Sublabel–Accurate Relaxation of Nonconvex Energies

Thomas Möllenhoff *1, Emanuel Laude *1, Michael Moeller1, Jan Lellmann2 and Daniel Cremers1

1Technical University of Munich, 2University of Lübeck

Abstract

We propose a novel spatially continuous framework for convex relaxations based on functional lifting. Our method can be interpreted as a sublabel–accurate solution to multilabel problems. We show that previously proposed functional lifting methods optimize an energy which is linear between two labels and hence require (often infinitely) many labels for a faithful approximation. In contrast, the proposed formulation is based on a piecewise convex approximation and therefore needs far fewer labels – see Fig. 1. In comparison to recent MRF-based approaches, our method is formulated in a spatially continuous setting and shows less grid bias. Moreover, in a local sense, our formulation is the tightest possible convex relaxation. It is easy to implement and allows an efficient primal-dual optimization on GPUs. We show the effectiveness of our approach on several computer vision problems.

1. Introduction

Energy minimization methods have become the central paradigm for solving practical problems in computer vision. The energy functional can often be written as the sum of a data fidelity and a regularization term. One of the most popular regularizers is the total variation (TV) due to its many favorable properties [4]. Hence, an important class of optimization problems is given as

\[
\min_{u: \Omega \rightarrow \Gamma} \int_{\Omega} \rho(x, u(x)) \, dx + \lambda \, TV(u), \tag{1}
\]

defined for functions \(u\) with finite total variation, arbitrary, possibly nonconvex dataterms \(\rho : \Omega \times \Gamma \rightarrow \mathbb{R}\), label spaces \(\Gamma\) which are closed intervals in \(\mathbb{R}\), \(\Omega \subset \mathbb{R}^d\), and \(\lambda \in \mathbb{R}^+\). The multilabel interpretation of the datateme is that \(\rho(x, u(x))\) represents the costs of assigning label \(u(x)\) to point \(x\). For (weakly) differentiable functions \(TV(u)\) equals the integral over the norm of the derivative, and therefore favors a spatially coherent label configuration. The difficulty of minimizing the nonconvex energy (1) has motivated researchers to develop convex reformulations.

Convex representations of (1) and more general related energies have been studied in the context of the calibration method for the Mumford-Shah functional [1]. Based on these works, relaxations for the piecewise constant [15] and piecewise smooth Mumford-Shah functional [16] have been proposed. Inspired by Ishikawa’s graph-theoretic globally optimal solution to discrete variants of (1), continuous analogues have been considered by Pock et al. in [17][18]. Con-
tinuous relaxations for multilabeling problems with finite label spaces \( \Gamma \) have also been studied in [11].

Interestingly, the discretization of the aforementioned continuous relaxations is very similar to the linear programming relaxations proposed for MAP inference in the Markov Random Field (MRF) community [10,22,24,26]. Both approaches ultimately discretize the range \( \Gamma \) into a finite set of labels. A closer analysis of these relaxations reveals, however, that they are not well-suited to represent the continuous valued range that we face in most computer vision problems such as stereo matching or optical flow. More specifically, the above relaxations are not designed to assign meaningful cost values to non-integral configurations. As a result, a large number of labels is required to achieve a faithful approximation. Solving real-world vision problems therefore entails large optimization problems with high memory and runtime requirement. To address this problem, Zach and Kohli [27], Zach [25] and Fix and Agarwal [7] introduced MRF-based approaches which retain continuous label spaces after discretization. For manifold-valued labels, this issue was addressed by Lellmann et al. [12], however with the sole focus on the regularizer.

1.1. Contributions

We propose the first sublabel–accurate convex relaxation of nonconvex problems in a spatially continuous setting. It exhibits several favorable properties:

- In contrast to existing spatially continuous lifting approaches [17,18], the proposed method provides substantially better solutions with far fewer labels – see Fig. 1. This provides savings in runtime and memory.
- In Sec. 3 we show that the functional lifting methods [17,18] are a special case of the proposed framework.
- In Sec. 3 we show that, in a local sense, our formulation is the tightest convex relaxation which takes dataterm and regularizer into account separately. It is unknown whether this “local convex envelope” property also holds for the discrete approach [27].
- Our formulation is compact and requires only half the amount of variables for the dataterm than the formulation in [27]. We prove that the sublabel–accurate total variation can be represented in a very simple way, introducing no overhead compared to [17,18]. In contrast, the regularizer in [27] is much more involved.
- Since our method is derived in a spatially continuous setting, the proposed approach easily allows different gradient discretizations. In contrast to [25,27] the regularizer is isotropic leading to noticeably less grid bias.

2. Notation and Mathematical Preliminaries

We make heavy use of the convex conjugate, which is given as \( f^*(y) = \sup_{x \in \mathbb{R}^n} \langle y, x \rangle - f(x) \) for functions \( f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \). The biconjugate \( f^{**} \) denotes its convex envelope, i.e. the largest lower-semicontinuous convex under-approximation of \( f \). For a set \( C \) we denote by \( \delta_C \) the function which maps any element from \( C \) to 0 and is \( \infty \) otherwise. For a comprehensive introduction to convex analysis, we refer the reader to [19]. Vector valued functions \( \mathbf{u} : \Omega \to \mathbb{R}^k \) are written in bold symbols. If it is clear from the context, we will drop the \( x \in \Omega \) inside the functions, e.g., we write \( \rho(\mathbf{u}) \) for \( \rho(x, \mathbf{u}(x)) \).

