Sublabel-Accurate Multilabeling Meets Product Label Spaces

Zhenzhang Ye[1(✉)], Bjoern Haefner[1], Yvain Quéau[2], Thomas Möllenhoff[3], and Daniel Cremers[1]

[1] TU Munich, Garching, Germany
`yez@in.tum.de`
[2] GREYC, UMR CNRS 6072, Caen, France
[3] RIKEN Center for AI Project, Tokyo, Japan

Abstract. Functional lifting methods are a promising approach to determine optimal or near-optimal solutions to difficult nonconvex variational problems. Yet, they come with increased memory demands, limiting their practicability. To overcome this drawback, this paper presents a combination of two approaches designed to make liftings more scalable, namely product-space relaxations and sublabel-accurate discretizations. Our main contribution is a simple way to solve the resulting semi-infinite optimization problem with a sampling strategy. We show that despite its simplicity, our approach significantly outperforms baseline methods, in the sense that it finds solutions with lower energies given the same amount of memory. We demonstrate our empirical findings on the nonconvex optical flow and manifold-valued denoising problems.

Keywords: Variational methods · Manifold-valued problems · Convex relaxation · Global optimization

1 Introduction

Many tasks in imaging and low-level computer vision can be transparently modeled as a variational problem. In practice, the resulting energy functionals are often nonconvex, for example due to data terms based on image-matching costs or manifold-valued constraints. The goal of this work is to develop a convex optimization approach to total variation-regularized problems of the form

$$\inf_{u:\Omega \to \Gamma} \int_{\Omega} c(x, u_1(x), \ldots, u_k(x)) \, \mathrm{d}x + \sum_{i=1}^{k} \lambda_i \mathrm{TV}(u_i). \tag{1}$$

Here, $\Gamma = \{(\gamma_1, \ldots, \gamma_k) \in \mathbf{R}^N : \gamma_i \in \Gamma_i, i = 1 \ldots k\}$ is based on compact, embedded manifolds $\Gamma_i \subset \mathbf{R}^{N_i}$ with $N = N_1 + \ldots + N_k$. Throughout this paper we only consider imaging applications and pick $\Omega \subset \mathbf{R}^2$ to be a rectangular image domain. The cost function $c : \Omega \times \Gamma \to \mathbf{R}_{\geq 0}$ in (1) can be a general *nonconvex* function. Notably, we only assume that we can *evaluate* the cost function

$c(x, u(x))$ but no gradient information or projection operators are available. This allows us to consider degenerate costs that are out of reach for gradient-based approaches.

As a regularization term in (1) we consider a simple separable *total variation regularization* $\mathrm{TV}(u_i)$ on the individual components $u_i : \Omega \to \mathbf{R}^{N_i}$ weighted by a tunable hyper-parameter $\lambda_i > 0$. The total variation (TV) encourages a spatially smooth but edge-preserving solution. It is defined as

$$\mathrm{TV}(u_i) := \sup_{\substack{p:\Omega\to\mathbf{R}^{N_i\times 2} \\ \|p(x)\|_* \leq 1}} \int_\Omega \langle \mathrm{Div}_x\, p(x), u_i(x) \rangle \, \mathrm{d}x = \int_\Omega \|\nabla u_i(x)\| \, \mathrm{d}x, \qquad (2)$$

where the last equality holds for sufficiently smooth u_i. We denote by $\nabla u_i(x) \in \mathbf{R}^{N_i\times 2}$ the Jacobian matrix in the Euclidean sense and by $\|\cdot\|_*$ the dual norm. Since our focus is on the data cost c, we consider only this separable TV case.

Problems of the form (1) find applications in low-level vision and signal processing. An example is the *optical flow* estimation between two RGB images $I_1, I_2 : \Omega \to \mathbf{R}^3$, where $\Gamma_1 = \Gamma_2 = [a, b] \subset \mathbf{R}$ are intervals and the cost function is given by $c(x, u_1(x), u_2(x)) = |I_1(x + (u_1(x), u_2(x))) - I_2(x)|$. In many applications, Γ_i is a curved manifold, see [24,37]. Examples include $\Gamma_i = \mathbb{S}^2$ in the case of normal field processing [24], SO(3) in the case of motion estimation [16] or the circle $\mathbb{S}^1$ for processing of cyclic data [11,33].

As one often wishes to estimate multiple quantities in a joint fashion, one naturally arrives at the product space formulation as considered in (1). A popular approach to address such joint optimization problems are expectation maximization procedures [12] or block-coordinate descent and alternating direction-type methods [6], where one estimates a single quantity while holding the other ones fixed. Sometimes, such approaches depend on a good initialization and can be prone to getting stuck in bad local minima. Our goal is to devise a convex relaxation of Problem (1) that can be directly solved to global optimality with standard proximal methods (possibly implemented on GPUs) such as the primal dual algorithm [29]. To achieve this, we offer the following contributions:

- To tackle relaxations of (1) in a memory-efficient manner, we propose a sublabel-accurate implementation of the product-space lifting [15]. This implementation is enabled by building on ideas from [26], which views sublabel-accurate multilabeling as a finite-element discretization.
- Our main contribution presented in Sect. 4 is a simple way to implement the resulting optimization problem with a sampling strategy. Unlike previous liftings [25,26,36], our approach does not require epigraphical projections and can therefore be applied in a black-box fashion to any cost $c(x, u(x))$.
- We show that our sublabel-accurate implementation attains a lower energy than the product-space lifting [15] on optical flow estimation and manifold-valued denoising problems.

The following Sect. 2 is aimed to provide an introduction to relaxation methods for (1) while also reviewing existing works and our contributions relative to them. We present the relaxation for (1) and its discretization in Sect. 3. In Sect. 4,

we show how to implement the discretized relaxation with the proposed sampling strategy. Section 5 presents numerical results on optical flow and manifold-valued denoising and our conclusions are drawn in Sect. 6.

