chi_E) && (5.109)\\
\overset{w(x), w(y) \in \Delta_l}{\leqslant}\;& \sqrt{2}\,\mathrm{TV}(\chi_E) && (5.110)\\
=\;& \sqrt{2}\,\mathrm{Per}(E) < \infty. && (5.111)
\end{aligned}
$$

The inequality $(*)$ is a consequence of the definition of $w^{\pm}_{\mathcal{F}E}$ and [AFP00, Thm. 3.59]. We may therefore apply [AFP00, Thm. 3.84] on w, which provides $w \in \mathrm{BV}(\Omega)^l$ and (5.105). Due to [AFP00, Prop. 3.61] (Thm. A.15), the sets $(E)^0, (E)^1$ and $\mathcal{F}E$ form a (pairwise disjoint) partition of Ω, up to an $\mathcal{H}^{d-1}$-zero set. Moreover, since $\Psi(Du) \ll |Du| \ll \mathcal{H}^{d-1}$ by construction, we have, for some Borel set A,

$$\begin{aligned}
\int_A \Psi(Dw) \quad &= \quad \int_{A\cap(E)^1} d\Psi(Dw) + \int_{A\cap(E)^0} d\Psi(Dw) + \\
&\qquad \int_{A\cap\mathcal{F}E\cap\Omega} \Psi(\nu_E\,(w^+_{\mathcal{F}E}(x) - w^-_{\mathcal{F}E}(x))^\top)\, d\mathcal{H}^{d-1} && (5.112)\\
&\overset{(**)}{\leqslant} \quad \int_{A\cap(E)^1} d\Psi(Dw) + \int_{A\cap(E)^0} d\Psi(Dw) + \\
&\qquad \int_{A\cap\mathcal{F}E\cap\Omega} \lambda_u\, d\mathcal{H}^{d-1} && (5.113)\\
&\overset{(5.111)}{\leqslant} \quad \int_{A\cap(E)^1} d\Psi(Dw) + \int_{A\cap(E)^0} d\Psi(Dw) + \lambda_u \mathrm{Per}(E). && (5.114)
\end{aligned}$$

The inequality $(**)$ holds due to the upper boundedness and Prop. 5.6. From [AFP00, Prop. 2.37] we obtain that Ψ is additive on mutually singular Radon measures μ, ν, i.e.

$$|\mu| \perp |\nu| \;\Rightarrow\; \int_B d\Psi(\mu+\nu) = \int_B d\Psi(\mu) + \int_B d\Psi(\nu) \quad \forall B \subseteq \Omega, B \text{ Borel set.} \tag{5.115}$$

Substituting Dw in (5.114) according to (5.105) and using the fact that the three measures that form Dw in (5.105) are mutually singular, the additivity property (5.115) leads to the remaining assertion,

$$\int_A d\Psi(Dw) \;\leqslant\; \int_{A\cap(E)^1} d\Psi(Du) + \int_{A\cap(E)^0} d\Psi(Dv) + \lambda_u \mathrm{Per}(E). \tag{5.116}$$

□

Proposition 5.9. *Let $u, v \in \mathrm{BV}(\Omega, \Delta_l)$, $E \subseteq \Omega$ such that $\mathrm{Per}(E) < \infty$, and*

$$u|_{(E)^1} = v|_{(E)^1} \quad \mathcal{L}^d\text{-a.e.} \tag{5.117}$$

Then $(Du)\llcorner(E)^1 = (Dv)\llcorner(E)^1$, and $\Psi(Du)\llcorner(E)^1 = \Psi(Dv)\llcorner(E)^1$. In particular,

$$\int_{(E)^1} d\Psi(Du) \;=\; \int_{(E)^1} d\Psi(Dv). \tag{5.118}$$

The result also holds when $(E)^1$ is replaced by $(E)^0$. Moreover, the condition (5.117) is equivalent to

$$u|_E = v|_E \quad \mathcal{L}^d\text{-a.e.} \tag{5.119}$$

Remark 5.10. Note that taking the measure-theoretic interior $(E)^1$ is of central importance. The corollary does not hold when replacing the integral over $(E)^1$ with the integral over E, as can be seen from the example of the closed unit ball, i.e. $E = \mathcal{B}_1(0)$, $u = \chi_E$ and $v \equiv 1$.

Proof. We first show (5.119). It suffices to show that

$$\{x \in (E)^1 \;\Leftrightarrow\; x \in E\} \quad \text{for } \mathcal{L}^d\text{-a.e. } x \in \Omega. \tag{5.120}$$

This can be seen by considering the precise representative $\widetilde{\chi_E}$ of χ_E [AFP00, Def. 3.63]

$$
\begin{aligned}
x \in (E)^1 \quad &\Leftrightarrow \quad \lim_{\delta \searrow 0} \frac{|E \cap \mathcal{B}_\delta(x)|}{|\mathcal{B}_\delta(x)|} = 1 && (5.121)\\
&\overset{\lim_{\delta\searrow 0}\frac{|\Omega\cap\mathcal{B}_\delta(x)|}{|\mathcal{B}_\delta(x)|}=1}{\Leftrightarrow} \quad \lim_{\delta \searrow 0} \frac{|(\Omega \setminus E) \cap \mathcal{B}_\delta(x)|}{|\mathcal{B}_\delta(x)|} = 0 && (5.122)\\
&\Leftrightarrow \quad \lim_{\delta \searrow 0} \frac{1}{|\mathcal{B}_\delta(x)|} \int_{\mathcal{B}_\delta(x)} |\chi_E - 1|\, dy = 0 && (5.123)\\
&\Leftrightarrow \quad \widetilde{\chi_E}(x) = 1. && (5.124)
\end{aligned}
$$

Substituting E by $\Omega \setminus E$, the same equivalence shows that $x \in (E)^0 \Leftrightarrow \widetilde{\chi_{\Omega\setminus E}}(x) = 1 \Leftrightarrow \widetilde{\chi_E}(x) = 0$. As $\mathcal{L}^d(\Omega \setminus ((E)^0 \cup (E)^1)) = 0$, this shows that $\chi_{E^1} = \widetilde{\chi_E}$ $\mathcal{L}^d$-a.e. Using the fact that $\widetilde{\chi_E} = \chi_E$ [AFP00, Prop. 3.64 a)], we conclude that $\chi_{(E)^1} = \chi_E$ $\mathcal{L}^d$-a.e., which proves (5.120) and therefore the assertion (5.119).

Since the measure-theoretic interior $(E)^1$ is defined over $\mathcal{L}^d$-integrals, it is invariant under $\mathcal{L}^d$-negligible modifications of E. Together with (5.120) this implies

$$
((E)^1)^1 = (E)^1, \quad \mathcal{F}\,(E)^1 = \mathcal{F}E \quad \text{and} \quad ((E)^1)^0 = (E)^0. \tag{5.125}
$$

To show the relation $(Du)\llcorner(E)^1 = (Dv)\llcorner(E)^1$, consider

$$
\begin{aligned}
Du\llcorner(E)^1 &= D(\chi_{\Omega\setminus(E)^1}\, u + \chi_{(E)^1}\, u)\llcorner(E)^1 && (5.126)\\
&\overset{(*)}{=} D(\chi_{\Omega\setminus(E)^1}\, u + \chi_{(E)^1}\, v)\llcorner(E)^1. && (5.127)
\end{aligned}
$$

The equality $(*)$ holds due to the assumption (5.117), and due to the fact that $Df = Dg$ if $f = g$ $\mathcal{L}^d$-a.e. (see e.g. [AFP00, Prop. 3.2]). We continue from (5.127) via

$$
\begin{aligned}
Du\llcorner(E)^1 &\overset{\text{Prop. 5.8}}{=} \{Du\llcorner((E)^1)^0 + Dv\llcorner((E)^1)^1 + && (5.128)\\
&\qquad \nu_{(E)^1}\,(u^+_{\mathcal{F}E^1} - v^-_{\mathcal{F}E^1})^\top \mathcal{H}^{d-1}\llcorner(\mathcal{F}\,(E)^1 \cap \Omega)\}\llcorner(E)^1\\
&\overset{(5.125)}{=} (Du\llcorner(E)^0 + Dv\llcorner(E)^1)\llcorner(E)^1 + && (5.129)\\
&\qquad \big(\nu_{(E)^1}(u^+_{\mathcal{F}E^1} - v^-_{\mathcal{F}E^1})^\top \mathcal{H}^{d-1}\llcorner(\mathcal{F}E \cap \Omega)\big)\llcorner(E)^1\\
&= Du\llcorner((E)^0 \cap (E)^1) + Dv\llcorner((E)^1 \cap (E)^1) + && (5.130)\\
&\qquad \nu_{(E)^1}(u^+_{\mathcal{F}E^1} - v^-_{\mathcal{F}E^1})^\top \mathcal{H}^{d-1}\llcorner(\mathcal{F}E \cap \Omega \cap (E)^1)\\
&= Dv\llcorner(E)^1. && (5.131)
\end{aligned}
$$

Therefore $Du\llcorner(E)^1 = Dv\llcorner(E)^1$. Then,

$$
\begin{aligned}
\Psi(Du)\llcorner(E)^1 &= \Psi(Du\llcorner(E)^1 + Du\llcorner(\Omega \setminus (E)^1)\,)\llcorner(E)^1 && (5.132)\\
&\overset{(*)}{=} \Psi(Du\llcorner(E)^1)\llcorner(E)^1 + \Psi(Du\llcorner(\Omega \setminus (E)^1))\llcorner(E)^1. && (5.133)
\end{aligned}
$$

In the equality $(*)$ we used the additivity of Ψ on mutually singular Radon measures [AFP00, Prop. 2.37]. By definition of the total variation, $|\mu\llcorner A| = |\mu|\llcorner A$ holds for any measure μ, therefore $|Du\llcorner(\Omega \setminus (E)^1)| = |Du|\llcorner(\Omega \setminus (E)^1)$ and $|Du\llcorner(\Omega \setminus (E)^1)|((E)^1) = 0$, which together with (again by definition) $\Psi(\mu) \ll |\mu|$ implies that the second term in (5.133) vanishes. Since all observations equally hold for v instead of u, we conclude

$$
\Psi(Du)\llcorner(E)^1 = \Psi(Du\llcorner(E)^1)\llcorner(E)^1 \overset{(5.131)}{=} \Psi(Dv\llcorner(E)^1)\llcorner(E)^1 = \Psi(Dv)\llcorner(E)^1. \tag{5.134}
$$

Equation (5.118) follows immediately. □

5.5.4 A Probabilistic A Priori Optimality Bound

In Sect. 5.5.2 we have shown that the rounding process induced by Alg. 5.2 is well-defined in the sense that it returns an integral solution $\bar{u}_\gamma \in \mathrm{BV}(\Omega)^l$ almost surely. We now return to proving an upper bound for the expectation of $f(\bar{u})$ as in the approximate coarea formula (5.7). We first show that the expectation of the *linear part* (data term) of f is invariant under the rounding process.

Proposition 5.11. *The sequence (u^k) generated by Alg. 5.2 satisfies*

$$\mathbb{E}(\langle u^k, s\rangle) \;=\; \langle u, s\rangle \quad \forall k \in \mathbb{N}. \tag{5.135}$$

Proof. In Alg. 5.2, instead of step 5 we consider the simpler update

$$u^k \;\leftarrow\; e^{i^k}\chi_{\{u_{i^k}^{k-1} > \alpha^k\}} + u^{k-1}\chi_{\{u_{i^k}^{k-1} \leqslant \alpha^k\}}. \tag{5.136}$$

This yields exactly the same sequence (u^k), since $u_{i^k}^{k-1}(x) > \alpha^k$ for any $\alpha^k \geqslant 0$ implies that either $x \in U^{k-1}$, or $u_{i^k}^{k-1}(x) = 1$. In both algorithms, points that are assigned a label e^{i^k} at some point in the process will never be assigned a *different* label at a later point. This is made explicit in Alg. 5.2 by keeping track of the set U^k of yet unassigned points. In contrast, using the step (5.136), a point may formally be assigned the same label multiple times.

Denote $\gamma' := (\gamma^1, ..., \gamma^{k-1})$ and $u^{\gamma'} := u_\gamma^{k-1}$. We apply induction on k: For $k \geqslant 1$,

$$\mathbb{E}_\gamma\langle u_\gamma^k, s\rangle \;=\; \mathbb{E}_{\gamma'}\frac{1}{l}\sum_{i=1}^{l}\int_0^1 \sum_{j=1}^{l} s_j \cdot \Big(e^i\chi_{\{u_i^{\gamma'} > \alpha\}} + u^{\gamma'}\chi_{\{u_i^{\gamma'} \leqslant \alpha\}}\Big)_j d\alpha \tag{5.137}$$

$$= \mathbb{E}_{\gamma'}\frac{1}{l}\sum_{i=1}^{l}\int_0^1 \Big(s_i \cdot \chi_{\{u_i^{\gamma'} > \alpha\}} + u^{\gamma'}\chi_{\{u_i^{\gamma'} \leqslant \alpha\}}\,\langle u^{\gamma'}, s\rangle\Big)\, d\alpha \tag{5.138}$$

$$= \mathbb{E}_{\gamma'}\frac{1}{l}\sum_{i=1}^{l}\int_0^1 \Big(s_i \cdot \chi_{\{u_i^{\gamma'} > \alpha\}} + \Big(1 - \chi_{\{u_i^{\gamma'} > \alpha\}}\Big)\langle u^{\gamma'}, s\rangle\Big)\, d\alpha. \tag{5.139}$$

Now we take into account the property [AFP00, Prop. 1.78], which is a direct consequence of Fubini's theorem, and also used in the proof of the thresholding theorem for the two-class case (Thm. 2.8):

$$\int_0^1 \int_\Omega s_i(x) \cdot \chi_{\{u_i > \alpha\}}(x) dx\, d\alpha = \int_\Omega s_i(x) u_i(x) d\, x = \langle u_i, s_i\rangle. \tag{5.140}$$

This leads to

$$\mathbb{E}_\gamma\langle u_\gamma^k, s\rangle \;=\; \mathbb{E}_{\gamma'}\frac{1}{l}\sum_{i=1}^{l}\Big(s_i u_i^{\gamma'} + \langle u^{\gamma'}, s\rangle - u_i^{\gamma'}\langle u^{\gamma'}, s\rangle\Big)d\alpha \tag{5.141}$$

$$\overset{u^{\gamma'}(x) \in \Delta_l}{=} \mathbb{E}_{\gamma'}\langle u^{\gamma'}, s\rangle \;=\; \mathbb{E}_\gamma\langle u_\gamma^{k-1}, s\rangle. \tag{5.142}$$

Since $\langle u^0, s\rangle = \langle u, s\rangle$, the assertion follows by induction. □

Remark 5.12. Prop. 5.11 shows that the data term is – in the mean – not affected by the probabilistic rounding process, i.e. it satisfies an *exact* coarea-like formula.

Bounding the regularizer is more involved: For $\gamma^k = (i^k, \alpha^k)$, define

$$\begin{aligned} U_{\gamma^k} &:= \{x \in \Omega | u_{i^k}(x) \leqslant \alpha^k\}, && (5.143)\\ V_{\gamma^k} &:= (U_{\gamma^k})^1, && (5.144)\\ V^k &:= (U^k)^1. && (5.145) \end{aligned}$$

As the measure-theoretic interior is invariant under $\mathcal{L}^d$-negligible modifications, given some fixed sequence γ the sequence (V^k) is invariant under $\mathcal{L}^d$-negligible modifications of $u = u^0$, i.e. it is uniquely defined when viewing u as an element of $L^1(\Omega)^l$. Some calculations yield

$$\begin{aligned} U^k &= U_{\gamma^1} \cap \ldots \cap U_{\gamma^k}, \quad k \geqslant 1, && (5.146)\\ U^{k-1} \setminus U^k &= U_{\gamma^1} \cap ((U_{\gamma^2} \cap \ldots \cap U_{\gamma^{k-1}}) \setminus (U_{\gamma^2} \cap \ldots \cap U_{\gamma^k})), \quad k \geqslant 2. && (5.147) \end{aligned}$$

From these observations and Prop. 5.7,

$$\begin{aligned} V^k &= V_{\gamma^1} \cap \ldots \cap V_{\gamma^k}, \quad k \geqslant 1, && (5.148)\\ V^{k-1} \setminus V^k &= V_{\gamma^1} \cap ((V_{\gamma^2} \cap \ldots \cap V_{\gamma^{k-1}}) \setminus (V_{\gamma^2} \cap \ldots \cap V_{\gamma^k})), \quad k \geqslant 2, && (5.149)\\ \Omega \setminus V^k &= \bigcup_{k'=1}^{k} \left(V^{k'-1} \setminus V^{k'}\right), \quad k \geqslant 1. && (5.150) \end{aligned}$$

Moreover, since V^k is the measure-theoretic interior of U^k, both sets are equal up to an $\mathcal{L}^d$-negligible set (cf. (5.120)).

We now prepare for an induction argument on the expectation of the regularizing term when restricted to the sets $V^{k-1} \setminus V^k$. The following proposition provides the initial step ($k = 1$).

Proposition 5.13. *Assume that $\Psi \colon \mathbb{R}^{d \times l} \to \mathbb{R}_{\geqslant 0}$ satisfies the lower and upper boundedness conditions (5.88) and (5.89). Then*

$$\mathbb{E} \int_{V^0 \setminus V^1} d\Psi(D\bar{u}) \leqslant \frac{2}{l} \frac{\lambda_u}{\lambda_l} \int_\Omega d\Psi(D\,u). \tag{5.151}$$

Proof. Denote $(i, \alpha) = \gamma^1$. Since $\chi_{U_{(i,\alpha)}} = \chi_{V_{(i,\alpha)}}$ $\mathcal{L}^d$-a.e., we have

$$\bar{u}_\gamma = \chi_{V_{(i,\alpha)}} e^i + \chi_{\Omega \setminus V_{(i,\alpha)}} \bar{u}_\gamma \quad \mathcal{L}^d\text{-a.e.} \tag{5.152}$$

Therefore, since $V^0 = (U^0)^1 = (\Omega)^1 = \Omega$,

$$\int_{V^0 \setminus V^1} d\Psi(D\bar{u}_\gamma) = \int_{\Omega \setminus V_{(i,\alpha)}} d\Psi(D\bar{u}_\gamma) = \int_{\Omega \setminus V_{(i,\alpha)}} d\Psi\big(D\big(\chi_{V_{(i,\alpha)}} e^i + \chi_{\Omega \setminus V_{(i,\alpha)}} \bar{u}_\gamma\big)\big).$$

Since $u \in \mathrm{BV}(\Omega)^l$, we know that $\mathrm{Per}(V_{(i,\alpha)}) < \infty$ holds for $\mathcal{L}^1$-a.e. α and any i [AFP00, Thm. 3.40]. Therefore we conclude from Prop. 5.8 that for $\mathcal{L}^1$-a.e. α,

$$\begin{aligned} \int_{\Omega \setminus V_{(i,\alpha)}} d\Psi(D\bar{u}_\gamma) &\leqslant \lambda_u \mathrm{Per}(V_{(i,\alpha)}) + \\ &\int_{(\Omega \setminus V_{(i,\alpha)}) \cap (\Omega \setminus V_{(i,\alpha)})^1} d\Psi(D\,e^i) + \int_{(\Omega \setminus V_{(i,\alpha)}) \cap (\Omega \setminus V_{(i,\alpha)})^0} d\Psi(D\bar{u}_\gamma). \end{aligned} \tag{5.153}$$

Both of the integrals are zero, since $De^i = 0$ and $(\Omega \setminus V_{(i,\alpha)})^0 = (V_{(i,\alpha)})^1 = V_{(i,\alpha)}$, therefore $\int_{\Omega\setminus V_{(i,\alpha)}} d\Psi(D\bar{u}_\gamma) \leqslant \lambda_u \operatorname{Per}(V_{(i,\alpha)})$. Carrying the bound over to the expectation yields

$$\mathbb{E}_\gamma \int_{\Omega\setminus V_{(i,\alpha)}} d\Psi(D\bar{u}_\gamma) \;\leqslant\; \frac{1}{l}\sum_{i=1}^{l} \int_0^1 \lambda_u \operatorname{Per}(V_{(i,\alpha)}) d\alpha. \tag{5.154}$$

Also, $\operatorname{Per}(V_{(i,\alpha)}) = \operatorname{Per}(U_{(i,\alpha)})$ since the perimeter is invariant under $\mathcal{L}^d$-negligible modifications. The assertion then follows using $V^0 = \Omega$, $V^1 = V_{(i,\alpha)}$ and the coarea formula:

$$\begin{aligned}
\mathbb{E}_\gamma \int_{V^0\setminus V^1} d\Psi(D\bar{u}_\gamma) \quad &\leqslant \quad \frac{1}{l}\sum_{i=1}^{l} \int_0^1 \lambda_u \operatorname{Per}(U_{(i,\alpha)}) d\alpha && (5.155)\\
&\overset{\text{coarea}}{=} \quad \frac{\lambda_u}{l}\sum_{i=1}^{l} \operatorname{TV}(u_i) = \frac{\lambda_u}{l}\int_\Omega \sum_{i=1}^{l} d\|D\,u_i\|_2 && (5.156)\\
&\overset{(5.88)}{\leqslant} \quad \frac{2}{l}\frac{\lambda_u}{\lambda_l}\int_\Omega d\Psi(D\,u). && (5.157)
\end{aligned}$$

□

We now take care of the induction step for the regularizer bound.