3. Functional Lifting

To derive a convex representation of [1], we rely on the framework of functional lifting. The idea is to reformulate the optimization problem in a higher dimensional space, in which the convex envelope approximates the nonconvex energy better than the one of the original low dimensional energy. We start by sampling the range \( \Gamma \) at \( L = k + 1 \) labels \( \gamma_1 < \ldots < \gamma_L \in \Gamma \). This partitions the range into \( k \) intervals \( \Gamma_i = [\gamma_i, \gamma_{i+1}] \) so that \( \Gamma = \Gamma_1 \cup \ldots \cup \Gamma_k \). Clearly, any
value in the range of \( u : \Omega \rightarrow \Gamma \) can be written as
\[
u(x) = \gamma_i^\alpha := \gamma_i + \alpha(\gamma_{i+1} - \gamma_i),
\]
(2)
for \( \alpha \in [0, 1] \) and some label index \( 1 \leq i \leq k \). We represent such a value in the range \( \Gamma \) by a \( k \)-dimensional vector
\[
u(x) = 1_i^\alpha := \alpha 1_i + (1 - \alpha)1_{i-1},
\]
(3)
where \( 1_i \) denotes a vector starting with \( i \) ones followed by \( k - i \) zeros. We call \( u : \Omega \rightarrow \mathbb{R}^k \) the lifted representation of \( u \), representing the graph of \( u \). This notation is depicted in Fig. 2 for \( k = 4 \). Back-projecting the lifted \( u(x) \) to the range of \( u \) using the layer cake formula yields a one-to-one correspondence between \( u(x) = \gamma_i^\alpha \) and \( u(x) = 1_i^\alpha \) via
\[
u(x) = \gamma_1 + \sum_{i=1}^{k} u_i(x)(\gamma_{i+1} - \gamma_i).
\]
(4)
We now formulate problem (1) in terms of such graph functions, a technique that is common in the theory of Cartesian currents \[8\].

3.1. Convexification of the Dataterm

For now, we consider a fixed \( x \in \Omega \). Then the dataterm from (1) is a possibly nonconvex real-valued function (cf. Fig. 3) that we seek to minimize over a compact interval \( \Gamma \):
\[
\min_{u \in \Gamma} \rho(u).
\]
(5)
Due to the one-to-one correspondence between \( \gamma_i^\alpha \) and \( 1_i^\alpha \) it is clear that solving problem (5) is equivalent to finding a minimizer of the lifted energy:
\[
\rho(u) = \min_{1 \leq i \leq k} \rho_i(u),
\]
(6)
\[
\rho_i(u) = \begin{cases}
\rho(\gamma_i^\alpha), & \text{if } u = 1_i^\alpha, \; \alpha \in \{0, 1\}, \\
\infty, & \text{else}.
\end{cases}
\]
(7)
Note that the constraint in (7) is essentially the nonconvex special ordered set of type 2 (SOS2) constraint \[3\]. More precisely, we demand that the derivative \( \partial_\gamma u \) is zero, except for two neighboring elements, which add up to one. In the following proposition, we derive the tightest convex relaxation of \( \rho \).

Proposition 1. The convex envelope of \( \rho(u) \) is given as:
\[
\rho^*(u) = \sup_{v \in \mathbb{R}^k} (u, v) - \max_{1 \leq i \leq k} \rho_i^*(v),
\]
(8)
where the conjugate of the individual \( \rho_i \) is
\[
\rho_i^*(v) = c_i(v) + \rho_i^*(\frac{v_i}{\gamma_{i+1} - \gamma_i}),
\]
(9)
with \( c_i(v) = (1_{i-1}, v) - \frac{\gamma_i}{\gamma_{i+1} - \gamma_i} \) and \( \rho_i = \rho + \delta_{\gamma_i} \).

Proof. See appendix.

The above proposition reveals that the convex relaxation implicitly convexifies the dataterm \( \rho \) on each interval \( \Gamma_i \). The equality \( \rho_i^* = \rho_i^{**} \) implies that starting with \( \rho_i \) yields exactly the same convex relaxation as starting with \( \rho_i^* \).

Corollary 1. If \( \rho \) is linear on each \( \Gamma_i \), then the convex envelopes of \( \rho(u) \) and \( \sigma(u) \) coincide, where the latter is:
\[
\sigma(u) = \begin{cases}
\rho(\gamma_i^\alpha), & \text{if } \exists i : u = 1_i^\alpha, \; \alpha \in \{0, 1\}, \\
\infty, & \text{else}.
\end{cases}
\]
(10)

Proof. Consider an additional constraint \( \delta_{\gamma_i, \gamma_{i+1}} \) for each \( \rho_i \), which corresponds to selecting \( \alpha \in \{0, 1\} \) in (7). The fact that our relaxation is independent of whether we choose \( \rho_i \) or \( \rho_i^{**} \), along with the fact that the convex hull of two points is a line, yields the assertion.