2 Related Work: Convex Relaxation Methods

Let us first consider a simplified version of problem (1) where Ω consists only of a single point, i.e., the nonconvex minimization of one data term:

$$\min_{\gamma \in \Gamma} \ c(\gamma_1, \ldots, \gamma_k). \tag{3}$$

A well-known approach to the global optimization of (3) is a *lifting* or *stochastic relaxation* procedure, which has been considered in diverse fields such as polynomial optimization [19], continuous Markov random fields [3,13,28], variational methods [30], and black-box optimization [5,27,32]. The idea is to relax the search space in (3) from $\gamma \in \Gamma$ to *probability distributions*[1] $\mathbf{u} \in \mathcal{P}(\Gamma)$ and solve

$$\min_{\mathbf{u} \in \mathcal{P}(\Gamma)} \int_{\Gamma} c(\gamma_1, \ldots, \gamma_k) \, d\mathbf{u}(\gamma_1, \ldots, \gamma_k). \tag{4}$$

Due to linearity of the integral wrt. $\mathbf{u}$ and convexity of the relaxed search space, this is a convex problem for any c. Moreover, the minimizers of (4) concentrate at the optima of c and can hence be identified with solutions to (3). If Γ is a continuum, this problem is infinite-dimensional and therefore challenging.

Discrete/Traditional Multilabeling. In the context of Markov random fields [17,18] and multilabel optimization [9,21,22,39] one typically discretizes Γ into a finite set of points (called the *labels*) $\Gamma = \{\mathbf{v}_1, \ldots, \mathbf{v}_\ell\}$. This turns (4) into a finite-dimensional linear program $\min_{\mathbf{u} \in \Delta^\ell} \langle c', \mathbf{u} \rangle$ where $c' \in \mathbf{R}^\ell$ denotes the label cost and $\Delta^\ell \subset \mathbf{R}^\ell$ is the $(\ell - 1)$-dimensional unit simplex.. If we evaluate the cost at the labels, this program *upper bounds* the continuous problem (3), since instead of all possible solutions, one considers a restricted subset determined by the labels. Since the solution will be attained at one of the labels, typically a fine meshing is needed. Similar to black-box and zero-order optimization methods, this strategy suffers from the *curse of dimensionality*. When each Γ_i is discretized into ℓ labels, the overall number is ℓ^k which quickly becomes intractable since many labels are required for a smooth solution. Additionally, for pairwise or regularizing terms, often a large number of dual constraints has to be implemented. In that context, the work [23] considers a constraint pruning strategy as an offline-preprocessing.

Sublabel-Accurate Multilabeling. The discrete-continuous MRF [13,38,40] and lifting methods [20,25,26] attempt to find a more label-efficient convex formulation. These approaches can be understood through duality [13,26]. Applied to (3), the idea is to replace the cost $c : \Gamma \to \mathbf{R}$ with a dual variable $\mathbf{q} : \Gamma \to \mathbf{R}$:

---

[1] $\mathcal{P}(\Gamma)$ is the set of nonnegative Radon measures on Γ with total mass $\mathbf{u}(\Gamma) = 1$.

$$\min_{\mathbf{u}\in\mathcal{P}(\Gamma)} \sup_{\mathbf{q}:\Gamma\to\mathbf{R}} \int_\Gamma \mathbf{q}(\gamma_1,\ldots,\gamma_k)\,\mathrm{d}\mathbf{u}(\gamma_1,\ldots,\gamma_k), \text{ s.t. } \mathbf{q}(\gamma)\le c(\gamma)\ \forall\gamma\in\Gamma. \quad (5)$$

The inner supremum in the formulation (5) *maximizes* the lower-bound $\mathbf{q}$ and if the dual variable is sufficiently expressive, this problem is equivalent to (4).

Approximating $\mathbf{q}$, for example with piecewise linear functions on Γ, one arrives at a *lower-bound* to the nonconvex problem (3). It has been observed in a recent series of works [20, 25, 26, 36, 40] that piecewise linear dual variables can lead to smooth solutions even when $\mathbf{q}$ (and therefore also $\mathbf{u}$) is defined on a rather coarse mesh. As remarked in [13, 20, 25], for an *affine* dual variable this strategy corresponds to minimizing the convex envelope of the cost, $\min_{\gamma\in\Gamma} c^{**}(\gamma)$, where c^{**} denotes the Fenchel biconjugate of c.

The implementation of the constraints in (5) can be challenging even in the case of piecewise-linear $\mathbf{q}$. This is partly due to the fact that the problem (5) is a semi-infinite optimization problem [4], i.e., an optimization problem with *infinitely many constraints*. The works [25, 40] implement the constraints via projections onto the epigraph of the (restricted) conjugate function of the cost within a proximal optimization framework. Such projections are only available in closed form for some choices of c and expensive to compute if the dimension is larger than one [20]. This limits the applicability in a "plug-and-play" fashion.

Product-Space Liftings. The product-space lifting approach [15] attempts to overcome the aforementioned exponential memory requirements of labeling methods in an orthogonal way to the sublabel-based methods. The main idea is to exploit the product-space structure in (1) and optimize over k marginal distributions of the probability measure $\mathbf{u}\in\mathcal{P}(\Gamma)$, which we denote by $\mathbf{u}_i\in\mathcal{P}(\Gamma_i)$. Applying [15] to the single data term (3) one arrives at the following relaxation:

$$\min_{\{\mathbf{u}_i\in\mathcal{P}(\Gamma_i)\}} \sup_{\{\mathbf{q}_i:\Gamma_i\to\mathbf{R}\}} \sum_{i=1}^k \int_{\Gamma_i} \mathbf{q}_i(\gamma)\,\mathrm{d}\mathbf{u}_i(\gamma)\ \text{ s.t. } \sum_{i=1}^k \mathbf{q}_i(\gamma)\le c(\gamma)\ \forall\gamma\in\Gamma. \quad (6)$$

Since one only has to discretize the individual Γ_i this substantially reduces the memory requirements from $\mathcal{O}(\ell^N)$ to $\mathcal{O}(\sum_{i=1}^k \ell^{N_i})$. While at first glance it seems that the curse of dimensionality is lifted, the difficulties are moved to the dual, where we still have a large (or even infinite) number of constraints. A global implementation of the constraints with Lagrange multipliers as proposed in [15] again leads to the same exponential dependancy on the dimension.