Proposition 5.14. *Let Ψ satisfy the upper boundedness (5.89). Then, for any $k \geqslant 2$,*

$$F := \mathbb{E}\int_{V^{k-1}\setminus V^k} d\Psi(D\bar{u}) \;\leqslant\; \frac{(l-1)}{l}\mathbb{E}\int_{V^{k-2}\setminus V^{k-1}} d\Psi(D\bar{u}). \tag{5.158}$$

Proof. Define the shifted sequence $\gamma' = (\gamma'^k)_{k=1}^\infty$ by $\gamma'^k := \gamma^{k+1}$, and let

$$W_{\gamma'} \;:=\; V_{\gamma'}^{k-2} \setminus V_{\gamma'}^{k-1} = (V_{\gamma^2} \cap \ldots \cap V_{\gamma^{k-1}}) \setminus (V_{\gamma^2} \cap \ldots \cap V_{\gamma^k}). \tag{5.159}$$

By Prop. 5.5 we may assume that, under the expectation, $\bar{u}_\gamma$ exists and is an element of $\operatorname{BV}(\Omega)^l$. We denote $\gamma^1 = (i,\alpha)$, then $V^{k-1}\setminus V^k = V_{(i,\alpha)} \cap W_{\gamma'}$ due to (5.149). For each pair (i,α) we denote by $((i,\alpha),\gamma')$ the sequence obtained by prepending (i,α) to the sequence γ'. Then

$$F \;=\; \frac{1}{l}\sum_{i=1}^{l} \int_0^1 \left(\mathbb{E}_{\gamma'} \int_{V_{(i,\alpha)}\cap W_{\gamma'}} d\Psi(D\bar{u}_{((i,\alpha),\gamma')}) \right) d\alpha. \tag{5.160}$$

Since in the first iteration of the algorithm no points in $U_{(i,\alpha)}$ are assigned a label, $\bar{u}_{((i,\alpha),\gamma')} = \bar{u}_{\gamma'}$ holds on $U_{(i,\alpha)}$, and therefore $\mathcal{L}^d$-a.e. on $V_{(i,\alpha)}$. Therefore we may apply Prop. 5.9 and substitute $D\bar{u}_{((i,\alpha),\gamma')}$ by $D\bar{u}_{\gamma'}$ in (5.160):

$$\begin{aligned}
F \;&=\; \frac{1}{l}\sum_{i=1}^{l} \int_0^1 \left(\mathbb{E}_{\gamma'} \int_{V_{(i,\alpha)}\cap W_{\gamma'}} d\Psi(D\bar{u}_{\gamma'}) \right) d\alpha && (5.161)\\
&=\; \frac{1}{l}\sum_{i=1}^{l} \int_0^1 \left(\mathbb{E}_{\gamma'} \int_{W_{\gamma'}} \chi_{V_{(i,\alpha)}}\, d\Psi(D\bar{u}_{\gamma'}) \right) d\alpha. && (5.162)
\end{aligned}$$

By definition of the measure-theoretic interior (Def. A.13), $\chi_{V_{(i,\alpha)}}$ is bounded from above by the density function $\Theta_{U_{(i,\alpha)}}$ of $U_{(i,\alpha)}$,

$$\chi_{V_{(i,\alpha)}}(x) \;\leqslant\; \Theta_{(i,\alpha)}(x) := \lim_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(x) \cap U_{(i,\alpha)}|}{|\mathcal{B}_\delta(x)|}, \tag{5.163}$$

which exists $\mathcal{H}^{d-1}$-a.e. on Ω by [AFP00, Prop. 3.61] (Thm. A.15). Therefore, denoting by $\mathcal{B}_\delta(\cdot)$ the mapping $x \in \Omega \mapsto \mathcal{B}_\delta(x)$,

$$F \;\leqslant\; \frac{1}{l} \sum_{i=1}^{l} \int_0^1 \left(\mathbb{E}_{\gamma'} \int_{W_{\gamma'}} \left(\lim_{\delta \searrow 0} \frac{|\mathcal{B}_\delta(\,\cdot\,) \cap U_{(i,\alpha)}|}{|\mathcal{B}_\delta(\,\cdot\,)|} \right) d\Psi(D\bar{u}_{\gamma'}) \right) d\alpha. \tag{5.164}$$

Rearranging the integrals and the limit, which can be justified by dominated convergence using (5.89) and $\mathrm{TV}(\bar{u}_{\gamma'}) < \infty$ almost surely, we get

$$F \;\leqslant\; \frac{1}{l} \mathbb{E}_{\gamma'} \lim_{\delta \searrow 0} \int_{W_{\gamma'}} \left(\sum_{i=1}^{l} \int_0^1 \left(\frac{|\mathcal{B}_\delta(\,\cdot\,) \cap U_{(i,\alpha)}|}{|\mathcal{B}_\delta(\,\cdot\,)|} \right) d\alpha \right) d\Psi(D\bar{u}_{\gamma'}) \tag{5.165}$$

$$= \;\frac{1}{l} \mathbb{E}_{\gamma'} \lim_{\delta \searrow 0} \int_{W_{\gamma'}} \frac{1}{|\mathcal{B}_\delta(\,\cdot\,)|} \left(\sum_{i=1}^{l} \int_0^1 \int_{\mathcal{B}_\delta(\cdot)} \chi_{\{u_i(y) \leqslant \alpha\}} d\,y\, d\alpha \right) d\Psi(D\bar{u}_{\gamma'}). \tag{5.166}$$

We again apply [AFP00, Prop. 1.78] to the two innermost integrals (alternatively, use Fubini's theorem), which leads to

$$F \;\leqslant\; \frac{1}{l} \mathbb{E}_{\gamma'} \lim_{\delta \searrow 0} \int_{W_{\gamma'}} \frac{1}{|\mathcal{B}_\delta(\,\cdot\,)|} \left(\sum_{i=1}^{l} \int_{\mathcal{B}_\delta(\cdot)} (1 - u_i(y)) d\,y \right) d\Psi(D\bar{u}_{\gamma'}). \tag{5.167}$$

Using the fact that $u(y) \in \Delta_l$, it turns out that

$$F \;\leqslant\; \frac{1}{l} \mathbb{E}_{\gamma'} \lim_{\delta \searrow 0} \int_{W_{\gamma'}} \frac{1}{|\mathcal{B}_\delta(\,\cdot\,)|} \left(\int_{\mathcal{B}_\delta(\cdot)} (l-1)\, d\,y \right) d\Psi(D\bar{u}_{\gamma'}) \tag{5.168}$$

$$= \;\frac{1}{l} \mathbb{E}_{\gamma'} \lim_{\delta \searrow 0} \int_{W_{\gamma'}} (l-1)\, d\Psi(D\bar{u}_{\gamma'}) \tag{5.169}$$

$$= \;\frac{l-1}{l} \mathbb{E}_{\gamma'} \int_{W_{\gamma'}} d\Psi(D\bar{u}_{\gamma'}) \tag{5.170}$$

$$= \;\frac{l-1}{l} \mathbb{E}_{\gamma'} \int_{V_{\gamma'}^{k-2} \setminus V_{\gamma'}^{k-1}} d\Psi(D\bar{u}_{\gamma'}). \tag{5.171}$$

Reverting the index shift and using the fact that $\bar{u}_{\gamma'} = \bar{u}_\gamma$ concludes the proof:

$$F \;\leqslant\; \frac{l-1}{l} \mathbb{E}_{\gamma} \int_{V_{\gamma}^{k-1} \setminus V_{\gamma}^{k}} d\Psi(D\bar{u}_{\gamma}). \tag{5.172}$$

□

The following theorem is the main result of this work, and provides an approximate coarea formula in the sense of (5.41).

Theorem 5.15. *Let $s \in L^\infty(\Omega)^l$, $s \geqslant 0$, $\Psi\colon \mathbb{R}^{d\times l} \to \mathbb{R}_{\geqslant 0}$ 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 (5.88) and (5.89) are satisfied. Then Alg. 5.2. generates an integral labeling $\bar{u} \in \mathcal{C}_\mathcal{E}$ almost surely, and*

$$\mathbb{E} f(\bar{u}) \;\leqslant\; 2\frac{\lambda_u}{\lambda_l} f(u). \tag{5.173}$$

Proof. The fact that the algorithm provides $\bar{u} \in \mathcal{C}_\mathcal{E}$ almost surely follows from Thm. 5.5. Therefore there almost surely exists $k' := k'(\gamma) \geqslant 1$ such that $U^{k'} = \emptyset$ and $\bar{u}_\gamma = u_\gamma^{k'}$. On one hand, this implies

$$\langle \bar{u}_\gamma, s\rangle = \langle u_\gamma^{k'}, s\rangle = \lim_{k\to\infty} \langle u_\gamma^k, s\rangle \tag{5.174}$$

almost surely. On the other hand, we have $V^{k'} = (U^{k'})^1 = \emptyset$ and therefore

$$\bigcup_{k=1}^{k'} V^{k-1} \setminus V^k \;\overset{(*)}{=}\; \Omega \setminus V^{k'} = \Omega \tag{5.175}$$

almost surely. The equality $(*)$ can be shown by induction: For the base case $k' = 1$, we have $V^0 = (U^0)^1 = (\Omega)^1 = \Omega$, since Ω is the open unit box. For $k' \geqslant 2$,

$$\begin{aligned}
\bigcup_{k=1}^{k'} V^{k-1} \setminus V^k \quad &= \quad \big(V^{k'-1} \setminus V^{k'}\big) \cup \bigcup_{k=1}^{k'-1} \big(V^{k-1} \setminus V^k\big) && (5.176)\\
&= \quad \big(V^{k'-1} \setminus V^{k'}\big) \cup \big(\Omega \setminus V^{k'-1}\big) && (5.177)\\
&\overset{V^{k'-1}\subseteq\Omega}{=} \; \Omega \setminus V^{k'-1}. && (5.178)
\end{aligned}$$

almost surely (cf. (5.150)). From (5.174) and (5.175) we obtain

$$\begin{aligned}
\mathbb{E}_\gamma f(\bar{u}_\gamma) \;&=\; \mathbb{E}_\gamma\Big(\lim_{k\to\infty} \langle u_\gamma^k, s\rangle\Big) + \mathbb{E}_\gamma\Bigg(\sum_{k=1}^{\infty} \int_{V^{k-1}\setminus V^k} d\Psi(D\bar{u}_\gamma)\Bigg) && (5.179)\\
&=\; \lim_{k\to\infty} \big(\mathbb{E}_\gamma \langle u_\gamma^k, s\rangle\big) + \sum_{k=1}^{\infty} \mathbb{E}_\gamma \int_{V^{k-1}\setminus V^k} d\Psi(D\bar{u}_\gamma) && (5.180)
\end{aligned}$$

The first term in (5.180) is equal to $\langle u, s\rangle$ due to Prop. 5.11. An induction argument using Prop. 5.13 and Prop. 5.14 shows that the second term can be bounded according to

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}_\gamma \int_{V^{k-1}\setminus V^k} \Psi(D\,\bar{u}_\gamma) \;&\leqslant\; \sum_{k=1}^{\infty} \left(\frac{l-1}{l}\right)^{k-1} \frac{2}{l} \frac{\lambda_u}{\lambda_l} \int_\Omega d\Psi(D\,u) && (5.181)\\
&=\; 2\frac{\lambda_u}{\lambda_l} \int_\Omega d\Psi(D\,u)\,, && (5.182)
\end{aligned}$$

therefore

$$\mathbb{E}_\gamma f(\bar{u}_\gamma) \;\leqslant\; \langle u, s\rangle + 2\frac{\lambda_u}{\lambda_l} \int_\Omega d\Psi(D\,u)\,. \tag{5.183}$$

Since $s \geqslant 0$ and $\lambda_u \geqslant \lambda_l$, and therefore $\langle u, s\rangle \leqslant 2(\lambda_u/\lambda_l)\langle u, s\rangle$, this proves the assertion. Swapping the integral and limits in (5.180) can be justified retrospectively by the dominated convergence theorem, using $0 \leqslant \langle u, s\rangle \leqslant \infty$ and Prop. 5.6. □

Corollary 5.16. *Under the conditions of Thm. 5.15, if u^* minimizes f over $\mathcal{C}$, $u^*_{\mathcal{E}}$ minimizes f over $\mathcal{C}_{\mathcal{E}}$, and $\bar{u}^*$ is the output of Alg. 5.2 when applied to u^*, then*

$$\mathbb{E} f(\bar{u}^*) \;\leqslant\; 2\frac{\lambda_u}{\lambda_l} f(u^*_{\mathcal{E}}). \tag{5.184}$$

Proof. This follows immediately from Thm. 5.15, since $\mathcal{C}_{\mathcal{E}} \subseteq \mathcal{C}$ implies $f(u^*) \leqslant f(u^*_{\mathcal{E}})$, cf. (5.42). □

We have demonstrated that the proposed approach allows to recover, from the solution u^* of the convex *relaxed* problem (5.2), an approximate *integral* solution $\bar{u}^*$ of the nonconvex *original* problem (5.1) with an upper bound on the objective.

For the metric embedding regularizer we obtain the following, less tight bounds. For simplicity, we assume that $A \in \mathbb{R}^{l\times l}$ is regular.

Proposition 5.17. *Assume that $A = (a^1|...|a^l) \in \mathbb{R}^{l\times l}$ is regular and let $\Psi = \Psi_A$. Then the definition*

$$\lambda_l = \frac{2}{\sqrt{l}\|A^{-1}\|} \quad \textit{and} \quad \lambda_u = \max_{i,j\in\{1,...,l\}} \|a^i - a^j\|_2, \tag{5.185}$$

where $\|A^{-1}\|$ denotes the operator norm of A^{-1} with respect to $\|\cdot\|_2$, fulfills the lower- and upper-boundedness conditions (5.88) and (5.89).

Proof. For the lower-boundedness condition (5.88), we compute:

$$\begin{aligned}
\frac{\lambda_l}{2}\sum_{i=1}^{l}\|z^i\|_2 &= \frac{1}{\sqrt{l}\,\|A^{-1}\|}\sum_{i=1}^{l}\|z^i\|_2 && (5.186)\\
&\leqslant \frac{1}{\sqrt{l}\,\|A^{-1}\|}\sqrt{l}\,\|z\|_2 && (5.187)\\
&= \|A^{-1}\|^{-1}\,\|z\,A^\top (A^\top)^{-1}\|_2 && (5.188)\\
&\overset{(*)}{\leqslant} \|A^{-1}\|^{-1}\,\|z\,A^\top\|_2\|A^{-1}\| && (5.189)\\
&= \|z\,A^\top\|_2 = \Psi_A(z). && (5.190)
\end{aligned}$$

The inequality $(*)$ relies on the fact that $\|MN\|_2 \leqslant \|M\|\,\|N\|_2$ for compatible matrices M, N, as can be seen from the definitions of the norms. The upper-boundedness follows directly from (5.89) and the definition of Ψ_A. □

The above proposition results in an optimality bound factor for Ψ_A depending on the condition of A:

$$2\,\frac{\lambda_u}{\lambda_l} \;=\; \sqrt{l}\,\|A^{-1}\| \max_{i,j\in\{1,...,l\}}\|a^i - a^j\|_2 \leqslant \sqrt{2\,l}\,\|A^{-1}\|\,\|A\|. \tag{5.191}$$

Note however that this estimate is rather loose, in particular it does not take the constraint on z in (5.88) into account. The corresponding result for the envelope regularizer $\Psi = \Psi_d$ is much tighter and more elegant:

Proposition 5.18. *Let $d\colon \mathcal{E}^2 \to \mathbb{R}_{\geqslant 0}$ be a metric and $\Psi = \Psi_d$. Then the definition*

$$\lambda_l = \min_{i \neq j} d(i,j) \quad \textit{and} \quad \lambda_u = \max_{i,j \in \{1,\ldots,l\}} d(i,j) \tag{5.192}$$

fulfills the lower- and upper-boundedness conditions (5.88) and (5.89).

Proof. From Prop. 2.6 we obtain, for any $y \in \mathbb{R}^d$ with $\|y\|_2 = 1$,

$$\Psi_d(y\,(e^i - e^j)) \;=\; d(i,j), \tag{5.193}$$

which shows the upper bound (5.89). For the lower bound (5.88), set

$$c := \min_{i \neq j} d(i,j), \quad v'^i := \frac{c}{2}\frac{w^i}{\|w^i\|_2}, \quad \text{and} \quad v := v'\,(I - \frac{1}{l}\,e\,e^\top)\,. \tag{5.194}$$

Then $\|v^i - v^j\|_2 = \|v'^i - v'^j\|_2 \leqslant c$ and $v\,e = v'\,(I - \frac{1}{l}\,e\,e^\top)\,e = 0$. Therefore $v \in \mathcal{D}^d_{\mathrm{loc}}$, which implies, for $w \in \mathbb{R}^{d \times l}$ satisfying $w\,e = 0$,

$$\Psi_d(w) \;\geqslant\; \langle w, v\rangle = \langle w, v'\rangle = \sum_{i=1}^{l} \langle w^i, \frac{c}{2}\frac{w^i}{\|w^i\|_2}\rangle) = \frac{c}{2}\sum_{i=1}^{l} \|w^i\|_2, \tag{5.195}$$

proving the lower bound. □

Finally, for Ψ_d we obtain the optimality bound factor

$$2\,\frac{\lambda_u}{\lambda_l} \quad = \quad 2\,\frac{\max_{i,j} d(i,j)}{\min_{i \neq j} d(i,j)}, \tag{5.196}$$

which is exactly the same as has been proven for the finite-dimensional metric labeling [KT99] and α-expansion [BVZ01] methods. The above considerations extend these results to problems on continuous domains for a broad class of regularizers.

5.6 Experimental Comparison

Although the main purpose of Alg. 5.2 is to provide a basis for deriving the bound in Thm. 5.15, we will briefly point out some of its empirical characteristics. It is important to keep in mind that for the *discretized* problem, analog bounds to those provided by Thm. 5.15 are only valid if the discretization respects the coarea formula. However, for these energies the original finite-dimensional proof [KT99] already applies. For the finite-differences discretization, a comparison of the *a posteriori* bounds computed via the primal-dual gap must be taken with a grain of salt, 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 is it not only caused by the discretization, and that the underlying integral solution may be suboptimal due to the relaxation. With this in mind, the

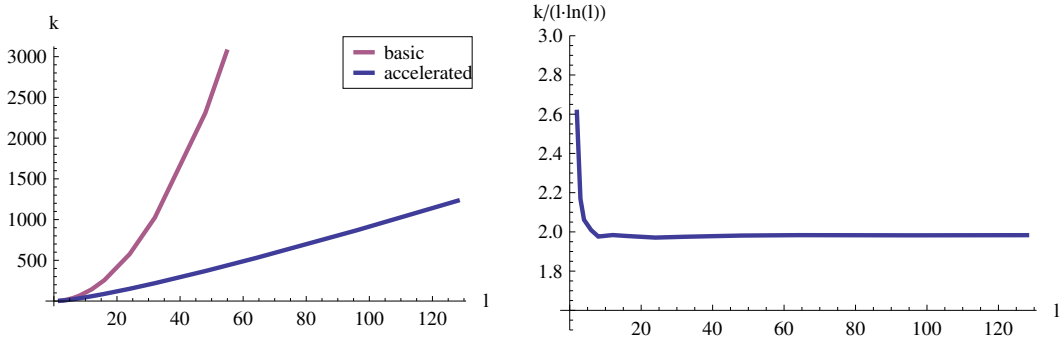

Figure 5.1. Left: Label count l vs. mean number of iterations k of the probabilistic rounding algorithm. The improved sampling of α^k greatly accelerates the method. **Right:** Empirically, $k \approx 2l \ln(l)$ for the accelerated method. As a result, the total runtime is comparable to the deterministic rounding methods for a moderate number of labels.