For the piecewise linear case, it is possible to find an explicit form of the biconjugate.
Proposition 2. Let us denote by \( r \in \mathbb{R}^k \) the vector with
\[
  r_i = \rho(\gamma_{i+1}) - \rho(\gamma_i), \quad 1 \leq i \leq k.
\]
Under the assumptions of Prop. 7 one obtains:
\[
  \sigma^{**}(u) = \begin{cases} 
    \rho(\gamma_1) + \langle u, r \rangle, & \text{if } u_i \geq u_{i+1}, u_i \in [0, 1], \\
    \infty, & \text{else}.
  \end{cases}
\]
(12)

Proof. See appendix. \( \square \)

Up to an offset (which is irrelevant for the optimization), one can see that (12) coincides with the dataterm of (15), the discretizations of [17, 18], and \( - \) after a change of variable \( - \) with [11]. This not only proves that the latter is optimizing a convex envelope, but also shows that our method naturally generalizes the work from piecewise linear to arbitrary piecewise convex energies. Fig. 3a and Fig. 3b illustrate the difference of \( \sigma^{**} \) and \( \rho^{**} \) on the example of a nonconvex stereo matching cost.

Because our method allows arbitrary convex functions on each \( \Gamma_i \), we can prove that, for the two label case, our approach optimizes the convex envelope of the dataterm.

Proposition 3. In the case of binary labeling, i.e., \( L = 2 \), the convex envelope of (6) reduces to
\[
  \rho^{**}(u) = \rho^{**}(\gamma_1 + u(\gamma_2 - \gamma_1)), \quad \text{with } u \in [0, 1].
\]
(13)

Proof. See appendix. \( \square \)

3.2. A Lifted Representation of the Total Variation

We now want to find a lifted convex formulation that emulates the total variation regularization in [1]. We follow [5] and define an appropriate integrand of the functional
\[
  TV(u) = \int_{\Omega} \Phi(x, Dv).
\]
(14)

where the distributional derivative \( Dv \) is a finite \( \mathbb{R}^{k\times d} \)-valued Radon measure [2]. We define
\[
  \Phi(g) = \min_{1 \leq i \leq j \leq k} \Phi_{i,j}(g).
\]
(15)

The individual \( \Phi_{i,j} : \mathbb{R}^{k\times d} \to \mathbb{R} \cup \{\infty\} \) are given by:
\[
  \Phi_{i,j}(g) = \begin{cases} 
    |\gamma_i^\alpha - \gamma_j^\beta| \cdot |\nu|_2, & \text{if } g = (1_i^\alpha - 1_j^\beta) \nu^T, \\
    \infty, & \text{else}.
  \end{cases}
\]
(16)

![Figure 4. Illustration of the epigraph projection. In the left subfigure the projection onto the epigraph of the conjugate of a convex quadratic \( \rho_i \), is shown. In the right subfigure the piecewise linear case is illustrated. In both cases all points that lie in the gray sets are orthogonally projected onto the respective linear parts whereas the points that lie in the green sets are projected onto the parabolic part (in the quadratic case) respectively the kinks (in the piecewise linear case). In the piecewise linear case the green sets are normal cones. The red dashed lines correspond to the boundary cases. \( \gamma_i, \gamma_{i+1}, \mu_1, \mu_2 \) are the slopes of the segments of \( \rho_i \), respectively the (sub-)label positions of \( \rho_i \).](image)

for some \( \alpha, \beta \in [0, 1] \) and \( \nu \in \mathbb{R}^d \). The intuition is that \( \Phi_{i,j} \) penalizes a jump from \( \gamma_i^\alpha \) to \( \gamma_j^\beta \) in the direction of \( \nu \). Since \( \Phi \) is nonconvex we compute the convex envelope.

Proposition 4. The convex envelope of (15) is
\[
  \Phi^{**}(g) = \sup_{q \in K} \langle q, g \rangle,
\]
(17)

where \( K \subset \mathbb{R}^{k\times d} \) is given as:
\[
  K = \left\{ q \in \mathbb{R}^{k\times d} \mid \left| q^T (1_i^\alpha - 1_j^\beta) \right|_2 \leq |\gamma_i^\alpha - \gamma_j^\beta|, \quad \forall 1 \leq i \leq j \leq k, \quad \forall \alpha, \beta \in [0, 1] \right\}.
\]
(18)

Proof. See appendix. \( \square \)

The set \( K \) from Eq. (16) involves infinitely many constraints which makes numerical optimization difficult. As the following proposition reveals, the infinite number of constraints can be reduced to only linearly many, allowing to enforce the constraint \( q \in K \) exactly.

Proposition 5. In case the labels are ordered, i.e., \( \gamma_1 < \gamma_2 < \ldots < \gamma_L \), then the constraint set \( K \) from Eq. (16) is equal to
\[
  K = \{ q \in \mathbb{R}^{k\times d} \mid |q_i|_2 \leq \gamma_{i+1} - \gamma_i, \quad \forall i \}.
\]
(19)

Proof. See appendix. \( \square \)
Proposed \( \Omega \rightarrow TV \) to unlifted \( L = 256 \) the quadratic dataterm, it fails to reach that optimality gap even for the baseline method optimizes a piecewise linear approximation of a solution within a certain energy gap to the optimal solution. As we show the time in seconds required for each method to produce method to the baseline method \([17]\) on the convex ROF problem.

Figure 5. Denoising comparison. We compare the proposed with Prop. 3 show that for \( k \) on (16) for \( \alpha \) the total variation from \([5]\), where it has been derived based lifting method can solve a large class of non-convex problems, it is almost as efficient as direct methods on convex problems.