As a side note, readers familiar with optimal transport may notice that the supremum in (6) is a multi-marginal transportation problem [8, 35] with transportation cost c. This view is mentioned in [1] where relaxations of form (6) are analyzed under submodularity assumptions.

In summary, the sublabel-accurate lifting methods, discrete-continuous MRFs [25, 40] and product-space liftings [15] all share a common difficulty: *implementation of an exponential or even infinite number of constraints on the dual variables.*

Summary of Contribution. Our main contribution is a simple way to implement the dual constraints in an online fashion with a random sampling strategy

which we present in Sect. 4. This allows a black-box implementation, which only requires an *evaluation* of the cost c and no epigraphical projection operations as in [25,40]. Moreover, the sampling approach allows us to propose and implement a sublabel-accurate variant of the product-space relaxation [15] which we describe in the following section.

3   Product-Space Relaxation

Our starting point is the convex relaxation of (1) presented in [15,34]. In these works, $\Gamma_i \subset \mathbf{R}$ is chosen to be an interval. Following [36] we consider a generalization to manifolds $\Gamma_i \subset \mathbf{R}^{N_i}$ which leads us to the following relaxation:

$$\min_{\{\mathbf{u}_i:\Omega \to \mathcal{P}(\Gamma_i)\}} \sup_{\substack{\{\mathbf{q}_i:\Omega \times \Gamma_i \to \mathbf{R}\} \\ \{\mathbf{p}_i:\Omega \times \Gamma_i \to \mathbf{R}^2\}}} \sum_{i=1}^{k} \int_{\Omega} \int_{\Gamma_i} \mathbf{q}_i(x,\gamma_i) - \mathrm{Div}_x\, \mathbf{p}_i(x,\gamma_i) \, \mathrm{d}\mathbf{u}_i^x(\gamma) \, \mathrm{d}x, \quad (7)$$

$$\text{s.t.} \quad \|P_{T_{\gamma_i}} \nabla_{\gamma_i} \mathbf{p}_i(x,\gamma_i)\|_* \leq \lambda_i, \text{ for all } 1 \leq i \leq k, \ (x,\gamma_i) \in \Omega \times \Gamma_i, \quad (8)$$

$$\sum_{i=1}^{k} \mathbf{q}_i(x,\gamma_i) \leq c(x,\gamma), \text{ for all } (x,\gamma) \in \Omega \times \Gamma. \quad (9)$$

This cost function appears similar to (6) explained in the previous section, but with two differences. First, we now have marginal distributions $\mathbf{u}_i(x)$ for every $x \in \Omega$ since we do not consider only a single data term anymore. The notation $\mathrm{d}\mathbf{u}_i^x$ in (7) denotes the integration against the probability measure $\mathbf{u}_i(x) \in \mathcal{P}(\Gamma_i)$. The variables $\mathbf{q}_i$ play the same role as in (6) and lower-bound the cost under constraint (9). The second difference is the introduction of additional dual variables $\mathbf{p}_i$ and the term $-\mathrm{Div}_x\, \mathbf{p}_i$ in (7). Together with the constraint (8), this can be shown to implement the total variation regularization [24,36]. Following [36], the derivative $\nabla_{\gamma_i} \mathbf{p}_i(x,\gamma_i)$ in (8) denotes the $(N_i \times 2)$-dimensional Jacobian considered in the Euclidean sense and $P_{T_{\gamma_i}}$ the projection onto the tangent space of Γ_i at the point γ_i. Next, we describe a finite-element discretization of (7).

3.1   Finite-Element Discretization

We approximate the infinite-dimensional problem (7) by restricting $\mathbf{u}_i$, $\mathbf{p}_i$ and $\mathbf{q}_i$ to be piecewise functions on a discrete meshing of $\Omega \times \Gamma_i$. The considered discretization is a standard finite-element approach and largely follows [36]. Unlike the forward-differences considered in [36] we use lowest-order Raviart-Thomas elements (see, e.g., [7, Section 5]) in Ω, which are specifically tailored towards the considered total variation regularization.

Discrete Mesh. We approximate each d_i-dimensional manifold $\Gamma_i \subset \mathbf{R}^{N_i}$ with a simplicial manifold Γ_i^h, given by the union of a collection of d_i-dimensional simplices $\mathcal{T}_i$. We denote the number of vertices ("labels") in the triangulation of Γ_i as ℓ_i. The set of labels is denoted by $\mathcal{L}_i = \{\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,\ell_i}\}$. As assumed,

$\Omega \subset \mathbf{R}^2$ is a rectangle which we split into a set of faces $\mathcal{F}$ of edge-length h_x with edge set $\mathcal{E}$. The number of faces and edges are denoted by $F = |\mathcal{F}|$, $E = |\mathcal{E}|$.

Data Term and the $\mathbf{u}_i$, $\mathbf{q}_i$ *Variables.* We assume the cost $c : \Omega \times \Gamma \to \mathbf{R}_{\geq 0}$ is constant in $x \in \Omega$ on each face and denote its value as $c(x(f), \gamma)$ for $f \in \mathcal{F}$, where $x(f) \in \Omega$ denotes the midpoint of the face f. Similarly, we also assume the variables $\mathbf{u}_i$ and $\mathbf{q}_i$ to be constant in $x \in \Omega$ on each face but continuous piecewise linear functions in γ_i. They are represented by coefficient functions $\mathbf{u}_i^h, \mathbf{q}_i^h \in \mathbf{R}^{F \cdot \ell_i}$, i.e., we specify the values on the labels and linearly interpolate inbetween. This is done by the interpolation operator $\mathbf{W}_{i,f,\gamma_i} : \mathbf{R}^{F \cdot \ell_i} \to \mathbf{R}$ which given an index $1 \leq i \leq k$, face f, and (continuous) label position $\gamma_i \in \Gamma_i$ computes the function value: $\mathbf{W}_{i,f,\gamma_i} \mathbf{u}_i^h = \mathbf{u}_i(x(f), \gamma_i)$. Note that after discretization, $\mathbf{u}_i$ is only defined on Γ_i^h but we can uniquely associate to each $\gamma_i \in \Gamma_i^h$ a point on Γ_i.