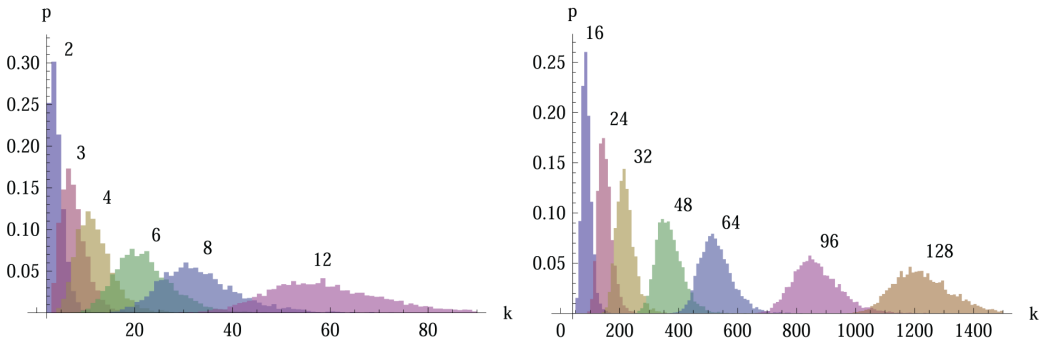

Figure 5.2. Histogram (probability density scale) of the number of required iterations k; sampled over 5000 runs for $2-128$ labels.

observations in the following subsections should be seen only as indicators of what results can be expected qualitatively.

Expected Number of Iterations. In practice, choosing $\alpha^k \in [0, 1]$ leads to an unnecessary large number of iterations, as no point is assigned a label in iteration k unless $\alpha^k < c_{i^k}^{k-1}$. The method can be accelerated without affecting the derived bounds by choosing $\alpha^k \in [0, c_{i^k}^{k-1}]$ instead, thereby skipping the redundant iterations.

Fig. 5.1 shows the mean number of iterations k until the condition $e^\top c^k < 1$ was satisfied, sampled over 5000 runs per label count; see Fig. 5.2 for the corresponding histograms. From the proof of Thm. 5.5 it can be seen that this provides a worst-case upper bound for the expected number of iterations until the algorithm can be stopped and $\bar{u}_\gamma$ is obtained. For the accelerated method, k is almost perfectly proportional to $l \ln(l)$; we conjecture that asymptotically $k = 2l \ln(l)$.

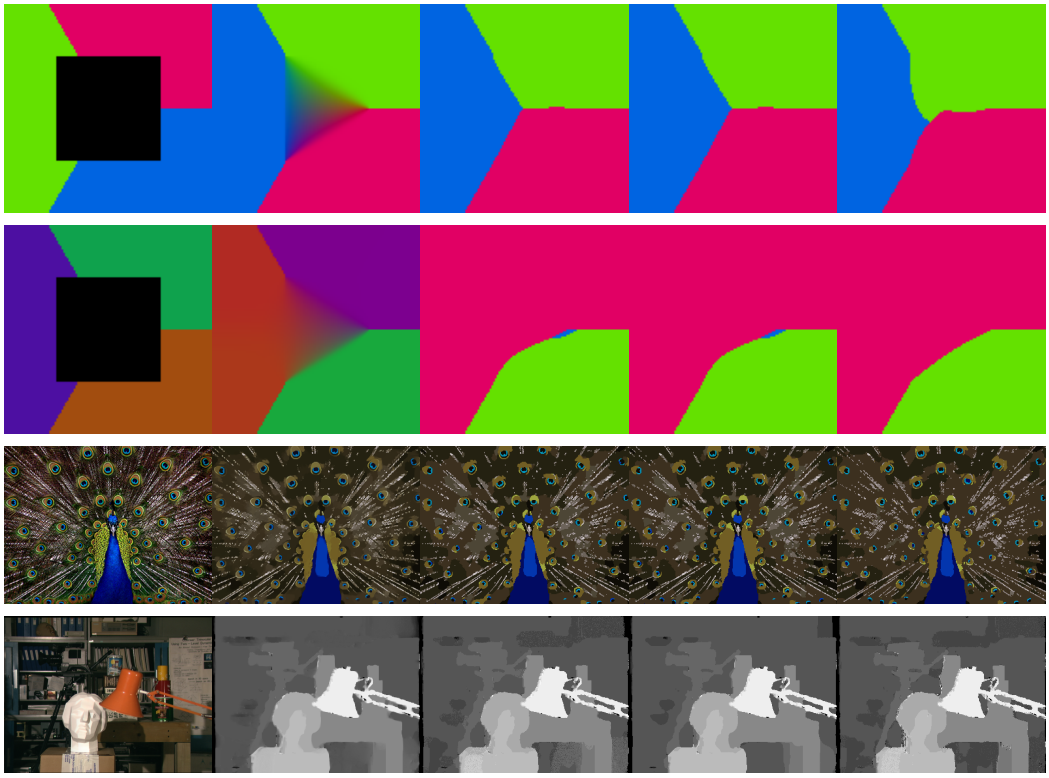

Figure 5.3. Top to bottom: Problems 2,3,8,11 of the test set. **Left to right:** Original input; relaxed solution; integral solutions obtained by deterministic "first-max" and "modified" rounding (Sect 5.3); result of the probabilistic rounding. In specially crafted situations, the probabilistic method may perform slightly worse (first row) or better (second row) than the deterministic approaches. On real-world data, results are very similar (rows 3-4). In contrast to the deterministic approaches, the probabilistic method provides true *a priori* optimality bounds.

A Priori and A Posteriori Bounds. In order to evaluate the tightness of the bound (5.184) in Thm. 5.15 in practice, we selected 12 prototypical multiclass labeling problems with $3-64$ labels each. For each we computed the relaxed solution u^* and the mean as well as the best objective of the rounded solution $\bar{u}^*$ during 10000 iterations of Alg. 5.2, see Fig. 5.3 for some exemplary results.

The primal-dual optimization approach provides an (approximate) *a posteriori* bound ε' as outlined in Sect. 4.3, in contrast to the theoretical *a priori* upper bound $\varepsilon = 2\lambda_u/\lambda_l - 1$ derived from Cor. 5.16. In practice, the *a posteriori* bound stayed well below the theoretical bound (Table 5.1), which is consistent with the good practical performance of the α-expansion method that has a similar *a priori* bound.

However, the experiments are based on the metric embedding approach, since otherwise the *a posteriori* gap could not be accurately computed. Therefore the regularizer bound is quite loose compared to what can be expected for the envelope approach. 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 Chap. 3. However, a large energy increase indicates that it 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.

problem	1	2	3	4	5	6	7	8	9	10	11	12
N	76800	14400	14400	129240	76800	86400	86400	76800	86400	76800	110592	21838
l	3	3	3	4	8	12	12	12	12	12	16	64
k	7.1	6.9	5.0	11.0	27.2	47.5	47.0	43.6	46.5	46.0	70.7	335.0
a priori ε	1.45	1.45	1.45	1.83	3	3.90	3.90	3.90	3.90	3.90	4.90e5	4.79e6
a posteriori												
- first-max	0.0008	0.0098	0.0083	0.0038	0.0266	0.0277	0.0277	0.1051	0.0666	0.0515	0.7330	0.0943
- modified	0.0008	0.0098	0.0083	0.0038	0.0266	0.0277	0.0277	0.1051	0.0666	0.0515	0.1424	0.0275
- prob. best	0.0008	0.0102	0.0048	0.0038	0.0282	0.0312	0.0312	0.1228	0.0812	0.0600	0.5888	0.1684
- prob. mean	0.0014	0.0186	0.0102	0.0106	0.0510	0.0591	0.0722	0.2140	0.1382	0.1173	1.4072	0.2772

Table 5.1. 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 bound predicted by Thm. 5.15. Problems $1-10$ are color segmentation/inpainting problems with $\Psi = \|\cdot\|_2$. The depth-from-stereo and inpainting problems 11 and 12 use an approximated truncated-linear metric (Fig. 2.7). For these nonstandard distances, the modified deterministic rounding method provides much better results than the other methods.

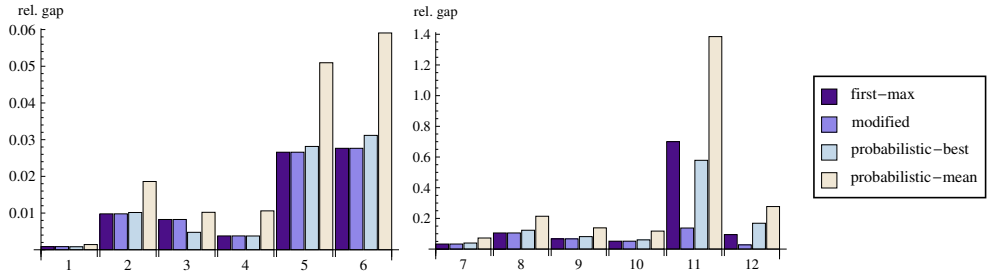

Figure 5.4. Relative gap (*a posteriori* bound) ε' of the rounded solution for the test problems 1–12 using deterministic "first-max" and "modified" rounding, and best and mean gap obtained using the proposed probabilistic method. While the energy increase through probabilistic rounding is usually slightly larger than for the deterministic methods, it is well below the *a priori* bound of $\varepsilon = 2\lambda_u/\lambda_l - 1$ derived in Cor. 5.16 (Table 5.1).

Deterministic and Probabilistic Methods. Compared to the two deterministic rounding methods, Alg. 5.2 usually leads to a slightly larger energy increase (Fig. 5.4). For problems 11 and 12, where λ_u/λ_l is large, the solution is clearly inferior to the one obtained using the "modified" rounding. This can be attributed to the fact that the latter takes into account the detailed structure of Ψ, which is neither required nor used in order to obtain the bounds in Thm. 5.15.

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. 5.3), 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 [KT99].

We would like to emphasize that the purpose of these experiments is not to demonstrate a practical superiority of the probabilistic method compared to other techniques, but rather to provide an illustration on what bounds can be *expected in practice* compared to the *a priori* bounds in Thm. 5.15.

In fact, the results in Table 5.1 show that for problems with nonstandard regularizers, the improved deterministic rounding technique from Sect. 5.3.2 consistently provided better bounds than the other methods. We observed that often this also

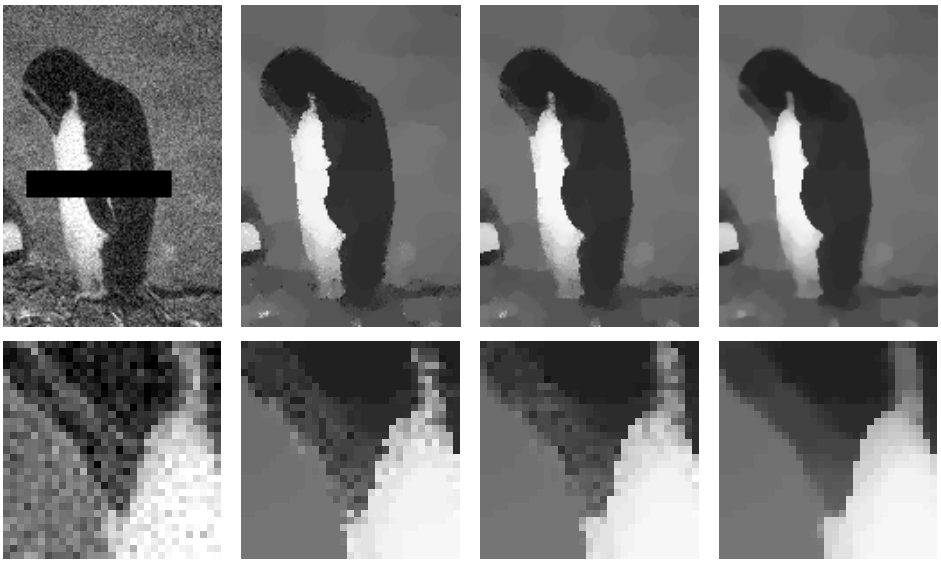

Figure 5.5. Improved deterministic rounding on the "penguin" denoising/inpainting problem (problem 12 in Table 5.1) with 64 classes. **Left to right:** Noisy input image with inpainting region marked black [SZS+06]; result with randomized rounding; result first-max rounding; result with the improved rounding scheme (5.30). The randomized and first-max method introduce noticeable noise in the rounding step. The improved deterministic method takes the non-uniformity of the used "truncated-linear" potential (Fig. 2.7) explicitly into account, resulting in a clean labeling and an *a posteriori* gap of only 2.75% vs. 9.4% for the first-max method and 16.8% for the probabilistic method.

translates to an improved visual quality, in particular for regularizers where λ_u/λ_l is large. An example where the difference is clearly visible is the "penguin" inpainting problem from Sect. 4.5.4. As opposed to the first-max scheme, the improved scheme generates considerably less noise, and an *a posteriori* optimality of $\varepsilon' = 0.0275$ compared to $\varepsilon' = 0.0943$ for the first-max approach and $\varepsilon' = 0.1684$ for the probabilistic method (Fig. 5.5).

5.7 Summary and Further Work

In this chapter we presented deterministic and probabilistic rounding methods for recovering approximate solutions of multiclass labeling or image partitioning problems from solutions of convex relaxations in the spatially continuous framework.

We provided an improved deterministic rounding technique, which – while it is a heuristic and provides only *a posteriori* bounds – considerably improves the results for non-standard potentials. In order to derive true *a priori* bounds, we presented a probabilistic approach. To our knowledge, this is the first fully convex approach that is both formulated in the spatially continuous setting and provides an *a priori* bound on the optimality of the generated integral solution. We showed that the approach can also be interpreted as an approximate variant of the coarea formula. Numerical experiments confirm the theoretical bounds.

Future work may include extending the results to non-homogeneous regularizers and improving the tightness of the bound. In particular, the *a priori* bounds could be improved by adapting further arguments from [KT99]: 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 [Bar98, 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 [KT99] provide a derandomized algorithm with $\varepsilon = 1 + 4/(r-2)$. By construction, tree metrics can be isometrically embedded into ℓ^1, which also yields a connection to the embedding technique from Sect. 2.5.2.

A probabilistic result [Bar98, Thm. 9], later derandomized in [CCG+98], shows that for any metric d, an r-hierarchically well-separated tree metric d_r approximation can be constructed such that

$$d(i,j) \leqslant d_r(i,j) \leqslant \alpha\, d(i,j), \tag{5.197}$$

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 [FRT04, 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. On a larger scale, the connection to the recent lifting/relaxation techniques for solving nonconvex variational problems (Sect. 2.7.3) should be further explored.

Chapter 6
Sparse Representation of Shape

6.1 Introduction and Overview

In the previous chapters we mainly considered regularizers with length-based terms. While they can be thoroughly analyzed theoretically, their flexibility is limited: knowledge about the specific *shape* of objects – i.e. about the appearance of the interface separating the class regions – cannot be easily introduced. In this chapter, we present a variational approach to image segmentation that implicitly takes into account prior knowledge about the specific shape in terms of a dictionary of shape templates.

The results in this chapter should be seen less an in-depth analysis, but rather as an extended outlook on possible further enhancements for introducing higher-order knowledge, departing from the length-based regularizers in the previous chapters.

Within this chapter we exclusively consider the two-class case, where the task is to segment a given image into foreground and background. We propose to represent the foreground region as the *union* of a small set of templates. This is motivated by the observation that complex real-world objects are often composed of a relatively small number of simpler geometrical shapes, such as boxes and roughly ellipsoidal shapes. The observed shape is then dictated by the union of the regions for the individual parts.

To further motivate our approach, we briefly outline the idea of the basis pursuit/sparse representation framework [CDS01], which recently has been very successful. Basis pursuit problems are generally formulated on finite-dimensional spaces, i.e. $I \in \mathbb{R}^n$ (equivalently, $I : \Omega^h \to \mathbb{R}$) represents a grayscale image with n pixels.

Basis pursuit methods are characterized by the assumption that such images – or more generally signals – I are additively composed of a small number of basis functions drawn from an overcomplete basis of $K \gg n$ vectors, $A = (a^1|...|a^K) \in \mathbb{R}^{n \times K}$. Precisely,

$$I = A\,w, \quad w \in \mathbb{R}^K, \tag{6.1}$$

for some *sparse* vector w, i.e. w contains only relatively few nonzero entries. Assuming that there is at least one representation (6.1) for a given I, finding the *sparsest* basis representation can then be posed as an optimization problem,

$$\min_{w \in \mathbb{R}^K} \|w\|_0 \text{ subject to } A\,w = I, \tag{6.2}$$

where $\|w\|_0$ refers to the ℓ^0 pseudo-norm, i.e. the number of non-zero entries of w.

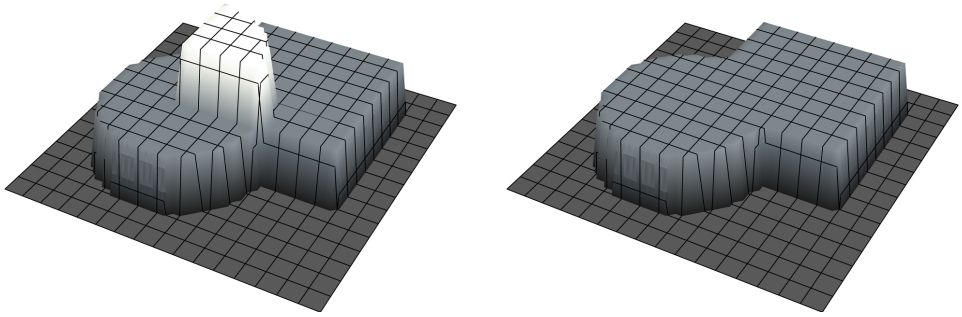

Figure 6.1. Illustration of the "union of basis shapes" principle. **Left:** In the classical sparse representation framework, images are assumed to be *additively* composed of a small number of basis functions. **Right:** In order to properly represent shapes as the *union* of a set of basis shapes, the additivity assumption is inappropriate: the indicator function for the union of the basis shapes is given by the *maximum* of the individual indicator functions, rather than their sum. However, this behavior can be approximated using a convex relaxation approach.

While solving (6.2) directly is a difficult combinatorial problem unless the matrix A comprises an orthogonal basis, it can be shown that under some circumstances one may replace $\|\cdot\|_0$ by $\|\cdot\|_1$ [CRTV05]. Moreover, in order to account for noise in the image I, the constraint is in practice usually enforced approximately by a penalty term with weight $\mu > 0$, resulting in the well-known *convex* problem

$$\min_{w \in \mathbb{R}^K} \{\mu \|w\|_1 + \|A\,w - I\|_2^2\}. \tag{6.3}$$

In the following, we will examine how this approach can be extended to sparse shape representation. The basic idea is to transfer the sparse representation method from the *image* domain to the *shape/segmentation* domain via the *characteristic function* representation, as done in the previous chapters: A segmentation of the image is encoded as a vector $u \in \{0,1\}^n$, with 1 representing the foreground – on which the shape prior should be applied – and 0 representing the background.

Consequently, we assume that the basis functions a^i are characteristic functions of some prototypical shapes, i.e. $a^i \in \{0,\,1\}^n$, and we replace the image I with the characteristic function of some local, and therefore noisy, segmentation $u \in \{0,1\}^n$.

From these definitions it becomes clear that directly using (6.3) to recover the basis shapes is bound to fail: Current sparse representation methods are based on the assumption that the basis functions overlay in an *additive* fashion. In contrast, in the shape context the basis functions are in a sense *opaque*, since the characteristic function of a union of sets is not the *sum* of the individual characteristic functions, but rather their pointwise *maximum* (Fig. 6.1).

We therefore propose to replace the additivity assumption (6.1) by the concept

$$u_i = \max\{(a^1)_i\, w_1, \ldots, (a^K)_i\, w_K\}, \quad w \in \{0,1\}^K. \tag{6.4}$$

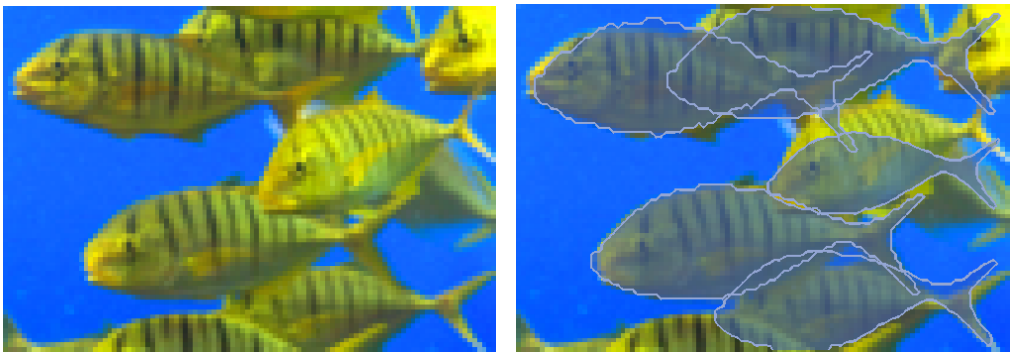

Figure 6.2. Separating fishes from the background and from each other by convex optimization, using a sparse covering of the image by shape templates. The dictionary of shape templates was generated from a single fish template by translation, rotation and scaling. The approach copes with a significant amount of overlapping templates and occlusion.