This shows that the proposed regularizer coincides with the total variation from \([5]\), where it has been derived based on \([16]\) for \( \alpha \) and \( \beta \) restricted to \([0, 1]\). Prop. 3 together with Prop. 5 show that for \( k = 1 \) our formulation amounts to unlifted \( TV \) optimization with a convexified dataterm.

Figure 6. Denoising using a robust truncated quadratic dataterm. The top row shows the input image along with the result obtained given below each image. The energy of the final solution as well as the total runtime are given below each image.

\[ \rho^*_i \] with a maximum over a real number \( q \in \mathbb{R} \) and obtain the following saddle point formulation of problem (20):

\[
\min_{u : \Omega \to \mathbb{R}^k} \max_{(v, q) \in C} \left\langle u, v \right\rangle - \sum_{x \in \Omega} q(x) + \left\langle p, \nabla u \right\rangle, \quad (21)
\]

\[
C = \{ (v, q) : \Omega \to \mathbb{R}^k \times \mathbb{R} \mid q(x) \geq \rho^*_i(v(x)), \forall x, \forall i \}. \quad (22)
\]

We numerically compute a minimizer of problem (21) using a first-order primal-dual method \([6] [10]\) with diagonal preconditioning \([14]\) and adaptive steps \([9]\). It basically alternates between a gradient descent step in the primal variable and a gradient ascent step in the dual variable. Subsequently the dual variables are orthogonally projected onto the sets \( C \) respectively \( K \). The projection onto the set \( K \) is a simple \( \ell_2 \)-ball projection. To simplify the projection onto \( C \), we transform the \( k \)-dimensional epigraph constraints in (22) into 1-dimensional scaled epigraph constraints by introducing an additional variable \( z : \Omega \to \mathbb{R}^k \) with:

\[
z_i(x) = [q(x) - c_i(v(x))](\gamma_{i+1} - \gamma_i). \quad (23)
\]
Using equation (9) we can now rewrite the constraints in (22) as

$$z_i(x) \geq \frac{\nu_i(x)}{\gamma_i+1 - \gamma_i} \cdot \rho_i^* \left( \frac{\nu_i(x)}{\gamma_i+1 - \gamma_i} \right).$$

We implement the newly introduced equality constraints (23) introducing a Lagrange multiplier $s : \Omega \to \mathbb{R}^k$. It remains to discuss the orthogonal projections onto the epigraphs of the conjugates $\rho_i^*$. Currently we support quadratic and piecewise linear convex pieces $\rho_i$. For the piecewise linear case, the conjugate $\rho_i^*$ is a piecewise linear function with domain $\mathbb{R}$. The slopes correspond to the $x$-positions of the sublabels and the intercepts correspond to the function values at the sublabel positions.

The conjugates as well as the epigraph projections of both, a quadratic and a piecewise linear convex piece are depicted in Fig. 4. For the quadratic case, the projection onto the epigraph of a parabola is computed using [23, Appendix B.2].

### 5. Experiments

We implemented the primal-dual algorithm in CUDA to run on GPUs. For $d = 2$, our implementation of the functional lifting framework [17], which will serve as a baseline method, requires $4N(L-1)$ optimization variables, while the proposed method requires $6N(L-1) + N$ variables, where $N$ is the number of points used to discretize the domain $\Omega \subset \mathbb{R}^d$. As we will show, our method requires much fewer labels to yield comparable results, thus, leading to an improvement in accuracy, memory usage, and speed.

#### 5.1. Rudin-Osher-Fatemi Model

As a proof of concept, we first evaluate the novel relaxation on the well-known Rudin-Osher-Fatemi (ROF) model [20]. It corresponds to (1) with the following dataterm:

$$\rho(x,u(x)) = (u(x) - f(x))^2,$$

where $f : \Omega \to \mathbb{R}$ denotes the input data. While there is no practical use in applying convex relaxation methods to an already convex problem such as the ROF model, the purpose of this is two-fold. Firstly, it allows us to measure the overhead introduced by our method by comparing it to standard convex optimization methods which do not rely on functional lifting. Secondly, we can experimentally verify that the relaxation is tight for a convex dataterm.

In Fig. 5 we solve (25) directly using the primal-dual algorithm [9], using the baseline functional lifting method [17] and using our proposed algorithm. First, the globally optimal energy was computed using the direct method with a very high number of iterations. Then we measure how long each method took to reach this global optimum to a fixed tolerance.

The baseline method fails to reach the global optimum even for 256 labels. While the lifting framework introduces a certain overhead, the proposed method finds the same globally optimal energy as the direct unlifted optimization approach and generalizes to nonconvex energies.
5.2. Robust Truncated Quadratic Dataterm

The quadratic dataterm in (25) is often not well suited for real-world data as it comes from a pure Gaussian noise assumption and does not model outliers. We now consider a robust truncated quadratic dataterm:

$$\rho(x, u(x)) = \frac{\alpha}{2} \min \left\{ (u(x) - f(x))^2, \nu \right\}. \tag{26}$$

To implement (26), we use a piecewise polynomial approximation of the dataterm. In Fig. 6 we degraded the input image with additive Gaussian and salt and pepper noise. The parameters in (26) were chosen as $\alpha = 25$, $\nu = 0.025$, and $\lambda = 1$. The proposed method requires significantly less labels to find lower energies than the baseline method.