Divergence and $\mathbf{p}_i$ *variables.* Our variable $\mathbf{p}_i$ is represented by coefficients $\mathbf{p}_i^h \in \mathbf{R}^{E \cdot \ell_i}$ which live on the edges in Ω and the labels in Γ_i. The vector $\mathbf{p}_i(x, \gamma_i) \in \mathbf{R}^2$ is obtained by linearly interpolating the coefficients on the vertical and horizontal edges of the face and using the interpolated coefficients to evaluate the piecewise-linear function on Γ_i^h. Under this approximation, the discrete divergence $\mathrm{Div}_x^h : \mathbf{R}^{E \cdot \ell_i} \to \mathbf{R}^{F \cdot \ell_i}$ is given by $(\mathrm{Div}_x^h \mathbf{p}_i^h)(f) = \left(\mathbf{p}_i^h(e_r) + \mathbf{p}_i^h(e_t) - \mathbf{p}_i^h(e_l) - \mathbf{p}_i^h(e_b)\right)/h_x$ where e_r, e_t, e_l, e_b are the right, top, left and bottom edges of f, respectively.

Total Variation Constraint. Computing the operator $P_{T_{\gamma_i}} \nabla_{\gamma_i}$ is largely inspired by [36, Section 2.2]. It is implemented by a linear map $\mathbf{D}_{i,f,\alpha,t} : \mathbf{R}^{E \cdot \ell_i} \to \mathbf{R}^{d_i \times 2}$. Here, $f \in \mathcal{F}$ and $\alpha \in [0,1]^2$ correspond to a point $x \in \Omega$ while $t \in \mathcal{T}_i$ is the simplex containing the point corresponding to $\gamma_i \in \Gamma_i$. First, the operator computes coefficients in $\mathbf{R}^{\ell_i}$ of two piecewise-linear functions on the manifold by linearly interpolating the values on the edges based on the face index $f \in \mathcal{F}$ and $\alpha \in [0,1]^2$. For each function, the derivative in simplex $t \in \mathcal{T}_i$ on the triangulated manifold is given by the gradient of an affine extension. Projecting the resulting vector into the d_i-dimensional tangent space for both functions leads to a $d_i \times 2$-matrix which approximates $P_{T_{\gamma_i}} \nabla_{\gamma_i} \mathbf{p}_i(x, \gamma_i)$.

Final Discretized Problem. Plugging our discretized $\mathbf{u}_i$, $\mathbf{q}_i$, $\mathbf{p}_i$ into (7), we arrive at the following finite-dimensional optimization problem:

$$\min_{\{\mathbf{u}_i^h \in \mathbf{R}^{F \cdot \ell_i}\}} \max_{\substack{\{\mathbf{p}_i^h \in \mathbf{R}^{E \cdot \ell_i}\}, \\ \{\mathbf{q}_i^h \in \mathbf{R}^{F \cdot \ell_i}\}}} h_x^2 \cdot \sum_{i=1}^{k} \langle \mathbf{u}_i^h, \mathbf{q}_i^h - \mathrm{Div}_x^h \mathbf{p}_i^h \rangle + \sum_{f \in \mathcal{F}} \mathbf{i}\{\mathbf{u}_i^h(f) \in \Delta^{\ell_i}\}, \quad (10)$$

$$\text{s.t.} \quad \|\mathbf{D}_{i,f,\alpha,t} \mathbf{p}_i^h\|_* \leq \lambda_i, \ \forall 1 \leq i \leq k, f \in \mathcal{F}, \alpha \in \{0,1\}^2, t \in \mathcal{T}_i, \quad (11)$$

$$\sum_{i=1}^{k} \mathbf{W}_{i,f,\gamma_i} \mathbf{q}_i^h \leq c\left(x(f), \gamma\right), \ \forall f \in \mathcal{F}, \gamma \in \Gamma, \quad (12)$$

where $\mathbf{i}\{\cdot\}$ is the indicator function. In our applications, we found it sufficient to enforce the constraint (11) at the corners of each face which corresponds to choosing $\alpha \in \{0,1\}^2$. Apart from the infinitely many constraints in (12), this is a finite-dimensional convex-concave saddle-point problem.

3.2   Solution Recovery

Before presenting in the next section our proposed way to implement the constraints (12), we briefly discuss how a primal solution $\{\mathbf{u}_i^h\}$ of the above problem is turned into an approximate solution to (1). To that end, we follow [24, 36] and compute the Riemannian center of mass via an iteration $\tau = 1, \ldots, T$:

$$V_j^\tau = \log_{u_i^\tau}(\mathbf{v}_{i,j}), \quad v^\tau = \sum_{j=1}^{\ell_i} \mathbf{u}_i^h(f, j) V_j^\tau, \quad u_i^{\tau+1} = \exp_{u_i^\tau}(v^\tau). \tag{13}$$

Here, $u_i^0 \in \Gamma_i$ is initialized by the label with the highest probability according to $\mathbf{u}_i^h(f, \cdot)$. $\log_{u_i^\tau}$ and $\exp_{u_i^\tau}$ denote the logarithmic and exponential mapping between Γ_i^h and it's tangent space at $u_i^\tau \in \Gamma_i$, which are both available in closed-form for the manifolds we consider here. In our case $T = 20$ was enough to reach convergence. For flat manifolds, $T = 1$ is enough, as both mappings boil down to the identity and (13) computes a weighted Euclidean mean.

In general, there is no theory which shows that $u^T(x) = (u_1^T(x), \ldots, u_k^T(x))$ from (13) is a global minimizer of (1). Tightness of the relaxation in the special case $k = 1$ and $\Gamma \subset \mathbf{R}$ is shown in [31]. For higher dimensional Γ, the tightness of related relaxations is ongoing research; see [14] for results on the Dirichlet energy. By computing a-posteriori optimality gaps, solutions of (7) were shown to be typically near the global optimum of the problem (1); see, e.g., [15].

