Discrete-Continuous ADMM for Transductive Inference in Higher-Order MRFs

Emanuel Laude [1] Jan-Hendrik Lange [2] Jonas Schüpfer [1] Csaba Domokos [1]
Laura Leal-Taixé [1] Frank R. Schmidt [1 3] Bjoern Andres [2 3 4] Daniel Cremers [1]

[1] Technical University of Munich [2] Max Planck Institute for Informatics, Saarbrücken
[3] Bosch Center for Artificial Intelligence [4] University of Tübingen

Abstract

This paper introduces a novel algorithm for transductive inference in higher-order MRFs, where the unary energies are parameterized by a variable classifier. The considered task is posed as a joint optimization problem in the continuous classifier parameters and the discrete label variables. In contrast to prior approaches such as convex relaxations, we propose an advantageous decoupling of the objective function into discrete and continuous subproblems and a novel, efficient optimization method related to ADMM. This approach preserves integrality of the discrete label variables and guarantees global convergence to a critical point. We demonstrate the advantages of our approach in several experiments including video object segmentation on the DAVIS data set and interactive image segmentation.

1. Introduction

Various problems in computer vision, computer graphics and machine learning can be formulated as MAP inference in a (possibly higher order) Markov random field (MRF) [33, 26, 40, 28, 30, 12, 32, 43]. The resulting optimization problem is defined over a hypergraph $(\mathcal{V}, \mathcal{C})$ and a finite label set $\mathcal{L}$ as:

$$\min_{y \in \mathcal{Y}} \sum_{i \in \mathcal{V}} E_i(y_i) + \sum_{\substack{C \in \mathcal{C} \\ |C|>1}} E_C(y_C). \tag{1}$$

The optimization variable $y \in \mathcal{Y} := \mathcal{L}^{|\mathcal{V}|}$ corresponds to a labeling of the vertices $\mathcal{V}$ and assigns a label $y_i \in \mathcal{L}$ to each vertex $i \in \mathcal{V}$. For convenience, we make a distinction between the singleton clique energies (unaries) $E_i(y_i)$ and the higher order energies $E_C(y_C)$, $|C| > 1$.

In computer vision tasks, where the image is interpreted as a (higher order) pixelgrid $(\mathcal{V}, \mathcal{C})$, the higher order potentials often correspond to priors favoring spatially smooth solutions. In semantic image segmentation, for instance, inference in MRFs is widely used as a post-processing step to introduce spatial smoothness on the labeling y [10]. In this sense, the overall task of semantic segmentation is subdivided into two tasks: First, a classifier, parameterized by W, is trained in a supervised fashion on a sufficiently large labeled training set, which assigns to each pixel in a test image a class (probability) score. Second, to enforce spatial smoothness of the labeling of the test image, MAP-inference in an MRF is performed in a post-processing, in which the class (probability) scores are interpreted as the unary energies $E_i(y_i; W)$ for each y_i. We argue that it is advantageous to merge such a two-step approach into a joint approach, as the training of the classifier profits from both, (i) the distribution of the unlabeled pixels in the color or feature space, and (ii) the available structural information

Figure 1: A pixel-classifier trained to predict an object mask (**top row**) performs well when the distribution of object pixels in the training image is similar to the test image (**left**), but often fails if it is dissimilar (**right**). In a transductive inference approach (**bottom row**), we optimize jointly for the test labels and classifier parameters, which successfully prevents the hallucination of object pixels in difficult scenes, such as in the case of occlusion (**right**), cf. Sec. 4.2.

about the unlabeled pixels, namely the spatial smoothness prior. Conversely, the segmentation will also profit from an improved classifier.

A joint formulation has two interpretations; on the one hand, it is a semi-supervised learning method that makes use of structural knowledge about the training data to learn a classifier. Such knowledge may take the form of higher-order clique energies E_C [57, 2] on the labeling and acts as weak supervision in the training process. This approach helps to mitigate the need for large amounts of annotated training data in typical modern machine learning applications. We, on the other hand, focus on its interpretation as a *transductive* inference method [56], i.e. the approach to directly infer the labels of specific test data given specific training data, which we accomplish by incorporating a variable classifier in the inference process. Transductive inference stands in contrast to *inductive* inference, which refers to first learning a general model from training data and subsequently applying the model to predict labels of a-priori unknown test data. We show the benefits of using transductive inference for the tasks of video object segmentation (cf. Fig. 1) as well as scribble-based segmentation (cf. Fig. 4).