5.3. Comparison to the Method of Zach and Kohli

We remark that Prop. 4 and Prop. 5 hold for arbitrary convex one-homogeneous functionals $\phi(\nu)$ instead of $|\nu|_2$ in equation (16). In particular, they hold for the anisotropic total variation $\phi(\nu) = |\nu|_1$. This generalization allows us to directly compare our convex relaxation to the MRF approach of Zach and Kohli [27]. In Fig. 7, we show the results of optimizing the two models entitled “DC-Linear” and “DC-MRF” proposed in [27], and of our proposed method with anisotropic regularization on the robust truncated denoising energy (26). We picked the parameters as $\alpha = 0.2$, $\nu = 500$, and $\lambda = 1$. The label space is also chosen as $\Gamma = [0, 256]$ as described in [27]. Overall, all the energies are better than the ones reported in [27]. It can be seen from Fig. 7 that the proposed relaxation is competitive to the one proposed in [27]. In addition, the proposed relaxation uses a more compact representation and extends to isotropic regularizers. To illustrate the advantages of isotropic regularizations, Fig. 8a and Fig. 8b show a comparison of our proposed method for isotropic and anisotropic regularization in the next section.

5.4. Stereo Matching

Given a pair of rectified images, the task of finding a correspondence between the two images can be formulated as an optimization problem over a scalar field $u: \Omega \rightarrow \Gamma$ where each point $u(x) \in \Gamma$ denotes the displacement along the epipolar line associated with each $x \in \Omega$. The overall cost functional fits Eq. (1). In our experiments, we computed $\rho(x, u(x))$ for 270 disparities on the Middlebury stereo benchmark [21] in a $4 \times 4$ patch using a truncated sum of absolute gradient differences. For the dataterm of the proposed relaxation, we convexify the matching cost $\rho$ in each range $\Gamma_i$ by numerically computing the convex envelope using the gift wrapping algorithm.

The first row in Fig. 9 shows the result of the proposed relaxation using the convexified energy between two labels. The second row shows the baseline approach using the same amount of labels. Even for $L = 2$, the proposed method produces a reasonable depth map while the baseline approach basically corresponds to a two region segmentation.

5.5. Phase Unwrapping

Many sensors such as time-of-flight cameras or interferometric synthetic aperture radar (SAR) yield cyclic data ly-
5.6. Depth From Focus

In depth from focus the task is to recover the depth of a scene, given a stack of images each taken from a constant position but in a different focal setting, so that in each image only the objects of a certain depth are sharp. We achieve this by estimating the depth of a point by locally maximizing its contrast over the set of images. We compute the cost by using the modified Laplacian function \([13]\) as a contrast measure. Similar to the stereo experiments, we convexify the cost on each label range by computing the convex hull. The results are shown in Fig. 10. While the baseline method clearly shows the label space discretization, the proposed approach yields a smooth depth map. Since the proposed method uses a convex lower bound of the lifted energy, the regularizer has slightly more influence on the final result. This explains why the resulting depth maps in Fig. 10 and Fig. 9 look overall less noisy.

For the experiments, we approximated the nonconvex energy by quadratic pieces as depicted in Fig. 11. The label space is chosen as \(\Gamma = [0, 4\pi]\) and the regularization parameter was set to \(\lambda = 0.005\). Again, it is visible in Fig. 11 that the baseline method shows label space discretization and fails to unwrap the depth map correctly if the number of labels is chosen too low. The proposed method yields a smooth unwrapped result using only 8 labels.

6. Conclusion

In this work we proposed a tight convex relaxation that can be interpreted as a sublabel–accurate formulation of classical multilabel problems. We showed that the local convex envelope involves infinitely many constraints, however we proved that it suffices to consider linearly many of those. The final formulation is a simple saddle-point problem that admits fast primal-dual optimization. Our method maintains sublabel accuracy even after discretization and for that reason outperforms existing spatially continuous methods. Interesting directions for future work include higher dimensional label spaces, manifold valued data and more general regularizers.
A. Appendix

Proof of Proposition 1. The proof follows from a direct calculation. We start with the definition of the biconjugate:

\[
\rho^{**}(u) = \sup_{v \in \mathbb{R}^k} \langle u, v \rangle - \left( \min_{1 \leq i \leq k} \rho_i(u) \right)^* = \sup_{v \in \mathbb{R}^k} \langle u, v \rangle - \max_{1 \leq i \leq k} \rho_i^*(u). 
\] (28)

This shows the first equation inside the proposition. For the individual \( \rho_i^* \) we again start with the definition of the convex conjugate:

\[
\rho_i^*(v) = \sup_{\alpha \in [0,1]} \langle \alpha 1_i + (1 - \alpha) 1_{i-1}, v \rangle - \rho(\alpha \gamma_i + (1 - \alpha) \gamma_i) 
= \sup_{\alpha \in [0,1]} \langle 1_{i-1}, v \rangle + \alpha v_i - \rho(\gamma_i^\alpha). 
\] (29)