4   Implementation of the Constraints

Though the optimization variables in (10) are finite-dimensional, the energy is still difficult to optimize because of the infinite constraints in (12).

Before we present our approach, let us first describe what we refer to as the *baseline method* in the rest of this paper. For the baseline approach, we consider the direct solution of (10) where we implemented the constraints only at the label/discretization points $\mathcal{L}_1 \times \ldots \times \mathcal{L}_k$ via Lagrange multipliers. This strategy is also employed by the (global variant) of the product-space approach [15].

We aim for a framework that allows for solving a better approximation of (12) than the above baseline while being of similar memory complexity. To achieve this, our algorithm alternates the following two steps in an iterative way.

1) Sampling. Based on the current solution we prune previously considered but feasible constraints and sample a new subset of the infinite constraints in (12). From all current sampled constraints, we consider the most violated constraints for each face, add one sample at the current solution and discard the rest.

2) Solving the subsampled problem. Considering the current finite subset of constraints, we solve problem (10) using a primal-dual algorithm.

These two phases are performed alternatingly, with the aim to eventually approach the solution of the continuous problem (10). The details of our constraint sampling strategy are shown in Algorithm 1. For each face in $\mathcal{F}$, the

---

Algorithm 1: Sampling strategy at face $f \in \mathcal{F}$.

Inputs: Solution $u = (u_1, \ldots, u_k)$ at face f, sublabel-set $\mathcal{S}_f$, n, δ, r

1 $\mathcal{S}'_f \leftarrow \text{uniformSample}(\Gamma, n)$ /* global exploration */
2 $\mathcal{S}'_f \leftarrow \mathcal{S}'_f \cup \text{localPerturb}(u, \delta, n)$ /* local exploration around sol. */
3 $\mathcal{S}_f \leftarrow \left\{ \gamma \in \mathcal{S}_f : \sum_{i=1}^{k} \mathbf{q}_i(f, \gamma) > c(f, \gamma) \right\}$ /* remove feas. cons. */
4 $\mathcal{S}_f \leftarrow \text{top-k}(\mathcal{S}'_f, r) \cup \mathcal{S}_f$ /* add the most violated r samples */
5 $\mathcal{S}_f \leftarrow \mathcal{S}_f \cup \{u\}$.   /* have one sample at cur. sol. */
6 **return** $\mathcal{S}_f$

Algorithm 2: Proposed algorithm for problem (1).

Inputs: $c : \Omega \times \Gamma \to \mathbf{R}$, $\lambda_i > 0$, $N_{it} > 0$, $M_{it} > 0$, $n > 0$, $\delta > 0$, $r > 0$

1 $\mathbf{u}_i^{h,0} = \mathbf{1}/\ell_i$, $\mathbf{q}_i^{h,0} = \mathbf{0}$, $\mathbf{p}_i^{h,0} = \mathbf{0}$, $\mathcal{S}_f^0 = \mathcal{L}_1 \times \ldots \times \mathcal{L}_k$.
2 **for** $it = 0$ **to** N_{it} **do**
3   Obtain $\mathbf{u}_i^{h,it+1}, \mathbf{q}_i^{h,it+1}, \mathbf{p}_i^{h,it+1}$ by running M_{it} iterations of the primal-dual method [29] on (10), with constraints (12) implemented at $\mathcal{S}_f^{it}$ for each $f \in \mathcal{F}$ via Lagrange multipliers; warmstart at $\mathbf{u}_i^{h,it}, \mathbf{q}_i^{h,it}, \mathbf{p}_i^{h,it}$. Compute current approximate solution u^{it+1} to (1) via (13).
4   Get $\mathcal{S}_f^{it+1}$ by calling Alg. 1 with $(u^{it+1}(f), \mathcal{S}_f^{it}, n, \delta, r)$ for each $f \in \mathcal{F}$.
5 **end**

---

algorithm generates a finite set of "sublabels" $\mathcal{S}_f \subset \Gamma$ at which we implement the constraints (12). In the following, we provide the motivation behind each line in the algorithm.

Random Uniform Sampling (Line 1). To have a global view of the cost function, we consider a uniform sampling on the label space Γ. The parameter $n > 0$ determines the number of the samples for each face.

Local Perturbation Around the Mean (Line 2). Besides the global information, we apply local perturbation around the current solution u. In case the current solution is close to the optimal one, this strategy allows us to refine it with these samples. The parameter $\delta > 0$ determines the size of the local neighbourhood. In experiments, we always used a Gaussian perturbation with $\delta = 0.1$.

Pruning Strategy (Lines 3–4). Most samples from previous iterations are discarded because the corresponding constraints are already satisfied. We prune all current feasible constraints as in [4]. Similarly, the two random sampling strategies (Lines 1 and 2) might return some samples for which the constraints are already fulfilled. Therefore, we only consider the samples with violated constraints and pick the r most violated from them. This pruning strategy is essential for a memory efficient implementation as shown later.

Sampling at u (Line 5). Finally, we add one sample which is exactly at the current solution $u \in \Gamma$ to have at least one guaranteed sample per face.
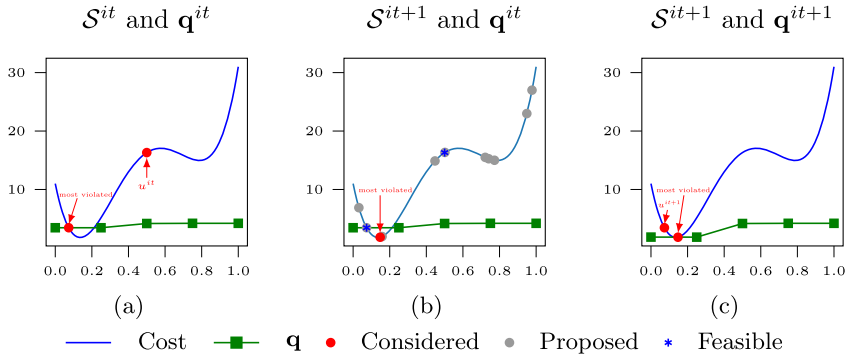
Fig. 1. Illustration of sampling strategies. (a) Two samples (red dots) are considered leading to the shown optimal dual variable $\mathbf{q}$ after running primal-dual iterations. (b) The two samples are pruned because the constraints are feasible. Several random samples are proposed (gray dots) and only one of them is picked (red dot). (c) One more sample on u^{it} is added and the $\mathbf{q}$ is refined. (Color figure online)