In order to concisely represent the problem, we denote the elementwise (Hadamard) product between two equally-sized vectors or matrices $A, B \in \mathbb{R}^{p\times q}$ by $A \odot B \in \mathbb{R}^{p\times q}$. Additionally, we define $A \odot x := A \odot (e\, x^\top)$ and $x \odot A := (x\, e^\top) \odot A$. For the matrix $A \in \mathbb{R}^{p\times q}$ (possibly a vector if $p = 1$), we denote by $\mathrm{vecmax}(A)$ the row-wise maximum, i.e. the vector $v \in \mathbb{R}^p$ such that $v_k = \max\{A_{k,1}, \ldots, A_{k,q}\}$.

Using these definitions, problem (6.4) admits the concise representation

$$u = \mathrm{vecmax}(A \odot w). \tag{6.5}$$

The overall objective derived from (6.3) then reads

$$\min_{w\in\{0,1\}^K} \{\mu\|w\|_1 + \|\mathrm{vecmax}(A \odot w) - u\|_2^2\}. \tag{6.6}$$

In contrast to the sparse representation objective (6.3), this problem is no longer convex, even if the constraint set is relaxed to $w \in [0,1]^K$. However, as shown below, it turns out that by switching to an appropriate convex relaxation, good solutions can be obtained by finding the global optimum of a *convex* problem. For an illustration, see Fig. 6.2.

6.2 Related Work

Variational approaches to image segmentation that utilize shape prior knowledge include statistical models of parametrized contours [CWS02, CKS03], level-set based segmentation [CSS06, CRD07, CS05], and discrete combinatorial approaches in terms of Markov Random Field (MRF) models [KRBT08, BKSS10]. A common property of these approaches is the inherent *nonconvexity* introduced by the respective shape

prior model. Instead, and in view of the models discussed in the previous chapters, we focus on a *convex* variational approach to foreground/background separation based on shape prior knowledge.

Our work is also motivated by the basis pursuit framework [CDS01] and the striking performance of "ℓ^1-decoding" by convex programming in underdetermined compressed sensing scenarios [CRTV05, CDD09]. We depart from these linear models by introducing non-additivity. We use shape dictionaries generated from a few basis templates by translation, rotation and scaling, that generally do not satisfy the strong mathematical properties that rigorously justify replacing the nonconvex problem (6.2) by the relaxation (6.3). However, numerical experiments reveal promising performance and a significant potential for real-world applications.

Related work in the field of computer vision includes the work of Borenstein and Ullman on segmentation using image fragments [BU02, BU08]. Unlike shape templates, image fragments model not only shape but also the image intensity function and image features for a particular object class. Accordingly, a strong focus lies on corresponding image features. The variational inference process is based on a Markov Random Field (MRF) model, and simplifications are made to keep it computationally tractable.

In contrast, we focus on the variational model from the optimization point of view and largely ignore the issue of feature extraction. This is in line with the models discussed in the previous chapters, which do not impose any restrictions on the choice of the (local) feature vector.

Finally, we also refer to [AEB06, YSM10] for recent work on dictionary-based image processing. Though the scope of their work is confined to shape denoising without a semantical interpretation such as image segmentation and object recognition, recent extensions towards *learning* of task-specific dictionaries [RZE10] constitute a highly relevant research direction. We present some numerical experiments indicating that adopting the dictionary learning idea for shape templates is promising indeed.

Organization. In this chapter we investigate the model (6.6) for the sparse representation of shapes:

- We derive a sufficiently tight convex relaxation, leading to a variational approach that is amenable to large-scale convex optimization (Sect. 6.3). The relaxation is based on a similar principle as the "local envelope" approach in Sect. 2.5.1.

- We demonstrate how the optimization methods from Chap. 4 can be applied to the relaxed problem (Sect. 6.4).

- We empirically evaluate and validate our approach on a range of numerical examples, and briefly address the issue of knowledge acquisition by learning from examples (Sect. 6.5).

This chapter is intended less as a comprehensive evaluation of the specific model, but more as a collection of possible directions for further work, departing from the rigorous – but restricted – formulation in the previous chapters.

6.3 A Convex Model for Sparse Shape

In order to relax (6.6) to a convex problem, we first note that the restriction of u to characteristic functions allows to rewrite the objective in a linear fashion: Denoting by $A_{k,\cdot}$ the k-th row of the basis shape matrix A,

$$\|\mathrm{vecmax}(A \odot w) - u\|_2^2 = \sum_k \left(\mathrm{vecmax}(A_{k,\cdot} \odot w) - u_k\right)^2 \tag{6.7}$$

$$= \sum_{k,u_k=1} \left(1 - \mathrm{vecmax}(A_{k,\cdot} \odot w)\right) + \sum_{k,u_k=0} \mathrm{vecmax}(A_{k,\cdot} \odot w) \tag{6.8}$$

$$= \sum_k \{u_k \underbrace{\left(1 - \mathrm{vecmax}(A_{k,\cdot} \odot w)\right)}_{=:g(A_{k,\cdot} \odot w)} + (1 - u_k) \underbrace{\mathrm{vecmax}(A_{k,\cdot} \odot w)}_{=:h(A_{k,\cdot} \odot w)}\}. \tag{6.9}$$

The functions g and h in (6.9) are defined solely on integral vectors in $\{0, 1\}^n$. The optimal convex relaxation of (6.9) to $[0, 1]^n$, ensuring a minimal amount of artificial non-integral solutions, is given by its convex envelope, i.e. the largest closed convex function majorized by (6.9).

However, finding the global convex envelope is generally a very difficult problem. Instead, we approximate the true convex envelope by *locally* computing the envelope of the individual terms involving g and h. This approach is very similar to the one in Sect. 2.5.1, where the local convex envelope was used to construct the tight multiclass regularizer Ψ_d for a given metric. Again, we compute the convex envelopes of g and h using their biconjugates $(g^*)^*$ and $(h^*)^*$:

Proposition 6.1. *Let*

$$g(p) = \begin{cases} 1, & p = 0, \\ 0, & p \in \{0,1\}^K, p \neq 0, \\ +\infty, & p \notin \{0,1\}^K, \end{cases} \qquad h(p) = \begin{cases} 0, & p = 0, \\ 1, & p \in \{0,1\}^K, p \neq 0, \\ +\infty, & p \notin \{0,1\}^K. \end{cases} \tag{6.10}$$

Then the convex envelopes are given by

$$g^{**}(p) = \begin{cases} \max\{0, 1 - e^\top p\}, & p \in [0,1]^K, \\ +\infty, & \textit{otherwise}, \end{cases} \quad h^{**}(p) = \begin{cases} \mathrm{vecmax}(p), & p \in [0,1]^K, \\ +\infty, & \textit{otherwise}. \end{cases} \tag{6.11}$$

We omit the proof since it is standard and quite technical. The full local relaxation of problem (6.6) according to Prop. 6.1 is thus

$$\min_{w \in \{0,1\}^K} \left\{ \mu \|w\|_1 + \sum_{k=1}^n \left(u_k\, g^{**}(A_{k,\cdot} \odot w) + (1 - u_k)\, h^{**}(A_{k,\cdot} \odot w)\right) \right\}. \tag{6.12}$$

The nonsmooth relaxed function h^{**} is considerably more difficult to handle numerically than g^{**}, as it requires to introduce a large amount of KKT multipliers and therefore dual variables. However, experiments showed that very good results can be obtained by

replacing $h^{**}(p)$ with the upper bound $e^\top p$, yielding our final relaxation

$$\min_{w\in\{0,1\}^K}\left\{\mu\|w\|_1+\sum_{k=1}^{n}\left(u_k\max\{0,1-\langle A_{k,\cdot},w\rangle\}+(1-u_k)\langle A_{k,\cdot},w\rangle\right)\right\}. \tag{6.13}$$

All evaluations in the following sections are based on the reduced model (6.13).

6.4 Optimization

While problem (6.13) permits to compute approximate solutions of the combinatorial problem (6.6), it remains to cope with the large problem size and nonsmoothness of the objective. As in Chap. 4, we deal with the nonsmoothness by introducing dual variables v, and transforming the problem to a bilinear saddle-point problem:

$$\min_{w\in\mathcal{C}}\max_{v\in\mathcal{D}}\{\langle t,w\rangle+\langle v,L\,w\rangle-\langle b,v\rangle\}. \tag{6.14}$$

Using this notation, the "max" term in (6.13) can be rewritten as

$$\max_{v_k\in[0,1]}(v_k(1-A_{k,\cdot}x)). \tag{6.15}$$

Thus, the optimization problem (6.13) can be represented according to (6.14) by setting

$$\mathcal{C} = [0,1]^K, \tag{6.16}$$
$$\mathcal{D} = [0,1]^n, \tag{6.17}$$
$$t = A^\top(1-u)+\mu\,e, \tag{6.18}$$
$$b = -f, \tag{6.19}$$
$$L = -f\odot A. \tag{6.20}$$

As $\mathcal{C}$ and $\mathcal{D}$ are bounded, it follows from [Roc70, Cor. 37.6.2] that (6.14) has a saddle point (w^*,v^*), and from [Roc70, Lemma 36.2] ensures strong duality.

Although $L\in\mathbb{R}^{n\times K}$ with $K\gg n$ is sparse in general, the columns are usually non-orthogonal, i.e. the sets of indices corresponding to nonzero entries overlap significantly, and L does not necessarily have much structure, depending on the chosen set of shape templates. This prohibits a direct application of the Douglas-Rachford optimization technique, which relies on a fast computation of resolvents involving $L^\top L$.

However, the sets $\mathcal{C}$ and $\mathcal{D}$ encode simple box constraints, which allows to compute the primal objective as well as the orthogonal projections $\Pi_\mathcal{C}$ and $\Pi_\mathcal{D}$ efficiently. We therefore applied the FPD and Nesterov optimization approaches as outlined Chap. 4, which is straightforward given the saddle point formulation (6.14).

The Nesterov approach was again much slower, therefore we computed the examples using FPD, with the primal and dual step size parameters set to equal values such that they fulfill the convergence condition $\tau_P\tau_D\le\|L\|^{-2}$. The algorithm was implemented in MATLAB R2009a. Matrix-vector multiplications involving the template matrix L were performed by C++ subroutines that store L implicitly and compute the shape shape templates on the fly by transforming a number of basis templates.

Figure 6.3. Synthetical example demonstrating the shape reconstruction capabilities of the proposed approach. **Top left:** Original image consisting of a circle and four squares. **Top center:** Local segmentation based on the gray value with heavy overlaid noise. This local pre-segmentation constitutes the input to the proposed method. **Top right:** Result obtained by our approach. The individual shapes are correctly identified with the aid of a dictionary of square and circular templates. In this case the output vector of the relaxed problem is integral, indicating that it is a minimizer of the original combinatorial problem, with each of the five nonzero entries corresponding to one of the five visible shape parts.

6.5 Experimental Evaluation

Basic Approach. As an illustration of the complete approach, consider the synthetical example in Fig. 6.3. The image consists of a centered circle that has considerable overlap with four equally-sized squares. We supplemented the image with noise to simulate a real-world scenario with imperfect features and noisy local segmentation. The shape dictionary consisted of square and circle templates translated to all possible image locations. Although the initial local segmentation contains a large amount of noise, the five basis shapes are accurately found by solving the relaxed problem. The method automatically selects the correct number of shapes and their locations.

Real-World Images. In order to see how the proposed approach can be applied to real images, consider the color images in Fig. 6.2, Fig. 6.4, and Fig. 6.5. For each of the images, we extracted an initial pre-segmentation by computing local features from histograms over regions preselected by the user (Fig. 6.2), or by inspecting the distance to the background/foreground color (Fig. 6.4, Fig. 6.5) and a simple local thresholding operation. The latter initial segmentations are depicted in Fig. 6.4 and Fig. 6.5.

The regularization parameter μ is set by hand and varies between the different experiments. It roughly reflects the minimal amount of pixels that have to be exclusively covered by a certain basis function in order to justify its presence. Again, we created the shape dictionary by shifting a set of basis templates to all possible image locations.

Shape Decomposition. We conducted two further experiments in order to illustrate the potential of our template-based representation of image segmentations for further processing. Figure 6.6 shows the decomposition of a horse shape, obtained using circular templates of different sizes. By restricting the resulting sparse covering to templates with a certain size, the original shape can be decomposed into parts with a certain scale, such as the torso and the limbs.

Figure 6.4. Application of the proposed approach to an image containing overlapping coins. Although not all objects are detected our approach reveals promising results and recognizes even highly occluded objects. **Left:** Input image. **Center:** Initial, local pre-segmentation obtained by inspecting the distance to the dark background color and local thresholding; used as input to the proposed method. **Right:** Final segmentation result of the proposed approach, computed solely from the pre-segmentation in the center. Although there are some mistakes, the method correctly detects most of the heavily-occluded objects on the lower levels.

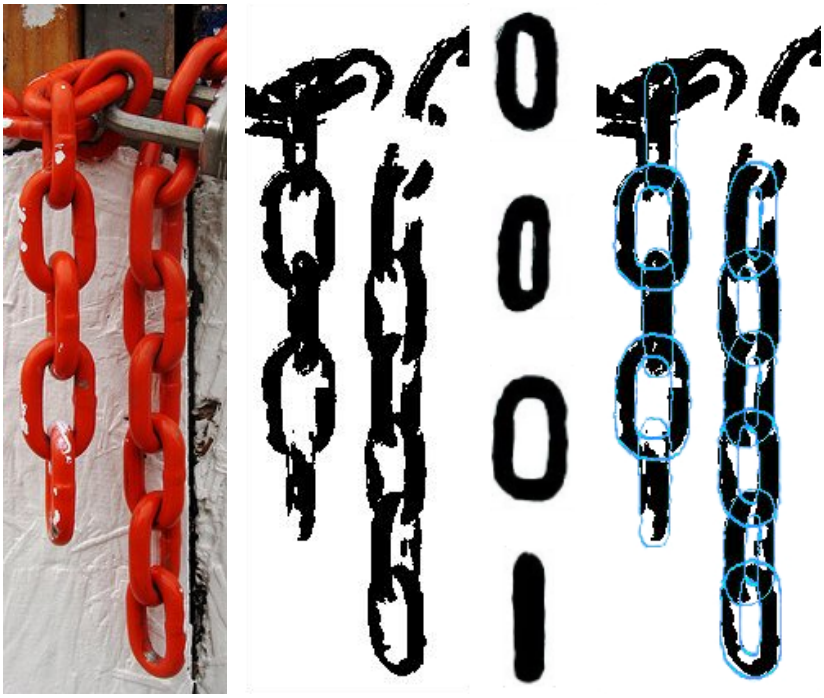

Figure 6.5. Application of the proposed approach to a real-world image involving nontrivial shapes. **Left to right:** Input image; local pre-segmentation using a thresholded distance to the color red as foreground indicator; hand-generated coarse shape templates used as basis templates; final result. Even highly overlapping parts are labeled correctly, and the shape is explained using a very small number of basis templates from the overcomplete shape template dictionary.

In a related experiment, we first built a coarse shape dictionary of prototypical horse parts, such as head, torso, and legs, and applied the proposed technique on a series of horse images (Fig. 6.7). Under moderate variation of the underlying shape or the observer's viewpoint, the proposed approach can robustly identify the templates

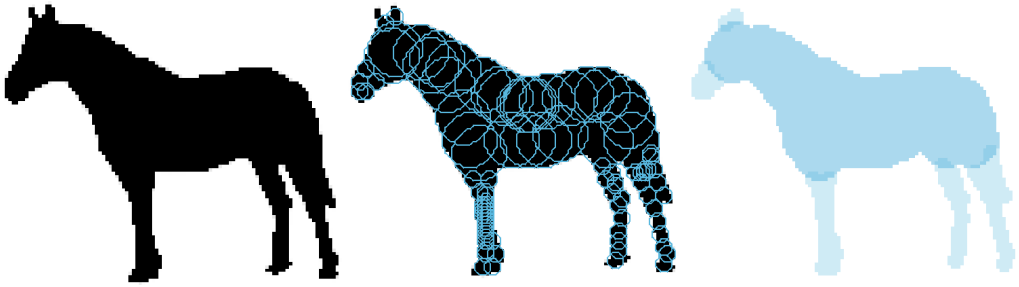

Figure 6.6. Shape decomposition by sparse covering with circular templates of different sizes. **Left:** Input shape. **Center:** Sparse reconstruction of the input shape using translated disc-shaped templates of different sizes. **Right:** Decomposition of the original shape into parts by restricting the reconstruction to discs with certain sizes. Using this method, the sparse representation approach can be used to decompose shapes into parts with a characteristic scale, such as torso and limbs in the horse example.

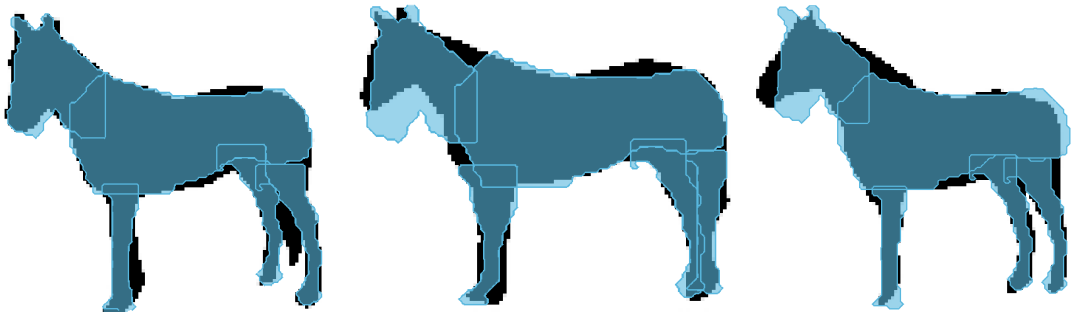

Figure 6.7. Image segmentation and shape decomposition with fixed templates from a pregenerated database of horse parts. While the method relies on only a few basis templates, it is fairly robust against moderate variation of the overall shape or the observer's viewpoint.

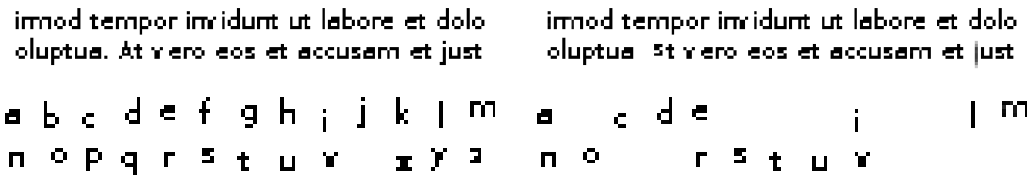

Figure 6.8. Learning shape templates from examples. **Top left:** Input image data. **Bottom left:** Initial dictionary. **Bottom right:** Learned dictionary. **Top right:** Segmentation (decomposition) of the input data using the learned dictionary. The learning process effectively reduces the initial number of templates by selecting those that are essential for explaining the image data. Removed templates are automatically replaced by superpositions of the remaining templates.

in different images and thus provides useful input for further template adjustment or contextual processing steps.

Shape Template Learning. Finally, we would like to point out the possibility of *learning* shape templates from a set of sample data, i.e. pairs of pre-segmentation and desired segmentation, in order to relieve the user from defining the templates (Fig. 6.8).

To this end, we another regularizer to the objective function that enforces sparsity of the template matrix A. Unfortunately, determining the optimal templates *simultaneously* with the optimal segmentation is a highly nonconvex problem. Therefore global optimality generally cannot be guaranteed and sufficiently accurate starting configurations are required.

Moreover, the problem becomes very large, since the complete matrix $A \in \mathbb{R}^{n \times K}$ has to be learned. Fig. 6.8 shows a scenario where the objective is to learn a small dictionary for recovering the text shown in the upper left panel. For the initial configuration, we used the dictionary visualized in the lower left of Fig. 6.8 together with the corresponding optimal representation w resulting from the text segmentation, i.e. the global optimum of problem (6.13) for fixed templates.