Applying the substitution \( \gamma_i^\alpha = \alpha \gamma_i + (1 - \alpha) \gamma_i \) and consequently \( \alpha = \frac{\gamma_i^\alpha - \gamma_i}{\gamma_{i+1} - \gamma_i} \) yields:

\[
\rho_i^*(v) = \sup_{\gamma_i^\alpha \in \Gamma_i} \langle 1_{i-1}, v \rangle + \frac{\gamma_i^\alpha - \gamma_i}{\gamma_{i+1} - \gamma_i} v_i - \rho(\gamma_i^\alpha) \\
= \langle 1_{i-1}, v \rangle - \frac{\gamma_i}{\gamma_{i+1} - \gamma_i} v_i + \sup_{\gamma_i^\alpha \in \Gamma_i} \frac{\gamma_i^\alpha}{\gamma_{i+1} - \gamma_i} v_i - \rho(\gamma_i^\alpha) \\
= \langle 1_{i-1}, v \rangle - \frac{\gamma_i}{\gamma_{i+1} - \gamma_i} v_i + (\rho + \delta_{\Gamma_i})^* \left( \frac{v_i}{\gamma_{i+1} - \gamma_i} \right) \\
= c_i(v) + \rho_i^* \left( \frac{v_i}{\gamma_{i+1} - \gamma_i} \right). 
\] (30)

Proof of Proposition 2. It is easy to see that

\[
\sigma^*(v) = \max_{i \in \{1, ..., L\}} \left( \sum_{i=1}^{L-1} v_i - \rho(\gamma_i) \right). 
\]

To compute the biconjugate, we write any input argument \( u = \sum_{i=1}^k u_i 1_{i+1} \), and use \( \sigma^{**} = \rho^{**} \) to obtain

\[
\rho^{**}(u) = \sup_{v} \langle u, v \rangle - \max_{i \in \{1, ..., L\}} \left( \sum_{i=1}^{L-1} v_i - \rho(\gamma_i) \right) \\
= \sup_{v} \sum_{i=1}^k \mu_i \sum_{l=1}^{L-1} v_{1l} - \max_{i \in \{1, ..., L\}} \left( \sum_{i=1}^{L-1} v_i - \rho(\gamma_i) \right). 
\]

Instead of taking the supremum of all \( v \), we might as well take the supremum over all vectors \( p \) with \( p_i = \sum_{l=1}^{L} v_{1l} \).

Care has to be taken of the first summmand in the second term of the above formulation. We obtain

\[
\sup_{v} \sum_{i=1}^k \mu_i \sum_{l=1}^{L-1} v_{1l} - \max_{i \in \{1, ..., L\}} \left( \sum_{i=1}^{L-1} v_i - \rho(\gamma_i) \right), \\
= \sup_{p} \sum_{i=1}^k \mu_i p_i - \max_{i \in \{2, ..., L\}} \max(p_{i-1} - \rho(\gamma_i), -\rho(\gamma_i)), \\
= \sum_{i=1}^k \mu_i p_i - \max_{i \in \{1, ..., k\}} \max(p_i - \rho(\gamma_{i+1}), -\rho(\gamma_i)), \\
= \sum_{i=1}^k \mu_i \rho(\gamma_{i+1}) \\
+ \sup_{p} \sum_{i=1}^k \mu_i p_i - \max_{i \in \{1, ..., k\}} \max(p_i, -\rho(\gamma_i)), 
\]

Note that for any \( \mu_i \) being negative, the supremum immediately yields infinity by taking \( p_i \to -\infty \). Similarly, if \( \sum_{i=1}^k \mu_i > 1 \) yields infinity by taking all \( p_i \to -\infty \).

For \( \mu_i \geq 0 \) for all \( i \), and \( \sum_{i=1}^k \mu_i \leq 1 \), we know that \( \sum_{i=1}^k \mu_i p_i \leq (\max_i p_i) \sum_{i=1}^k \mu_i \). Since equality can be obtained by choosing \( p_i = \max_i p_i \) for all \( i \), we can reduce the above supremum to

\[
\sup_{z} \left( z \sum_{i=1}^k \mu_i - \max(z, -\rho(\gamma_1)) \right) = \left( 1 - \sum_{i=1}^k \mu_i \right) \rho(\gamma_1), 
\]

where we used that the supremum over \( z \) is attained at \( z = -\rho(\gamma_1) \) (still assuming that \( \sum_{i=1}^k \mu_i \leq 1 \)). Let us now undo our change of variable. It is easy to see that \( \mu_k = u_k \), and \( \mu_i = u_i - u_{i+1} \) for \( i = 1, ..., k - 1 \). The latter leads to

\[
\sum_{i=1}^k \mu_i \rho(\gamma_{i+1}) + \left( 1 - \sum_{i=1}^k \mu_i \right) \rho(\gamma_1) \\
= \rho(\gamma_k) + \sum_{i=1}^{k-1} (u_i - u_{i+1}) \rho(\gamma_{i+1}) + (1 - u_1)\rho(\gamma_1) \\
= \rho(\gamma_1) + \langle u, r \rangle, 
\]

for \( r_i = \rho(\gamma_{i+1}) - \rho(\gamma_i) \). Considering the aforementioned constraints of \( \mu_i \geq 0 \), and \( \sum_{i=1}^k \mu_i \leq 1 \), we finally find

\[
\rho^{**}(u) = \begin{cases} 
\rho(\gamma_1) + \langle u, r \rangle & \text{if } 1 \geq u_1 \geq \ldots \geq u_k \geq 0, \\
-\infty & \text{else}.
\end{cases} 
\]