Overall Algorithm. After implementing the constraints at the finite set determined by Algorithm 1, we apply a primal-dual method [10] with diagonal preconditioning [29] to solve (10). Both constraints (11) and (12) are implemented using Lagrange multipliers. Based on the obtained solution, a new set of samples is determined.

This scheme is alternated for a fixed number of outer iterations N_{it} and we have summarized the overall algorithm in Algorithm 2. While we do not prove convergence of the overall algorithm, convergence results for related procedures exist; see, e.g., [4, Theorem 2.4].

Finally, let us note that a single outer iteration of Algorithm 2 with large number of M_{it} corresponds to the baseline method.

5 Numerical Validation

Our approach and the baseline are implemented in PyTorch. Code for reproducing the following experiments can be found here: https://github.com/zhenzhangye/sublabel_meets_product_space. Note that a specialized implementation as in [15] will allow the method to scale by factor $10 - 100\times$.

5.1 Illustration of Our Sampling Idea

To illustrate the effect of the sampling strategies, we consider the minimization of a single nonconvex data term. The cost c and the corresponding dual variable $\mathbf{q}$ are plotted in Fig. 1. As shown in (a), the primal-dual method can obtain the optimal $\mathbf{q}^h$ for the sampled subproblem only. Our sampling strategy can provide necessary samples and prune the feasible ones as, cp. (b). These few but necessary samples lead the $\mathbf{q}^h$ to achieve global optimality, cp. (c). If one more iteration is performed, the sampling at u^{it} can stabilize the optimal $\mathbf{q}$.

Table 1. Ablation study indicating the effect of individual lines in Algorithm 1. Numbers in parentheses indicate the standard deviation across 20 runs.

	Labels	Baseline	+Line 1&4	+Line 2	+Line 3	+Line 5
Energy	3	4589 ($\pm$0.00)	2305 ($\pm$3.73)	2291 ($\pm$3.6)	8585 ($\pm$130.4)	2051 ($\pm$10.7)
Time [s]		8.98	22.77	23.22	23.22	23.33
Mem. [Mb]		11.21	13.94	15.53	11.65	12.05
Energy	7	2582 ($\pm$0.00)	2020 ($\pm$2.68)	2012 ($\pm$1.3)	7209 ($\pm$116.7)	1969 ($\pm$3.6)
Time [s]		74.13	16.02	16.61	15.56	18.38
Mem. [Mb]		28.35	32.96	33.49	28.356	28.68
Energy	13	2029 ($\pm$0.00)	1935 ($\pm$1.14)	1926 ($\pm$0.7)	5976 ($\pm$75.7)	1901 ($\pm$3.7)
Time [s]		183.80	37.65	38.84	38.29	38.22
Mem. [Mb]		52.85	60.55	60.94	54.35	54.73
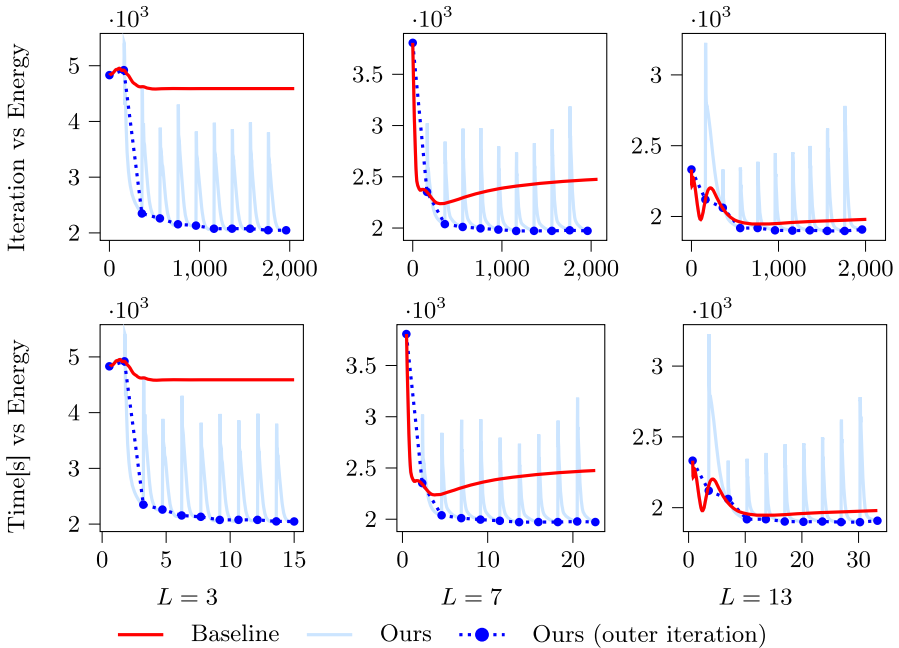
Fig. 2. Comparison between the baseline and our approach on a 64×64 image, degraded with Gaussian and salt-and-papper noise. Our approach finds lower energies in fewer iterations and time, which implements the constraints only at the label points.