1.1. Contributions

We propose a general joint model that assumes the unaries not be fixed for inference in the MRF, but rather optimized jointly with the labeling. Let for each $i \in \mathcal{V}$, $x_i \in \mathbb{R}^d$ denote the d-dimensional feature vector associated to the ith vertex and let $\varphi : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a feature map. Then, mathematically, such a task can be naturally formulated in terms of a bilevel optimization problem:

$$\begin{aligned} &\min_{\substack{y\in\mathcal{Y},\\ W\in\mathbb{R}^{|\mathcal{L}|\times d'}}} \quad \sum_{i\in\mathcal{V}} \ell(y_i; W\varphi(x_i)) + \sum_{C\in\mathcal{C}} E_C(y_C) \qquad (2)\\ &\text{subject to} \quad W = \operatorname*{arg\,min}_{W\in\mathbb{R}^{|\mathcal{L}|\times d'}} \sum_{i\in\mathcal{V}} \ell(y_i; W\varphi(x_i)) + g(W). \end{aligned}$$

Here, the upper-level task is inference in a MRF with additional unaries $E_i(y_i; W) := \ell(y_i; W\varphi(x_i))$, parameterized by linear classifier weights $W \in \mathbb{R}^{|\mathcal{L}|\times d'}$, and a loss function $\ell\colon \mathcal{L} \times \mathbb{R}^{|\mathcal{L}|} \to \mathbb{R}$. The lower-level task associates to each given set of labels y the optimal parameters W. For instance, if ℓ is the hinge loss and $g(W) = \|W\|_F^2$, then the lower-level optimization problem amounts to the training of a classical SVM. Note that in a semi-supervised learning context it might be more convenient to swap upper and lower level tasks, since the primary interest is the estimation of the classifier that minimizes the generalization error and not the inferred labeling. However, mathematically, both viewpoints are equivalent.

The model (2) suggests a simple alternating optimization scheme as in Lloyd's algorithm [37] for k-means to compute a local optimum. However, such an approach has two major drawbacks: (i) The lower-lever subproblems are expensive, which is prohibitive for large scale applications. (ii) The optimization is prone to poor local optima and therefore sensitive to initialization [61]. Motivated by the good practical performance of the alternating direction method of multipliers (ADMM) in nonconvex optimization, we propose to generalize vanilla ADMM (commonly applied in continuous optimization) to discrete-continuous problems of the form (2), while preserving integrality of the discrete variables. Since our method serves as a general algorithmic framework to tackle such problems, it is also relevant to semi-supervised and transductive learning in a broader sense.

The main contributions of this work can be summarized as follows:

- We devise a decomposition of the model into simple, purely discrete and purely continuous subproblems within the framework of proximal splitting. The subproblems can be solved in a distributed fashion.
- We devise a tailored ADMM-inspired algorithm, *discrete-continuous ADMM*, to compute a local optimum of (2). In contrast to vanilla ADMM, our algorithm allows us to obtain sub-optimal solutions of the MAP inference problem so that also computationally more challenging MRFs can be considered.
- We generalize the convergence of nonconvex vanilla ADMM to the presented inexact discrete-continuous ADMM.
- In diverse experiments we demonstrate the relevance and generality of our model and the efficiency of our method: In contrast to standard k-means, our model integrates well with deep features. In contrast to a tailored SDP relaxation approach for transductive logistic regression, our method produces more consistent results, while being more efficient in terms of both runtime and memory consumption.

1.2. Related work

To improve image segmentation results it is common practice to treat the unary terms E_i as additional variables in the optimization [4, 62, 7, 47, 53, 54, 55]. More recently, [53, 54] revealed the equivalence of k-means clustering with pairwise constraints [57, 2] and the Chan-Vese [7] approach, where the average foreground and background intensities (corresponding to the centroids in k-means) are not assumed to be fixed, but are rather treated as additional variables. The goal of this approach is to jointly cluster the pixels in the color space and regularize the cluster-assignment (the segmentation) in the image space. The clustering viewpoint suggests the application of the "kernel trick", which