Proof of Proposition 3. For the special case \( k = 1 \) the biconjugate from (28) is just:

\[
\rho^{**}(u) = \sup_{v \in \mathbb{R}} uv - \rho_i^*(v) = \rho_i^{**}(u). 
\] (31)
Now using the first line in (30), $\rho_1^* \Phi$ becomes:

$$
\rho_1^*(u) = \sup_{v \in \mathbb{R}} uv - \sup_{\gamma \in \Gamma} \frac{\gamma - \gamma_1}{\gamma_2 - \gamma_1} v - \rho(\gamma)
$$

$$
= \sup_{v \in \mathbb{R}} v \left( u + \frac{\gamma_1}{\gamma_2 - \gamma_1} \right) - \sup_{\gamma \in \Gamma} \frac{\gamma}{\gamma_2 - \gamma_1} v - \rho(\gamma)
$$

$$
= \sup_{v \in \mathbb{R}} v \left( u + \frac{\gamma_1}{\gamma_2 - \gamma_1} \right) - \rho^* \left( \frac{v}{\gamma_2 - \gamma_1} \right)
$$

$$
= \sup_{\tilde{v} \in \mathbb{R}} \tilde{v} (\gamma_1 + u(\gamma_2 - \gamma_1)) - \rho^* (\tilde{v})
$$

$$
= \rho^* (\gamma_1 + u(\gamma_2 - \gamma_1)),
$$

where we used $\text{dom}(\rho) = \Gamma$ as well as the substitution $v = (\gamma_2 - \gamma_1) \tilde{v}$.

**Proof of Proposition 4.** We compute the individual conjugate as:

$$
\Phi_{i,j}^* (q) = \sup_{\alpha, \beta \in [0,1]} \sup_{\nu \in \mathbb{R}^d} \langle q, (1_i^\alpha - 1_j^\beta) \nu^T \rangle - |\nu|_2
$$

$$
= \sup_{\alpha, \beta \in [0,1]} \sup_{\nu \in \mathbb{R}^d} \langle q^T, (1_i^\alpha - 1_j^\beta), \nu \rangle - |\nu|_2
$$

$$
= \sup_{\alpha, \beta \in [0,1]} \sup_{\nu \in \mathbb{R}^d} \langle q^T, (1_i^\alpha - 1_j^\beta), \nu \rangle - |\nu|_2.
$$

The inner maximum over $\nu$ is the conjugate of the $\ell_2$-norm scaled by $|\gamma_i^\alpha - \gamma_j^\beta|$ evaluated at $q^T \left( 1_i^\alpha - 1_j^\beta \right)$. This yields:

$$
\Phi_{i,j}^* (q) = \begin{cases} 
0, & \text{if } |q^T (1_i^\alpha - 1_j^\beta)|_2 \leq |\gamma_i^\alpha - \gamma_j^\beta|, \\
\infty, & \text{else}.
\end{cases}
$$

For the overall biconjugate we have:

$$
\Phi^{**} (g) = \sup_{q \in \mathbb{R}^{k \times d}} \langle g, q \rangle - \max_{1 \leq i, j \leq k} \Phi_{i,j}^* (q)
$$

$$
= \sup_{q \in \mathbb{R}^{k \times d}} \langle g, q \rangle.
$$

Since we have the max over all $1 \leq i, j \leq k$ conjugates, the set $\mathcal{K}$ is given as the intersection of the sets described by the individual indicator functions $\Phi_{i,j}$:

$$
\mathcal{K} = \left\{ q \in \mathbb{R}^{k \times d} \mid |q^T (1_i^\alpha - 1_j^\beta)|_2 \leq |\gamma_i^\alpha - \gamma_j^\beta|, \forall 1 \leq i \leq j \leq k, \forall \alpha, \beta \in [0,1] \right\}.
$$

**Proof of Proposition 5.** First we rewrite (36) by expanding the matrix-vector product into sums:

$$
\sum_{l=j}^{i-1} q_l + \alpha q_i - \beta q_j \leq |\gamma_i^\alpha - \gamma_j^\beta|, \forall 1 \leq j \leq i \leq k, \forall \alpha, \beta \in [0,1].
$$

(37)

Since the other cases for $1 \leq i \leq j \leq k$ in (36) are equivalent to (37), it is enough to consider (37) instead of (36).

Let $\gamma_1 < \gamma_2 < \ldots < \gamma_L$. We show the equivalences:

$$
\text{(37)} \iff
$$

$$
\sum_{l=j}^{i} q_l \leq \gamma_{i+1} - \gamma_j, \forall 1 \leq j \leq i \leq k.
$$

(38)

The direction “(37) $\Rightarrow$ (38)” follows by setting $\alpha = 1$ and $\beta = 0$ in (37), and “(38) $\Rightarrow$ (39)” follows by setting $i = j$ in (38).