5.2   Ablation Study

Next, we study the effect of each line in Algorithm 1. We evaluate our method on the truncated quadratic energy $c(x, u(x)) = \min\{(u(x) - f(x))^2, \nu\}$. where $f : \Omega \to \mathbf{R}$ is the input data. For this specific experiment, the parameters are chosen as $\nu = 0.025$, $\lambda = 0.25$, $N_{it} = 10$, $M_{it} = 200$, $n = 10$ and $r = 1$. To reduce

Baseline						
#Labels	3	7	11	15	19	23
Energy	56730.6	35902.5	25943.1	18923.1	16402.7	15297.1
Memory [Mb]	334.9	1166.6	2427.3	4111.3	6220.6	8748.51
aep [px]	3.08	2.43	2.19	2.10	2.09	2.07
aae [°]	0.79	0.51	0.37	0.32	0.31	0.30
Ours						
#Labels	3	7	11	15	19	23
Energy	**31917.5**	**18843.4**	**12891.6**	**11238.0**	**10451.8**	**9867.4**
Memory [Mb]	334.9	1167.3	2427.3	4111.2	6213.1	8748.51
aep [px]	2.58	2.32	2.09	2.08	2.07	2.07
aae [°]	0.65	0.44	0.31	0.30	0.30	0.29

Fig. 3. We compute the optical flow on Grove3 [2] using our method and our baseline for a varying amount of labels. Given an equal number of labels/memory, our sampling strategy performs favorably to an implementation of the constraints at the labels.

the effect of randomness, we run each algorithm 20 times and report mean and standard deviation of the final energy for different number of labels in Table 1.

Adding uniform sampling and picking the most violated constraint per face already decreases the final energy significantly, i.e. Line 1 and Line 4 of Algorithm 1. We also consider local exploration around the current solution, cf. Line 2, which helps to find better energies at the expense of higher memory requirements.

To circumvent that, we introduce our pruning strategy in Line 3 of Algorithm 1. However, the energy deteriorates dramatically because some faces could end up having no samples after pruning. Therefore, keeping the current solution as a sample (Line 5) per face prevents the energy from degrading.

Including all the sampling strategies, the proposed method can achieve the best energy and run-time, at comparable memory usage to the baseline method. We further illustrate the comparison on the number of iterations and time between the baseline and our proposed method in Fig. 2. Due to the replacement on the samples, we have a peak right after each sampling phase. The energy however converges immediately, leading to an overall decreasing trend.

5.3 Optical Flow

Given two input images I_1, I_2, we compute the optical flow $u : \Omega \to \mathbf{R}^2$. The label space $\Gamma = [a, b]^2$ in our case is chosen as $a = -2.5$ and $b = 7.5$. We use a simple ℓ_2-norm for the data term, i.e. $c(x, u(x)) = ||I_2(x) - I_1(x + u(x))||$
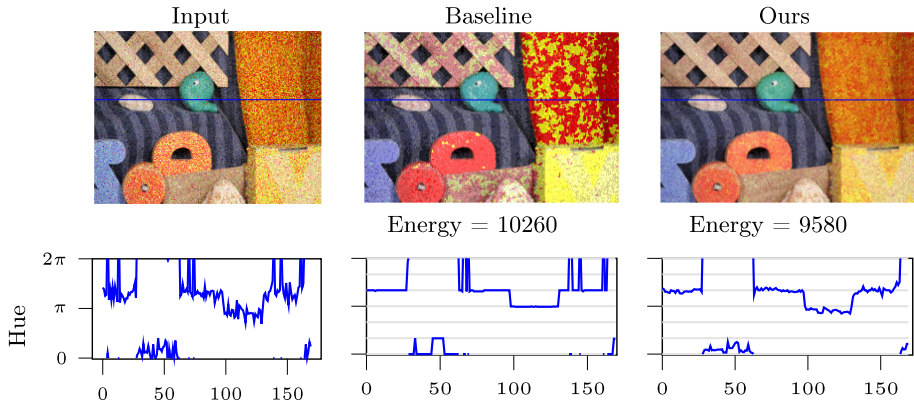
Fig. 4. Denoising of an image in HSV color space ($\Gamma_1 = \mathbb{S}^1$) using our method and the baseline. Since our approach implements the constraints adaptively inbetween the labels it reaches a lower energy with less label bias.

and set the regularization weight as $\lambda = 0.04$. The baseline approach runs for 50K iterations, while we set $N_{it} = 50$ and $M_{it} = 1000$ for a fair comparison. Additionally, we choose $n = 20$ and $r = 1$ in Algorithm 1.

The results are shown in Fig. 3. Our method outperforms the baseline approach regarding energy under the same number of labels and requires the same amount of memory. We can achieve lower energy with about half number of labels.

5.4   Denoising in HSV Color Space

In our final application, we evaluate on a manifold-valued denoising problem in HSV color space. The hue component of this space is a circle, i.e., $\Gamma_1 = \mathbb{S}^1$, $\Gamma_2, \Gamma_3 = [0, 1]$. The data term of this experiment is a truncated quadratic distance, where for the hue component the distance is taken on the circle $\mathbb{S}^1$.

Both the baseline and our method are implemented with 7 labels. 30K iterations are performed on the baseline and $N_{it} = 100$ outer iterations for our method with 300 inner primal-dual steps are used to get an equal number of total iterations. Other parameters are chosen as $\lambda = 0.015$, $n = 30$ and $r = 5$. As shown in Fig. 4, our method can achieve a lower energy than the baseline. Qualitatively, since our method implements the constraints not only at the labels but also inbetween, there is less bias compared to the baseline.

6   Conclusion

In this paper we made functional lifting methods more scalable by combining two advances, namely product-space relaxations [15] and sublabel-accurate discretizations [26,36]. This combination is enabled by adapting a cutting-plane

method from semi-infinite programming [4]. This allows an implementation of sublabel-accurate methods without difficult epigraphical projections.

Moreover, our approach makes sublabel-accurate functional-lifting methods applicable to any cost function in a simple black-box fashion. In experiments, we demonstrate the effectiveness of the approach over a baseline based on the product-space relaxation [15] and provided a proof-of-concept experiment showcasing the method in the manifold-valued setting.