Applying an alternating technique for optimizing A and w yields an improvement of the dictionary, while almost preserving reconstruction performance. Non-existent letters are quickly removed from the dictionary due to the sparsity term, whereas letters occurring rarely are approximated by the superposition of more frequent letters. For instance, the letters p and b are reconstructed using a combination of the letters i, o and l, with almost no decrease in reconstruction performance.

6.6 Summary and Further Work

In this chapter we presented an approach to include advanced shape knowledge into the segmentation by representing complex shapes as a union of basis shapes. The approach is highly flexible since there are no restrictions on the set of basis templates. Nevertheless, the proposed combinatorial objective permits a convex approximation using a local relaxation that appears to be sufficiently tight for practical use.

First numerical tests indicate that the approach equally deals with heavy noise and occlusion, and has several possible applications, such as extracting structures with a specific scale and improving/learning of shape dictionaries.

Regarding future work, an important question concerns a theoretically sound continuous formulation – with respect to the spatial domain, but also with respect to the *parametrization of the shape templates*. In view of Chap. 3, this could be in particular valuable in connection with shape dictionaries that are generated from basis templates by operations with a continuous parameter set, such as translation, rotation and scaling: As can be seen in Fig. 6.2, quantizing the continuous parameter sets in order to generate a finite directory of shape templates may lead to a relatively coarse representation. If one could derive a continuous formulation together with a suitable discretization, this might allow to estimate the parameters with "sub-template" accuracy, similar to the sub-pixel accuracy provided by the continuous labeling approach with finite-differences scheme.

In particular, a formulation similar to the following seems possible: We fix the parameter set $\mathcal{P} := \mathbb{R} \times \{1, \ldots, K\}$, where an element (p, i) represents the shape that is obtained by applying some transformation (e.g. rotation, translation) with continuous parameter p to the i-th of K basis templates. We replace the matrix A with a mapping $A\colon \Omega \times \mathcal{P} \to \{0, 1\}$, where $A(x, (p, i)) = 1$ indicates that the point x is contained in the shape template defined by the tuple (p, i).

We then constrain w to a space of positive *measures* on $\mathcal{P}$ with the property $\lim_{\rho\to 0} w(\mathcal{B}_\rho(x)) \leqslant 1$ for all $x \in \Omega$. Ideally, w should be a *finite* sum of Dirac measures, $w = \delta_{(p_1,i_1)} + \ldots + \delta_{(p_k,i_k)}$, indicating a segmentation into the union of the corresponding k shape templates. Similar techniques have recently been discussed in [SW09, BP10] for an L^2 data term. Using these notations, the relaxation (6.13) can be formally rewritten as

$$\min_w \left\{ \mu\, w(\mathcal{P}) + \int_\Omega u \max\left\{0, 1 - \int_{\mathcal{P}} A(x,p)\, dw(p)\right\} + (1-u) \int_{\mathcal{P}} A(x,p)\, dw(p)\, dx \right\},$$

where the proper function and measure spaces have yet to be determined. The most prominent question is whether there are conditions on A that are sufficient for w to be a sum of Dirac measures, indicating a finite selection of basis templates. Even in the finite-dimensional case, such a condition, paralleling the Restricted Isometry Property in compressed sensing, would be quite valuable and allow a deeper insight under which conditions shapes can be reconstructed *exactly* by solving the relaxed problem.

Chapter 7
Conclusion

Summary. With the present work, we hope to have provided a usable framework for continuous multilabeling approaches. Motivated by the two-class case, we extended the approach to multiple labels in the framework of functions of bounded variation and showed how to construct regularizers with specific properties. The *local envelope* approach provides tight regularizers and can be applied for any metric interaction potentials, while the *embedding* approach is less powerful, but also computationally much easier to handle (Chap. 2).

Under a finite-differences discretization, the discretized functionals can be shown to approximate the spatially continuous functional in the sense of Γ-convergence, which also implies convergence of the minimizers. This motivated to investigate the conceptual difference between *combinatorial* and *relaxed* problem formulations. We concluded that a relaxation technique in combination with local rounding can have a theoretically better justification than the – more obvious – combinatorial approach: for the same neighborhood size, the additional freedom in specifying continuous energies allows for much greater precision (Chap. 3).

In order to solve the nonsmooth discretized problem, we considered two first-order approaches, based on Nesterov's method and on the Douglas-Rachford operator splitting framework. Numerical experiments indicate that the latter is quite robust, handles difficult regularizers well, and returns moderate-accuracy solutions in comparable time to commercial interior-point solvers (Chap. 4).

From a theoretical viewpoint, we then proposed a probabilistic rounding method – which can also be viewed as an *approximate coarea formula* –, and showed that it allows to obtain integral solutions with an *a priori* optimality bound in the spatially continuous framework (Chap. 5).

In the last chapter we presented some further ideas for including more detailed shape knowledge into the labeling process, by modeling shape as a union of a sparse set of basis templates. Based on the sparse representation setting, the approach has several interesting applications which are derived by choosing different shape directories (Chap. 6).

Further Work. Finally, we would like to point out the – in our opinion – most promising main areas for further research:

- Improving and *extending the model*: In particular, including nonlocal regularizers into the framework and investigating the connection to transportation problems seem to be promising directions.
- Integrating the approach as an inference component into a learning framework, in particular automatic *adaptation of the regularizer* based on sample pairs.
- *Numerical improvements:* Apart from customizing higher-order methods, an intriguing thought is the possibility of adapting techniques from graph-cut and max-flow approaches in order to speed up the optimization for the non-combinatorial problem with the improved finite-differences discretization.
- Derivation of *tighter bounds* for the relaxation: The most promising direction seems to be the adaptation of finite-dimensional approaches based on tree metrics. For the sparse shape representation, a rigorous formulation together with a result similar to the restricted isometry property would represent a major step in understanding the continuous formulation, and could lead to similar improvements for discretizing sparse representation problems with continuous parameters as are now available for labeling problems.

In any case we think that the present framework unites continuous and discrete worlds in an appealing way, and provides several compelling reasons to accept the additional challenges that appear when leaving the finite-dimensional realm.

Appendix A
Mathematical Preliminaries

A.1 Functions of Bounded Variation

In the following sections we provide a brief introduction to the concept of functions of bounded variation and corresponding functionals. For more detailed expositions we refer to [AFP00, Zie89, Mey01].

A.1.1 Total Variation and BV

For a differentiable scalar-valued function u, the total variation is simply the integral over the norm of its gradient:

$$\mathrm{TV}(u) \;=\; \int_\Omega \|\nabla u\|_2 \, dx. \tag{A.1}$$

As u is the designated labeling function, which ideally should be piecewise constant, the differentiability and continuity assumptions have to be dropped. In the following we will shortly review some basic definitions and properties.

For simplicity, we will assume in the following that the image domain is the open unit box, $\Omega = (0,1)^d$. However, most results could be formulated on general bounded domains with sufficiently smooth boundary, such as bounded open domains with compact Lipschitz boundary. By passing on to local convergence, i.e. replacing $L^1(\Omega)$ by the locally absolutely integrable functions $L^1_{\mathrm{loc}}(\Omega)$, many results can also be formulated for unbounded Ω. However, this considerably complicates the notation, and since images are usually defined on bounded sets we restrict ourselves to bounded domains.

We consider general vector-valued functions $u\colon \Omega \to \mathbb{R}^l$ which are absolutely integrable, i.e. $u \in L^1(\Omega)^l$. For any such function u, the *total variation* $\mathrm{TV}(u)$ can be defined in a dual way:

Definition A.1. *[AFP00, Def. 3.4, Prop. 3.6] Let $u \in L^1(\Omega)^l$. Then the* total variation *(sometimes just called* variation*) of u is defined as*

$$\begin{aligned}
\mathrm{TV}(u) \;&:=\; \sup_{v \in \mathcal{D}^{\mathrm{TV}}} -\sum_{j=1}^{l} \int_\Omega u_j \operatorname{div} v^j d\,x = \sup_{v \in \mathcal{D}^{\mathrm{TV}}} -\int_\Omega \langle u, \operatorname{Div} v\rangle d\,x, && \text{(A.2)}\\
\mathcal{D}^{\mathrm{TV}} \;&:=\; \{v \in C_c^\infty(\Omega)^{d\times l} \,|\, \|v(x)\|_2 \leqslant 1 \;\forall x \in \Omega\}, && \text{(A.3)}\\
\operatorname{Div} v \;&:=\; (\operatorname{div} v^1, \ldots, \operatorname{div} v^l)^\top.
\end{aligned}$$

This definition can be derived for continuously differentiable u by extending (A.1) to vector-valued u (where ∇u now denotes the Jacobian matrix),

$$\mathrm{TV}(u) \;=\; \int_\Omega \|\nabla u\|_2 d\,x, \tag{A.4}$$

replacing the norm by its dual formulation, and subsequent partial integration. If u has finite total variation $\mathrm{TV}(u)<\infty$, it is said to be of *bounded variation*. The vector space of all such functions is denoted by $\mathrm{BV}(\Omega)^l$,

$$\mathrm{BV}(\Omega)^l \;:=\; \{u\in L^1(\Omega)^l | \mathrm{TV}(u)<\infty\}. \tag{A.5}$$

For some set $\mathcal{S}\subseteq\mathbb{R}^l$, we define the restriction

$$\mathrm{BV}(\Omega,\mathcal{S}) \;:=\; \{u\in \mathrm{BV}(\Omega)^l | u(x)\in\mathcal{S} \text{ for } \mathcal{L}^d\text{-a.e. } x\in\Omega\}. \tag{A.6}$$

Any $u\in\mathrm{BV}(\Omega,\mathcal{S})$ therefore has a representative (in the L^1 equivalence class) satisfying $u(x)\in\mathcal{S}$ for all $x\in\Omega$. The space BV can alternatively be defined in terms of distributional derivatives [AFP00, Def. 3.1]:

Proposition A.2. *[AFP00, Prop. 3.6] The following two conditions are equivalent:*

- $u\in\mathrm{BV}(\Omega)^l$,
- $u\in L^1(\Omega)^l$ *and its distributional derivative corresponds to a finite Radon measure; i.e.* $u_j\in L^1(\Omega)$ *and there exist* $\mathbb{R}^d$*-valued measures* $D\,u_j=(D_1u_j,\ldots,D_du_j)$ *on the Borel subsets* $\mathcal{B}(\Omega)$ *of* Ω *such that*

$$-\sum_{j=1}^{l}\int_\Omega u_j \operatorname{div} v^j d\,x = \sum_{j=1}^{l}\sum_{i=1}^{d}\int_\Omega v_i^j d\,D_iu_j, \quad \forall v\in C_c^\infty(\Omega)^{d\times l}. \tag{A.7}$$

In the following, measures are generally signed and possibly vector-valued in the sense of [AFP00, Def. 1.4]. In this sense, the measures in (A.7) form the distributional derivative $D\,u=(D\,u_1|\ldots|D\,u_l)$, which is again a measure that vanishes on any $\mathcal{H}^{(d-1)}$-negligible set. Moreover, it can be shown that the total variation of u is exactly the measure-theoretic total variation of its distributional gradient [AFP00, Prop. 3.6].

Definition A.3. *Let* μ *be a (possibly vector-valued) measure on some measure space* $(X,\mathcal{A})$. *Then the* total variation *of* μ *on a set* $A\subset\mathcal{A}$ *is defined as*

$$|\mu|(A) \;:=\; \sup\left\{\sum_{k=0}^{\infty}\|\mu(A_k)\|_2 \,|\, (A_k)\subseteq\mathcal{A}\,\text{pairwise disjoint}, A=\bigcup_{k=0}^{\infty}A_k\right\}. \tag{A.8}$$

Proposition A.4. *Let* $u\in L^1(\Omega)^l$. *Then*

$$\int_\Omega d\,|Du|(\Omega) = |Du|(\Omega) = \mathrm{TV}(u). \tag{A.9}$$

This shows that the definition of the total variation is compatible with its smooth counterpart (A.1), with the gradient of u replaced by its distributional generalization. Part of the popularity of the total variation can be attributed to the fact that it has an intuitive geometrical interpretation:

Definition A.5. *Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a Lebesgue-measurable set. Then the* perimeter $\mathrm{Per}(\mathcal{S})$ *is defined as the total variation of its characteristic function,*

$$\mathrm{Per}(\mathcal{S}) \;:=\; \mathrm{TV}(\chi_{\mathcal{S}}). \tag{A.10}$$

Assuming the boundary $\partial\mathcal{S}$ is sufficiently regular, $\mathrm{Per}(\mathcal{S})$ is exactly the classical length ($d = 2$) or area ($d = 3$) of the boundary. More precisely, for any set $\mathcal{S} \subset \Omega$ with C^1-boundary $\partial\mathcal{S}$, we have [AFP00, (3.30)]

$$\mathrm{Per}(\mathcal{S}) \;=\; \mathcal{H}^{d-1}(\partial\mathcal{S}). \tag{A.11}$$

A.1.2 Properties of TV and Compactness

In this section we briefly review the most important facts for proving existence of minimizers for variational problems involving TV and BV.

Proposition A.6. *The total variation has the following properties:*

1. TV *is convex:*

$$\mathrm{TV}(\alpha\, u + (1-\alpha)\, u') \leqslant \alpha\, \mathrm{TV}(u) + (1-\alpha)\, \mathrm{TV}(u') \quad \forall u, u' \in L^1(\Omega), \forall \alpha \in [0,1].$$

2. TV *is positively homogeneous:* $\mathrm{TV}(\alpha\, u) = \alpha\, \mathrm{TV}(u) \quad \forall u \in L^1(\Omega), \forall \alpha > 0.$

3. TV *is lower semicontinuous in* $\mathrm{BV}(\Omega)^l$ *with respect to the* $L^1(\Omega)^l$ *topology, i.e. for all sequences* $(u^{(k)}) \subseteq \mathrm{BV}(\Omega)^l$ *converging (in the* L^1 *sense) to some* $u \in \mathrm{BV}(\Omega)^l$*,*

$$\liminf_{k\to\infty} \mathrm{TV}(u^{(k)}) \;\geqslant\; \mathrm{TV}(u^{(k)}).$$

The first two properties directly follow from the fact that the total variation is the pointwise supremum of a family of linear functions as in Def. A.1. The last statement can be shown using the same definition and a continuity argument [AFP00, Rem. 3.5, Prop. 3.6]. From Prop. A.6 it follows that

$$\|u\|_{\mathrm{BV}} \;:=\; \int_\Omega \|u\|_2 \, d\, x + \mathrm{TV}(u) \tag{A.12}$$

defines a norm on $\mathrm{BV}(\Omega)^l$. However, the induced topology is often too strong, i.e. it does not provide suitable compactness properties that are required for showing existence of minimizers. Therefore one frequently uses the weak* convergence:

Definition A.7. *Define* $u^{(k)} \to u$ weakly* in BV *iff*

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

$$\forall v \in C_0(\Omega)\colon \lim_{k\to\infty} \int_\Omega v \, d\, D\, u^{(k)} = \int_\Omega v \, d\, D\, u. \tag{A.13}$$

Weak* convergence is equivalent to L^1 convergence with an additional condition on the uniform boundedness:

Proposition A.8. *[AFP00, 3.13] Let $u, u^{(k)} \in \mathrm{BV}(\Omega)^l$. Then $u^{(k)} \to u$ weakly* in BV if and only if both of the following conditions hold:*

1. *$u^{(k)} \to u$ in $L^1(\Omega)^l$ and*
2. *the sequence $(u^{(k)})$ is* uniformly bounded *in $\mathrm{BV}(\Omega)^l$, i.e. there exists $C < \infty$ such that $\|u^{(k)}\|_{\mathrm{BV}} \leqslant C \ \forall k \in \mathbb{N}$.*

For the weak* topology in BV, a compactness result holds:

Theorem A.9. *Let $(u^{(k)}) \subset \mathrm{BV}(\Omega)^l$ be uniformly bounded in $\mathrm{BV}(\Omega)^l$. Then $(u^{(k)})$ contains a subsequence weakly*-converging to some $u \in \mathrm{BV}(\Omega)^l$.*

This result is a special case of the [AFP00, 3.23], using the fact that Ω is a bounded extension domain. The general formulation has local assumptions and holds for any open (even unbounded) set Ω; consequently the subsequence can only be shown to converge with respect to local convergence, i.e. in $L^1_{\mathrm{loc}}(\Omega)^l$.

A.1.3 Decomposition and general functionals on BV

In important property of the measure that forms the distributional derivative Du is that it can be uniquely decomposed into three mutually singular measures

$$Du \;=\; D^a u + D^j u + D^c u, \tag{A.14}$$

that is: the *absolutely continuous* part D^a, the *jump* part D^j, and the *Cantor* part D^c. Mutual singularity refers to the fact that Ω can be partitioned into three subsets, such that each of the measures is concentrated on exactly one of the sets, i.e. each set is a zero set under two of the measures. This will be explained in detail in the following.

Definition A.10. *Let ν be a (possibly vector-valued) measure, and μ a positive measure on some measure space. Then ν is* absolutely continuous *with respect to μ,*

$$\nu \ll \mu, \tag{A.15}$$

if ν vanishes on any μ-zero set, i.e. $\{\mu(A) = 0 \Rightarrow |\nu|(A) = 0\}$ for all $A \subseteq \mathbb{R}^d$. The measures μ, ν are mutually singular*,*

$$\nu \perp \mu, \tag{A.16}$$

if they vanish on complementary subsets of Ω, i.e. if there exists a Borel subset $B \subseteq \mathbb{R}^d$ such that $|\nu|(B) = 0$ and $|\mu|(\Omega \setminus B) = 0$. The definition recursively extends to more than two measures.

Definition A.11. *Let μ be a measure on a measure space $(X, \mathcal{A})$, and $A \in \mathcal{A}$. Then the* restriction *of μ to A is defined as the measure*

$$(\mu \llcorner A)(B) \;:=\; \mu(A \cap B). \tag{A.17}$$

If μ is a Radon measure and A is a Borel set, then $\mu \llcorner A$ is also a Radon measure. To give a meaning to the individual parts in (A.14), we need several more definitions.

Definition A.12. *(Reduced boundary $\mathcal{F}E$) [AFP00, Def. 3.54] Let E be an $\mathcal{L}^d$-measurable subset of $\mathbb{R}^d$ and Ω the largest open set such that E is locally of finite perimeter in Ω. Then the* reduced boundary *$\mathcal{F}E$ is defined as the set of all points $x \in \Omega$ in the support of $|D\chi_E|$ such that*

$$\nu_E(x) \;:=\; \lim_{h \searrow 0} \frac{D\chi_E(\mathcal{B}_h(x))}{|D\chi_E(\mathcal{B}_h(x))|} \tag{A.18}$$

exists in $\mathbb{R}^d$ and satisfies $\|\nu_E(x)\|_2 = 1$. The function $\nu_E \colon \mathcal{F}E \to S^{d-1}$ is the generalized inner normal *to E.*

De Giorgi (cf. [AFP00, Thm. 3.59]) proved the central fact that the total variation of characteristic functions reduces to the $(d-1)$-dimensional Hausdorff measure,

$$|D\chi_E| \;=\; |D\chi_E| \llcorner \mathcal{F}E = \mathcal{H}^{d-1} \llcorner \mathcal{F}E, \tag{A.19}$$

and that $\mathcal{F}E$ is countably $\mathcal{H}^{d-1}$-rectifiable for any $\mathcal{L}^d$-measurable subset of $\mathbb{R}^d$.

Definition A.13. *(Points of density t and essential boundary) [AFP00, Def. 3.60] For $t \in [0,1]$ and $\mathcal{L}^d$-measurable $E \subseteq \mathbb{R}^d$, the set of* points of density t *is defined as*

$$(E)^t \;:=\; \left\{ x \in \mathbb{R}^d \middle| \lim_{h \searrow 0} \frac{|E \cap \mathcal{B}_h(x)|}{|\mathcal{B}_h(x)|} = t \right\}. \tag{A.20}$$

We denote by

$$\partial^* E \;:=\; \mathbb{R}^d \setminus ((E)^0 \cup (E)^1) \tag{A.21}$$

the essential boundary *of E. The sets $(E)^1$ and $(E)^0$ are called the* measure-theoretic interior *and* exterior *of E.*

Remark A.14. By definition, the measure-theoretic interior and exterior are invariant under modifications of E on $\mathcal{L}^d$-negligible sets, in contrast to the classical interior and exterior (consider for instance the sets $\mathcal{B}_1(0) \subseteq \mathbb{R}^2$ and $\mathcal{B}_1(0) \setminus \{(x,0) | x \in (-1,1)\}$).