The direction “(39) $\Rightarrow$ (38)” can be proven by a quick calculation:

$$
\sum_{l=j}^{i} q_l \leq \sum_{l=j}^{i} |q_l|_2 \leq \sum_{l=j}^{i} |\gamma_{i+1} - \gamma_l| = \gamma_{i+1} - \gamma_j.
$$

(40)

It remains to show “(38) $\Rightarrow$ (37)”. We start with the case $j = i$:

$$
|\alpha q_i - \beta q_i|_2 = |\alpha - \beta||q_i|_2
$$

$$
\leq |\alpha - \beta|(|\gamma_{i+1} - \gamma_i|
$$

$$
= |(\gamma_{i+1} - \gamma_i)\alpha - (\gamma_{i+1} - \gamma_i)\beta|
$$

$$
= |(\alpha - \beta)(\gamma_{i+1} - \gamma_i)| = |\gamma_i^\alpha - \gamma_i^\beta|.
$$

(41)

Now let $j < i$. Since $\gamma_j < \gamma_i$, it also holds that $\gamma_j^\beta \leq \gamma_i^\alpha$, thus it is equivalent to show (37) without the absolute value on the right hand side.
First we show that \( \alpha, \beta \in \{0, 1\} \) and \( \alpha, \beta \in [0, 1] \):
\[
\sum_{i=1}^{i-1} q_{i} + \alpha q_{i} + (1 - \beta)q_{j} \leq \sum_{i=1}^{i-1} q_{i} + (1 - \beta)q_{j} + \alpha|q_{j}|_{2}
\]
for \( \beta = 0 \) or \( \beta = 1 \)
\[
\leq \gamma_{i} + (1 - \beta)(\gamma_{j} - \beta + \alpha(\gamma_{i+1} - \gamma_{i})) \approx \gamma_{i}^{\alpha} - \gamma_{j}^{\beta}.
\]
Using a similar argument we show that, using the above, \( \beta \rightarrow \beta \) for all \( \alpha, \beta \in [0, 1] \).

\[
\sum_{i=1}^{i-1} q_{i} + \alpha q_{i} + (1 - \beta)q_{j} \leq \sum_{i=1}^{i-1} q_{i} + (1 - \beta)q_{j} + \alpha|q_{j}|_{2}
\]
using \( \beta = 1 \)
\[
\leq \gamma_{i}^{\alpha} - \gamma_{j}^{\beta}.
\]

\[
\square
\]

References

[1] G. Alberti, G. Bouchitté, and G. D. Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Calc. Var. Partial Dif.*, 3(16):299–333, 2003.

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

[3] E. Beale and J. Tomlin. Special facilities in a general mathematical programming system for nonconvex problems using ordered sets of variables. *Proceedings of the fifth international conference on operational research*, pages 447–454, 1970.

[4] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9:263–340, 2010.

[5] A. Chambolle, D. Cremers, and T. Pock. A convex approach to minimal partitions. *SIAM Journal on Imaging Sciences*, 5(4):1113–1158, 2012.

[6] E. Esser, X. Zhang, and T. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[7] A. Fix and S. Agarwal. Duality and the continuous graphical model. In *Computer Vision ECCV 2014*, volume 8691 of *Lecture Notes in Computer Science*, pages 266–281. Springer International Publishing, 2014.

[8] M. Giaquinta, G. Modica, and J. Souček. *Cartesian currents in the calculus of variations I, II.*, volume 37-38 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3*. Springer-Verlag, Berlin, 1998.

[9] T. Goldstein, E. Esser, and R. Baraniuk. Adaptive Primal-Dual Hybrid Gradient Methods for Saddle-Point Problems. *arXiv Preprint*, 2013.

[10] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(10):1333–1336, 2003.

[11] J. Lellmann and C. Schnörr. Continuous multiclass labeling approaches and algorithms. *SIAM J. Imaging Sciences*, 4(4):1049–1096, 2011.

[12] J. Lellmann, E. Strekalovskiy, S. Koetter, and D. Cremers. Total variation regularization for functions with values in a manifold. In *ICCV*, December 2013.

[13] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(8):824–831, Aug. 1994.

[14] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *ICCV*, pages 1762–1769, 2011.

[15] T. Pock, A. Chambolle, H. Bischof, and D. Cremers. A convex relaxation approach for computing minimal partitions. In *CVPR*, pages 810–817, 2009.

[16] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the piecewise smooth Mumford-Shah functional. In *ICCV*, 2009.

[17] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM J. Imaging Sci.*, 3(4):1122–1145, 2010.

[18] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. A convex formulation of continuous multi-label problems. In *European Conference on Computer Vision (ECCV)*, Marseille, France, October 2008.

[19] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[20] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
[21] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nei, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In German Conference on Pattern Recognition, volume 8753, pages 31–42. Springer, 2014.

[22] M. Schlesinger. Sintaksicheskiy analiz dvumernykh zritel-nikh signalov v usloviyakh pomekh (Syntactic analysis of two-dimensional visual signals in noisy conditions). Kibernetika, 4:113–130, 1976.

[23] E. Strekalovskiy, A. Chambolle, and D. Cremers. Convex relaxation of vectorial problems with coupled regularization. SIAM Journal on Imaging Sciences, 7(1):294–336, 2014.

[24] T. Werner. A linear programming approach to max-sum problem: A review. IEEE Trans. Pattern Anal. Mach. Intell., 29(7):1165–1179, July 2007.

[25] C. Zach. Dual decomposition for joint discrete-continuous optimization. In AISTATS, pages 632–640, 2013.

[26] C. Zach, C. Hane, and M. Pollefeys. What is optimized in tight convex relaxations for multi-label problems? In CVPR, pages 1664–1671, 2012.

[27] C. Zach and P. Kohli. A convex discrete-continuous approach for markov random fields. In ECCV, volume 7577, pages 386–399. Springer Berlin Heidelberg, 2012.