References

1. Bach, F.: Submodular functions: from discrete to continuous domains. Math. Program. **175**, 419–459 (2018). https://doi.org/10.1007/s10107-018-1248-6
2. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. (IJCV) **92**(1), 1–31 (2011)
3. Bauermeister, H., Laude, E., Möllenhoff, T., Moeller, M., Cremers, D.: Lifting the convex conjugate in Lagrangian relaxations: a tractable approach for continuous Markov random fields. arXiv:2107.06028 (2021)
4. Blankenship, J.W., Falk, J.E.: Infinitely constrained optimization problems. J. Optim. Theory Appl. **19**(2), 261–281 (1976)
5. de Boer, P., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**(1), 19–67 (2005)
6. Boyd, S.P., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
7. Caillaud, C., Chambolle, A.: Error estimates for finite differences approximations of the total variation. Preprint hal-02539136 (2020)
8. Carlier, G.: On a class of multidimensional optimal transportation problems. J. Convex Anal. **10**(2), 517–530 (2003)
9. Chambolle, A., Cremers, D., Pock, T.: A convex approach to minimal partitions. SIAM J. Imaging Sci. **5**(4), 1113–1158 (2012)
10. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**, 120–145 (2011)
11. Cremers, D., Strekalovskiy, E.: Total cyclic variation and generalizations. J. Math. Imaging Vis. **47**(3), 258–277 (2013)
12. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–22 (1977)
13. Fix, A., Agarwal, S.: Duality and the continuous graphical model. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 266–281. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_18
14. Ghoussoub, N., Kim, Y.H., Lavenant, H., Palmer, A.Z.: Hidden convexity in a problem of nonlinear elasticity. SIAM J. Math. Anal. **53**(1), 1070–1087 (2021)
15. Goldluecke, B., Strekalovskiy, E., Cremers, D.: Tight convex relaxations for vector-valued labeling. SIAM J. Imaging Sci. **6**(3), 1626–1664 (2013)
16. Görlitz, A., Geiping, J., Kolb, A.: Piecewise rigid scene flow with implicit motion segmentation. In: International Conference on Intelligent Robots and Systems (IROS) (2019)

17. Ishikawa, H.: Exact optimization for Markov random fields with convex priors. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **25**(10), 1333–1336 (2003)
18. Kappes, J., et al.: A comparative study of modern inference techniques for discrete energy minimization problems. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
19. Lasserre, J.B.: Global optimization with polynomials and the problem of moments. SIAM J. Optim. **11**(3), 796–817 (2000)
20. Laude, E., Möllenhoff, T., Moeller, M., Lellmann, J., Cremers, D.: Sublabel-accurate convex relaxation of vectorial multilabel energies. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 614–627. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_37
21. Lellmann, J., Schnörr, C.: Continuous multiclass labeling approaches and algorithms. SIAM J. Imaging Sci. **4**(4), 1049–1096 (2011)
22. Lellmann, J., Kappes, J., Yuan, J., Becker, F., Schnörr, C.: Convex multi-class image labeling by simplex-constrained total variation. In: Tai, X.-C., Mørken, K., Lysaker, M., Lie, K.-A. (eds.) SSVM 2009. LNCS, vol. 5567, pp. 150–162. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02256-2_13
23. Lellmann, J., Lellmann, B., Widmann, F., Schnörr, C.: Discrete and continuous models for partitioning problems. Int. J. Comput. Vis. (IJCV) **104**(3), 241–269 (2013)
24. Lellmann, J., Strekalovskiy, E., Koetter, S., Cremers, D.: Total variation regularization for functions with values in a manifold. In: International Conference on Computer Vision (ICCV) (2013)
25. Möllenhoff, T., Laude, E., Moeller, M., Lellmann, J., Cremers, D.: Sublabel-accurate relaxation of nonconvex energies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. Möllenhoff, T., Cremers, D.: Sublabel-accurate discretization of nonconvex free-discontinuity problems. In: International Conference on Computer Vision (ICCV) (2017)
27. Ollivier, Y., Arnold, L., Auger, A., Hansen, N.: Information-geometric optimization algorithms: a unifying picture via invariance principles. J. Mach. Learn. Res. **18**, 18:1–18:65 (2017)
28. Peng, J., Hazan, T., McAllester, D., Urtasun, R.: Convex max-product algorithms for continuous MRFs with applications to protein folding. In: International Conference on Machine Learning (ICML) (2011)
29. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: International Conference on Computer Vision (ICCV) (2011)
30. Pock, T., Schoenemann, T., Graber, G., Bischof, H., Cremers, D.: A convex formulation of continuous multi-label problems. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 792–805. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88690-7_59
31. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: Global solutions of variational models with convex regularization. SIAM J. Imaging Sci. **3**(4), 1122–1145 (2010)
32. Schaul, T.: Studies in continuous black-box optimization. Ph.D. thesis, Technische Universität München (2011)
33. Steinke, F., Hein, M., Schölkopf, B.: Nonparametric regression between general Riemannian manifolds. SIAM J. Imaging Sci. **3**(3), 527–563 (2010)
34. Strekalovskiy, E., Chambolle, A., Cremers, D.: Convex relaxation of vectorial problems with coupled regularization. SIAM J. Imaging Sci. **7**(1), 294–336 (2014)

35. Villani, C.: Optimal Transport: Old and New. Springer, Heidelberg (2008)
36. Vogt, T., Strekalovskiy, E., Cremers, D., Lellmann, J.: Lifting methods for manifold-valued variational problems. In: Grohs, P., Holler, M., Weinmann, A. (eds.) Handbook of Variational Methods for Nonlinear Geometric Data, pp. 95–119. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-31351-7_3
37. Weinmann, A., Demaret, L., Storath, M.: Total variation regularization for manifold-valued data. SIAM J. Imaging Sci. **7**(4), 2226–2257 (2014)
38. Zach, C.: Dual decomposition for joint discrete-continuous optimization. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2013)
39. Zach, C., Gallup, D., Frahm, J.M., Niethammer, M.: Fast global labeling for real-time stereo using multiple plane sweeps. In: Proceedings of the Vision, Modeling and Visualization Workshop (VMV) (2008)
40. Zach, C., Kohli, P.: A convex discrete-continuous approach for Markov random fields. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 386–399. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_28