The next theorem of Federer shows that $\partial^* E$ (equivalently $(E)^{1/2}$) can be seen as the proper definition of the boundary compatible with the measure-theoretic definition of the interior and exterior, and are basically the same as the reduced boundary $\mathcal{F}E$.

Theorem A.15. *(Reduced boundary and essential boundary) [AFP00, Thm. 3.61] Let E be a set of finite perimeter in Ω. Then*

$$\mathcal{F}E \cap \Omega \subseteq (E)^{1/2} \subseteq \partial^* E \tag{A.22}$$

and

$$\mathcal{H}^{d-1}(\Omega \setminus (E^0 \cup \mathcal{F}E \cup E^1)) \;=\; 0. \tag{A.23}$$

In particular, this shows that $\mathcal{F}E$, $E^{1/2}$ and $\partial^* E$ are equal up to $\mathcal{H}^{d-1}$-negligible sets, which implies

$$Du \llcorner \mathcal{F}E = Du \llcorner E^{1/2} = Du \llcorner \partial^* E. \tag{A.24}$$

Definition A.16. *(Approximate limit and approximate discontinuity set S_u) [AFP00, Def. 3.63] Let $u \in L^1(\Omega)^l$. For $x \in \Omega$, we say that u has an* approximate limit *at x if there is a vector $z \in \mathbb{R}^l$ such that*

$$\lim_{h \searrow 0} \frac{1}{|\mathcal{B}_h(x)|} \int_{\mathcal{B}_h(x)} \|u(y) - z\|_2 \, dy \;=\; 0. \tag{A.25}$$

We define the approximate discontinuity *set $S_u \subseteq \Omega$ to be the set of all points which do* not *have an approximate limit. For any $x \in \Omega \setminus S_u$, we denote the* approximate limit

$$\tilde{u}(x) \;:=\; z. \tag{A.26}$$

In particular, if u is continuous in x, it has an approximate limit in x. Therefore S_u is comprised of the points where x is discontinuous in a specific sense. To further classify the discontinuities, we consider the set of points where u essentially "jumps" between two values along a hypersurface [AFP00, Def. 3.67]:

Definition A.17. *(The set J_u of approximate jump points) Let $u \in L^1(\Omega)^l$. Define*

$$\begin{aligned} \mathcal{B}_h^+(x,\nu) &:= \{y \in \mathcal{B}_h(x) \mid \langle y - x, \nu \rangle > 0\}, \\ \mathcal{B}_h^-(x,\nu) &:= \{y \in \mathcal{B}_h(x) \mid \langle y - x, \nu \rangle < 0\}. \end{aligned} \tag{A.27}$$

For $x \in \Omega$, we say that x is an approximate jump point *if there exist vectors $a, b \in \mathbb{R}^l$, $a \neq b$, and a unit vector $\nu \in S^{d-1}$ such that*

$$\lim_{h \searrow 0} \frac{1}{|\mathcal{B}_h(x)|} \int_{\mathcal{B}_h^+(x,\nu)} \|u(y) - a\|_2 \, dy \;=\; 0, \tag{A.28}$$

$$\lim_{h \searrow 0} \frac{1}{|\mathcal{B}_h(x)|} \int_{\mathcal{B}_h^-(x,\nu)} \|u(y) - b\|_2 \, dy \;=\; 0. \tag{A.29}$$

We define $J_u \subseteq \Omega$ to be the set of all approximate jump points. At each approximate jump point x, we denote $u^+(x) := a, u^-(x) := b, \nu_u(x) := \nu$. The values are unique up to the transformation $(a, b, \nu) \leftrightarrow (b, a, -\nu)$.

The idea behind this definition is that *locally* u "looks like" a function that takes only two values $u^+(x)$ and $u^-(x)$, with the discontinuity concentrated on a hyperplane with normal $\nu_u(x)$. J_u is a Borel subset of S_u, and it can be shown that u^+, u^- and ν_u can be chosen as Borel functions on J_u [AFP00, Prop. 3.69].

An important intermediate result involving these definitions is the following:

Lemma A.18. *[AFP00, Lemma 3.76] Let $u \in \mathrm{BV}(\Omega)^l$ and $B \subseteq \Omega$ a Borel set. Then*

1. *$|Du|$ vanishes on $\mathcal{H}^{d-1}$-negligible sets, i.e. $|Du| \ll \mathcal{H}^{d-1}$. Specifically,*

$$\mathcal{H}^{d-1}(B) = 0 \;\Rightarrow\; |Du|(B) = 0. \tag{A.30}$$

2. *$|Du|$ vanishes on $\mathcal{H}^{d-1}$-dimensional sets that have empty intersection with S_u:*

$$\mathcal{H}^{d-1}(B) < \infty,\; B \cap S_u = \emptyset \;\Rightarrow\; |Du|(B) = 0. \tag{A.31}$$

We now state the full decomposition of Du [AFP00, Def. 3.91].

Definition A.19. *Let $u \in \mathrm{BV}(\Omega)^l$.*

- *We define the* absolutely continuous part *and* singular part *of Du as the Radon measures $D^a u$ and $D^s u$ such that*

$$Du \;=\; D^a u + D^s u \tag{A.32}$$

  is the Lebesgue decomposition of Du.

- *The* jump part *$D^j u$ is defined as*

$$D^j u \;:=\; D^s u \llcorner J_u. \tag{A.33}$$

- *The* Cantor part *$D^c u$ is defined as*

$$D^c u \;:=\; D^s u \llcorner (\Omega \setminus S_u). \tag{A.34}$$

The decomposition (A.32) is possible as Du is a Radon measure by the assumption [EG92, Thm. 2] [AFP00, Thm. 1.28]. By definition $D^a \ll \mathcal{L}^d$, i.e. $|D^a|$ vanishes on any $\mathcal{L}^d$-negligible set.

Proposition A.20. *[AFP00, (3.89)] Let $u \in \mathrm{BV}(\Omega)^l$. Then*

$$Du = D^a u + \underbrace{D^j u + D^c u}_{D^s u}. \tag{A.35}$$

Furthermore, D^a, D^j and D^c are mutually singular.

Definition A.21. *We denote by* $\mathrm{SBV}(\Omega)$ *the set of all functions $u \in \mathrm{BV}(\Omega)$ for which $Du = D^a u + D^j u$, i.e. for which the Cantor part of Du vanishes.*

Intuitively, the absolutely continuous part D^a captures the "smooth" variations of u: in any neighborhood where u has a classical Jacobian ∇u, the jump part and the Cantor part vanish and

$$D u = D^a u = \nabla u \mathcal{L}^d \tag{A.36}$$

(this can be generalized to *approximate differentials* [AFP00, Def. 3.70, Thm. 3.83]). The quantity $|D^a u|(\Omega)$ corresponds to integrating $\|\nabla u\|_2$ over the image domain.

The jump part D^j is concentrated on the set J_u of points where locally u jumps between two values u^- and u^+ along a hypersurface with normal $\nu_u \in S^{d-1}$ (Def. A.17). It can be expressed as [AFP00, (3.90)]

$$D^j u = D u \llcorner J_u = \nu_u (u^+ - u^-)^\top \mathcal{H}^{d-1} \llcorner J_u. \tag{A.37}$$

Therefore, its total variation $|D^j u|(\Omega)$ corresponds to integrating the magnitude of the jump along the jump set J_u. The Cantor part D^c captures anything that is left.

For the special case where u is the characteristic function of a set of finite perimeter, equation (A.19) shows that

$$D\chi_E = (D\chi_E) \llcorner \mathcal{F}E = D^j \chi_E. \tag{A.38}$$

As an important consequence of the decomposition (A.35) and the mutual singularity of the parts, the total variation decomposes into the individual total variation measures,

$$|Du| \;=\; |D^a u| + |D^j u| + |D^c u|. \tag{A.39}$$

Accordingly,

$$|Du|(\Omega) \;=\; \int_\Omega \|\nabla u(x)\|_2\,dx + \int_{J_u} \|\nu_u(u^+-u^-)^\top\|_2\,d\mathcal{H}^{d-1} + \int_\Omega 1\,d\,|D^c u|. \tag{A.40}$$

Using this idea, one can define functionals depending on the distributional gradient Du [AFP00, Prop. 2.34]. Let $u\in \mathrm{BV}(\Omega)^l$ and define, for some convex, lower semicontinuous $\Psi\colon \mathbb{R}^{d\times l}\to\mathbb{R}$,

$$\begin{aligned} J(u) := \int_\Omega d\Psi(Du) := & \int_\Omega \Psi(\nabla u(x))dx + \\ & \int_{J_u} \Psi_\infty(\nu_u(x)(u^+(x)-u^-(x))^\top)d\mathcal{H}^{d-1} + \\ & \int_\Omega \Psi_\infty\!\left(\frac{D^c u}{|D^c u|}\right)d|D^c u|. \end{aligned} \tag{A.41}$$

Here Ψ_∞ denotes the *recession function* $\Psi_\infty(x)=\lim_{t\to\infty}\frac{\Psi(t\,x)}{t}$ of Ψ, and $D^c u/|D^c u|$ denotes the *polar decomposition* of $D^c u$, which is the density of $D^c u$ with respect to its total variation measure $|D^c u|$. In case Ψ is positively homogeneous, $\Psi^\infty=\Psi$ holds and

$$J(u) \;=\; \int_\Omega \Psi\!\left(\frac{Du}{|Du|}\right)d|Du|. \tag{A.42}$$

Essentially, Ψ generates a new measure from $\Psi(Du) = (\Psi\circ(Du/|Du|))|Du|$ by transforming its density with respect to $|Du|$ [AFP00, Thm. 2.38]. This extends the meaning of Ψ as acting on the *Jacobian* of u to the jump set as acting on the *difference* of the left and right side limits of u at the discontinuity, and allows to uniformly handle discontinuities as well as regions where u is differentiable.

A.1.4 The Coarea Formula

One of the central and most useful properties of the total variation is the *coarea formula*. In order to specify the full theorem (Thm. A.29 and Thm. A.32), several additional definitions are required:

Definition A.22. *[AFP00, Def. 2.11] Let $E\subseteq\mathbb{R}^d$ and $f\colon E\to\mathbb{R}^m$. Then f is called a* Lipschitz function*, $f\in\mathrm{Lip}(E)^m$, if*

$$\mathrm{Lip}(f,E) \;:=\; \sup\left\{\frac{|f(x)-f(y)|}{|x-y|}\,|\,x,y\in E, x\neq y\right\} < \infty. \tag{A.43}$$

Similar to the classical definition of Lipschitz continuity, it holds that $|f(x)-f(y)|\leqslant \mathrm{Lip}(f,E)\,|x-y|$ for all $x,y\in E, x\neq y$, and $\mathrm{Lip}(f,E)$ is the smallest such constant.

Definition A.23. *[AFP00, Def. 2.57] Let $k\in\mathbb{N}_0$, $k\leqslant d$, and $E\subseteq\mathbb{R}^d$ be an $\mathcal{H}^k$-measurable set. Then E is said to be*

- countably k-rectifiable*, if there exist countably many $\mathbb{R}^d$-valued Lipschitz functions $f^{(i)}\colon\mathbb{R}^k\to\mathbb{R}^d$ such that*

$$E \;\subseteq\; \bigcup_{i=0}^{\infty} f^{(i)}(\mathbb{R}^k), \tag{A.44}$$

  i.e. E can be covered by the images of countably many Lipschitz functions,

- countable $\mathcal{H}^k$-rectifiable, *if there exist countably many* $\mathbb{R}^d$*-valued Lipschitz functions* $f^{(i)}\colon \mathbb{R}^k \to \mathbb{R}^d$ *such that*

$$\mathcal{H}^k\left(E \setminus \bigcup_{i=0}^{\infty} f^{(i)}(\mathbb{R}^k) \right) \;=\; 0, \tag{A.45}$$

  i.e. E can be covered by the images of countably many Lipschitz functions up to a $\mathcal{H}^k$-negligible set,
- $\mathcal{H}^k$-rectifiable, *if E is countable $\mathcal{H}^k$-rectifiable and $\mathcal{H}^k(E) < \infty$.*

Definition A.24. *[AFP00, Def. 2.68] Let V, W be Hilbert spaces such that* $\dim(V) = k \leqslant d = \dim(W)$ *and $L\colon V \to \Omega$ a linear map. We denote*

$$\boldsymbol{J}_k L \;:=\; \sqrt{\det(L^* L)}, \tag{A.46}$$

where $L^*\colon W^* \to V^*$ is the transpose of L. Analogously,

$$\boldsymbol{C}_k L \;:=\; \sqrt{\det(L\, L^*)}. \tag{A.47}$$

The determinant can be computed using any matrix representation of $L^*\, L$. For an orthonormal basis matrix representation M of L, $L^*\, L$ has the matrix representation $M^* M$, therefore for real Hilbert spaces $\boldsymbol{J}_k L = \sqrt{\det(M^\top M)}$.

Theorem A.25. *(Area formula) [AFP00, Thm. 2.71] Let $k \in \mathbb{N}_0$, $k \leqslant d$, $f\colon \mathbb{R}^k \to \mathbb{R}^d$ be a Lipschitz function, and $E \subseteq \mathbb{R}^k$ be an $\mathcal{L}^k$-measurable set. Then*

$$\int_{\mathbb{R}^d} \mathcal{H}^0(E \cap f^{-1}(y))\, d\mathcal{H}^k(y) \;=\; \int_E \boldsymbol{J}_k\, df_x\, dx, \tag{A.48}$$

where df_x denotes the (Fréchet-) differential of f in x.

This is well-defined since by Rademacher's theorem, the differential df_x exists $\mathcal{L}^k$-a.e. for Lipschitz functions and coincides $\mathcal{L}^k$-a.e. with the weak derivative [EG92, 3.1.2 Thm. 2] [Zie89, Sect. 2.2 , Prop. 2.2.1].

Definition A.26. *(Approximate tangent space to a set) [AFP00, Def. 2.79, Def. 2.86] Let $k \in \mathbb{N}_0$, $k \leqslant d$, and $S \subseteq \Omega \subseteq \mathbb{R}^d$ be a countably $\mathcal{H}^k$-rectifiable set. Let $(S^{(i)})$ be a partition of $\mathcal{H}^k$-almost all of S into $\mathcal{H}^k$-rectifiable sets.*

If for some $x \in S^{(i)}$ there exists a k-dimensional linear subspace U_x of $\mathbb{R}^d$, and a scalar $\theta \in \mathbb{R}$ such that

$$\lim_{\rho \searrow 0} \frac{1}{\rho^k} \int_{S^{(i)}} \varphi\left(\frac{y-x}{\rho}\right) d\mathcal{H}^k(y) \;=\; \theta \int_U \varphi(y)\, d\mathcal{H}^k(y) \quad \forall \varphi \in C_c^\infty(\mathbb{R}^d), \tag{A.49}$$

we define the approximate tangent space $\mathrm{Tan}^k(S, x) := U_x$.

The approximate tangent space $\mathrm{Tan}^k(S, x)$ is not well-defined pointwise in the sense of an $\mathcal{H}^k$-equivalence class of mappings from the set S to the space of k-dimensional linear subspaces of $\mathbb{R}^d$, that coincide $\mathcal{H}^k$-a.e. [AFP00, Prop. 2.85, Rem. 2.87]. If the set S is the image of an $\mathcal{L}^d$-measurable set $D \subseteq \mathbb{R}^k$ under an *injective* Lipschitz function $f\colon \mathbb{R}^k \to \mathbb{R}^d$, i.e. $S = f(D)$, then the approximate tangent space $\mathrm{Tan}^k(S, x)$ is the image of the differential of f at $f^{-1}(x)$ for $\mathcal{H}^k$-a.e. $x \in S$ [AFP00, Prop. 2.88].

Theorem A.27. *(Tangential differential) [AFP00, Thm. 2.90] Let $E \subseteq \mathbb{R}^m$ be a countably $\mathcal{H}^k$-rectifiable set and $f: \mathbb{R}^m \to \mathbb{R}^d$ be a Lipschitz function.*

Then for $\mathcal{H}^k$-a.e. $x \in E$, the restriction of f to the affine space $x + \mathrm{Tan}^k(E, x)$ is differentiable. The corresponding (Fréchet-) differential, mapping from $\mathrm{Tan}^k(E, x)$ to $\mathbb{R}^d$, is said to be the tangential differential*, and denoted by $d^E f_x$.*

Theorem A.28. *(Generalized area formula) [AFP00, Thm. 2.91] Let $k \in \mathbb{N}_0, k \leqslant N$, $f: \mathbb{R}^m \to \mathbb{R}^d$ be a Lipschitz function and $E \subseteq \mathbb{R}^m$ a countably $\mathcal{H}^k$-rectifiable set. Then*

$$\int_{\mathbb{R}^d} \mathcal{H}^0(E \cap f^{-1}(x))\, d\mathcal{H}^k(y) \;=\; \int_E \boldsymbol{J}_k\, d^E f_x\, d\mathcal{H}^k(x). \tag{A.50}$$

Theorem A.29. *(Coarea formula for Lipschitz functions) [Fed69, Prop. 3.2.11][Mor95, Prop. 3.8][AFP00, Thm. 2.93] Let $f: \mathbb{R}^m \to \mathbb{R}^k$ be a Lipschitz function and $E \subseteq \mathbb{R}^m$ be a countably $\mathcal{H}^d$-rectifiable set for some $d \geqslant k$. Then $E \cap f^{-1}(t)$ is countably $\mathcal{H}^{d-k}$-rectifiable for $\mathcal{L}^k$-a.e. $t \in \mathbb{R}^k$, and*

$$\int_E \boldsymbol{C}_k d^E f_x\, d\mathcal{H}^d(x) \;=\; \int_{\mathbb{R}^k} \mathcal{H}^{d-k}(E \cap f^{-1}(t))\, dt. \tag{A.51}$$

Remark A.30. There also exists an inhomogeneous generalization with an additional weighting function $g(x)$ [Fed69, Prop. 3.2.12].

Remark A.31. A particular case is $m = d$, $E = \mathbb{R}^d$ and $k = 1$, where $\mathrm{Tan}^k(S, x) = \mathbb{R}^{d \times 1}$, and the differential $d^E f_x$ (which exists $\mathcal{L}^d$-a.e.) maps from $\mathbb{R}^d$ to $\mathbb{R}$ and is described by the gradient $\nabla f(x) \in \mathbb{R}^d$. Therefore $\boldsymbol{C}_k d^E f_x = \|\nabla f\|_2$, and the coarea formula becomes

$$\int_{\mathbb{R}^d} \|\nabla f\|_2\, dx \;=\; \int_{\mathbb{R}} \mathcal{H}^{d-1}(f^{-1}(t))\, dt. \tag{A.52}$$

Theorem A.32. *(Coarea formula in* BV*) [FR60][AFP00, Thm. 3.40] Let $\Omega \subseteq \mathbb{R}^d$ open and $u \in \mathrm{BV}(\Omega)$. Then the set $\{u > t\} := \{x \in \Omega | u(x) \geqslant t\}$ has finite perimeter in Ω for $\mathcal{L}^1$-a.e. $t \in \mathbb{R}$ and, for any Borel set $B \subseteq \Omega$,*

$$Du(B) \;=\; \int_{\mathbb{R}} D\chi_{\{u>t\}}(B)\, dt, \tag{A.53}$$

$$|Du|(B) \;=\; \int_{\mathbb{R}} |D\chi_{\{u>t\}}|(B)\, dt. \tag{A.54}$$

Remark A.33. The latter formulation provides the analogon to the coarea formula for Lipschitz functions. By the theorem of De Giorgi (A.19), $|D\chi_E| = \mathcal{H}^{d-1} \llcorner \mathcal{F}E$, therefore equation (A.54) can be expressed as

$$|Du|(\Omega) \;=\; \int_{\mathbb{R}} \mathcal{H}^{d-1}(\mathcal{F}\,\{u > t\}). \tag{A.55}$$

The essential difference to the formulation for Lipschitz functions (A.52) consists in replacing the inverse image – which may not have a "nice" structure, since u is not required to be continuous and may contain jumps – by the reduced boundary of the superlevelset. Alternatively, the essential boundary or the set of points with density $1/2$ may be used, cf. Thm. A.15.

A.2 Γ-Convergence

In this section we state the necessary properties of Γ-converging sequences [DGF75, DM93, AK00, Bra02]. Γ-convergence is particularly useful for characterizing the convergence of functionals and their minimizers, and can be interpreted as set-convergence of the epigraphs [DM93, Chap. 4].