allows us to separate more complicated, possibly nonlinearly deformed color clusters [57, 2, 53, 54]. Experimentally, it has been shown that this approach integrates well with color or even depth pixel features [53, 54]. Due to the enormous success of deep convolutional neural networks on computer vision tasks, it is tempting to replace the color features by more sophisticated deep features that are capable of compactly representing complicated semantic information [31, 10]. However, high-dimensional deep features are in general not "k-means friendly" [59] and without further preprocessing of the features as in [59] the plain Chan-Vese k-means approach does not generalize very well to deep features, despite of (almost) linear separability of the data. Since deep neural network classifiers can be viewed as a linear model on top of a deep feature extractor, we propose to alter the approaches from related work by using a (multiclass) SVM or a (multinomial) logistic regression model along with deep features. Under the absence of general higher order terms (i.e. $E_C = 0$, for all $C \in \mathcal{C}$) our model (2) is closely related to transductive SVMs [56, 21, 3] and transductive logistic regression [22]. In such a setting, optimization schemes that alternately optimize w.r.t. to labels and model parameters as in Lloyd's algorithm [37] are ineffective [61]. Other approaches, such as SDP relaxations are computationally expensive [22].

Instead, we propose an algorithm related to ADMM, which has recently been successfully applied to many nonconvex continuous optimization problems [9, 42, 50, 39, 34]. ADMM appears similar in form to message passing and subgradient descent schemes applied to the Lagrangian dual problem (dual decomposition) [29, 5, 38, 60, 51]. The latter is a Lagrangian relaxation approach, so that in difficult nonconvex cases the linear equality constraints may remain violated in the limit [29]. In contrast, ADMM attempts to solve the problem exactly and enforce the linear equality constraints strictly via additional quadratic penalty terms. In order to make mixed discrete-continuous problems such as (2) amenable to ADMM, related approaches often relax the discrete variable and perform rounding operations [20, 52]. In contrast, we propose a generalization of vanilla ADMM that preserves the integrality of the label variable and admits a theoretical convergence guarantee under affordable conditions. In the traditional convex and continuous setting, ADMM [18, 17] converges under mild conditions [16, 13]. For more restrictive nonconvex problems, its convergence has only been established recently [19, 36]. In this case, however, the required assumptions are fairly strong.

2. Discrete-Continuous ADMM

The coupling of the discrete labeling variable y and the continuous variable W renders problem (2) hard to solve. This is not surprising since the related k-means clustering problem is known to be NP-hard. A common approach is to compute a local optimum by a simple discrete-continuous coordinate descent approach as in Lloyd's algorithm [37]. Instead, we propose an advantageous decoupling into purely discrete and purely continuous subproblems, which allows us to compute a local optimum by updating the continuous and discrete variables jointly and efficiently.

2.1. Variable decoupling via ADMM

To this end, we employ a change of representation to make the proposed problem amenable to the "kernel trick". Note that, for any fixed labeling y, the lower-level task in (2) amounts to supervised SVM training (resp. supervised logistic regression). Thus, we can apply the representer theorem [48]: Let $\Phi(X)$ be the feature matrix for a (possibly infinte-dimensional) matrix feature map $\Phi : \mathbb{R}^{d\times|\mathcal{V}|} \to \mathbb{R}^{d'\times|\mathcal{V}|}$ and let

$$g(W) = h(\|W\|_F), \tag{3}$$

for $h\colon [0,\infty) \to \mathbb{R}$ strictly monotonically increasing. Then, the weights $W^\top = \Phi(X)\alpha$ can be substituted via their representation $\alpha \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{L}|}$ in terms of the features. More precisely, we replace the scalar products $W\varphi(x_i)$, up to transposition, by $K_i\alpha = (W\varphi(x_i))^\top$ where $K := \Phi(X)^\top\Phi(X)$ denotes the Gram or kernel matrix.

For $f : \mathbb{R}^{|\mathcal{V}|\times|\mathcal{L}|} \to \mathbb{R}$, being defined as

$$f(\alpha) := h(\|\Phi(X)\alpha\|_F), \tag{4}$$

this substitution leaves us with the following equivalent mixed integer nonlinear program formulation of (2):

$$\min_{\substack{y\in\mathcal{Y},\\ \alpha\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{L}|}}} \sum_{i\in\mathcal{V}} \ell(y_i; K_i\alpha) + f(\alpha) + \sum_{C\in\mathcal{C}} E_C(y). \tag{5}$$

In order to decompose problem (5) into simple subproblems associated with each $i \in \mathcal{V}$, we introduce auxiliary variables $\beta_i = K_i\alpha$, which yields

$$\begin{aligned} \min_{\substack{y\in\mathcal{Y},\\ \alpha,\beta\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{L}|}}} \quad & \sum_{i\in\mathcal{V}} \ell(y_i;\beta_i) + f(\alpha) + \sum_{C\in\mathcal{C}} E_C(y) \\ \text{subject to} \quad & K\alpha = \beta. \end{aligned} \tag{6}$$