Definition A.34. *(Γ-convergence) [DM93, Prop. 8.1][Bra02, Def. 1.5] Let X be a topological space that satisfies the first axiom of countability. Let $(F^{(k)})_{k\in\mathbb{N}}, F^{(k)}: X \to \bar{\mathbb{R}}$ be a sequence of functions. Then $F = \Gamma - \lim_{k\to\infty} F^{(k)}$ (i.e. $F^{(k)}$ is said to Γ-converge to F in the topology of X) iff*

1. $\forall x \in X, \forall (x^{(k)}) \subseteq X$ *s.t.* $x^{(k)} \to x$ *it holds that* $F(x) \leqslant \liminf_{k\to\infty} F^{(k)}(x^{(k)})$,
2. $\forall x \in X \exists (x^{(k)}) \subseteq X$ *s.t.* $x^{(k)} \to x$ *and* $F(x) \geqslant \limsup_{k\to\infty} F^{(k)}(x^{(k)})$.

Remark A.35. The first axiom of countability refers to the condition that for each $x \in X$ there is a countable set $\{U_1, U_2, \ldots\}$ such that for any neighborhood U of x there exists an index i such that $U_i \subseteq U$. This is always satisfied in metric spaces such as L^1 (set $U_i = i^{-1}\mathcal{B}_1(x)$, i.e. scaled balls around x).

Remark A.36. The second condition is equivalent to

$$\forall x \in X \exists (x^{(k)}) \subseteq X, x^{(k)} \to x \text{ s.t. } F(x) = \lim_{k\to\infty} F^{(k)}(x^{(k)}). \tag{A.56}$$

Proposition A.37. *(Lower semicontinuity) [DM93, Def. 1.1, Def. 1.4, Rem. 1.5] Let X be a topological space, and $F: X \to \bar{\mathbb{R}}$. Then F is said to be*

- lower semicontinuous, *if for every $x \in X$ and every $t \in \mathbb{R}$ with $t < F(x)$ there exists a neighborhood U of x such that $t < F(y)$ for all $y \in U$,*
- sequentially lower semicontinuous, *if for every $x \in X$ and every sequence (x^k) converging to x,*

$$F(x) \leqslant \lim_{k\to\infty} \inf F(x^{(k)}) \tag{A.57}$$

holds.

If F is lower semicontinuous, then F is sequentially lower semicontinuous. If X satisfies the first axiom of countability (as is the case for metrizable X), and F is sequentially lower semicontinuous, then F is lower semicontinuous.

Proposition A.38. *(Coercivity) [DM93, Def. 1.12, Rem. 1.13] Let X be a topological space, and $F: X \to \bar{\mathbb{R}}$. Then F is said to be*

- coercive *on X, if the closure of $\{F \leqslant t\}$ is countably compact in X for every $t \in \mathbb{R}$,*
- sequentially coercive *on X, if the closure of $\{F \leqslant t\}$ is sequentially compact in X for every $t \in \mathbb{R}$.*

If F is sequentially coercive, then F is coercive. If X is metrizable, then F is coercive if and only if F is sequentially coercive.

Proposition A.39. *(Equicoercivity) [DM93, Def. 7.6, Prop. 7.7] Let $(F^{(k)})$ be a sequence in X. Then $(F^{(k)})$ is said to be* equicoercive *if for every $t \in \mathbb{R}$ there exists a closed countably compact subset K_t of X such that $\{F^{(k)} \leqslant t\} \subseteq K_t$ for every k.*

The sequence (F^k) is equicoercive if and only if there exists a lower semicontinuous, coercive *function $G\colon X \to \mathbb{R} \cup \{\pm\infty\}$ such that $F^{(k)} \geqslant G$ uniformly, i.e. for every k.*

Proposition A.40. *[DM93, Def. 7.10, Thm. 7.23] Let $(F^{(k)})$ be an equicoercive sequence Γ-converging to a function F. Define the* set of minimizers

$$M(F) \;:=\; \{x \in X \,|\, F(x) = \inf_{x' \in X} F(x')\}, \tag{A.58}$$

and for $\varepsilon > 0$ the set of approximate minimizers

$$M_\varepsilon(F) \;:=\; \{x \in X \,|\, F(x) \leqslant \max\{-1/\varepsilon,\, \inf_{x' \in X} F(x') + \varepsilon\}\}. \tag{A.59}$$

Then for every neighborhood U of $M(F)$ there exists $\varepsilon > 0$ and k_0 such that

$$M(F^{(k)}) \subseteq M_\varepsilon(F^{(k)}) \subseteq U \quad \forall k \geqslant k_0. \tag{A.60}$$

Prop. A.40 states that both the sets of exact minimizers and the sets of approximate minimizers of the functionals $F^{(k)}$ converge to the set $M(F)$ of minimizers of F, in the sense that they can be forced inside any neighborhood of $M(F)$ for k sufficiently large. If F has a unique minimizer x_0, Prop. A.40 implies that any sequence of minimizers (or $\varepsilon^{(k)}$-minimizers, with $\varepsilon^{(k)} \searrow 0$) converges to x_0 (cf. [DM93, Cor. 7.24]).

A.3 Set-Valued Operators and Proximal Steps

In this section we briefly review the most important concepts from the theory of set-valued mappings, monotone operators, proximal point and operator splitting methods, as required in particular for the Douglas-Rachford splitting (Sect. 4.3.2). For a more detailed analysis, we refer the reader to [Eck89, RW04]. While we restrict ourselves to the case of finite-dimensional vector spaces, most results transfer also to the case of general Hilbert spaces [Zl02, Mor06, Set09b, BC10, CP10b].

The need for set-valued mappings can be motivated by the fact that for any proper, convex and lower-semicontinuous (*lsc*) function $f\colon X \to \mathbb{R} \cup \{-\infty, +\infty\}$, where $X := \mathbb{R}^n$ for some $n \in \mathbb{N}$, $x \in X$ minimizes f if and only if 0 is a subgradient of f in x. Defining the *subdifferential* of f as the set of subgradients for each $x \in X$,

$$\partial f(x) \;:=\; \{z \in X \,|\, \langle y - x, z\rangle \leqslant f(y) - f(x)\; \forall y \in X\}, \tag{A.61}$$

this connection can be concisely expressed as the *generalized Fermat condition*,

$$f(x) = \min_{x' \in X} f(x') \;\Leftrightarrow\; 0 \in \partial f(x). \tag{A.62}$$

The minimization of f reduces to finding a *zero* of ∂f, i.e. an $x \in X$ such that $0 \in \partial f(x)$. This motivates the following definitions.

Definition A.41. *(set-valued mappings and operators) Let $X := \mathbb{R}^n$ and $Y := \mathbb{R}^m$ for some $n, m \in \mathbb{N}$. A* set-valued mapping $T: X \rightrightarrows Y$ *is defined by a subset of $X \times Y$, $T \subseteq X \times Y$. We denote*

- *(mapping)* $Tx := \{y' \in Y \,|\, (x, y') \in T\}, x \in X$,
- *(inverse)* $T^{-1} := \{(y, x) | (x, y) \in T\} \subseteq Y \times X$, *i.e.*

$$x \in T^{-1} y \;\Leftrightarrow\; y \in Tx,$$

- *(domain)* $\operatorname{dom} T := \{x \in X \,|\, Tx \neq \emptyset\}$,
- *(range)* $\operatorname{range} T := \operatorname{dom} T^{-1} = \{y \in Y \,|\, T^{-1} y \neq \emptyset\}$,
- T *is* single-valued *if* $Tx = \{y(x)\}$ *for all* $x \in X$.

An operator on X *is a set-valued mapping* $T: X \to X$.

The set-valued mappings associated to ordinary functions $g: X \to Y$ are single-valued and coincide with their graph, $\{(x, g(x)) | x \in X\}$. The subdifferential ∂f can be viewed as a set-valued mapping and is single-valued if and only if f is differentiable. A central property of operators is maximal monotonicity:

Definition A.42. *An operator $T: X \rightrightarrows X$ is*

- monotone *if*

$$\langle x - x', Tx - Tx' \rangle \geqslant 0 \quad \forall x, x' \in X, \tag{A.63}$$

  in the sense that

$$\langle x - x', y - y' \rangle \geqslant 0 \quad \forall x, x' \in X, y \in Tx, y' \in Tx', \tag{A.64}$$

- maximal monotone *if T is monotone and there is no other monotone operator S with $T \subseteq S$. Equivalently,*

$$T \text{ maximal monotone } \Leftrightarrow\; \forall (x,\, y) \notin T \,\exists (x',\, y') \in T \;\; s.t. \;\; \langle x - x',\, y - y' \rangle < 0, \tag{A.65}$$

- nonexpansive *if*

$$\|Tx - Tx'\|_2 \;\leqslant\; \|x - x'\|_2 \quad \forall x, x' \in X, \tag{A.66}$$

  in the sense that

$$\|y - y'\|_2 \leqslant \|x - x'\|_2 \quad \forall x, x' \in X, y \in Tx, y' \in Tx', \tag{A.67}$$

- firmly nonexpansive *if*

$$\|y - y'\|_2^2 \;\leqslant\; \|x - x'\|_2^2 - \|(x - x') - (y - y')\|_2^2 \;\; \forall x, x' \in X, y \in Tx, y' \in Tx'.$$

If T corresponds to a differentiable single-valued function g via $Tx = \{g(x)\}$, as in the case of $g(x) = \nabla f(x)$ for some $f \in C^2(X, \mathbb{R})$, monotonicity corresponds to the positive semidefiniteness of $\nabla g = \nabla^2 f$, i.e. the second-order convexity criterion for f. Any continuous monotone operator is maximally monotone [RW04, 12.7], therefore ∇f is maximal monotone for $f \in C^2(X, \mathbb{R})$. For nondifferentiable functions, the following result holds:

Proposition A.43. *(adapted from [RW04, Thm. 12.17]) If $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper and convex, then its subdifferential ∂f is monotone. If additionally f is lower semicontinuous, then ∂f is maximal monotone.*

A monotone operator T is maximal monotone if and only if it is surjective in the sense that range $T = X$ [Eck89, Thm. 3.5]. Nonexpansivity guarantees that T is a contraction in a generalized sense; T is firmly nonexpansive if and only if $2\,T - I$ is nonexpansive [Eck89, Prop. 3.8].

Theorem A.44. *[Eck89, Thm. 3.6] Let $T\colon X \rightrightarrows X$ and $\lambda > 0$. Define the* resolvent *of the operator T as*

$$J_{\lambda T} := (I + \lambda\, T)^{-1}. \tag{A.68}$$

Then the following relations hold:

1. *T is monotone $\Rightarrow$ $J_{\lambda T}$ is single-valued.*
2. *T is monotone $\Leftrightarrow$ $J_{\lambda T}$ is firmly nonexpansive.*
3. *T is maximal monotone $\Leftrightarrow$ $J_{\lambda T}$ is firmly nonexpansive and* $\operatorname{dom} J_{\lambda T} = X$.

The importance of resolvents becomes apparent when considering that for any maximal monotone operator T it holds that

$$0 \in Tx \;\Leftrightarrow\; J_{\lambda T}(x) = \{x\}, \tag{A.69}$$

i.e. finding a zero of a maximal monotone operators corresponds to finding a fixed point of the associated resolvent. Consider the fixed point approach for finding a zero of T,

$$0 \in Tx \;\Leftrightarrow\; x \in x - \lambda\, T(x). \tag{A.70}$$

Two possibilities for turning (A.70) into an iterative scheme are

$$\textit{(forward step)} \quad x^{k+1} \in x^k - \lambda\, Tx^k, \Leftrightarrow x^{k+1} \in (I - \lambda\, T)\, x^k, \tag{A.71}$$

$$\textit{(backward step)} \quad x^{k+1} \in x^k - \lambda\, Tx^{k+1} \Leftrightarrow x^{k+1} \in (I + \lambda\, T)^{-1}\, x^k. \tag{A.72}$$

Suitable combinations of these two steps form the basis of many first-order methods in nonsmooth optimization. The notation is in analogy to explicit/implicit integration methods such as forward/backward Euler steps. In optimization, we have $T = \partial f$, therefore the forward step amounts to subgradient descent, which has intrinsic difficulties such as nonuniqueness, bounds on the step size and numerical issues when computing the set of active constraints. These problems are avoided by using backward steps. From the definition of $J_{\lambda T}$,

$$x \in J_{\lambda T} y \;\Leftrightarrow\; (I + \lambda\, T)\, x \ni y \Leftrightarrow 0 \in (x - y) + \lambda\, T \tag{A.73}$$

$$\Leftrightarrow\; x = \arg\min_{x'} \{\frac{1}{2}\|x - y\|^2 + \lambda\, f(x)\} =: (P_{\lambda f})(y) \tag{A.74}$$

Thus the resolvent, i.e. the backward step, can be computed by solving a regularized optimization problem. This is also known as *prox*-step, as it involves minimizing f together with a *proximity* term that keeps x close to the previous iterate. However, solving (A.73) is generally as hard as the original problem.

This can be circumvented by applying splitting techniques. In the operator splitting framework, ∂f is assumed to be decomposable into the sum of two "simple" operators, $T = A + B$, of which forward and backward steps can practically be computed. Finding a zero of T is then reduced to a suitable combination of steps on the individual operators A and B. Operator splittings can be obtained by decomposing the function f into a sum of "simpler" functions f_i under a constraint on the relative interiors of their domains; see [Roc70, Thm. 23.8] or [RW04, Cor. 10.9] for the most general statement:

Proposition A.45. *Let $f = f_1 + ... + f_m$ for proper convex functions f_1, ..., f_m on $X = \mathbb{R}^n$. Then*

$$\bigcap_{i=1}^{m} \operatorname{ri}(\operatorname{dom} f_i) \neq \emptyset \implies \partial f = \partial f_1 + ... + \partial f_m. \tag{A.75}$$

Usually one requires that the f_i be proper, convex and lsc, such that by Prop. A.43 the subdifferentials ∂f_i are maximal monotone, which with Thm. A.44 provides uniqueness of the backward steps on the individual f_i. We now review some duality properties.

Proposition A.46. *(Properties of the subdifferential) [RW04, Prop. 11.3,11.4] Assume that $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, lsc and convex, and denote by f^* its Legendre-Fenchel transform, $f^*(y) = \sup_{x\in\mathbb{R}^n} \{\langle x, y\rangle - f(x)\}$. Then*

1. *$\partial f^* = (\partial f)^{-1}$.*
2. *$f(x) + f^*(y) \geqslant \langle x, y\rangle$ and*

$$f(x) + f^*(y) = \langle x, y\rangle \Leftrightarrow y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y). \tag{A.76}$$

3. *$\partial f(x) = \arg\max_y \{\langle x, y\rangle - f^*(y)\}$, $\partial f^*(y) = \arg\max_x \{\langle x, y\rangle - f(x)\}$.*
4. *For a closed convex set $\mathcal{C} \neq \emptyset$, we have $(\delta_{\mathcal{C}})^* = \sigma_{\mathcal{C}}$ and*

$$\partial\delta_{\mathcal{C}}(x) \;=\; N_{\mathcal{C}}(x) := \{y \in \mathbb{R}^n \,|\, \langle x' - x, y\rangle \leqslant 0 \;\forall x' \in \mathcal{C}\} \tag{A.77}$$

5. *For a closed convex set $\mathcal{D}$,*

$$y \in \partial\sigma_{\mathcal{D}}(x) \;\Leftrightarrow\; x \in N_{\mathcal{D}}(y) \Leftrightarrow \{y \in \mathcal{D} \text{ and } \langle x, y\rangle = \sigma_{\mathcal{D}}(x)\}. \tag{A.78}$$

Proposition A.47. *Let $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper, lsc and convex, and $z \in \mathbb{R}^n$. Then there is a unique decomposition*

$$z \;=\; x + y, \quad f(x) + f^*(y) = \langle x, y\rangle. \tag{A.79}$$

This decomposition is given by

$$x = P_f(z) = J_{\partial f}(z), \quad y = P_{f^*}(z) = J_{\partial f^*}(z). \tag{A.80}$$

In particular, by the identity $(\lambda f)^*(y) = \lambda f^*(y/\lambda)$ for $\lambda > 0$, we get

$$z = x + y, \quad f(x) + \lambda f^*(y/\lambda) = \langle x, y\rangle/\lambda \tag{A.81}$$

$$\Leftrightarrow\; x = P_{\lambda f}(z), \quad y = \lambda P_{\lambda^{-1} f^*}(z/\lambda). \tag{A.82}$$

An instant result of this theorem is the proximal step for support functions with closed, convex $\mathcal{D} \subseteq \mathbb{R}^n$,

$$z = P_{\lambda\sigma_{\mathcal{D}}}(z) + \lambda\, P_{\lambda^{-1}\delta_{\mathcal{D}}}(z/\lambda) \tag{A.83}$$

$$\Rightarrow \quad P_{\lambda\sigma_{\mathcal{D}}}(z) = z - \lambda\, \Pi_{\mathcal{D}}(z/\lambda). \tag{A.84}$$

For reference, we list the most common backward steps:

General Constraints. For $h(z) = \delta_{\mathcal{C}}(z)$, we have

$$P_{\lambda h}(z) \;=\; \Pi_{\mathcal{C}}(z), \tag{A.85}$$

where $\Pi_{\mathcal{C}}$ is the Euclidean projector on $\mathcal{C}$.

Affine Constraints. Often one has constraints of the form $h(z) = \delta_{Az=0}$. The backward step is then given by the projection onto the null space of A, which can be formulated in terms of the pseudoinverse A^+,

$$P_{\lambda h}(z) \;=\; (I - A^+ A)\, z. \tag{A.86}$$

Usually $A \in \mathbb{R}^{p\times q}$ with $p \leqslant q$ and A has full rank. Then $A^+ = A^\top (A\, A^\top)^{-1}$ and

$$P_{\lambda h}(z) \;=\; (I - A^\top (A\, A^\top)^{-1} A)\, z. \tag{A.87}$$

Support Functions. For $h(z) = \sigma_{\mathcal{D}}(z)$, we get from (A.84)

$$P_{\lambda h}(z) \;=\; z - \lambda\, \Pi_{\mathcal{D}}(z/\lambda), \tag{A.88}$$

i.e. for support functions the backward steps can be evaluated *exactly* if projections on the underlying dual sets can be performed.

Bibliography

[ABDM03] G. Alberti, G. Bouchitté, and G. Dal Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Calc. Var. Part. Diff. Eq.*, 16(3):299–333, 2003.

[ABKM10] C. Arora, S. Banerjee, P. Kalra, and S. N. Maheshwari. An efficient graph cut algorithm for computer vision problems. In *Europ. Conf. Comp. Vis.*, 2010.

[AEB06] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representations. Tech. rep., 2006.

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

[AHU64] K. J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and non-linear programming. With contributions by Hollis B. Chenery [and others]*. Stanford Univ. Press, 1964.

[AK00] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing*. Springer, 2000.

[AMT91] L. Ambrosio, S. Mortola, and V. M. Tortorelli. Functionals with linear growth defined on vector valued BV functions. *J. Math. pures et appl.*, 70:193–323, 1991.

[AT90] L. Ambrosio and V. M. Tortorelli. Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence. *Comm. Pure Appl. Math.*, 43(8):999–1036, 1990.

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

[Auj08] J.-F. Aujol. Some first-order algorithms for total variation based image restoration. *J. Math. Imaging Vis.*, 2008. published online.

[Bar98] Y. Bartal. On approximating arbitrary metrices by tree metrics. In *ACM Symposium on Theory of Computing*, 1998.

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

[BC00] B. Bourdin and A. Chambolle. Implementation of an adaptive finite-element approximation of the Mumford-Shah functional. *Num. Math.*, 85:610–646, 2000.

[BC10] H. H. Bauschke and P. L. Combettes. The Baillon-Haddad theorem revisited. *J. Convex Analysis*, 17(4):781–787, 2010.

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

[Ber98] D. P. Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific, 1998.

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

[Bes86] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Stat. Soc., Series B*, 48(3):259–302, 1986.

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

[BH02] E. Boros and P. L. Hammer. Pseudo-boolean optimization. *Discr. Appl. Math.*, 123:155–225, 2002.

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

[BKP10] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *J. Imaging Sci.*, 3(3):294–526, 2010.

[BKSS10] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *Int. J. Comp. Vis.*, 87(1-2):93–117, 2010.

[Blo98] P. V. Blomgren. *Total Variation Methods for Restoration of Vector Valued Images*. PhD thesis, UCLA, 1998.

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

[Boy04] S. Boyd. *Convex Optimization*. Cambridge Univ. Press, 2004.

[BP10] K. Bredies and H. K. Pikkarainen. Inverse problems in spaces of measures. SFB Report 2010-037, Univ. Graz, 2010.

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

[BT09a] E. Bae and X.-C. Tai. Global minimization for continuous multiphase partitioning problems using a dual approach. CAM report, UCLA, 2009.

[BT09b] E. Bae and X.-C. Tai. Graph cut optimization for the piecewise constant level set method applied to multiphase image segmentation. In *Scale Space and Var. Meth.*, volume 5567 of *LNCS*, pages 1–13, 2009.

[BT09c] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In *Int. Conf. on Acoustics, Speech and Signal Proc.*, 2009.

[BT10] A. Beck and M. Teboulle. *Signal Processing and Communications*, chapter Gradient-Based Algorithms with Applications in Signal Recovery Problems. Cambridge Univ. Press, 2010.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. MPS/SIAM, 2001.

[BU02] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Europ. Conf. Comp. Vis,*, volume 2351 of *LNCS*, pages 109–124. Springer, 2002.

[BU08] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *Patt. Anal. Mach. Intell.*, 30(12):2109 – 2125, 2008.

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

[BW05] D. Bertsimas and R. Weismantel. *Optimization over Integers*. Dynamic Ideas, 2005.

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

[Car01] J. L. Carter. *Dual Methods for Total Variation-Based Image Restoration*. PhD thesis, UCLA, 2001.

[CCG+98] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. Plotkin. Approximating a finite metric by a small number of tree metrics. In *ACM Found. Comp. Sci*, pages 379–388, 1998.

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

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

[CDD09] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. AMS*, 22(1):211–231, 2009.

[CDG+09] T.-H. H. Chan, K. Dhamdhere, A. Gupta, J. Kleinberg, and A. Slivkins. Metric embeddings with relaxed guarantees. *J. Computing*, 38(6):2303–2329, 2009.

[CDM99] A. Chambolle and G. Dal Maso. Discrete approximations of the Mumford-Shah functional in dimension two. *Model. Math. et Anal. Num.*, 33:651–672, 1999.

[CDS01] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

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

[CEPY05] T. Chan, S. Esedoḡlu, F. Park, and A. Yip. Total variation image restoration: Overview and recent developments. In *The Handbook of Mathematical Models in Computer Vision*. Springer, 2005.

[CGM99] T. F. Chan, G. H. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *J. Sci. Computing*, 20:1964–1977, 1999.

[Cha99] A. Chambolle. Finite-differences discretizations of the Mumford-Shah functional. *Model. Math. et Anal. Num.*, 33:261–288, 1999.

[Cha04] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[Cha05] A. Chambolle. Total variation minimization and a class of binary MRF models. In *Energy Min. Meth. Comp. Vis. Patt. Recogn.*, volume 3757, pages 136–152, 2005.

[Cho54] L. Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954.

[Cho56] G. Choquet. Existence des représentations intégrales au moyen des points extrémaux dans les cônes convexes. *C. R. Acad. Sci.*, 243:669–702 and 736–737, 1956.

[CKNZ01] C. Chekuri, S. Khanna, J. Naor, and L. Zosin. Approximation algorithms for the metric labeling problem via a new linear programming formulation. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 109–118, 2001.

[CKR98] G. Călinescu, H. Karloff, and Y. Rabani. An improved approximation algorithm for multiway cut. In *ACM Symposium on Theory of Computing*, 1998.

[CKS97] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *Int. J. Comp. Vis.*, 22:61–79, 1997.

[CKS03] D. Cremers, T. Kohlberger, and C. Schnörr. Shape statistics in kernel space for variational image segmentation. *Patt. Recogn.*, 36(9):1929–1943, 2003.

[CKSS96] V. Caselles, R. Kimmel, G. Sapiro, and C. Sbert. Three dimensional object modeling via minimal surfaces. In *Europ. Conf. Comp. Vis.*, 1996.

[CL97] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188, 1997.

[CMM00] T. Chan, A. Marquina, and P. Mulet. Higher-order total variation-based image restoration. *J. Sci. Computing*, 22(2):503–516, 2000.

[CN04] J. Chuzhoy and J. S. Naor. The hardness of metric labeling. In *Found. Comp. Sci.*, 2004.

[Com96] P. L. Combettes. The convex feasibility problem in image recovery. *Adv. in Imag. and Electron Phys.*, 95:155–270, 1996.

[Com04] P. L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53:475–504, 2004.

[CP07] P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Selected Topics Sign. Proc.*, 1(4):564–574, 2007.

[CP08] P. L. Combettes and J.-C. Pesquet. A proximal decomposition method for solving convex variational inverse problems. *Inverse Problems*, 24(6), 2008.

[CP10a] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 2010. published online.

[CP10b] P. L. Combettes and J.-C. Pesquet. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal splitting methods in signal processing. Springer New York, 2010.

[CRD07] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *Int. J. Comp. Vis.*, 72(2):195–215, 2007.

[CRTV05] E. Candes, M. Rudelson, T. Tao, and R. Vershynin. Error correcting via linear programming. In *46th Ann. IEEE Symp. Found. Comp. Science*, pages 295–308, 2005.

[CS05] T. F. Chan and J. Shen. *Image processing and analysis*. SIAM, 2005.

[CSS06] D. Cremers, N. Sochen, and C. Schnörr. Multiphase dynamic labeling for variational recognition-driven image segmentation. *Int. J. Comp. Vis.*, 66(1):67–81, 2006.

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

[CWS02] F. Cremers, D. and Tischhäuser, J. Weickert, and C. Schnörr. Diffusion snakes: Introducing statistical shape knowledge into the Mumford-Shah functional. *Int. J. Comp. Vis.*, 50(3):295–313, 2002.

[DAG09] V. Duval, J.-F. Aujol, and Y. Gousseau. The TVL1 model: a geometric point of view. Technical Report 00380195, HAL, 2009.

[Dav05] G. David. *Singular Sets of Minimizers for the Mumford-Shah Functional*. Birkhäuser, 2005.

[DAV08] V. Duval, J.-F. Aujol, and L. Vese. A projected gradient algorithm for color image decomposition. *CMLA Preprint*, (2008-21), 2008.

[DFN09] S. Durand, J. Fadili, and M. Nikolova. Multiplicative noise removal using L1 fidelity on frame coefficients. *J. Math. Imaging Vis.*, 36(3):201–226, 2009.

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

[DGF75] E. De Giorgi and T. Franzoni. Su un tipo di convergenza variazionale. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.*, 68:842–850, 1975.

[DJPS94] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, and P. D. Seymour. The complexity of multiterminal cuts. *J. Computing*, 23(4):864–894, 1994.

[DKP05] N. Dixit, R. Keriven, and N. Paragios. GPU-cuts: Combinatorial optimisation, graphic processing units and adaptive object extraction. Technical Report 05-07, CERTIS, ENPC, 2005.

[DM93] G. Dal Maso. *An Introduction to Γ-Convergence*. Birkhäuser, 1993.

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

[DS06a] J. Darbon and M. Sigelle. Image restoration with discrete constrained total variation part I: Fast and exact optimization. *J. Math. Imaging Vis.*, 26(3):261–276, 2006.

[DS06b] J. Darbon and M. Sigelle. Image restoration with discrete constrained total variation part II: Levelable functions, convex priors and non-convex cases. *J. Math. Imaging Vis.*, 26(3):277–291, 2006.

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

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

[EG92] L. C. Evans and F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992.

[EK72] J. Edmonds and R. M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, 19:248–264, 1972.

[EO04] S. Esedoḡlu and S. J. Osher. Decomposition of images by the anisotropic Rudin-Osher-Fatemi model. *Comm. Pure Appl. Math.*, 57(12):1609–1626, 2004.

[EZC10] E. Esser, X. Zhang, and T. F. Chan. 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.

[Fed69] H. Federer. *Geometric Measure Theory*. Springer, 1969.

[FF56] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Can. J. Math.*, 8:399–404, 1956.

[FF62] L. R. Ford and D. R. Fulkerson. *Flows in Networks*. Princeton Univ. Press, 1962.

[FL01] O. Faugeras and Quang-TuanFa Luong. *The Geometry of Multiple Images*. MIT Press, 2001.

[Fle57] W. H. Fleming. Functions with generalized gradient and generalized surfaces. *Annali di Mathematica Pura ed Applicata*, 44(1), 1957.

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

[FRT04] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *J. Comp. System Sci.*, 69(3):485–497, 2004.

[Gab83] D. Gabay. *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, chapter IX Applications of the Method of Multipliers to Variational Inequalities, pages 299–331. Elsevier, 1983.

[GBO09a] T. Goldstein, X. Bresson, and S. Osher. Geometric applications of the split Bregman method: Segmentation and surface reconstruction. CAM Report 09-06, UCLA, 2009.

[GBO09b] T. Goldstein, X. Bresson, and S. Osher. Global minimization of Markov random field with applications to optical flow. CAM Report 09-77, UCLA, 2009.

[GC10a] B. Goldluecke and D. Cremers. An approach to vectorial total variation based on geometric measure theory. In *Comp. Vis. Patt. Recogn.*, 2010.

[GC10b] S. Goldluecke and D. Cremers. Convex relaxation for multilabel problems with product label spaces. In *Europ. Conf. Comp. Vis.*, 2010.

[GG84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.*, 6:721–741, 1984.

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

[GM01] M. Gobbino and M. G. Mora. Finite-difference approximation of free-discontinuity problems. In *Proc. Royal Soc. Edinburgh*, volume 131, pages 567–595, 2001.

[GM09] D. Goldfarb and S. Ma. Fast multiple splitting algorithms for convex optimization. arXiv Preprint 0912.4570, 2009.

[GO09] T. Goldstein and S. Osher. The split Bregman method for L1-regularized problems. *J. Imaging Sci.*, 2:323–343, 2009.

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

[GPS89] D. M. Greig, B. T. Porteous, and H. Seheult. Exact maximum a posteriori estimation for binary images. *J. Royal Stat. Soc., Series B (Methodological)*, 51(2):271–279, 1989.

[Gra81] A. Graham. *Kronecker Products and Matrix Calculus with Applications*. J. Wiley and Sons, NY, 1981.

[GSS82] L. M. Goldschlager, R. A. Shaw, and J. Staples. The maximum flow problem is log space complete for P. *Theor. Comp. Sci.*, 21:105–111, 1982.

[GT88] A. B. Goldberg and R. E. Tarjan. A new approach to the maximum-flow problem. *J. ACM*, 35(4):921–940, 1988.

[HS06] M. Hintermüller and G. Stadler. An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. *J. Sci. Computing*, 28(1):1–23, 2006.

[HVD07] M. Hussein, A. Varshney, and L. Davis. On implementing graph cuts on CUDA. In *First Workshop on General Purpose Processing on Graphics Processing Units*, 2007.

[HZ00] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2000.

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

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

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

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[KKBC09] K. Kolev, M. Klodt, T. Brox, and D. Cremers. Continuous global optimization in multiview 3d reconstruction. *Int. J. Comp. Vis.*, 84(1), 2009.

[KKMR06] H. Karloff, S. Khot, A. Mehta, and Y. Rabani. On earthmover distance, metric labeling, and 0-extension. In *ACM symposium on Theory of computing*, pages 547–556, 2006.

[KMN93] R. M. Karp, R. Motwani, and N. Nisa. Probabilistic analysis of network flow algorithms. *Math. Op. Research*, 18(1):71–97, 1993.

[Kol06] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Patt. Anal. Mach. Intell.*, 28, 2006.

[Kom10] N. Komodakis. Towards more efficient and effective LP-based algorithms for MRF optimization. In *Europ. Conf. Comp. Vis.*, 2010.

[KPT07] N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *Int. Conf. Comp. Vis.*, 2007.

[KRBT08] P. Kohli, J. Rihan, M. Bray, and P. H. S. Torr. Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. *Int. J. Comp. Vis.*, 79(3):285–298, 2008.

[KSK+08] M. Klodt, T. Schoenemann, K. Kolev, M. Schikora, and D. Cremers. An experimental comparison of discrete and continuous shape optimization methods. In *Europ. Conf. Comp. Vis.*, Marseille, France, October 2008.

[KSSC03] J. Keuchel, C. Schnörr, C. Schellewald, and D. Cremers. Binary partitioning, perceptual grouping, and restoration with semidefinite programming. *Patt. Anal. Mach. Intell.*, 25:1364–1379, 2003.

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

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

[KTP08] N. Komodakis, G. Tziritas, and N. Paragios. Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal-dual strategies. *Comp. Vis. Image Underst.*, 112:14–29, 2008.

[KVT11] M. P. Kumar, O. Veksler, and P. H. S. Torr. Improved moves for truncated convex models. *J. Mach. Learn. Res.*, 12:31–67, 2011.

[KWT87] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contours models. *Int. J. Comp. Vis.*, 1(4):321–331, 1987.

[Lau96] S. Lauritzen. *Graphical Models*. Oxford Univ. Press, 1996.

[LBS10] J. Lellmann, D. Breitenreicher, and C. Schnörr. Fast and exact primal-dual iterations for variational problems in computer vision. In *Europ. Conf. Comp. Vis.*, volume 6312 of *LNCS*, pages 494–505, 2010.

[LKY+08] J. Lellmann, J. Kappes, J. Yuan, F. Becker, and C. Schnörr. Convex multi-class image labeling by simplex-constrained total variation. Tech. rep., Univ. of Heidelberg, 2008.

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

[LLT06] J. Lie, M. Lysaker, and X.-C. Tai. A variant of the level set method and applications to image segmentation. *Math. Comp.*, 75:1155–1174, 2006.

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

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

[LRB07] V. Lempitsky, C. Rother, and A. Blake. Logcut – efficient graph cut optimization for Markov random fields. In *Europ. Conf. Comp. Vis.*, pages 1–8, 2007.

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

[LS10] J. Lellmann and C. Schnörr. Continuous multiclass labeling approaches and algorithms. Tech. rep., Univ. of Heidelberg, Feb. 2010.

[LT06] M. Lysaker and X.-C. Tai. Iterative image restoration combining total variation minimization and a second-order functional. *Int. J. Comp. Vis.*, 66(1):5–18, 2006.

[Mey01] Y. Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, volume 22 of *Univ. Lect. Series*. AMS, 2001.

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

[Mit03] H. D. Mittelmann. An independent benchmarking of SDP and SOCP solvers. *Math. Prog.*, pages 407–430, 2003.

[Mit11] H. D. Mittelmann. Benchmarks for optimization software, 2011.

[Mor95] F. Morgan. *Geometric Measure Theory: A Beginner's Guide*. Academic Press, 2nd edition, 1995.

[Mor06] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I*. Springer, 2006.

[MPR98] J. E. Mitchell, P. M. Pardalos, and M. G. C. Resende. *Interior Point Methods for Combinatorial Optimization*. Kluwer Academic Publishers, 1998.

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

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

[Neg99] M. Negri. The anisotropy introduced by the mesh in the finite element approximation of the Mumford-Shah functional. *Numer. Funct. Anal. Optim*, 20:957–982, 1999.

[Nes04a] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[Nes04b] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Prog.*, 103(1):127–152, 2004.

[Nik01] M. Nikolova. Smoothing of outliers in image restoration by minimizing regularized objective functions with nonsmooth data-fidelity terms. In *Int. Conf. Image Proc.*, volume 1, pages 233–236, 2001.

[NN93] Y. Nesterov and A. S. Nemirovskii. *Interior Point Methods in Convex Optimization: Theory and Applications*. SIAM, 1993.

[OBOK09] C. Olsson, M. Byröd, N. C. Overgaard, and F. Kahl. Extending continuous cuts: Anisotropic metrics and expansion moves. In *Int. Conf. Comp. Vis.*, 2009.

[Øks03] B. Øksendal. *Stochastic Differential Equations*. Springer, 6th edition, 2003.

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

[OS88] S. Osher and J. A. Sethian. Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi equations. *J. Comp. Phys.*, pages 12–49, 1988.

[PCBC09] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *Int. Conf. Comp. Vis.*, 2009.

[PCBC10] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *J. Imaging Sci.*, 3(4):1122–1145, 2010.

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

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

[PSG+08] T. Pock, T. Schönemann, G. Graber, H. Bischof, and D. Cremers. A convex formulation of continuous multi-label problems. In *Europ. Conf. Comp. Vis.*, volume 3, pages 792–805, 2008.

[Ren01] J. Renegar. *A Mathematical View of Interior-Point Methods in Convex Optimization*. MPS/SIAM, 2001.

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

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

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

[RZE10] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *Trans. Sign. Proc.*, 58:1553–1564, 2010.

[Sch98] O. Scherzer. Denoising with higher order derivatives of bounded variation and an application to parameter estimation. *Computing*, 60:1–27, 1998.

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

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

[SGG+09] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, 2009.

[SKO09] P. Strandmark, F. Kahl, and N. C. Overgaard. Optimizing parametric total variation models. In *Int. Conf. Comp. Vis.*, 2009.

[SKSS11] B. Savchynskyy, J. Kappes, S. Schmidt, and C. Schnörr. A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling. In *Comp. Vis. Patt. Recogn.*, 2011.

[SR96] G. Sapiro and D. L. Ringach. Anisotropic diffusion of multivalued images with applications to color filtering. *IEEE Trans. Image Proc.*, 5(11):1582–1586, 1996.

[Sri99] A. Srinivasan. Approximation algorithms via randomized rounding: A survey. In *Series in Advanced Topics in Mathematics*. Polish Scientific Publishers, 1999.

[STC09] F. R. Schmidt, E. Töppe, and D. Cremers. Efficient planar graph cuts with applications in computer vision. In *Comp. Vis. Patt. Recogn.*, 2009.

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

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

[Sud06] E. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, MIT, 2006.

[SW09] O. Scherzer and B. Walch. Sparsity regularization for Radon measures. PAI Report 10, Univ. of Innsbruck, 2009.

[SY98] H. Stark and Y. Yang. *Vector Space Projections – A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics*. John Wiley & Sons, 1998.

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

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

[UMPB09] M. Unger, T. Mauthner, T. Pock, and H. Bischof. Tracking as segmentation of spatial-temporal volumes by anisotropic weighted TV. In *Energy Min. Meth. Comp. Vis. Patt. Recogn.*, volume 5681 of *Springer LNCS*, pages 193–206, 2009.

[VN08] V. Vineet and P. J. Narayanan. CUDA cuts: Fast graph cuts on the GPU. In *Comp. Vis. Patt. Recogn.*, 2008.

[VO96] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *J. Sci. Computing*, 17:227–238, 1996.

[WABF09] P. Weiss, G. Aubert, and L. Blanc-Féraud. Efficient schemes for total variation minimization under constraints in image processing. *J. Sci. Computing*, 31(6260):2047–2080, 2009.

[WJW05] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. MAP estimation via agreement on trees: Message-passing and linear programming. *Information Theory*, 51:3697–3717, 2005.

[WS01] J. Weickert and C. Schnörr. A theoretical framework for convex regularizers in PDE-based computation of image motion. *Int. J. Comp. Vis.*, 45(3):245–264, 2001.

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

[WZDT11] C. Wu, J. Zhang, Y. Duan, and X.-C. Tai. Augmented Lagrangian method for total variation based image restoration and segmentation over triangulated surfaces. *J. Sci. Computing*, published online, 2011.

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

[YBTB10] J. Yuan, E. Bae, X.-C. Tai, and Y. Boykov. A continuous max-flow approach to potts model. In *Europ. Conf. Comp. Vis.*, pages 379–392, 2010.

[You78] D. Youla. Generalized image restoration by the method of alternating orthogonal projections. *IEEE Trans. Circuits and Systems*, 25(9):694–702, 1978.

[YSM10] G. Yu, G. Sapiro, and S. Mallat. Image modeling and enhancement via structured sparse model selection. In *Proc. IEEE Int. Conf. Image Processing*, 2010.

[YYZW08] J. Yang, W. Yin, Y. Zhang, and Y. Wang. A fast algorithm for edge-preserving variational multichannel image restoration. Tech. Rep. 08-09, Rice Univ., 2008.

[Zl02] C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, 2002.

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

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

[ZHW10] Y. Zhang, R. Hartley, and L. Wang. Fast multi-labelling for stereo matching. In *Europ. Conf. Comp. Vis.*, pages 524–537, 2010.

[Zie89] W. P. Ziemer. *Weakly Differentiable Functions*. Springer, 1989.

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

[ZWC10] M. Zhu, S. Wright, and T. F. Chan. Duality-based algorithms for total-variation-regularized image restoration. *Comp. Opt. Appl.*, 47:377–400, 2010.