Note that the objective of (6) is a separable function over the β_i. This suggests to relax the linear constraint $K\alpha = \beta$ and consider the equivalent saddle point problem:

$$\min_{\substack{y\in\mathcal{Y},\\ \alpha,\beta\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{L}|}}} \max_{\lambda\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{L}|}} \mathfrak{L}_\rho(\alpha,\beta,\lambda,y), \tag{7}$$

where $\lambda \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{L}|}$ are the Lagrange multipliers corresponding to $K\alpha = \beta$ and $\mathfrak{L}_\rho$ denotes the "discrete-continuous" augmented Lagrangian, that for some penalty

parameter $\rho > 0$ is defined as

$$\mathfrak{L}_\rho(\alpha, \beta, \lambda, y) := \sum_{i \in \mathcal{V}} \ell(y_i; \beta_i) + f(\alpha) + \sum_{C \in \mathcal{C}} E_C(y) + \langle \lambda, K\alpha - \beta \rangle + \frac{\rho}{2} \|K\alpha - \beta\|_F^2. \tag{8}$$

We show in Sec. 2.2 that, for fixed λ and α, the function $\mathfrak{L}_\rho(\alpha, \cdot, \lambda, \cdot)$ can be minimized (not necessarily to global optimality) efficiently and jointly over y and β. This central observation and the good practical performance of ADMM in nonconvex optimization motivates the following generalization to discrete-continuous problems of the form (7).

We propose an algorithm that, similar to continuous ADMM, updates the discrete-continuous variable-pair (β^{t+1}, y^{t+1}) via joint (and possibly suboptimal) minimization of $\mathfrak{L}_\rho(\alpha^t, \cdot, \lambda^t, \cdot)$. Subsequently, it updates α^{t+1} via minimization of $\mathfrak{L}_\rho(\cdot, \beta^{t+1}, \lambda^t, y^{t+1})$ and the Lagrange multiplier λ^t by performing one iteration of gradient ascent on $\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \cdot, y^{t+1})$ with step size $\rho > 0$. In summary, the update steps at iteration t are given as

$$(\beta^{t+1}, y^{t+1}) = \arg\min_{\beta, y} \mathfrak{L}_\rho(\alpha^t, \beta, \lambda^t, y), \tag{9}$$

$$\alpha^{t+1} = \arg\min_{\alpha} \mathfrak{L}_\rho(\alpha, \beta^{t+1}, \lambda^t, y^{t+1}), \tag{10}$$

$$\lambda^{t+1} = \lambda^t + \rho(K\alpha^{t+1} - \beta^{t+1}). \tag{11}$$

In practice, we choose the step size adaptively, as this often leads to better solutions in terms of objective value: For finitely many iterations, the penalty parameter ρ is increased according to the schedule $\rho_{t+1} = \min\{\rho_{\max}, \tau\rho_t\}$ with $\tau > 1$ and some $\rho_{\max} > 0$ that guarantees theoretical convergence of the algorithm (cf. Sec. 3).

2.2. Distributed solution of the subproblems

In this section, we describe the implementation of update steps (9)–(11) in our algorithm. In principle, (9) could be solved by minimization over β for every feasible labeling $y \in \mathcal{Y}$. Obviously, this is not a viable approach, as it implies performing exhaustive search over the set $\mathcal{Y}$, which has size $|\mathcal{L}|^{|\mathcal{V}|}$. Instead, we pursue the following more efficient strategy.

Solution via lookup-tables. Assume first the absence of any higher order energies, i.e. $E_C = 0$, for all $C \in \mathcal{C}$. Then, since $\mathfrak{L}_\rho(\alpha^t, \beta, \lambda^t, y)$ is separable w.r.t. β_i and y_i, we can decompose problem (9) into $|\mathcal{V}|$ independent problems of the form

$$\arg\min_{\beta_i, y_i} \underbrace{\ell(y_i; \beta_i) + \frac{\rho}{2} \|\beta_i - K_i\alpha^t - \lambda_i^t/\rho\|_F^2}_{\psi_i(\beta_i, y_i; \alpha^t, \lambda_i^t)}, \tag{12}$$

which can thus be solved in parallel. In the presence of higher-order energies, however, the problems (12) are not completely independent, because the variables y_i are coupled via the energies E_C in which they appear. In this case, we first solve (12) w.r.t. only the continuous variables β_i for *every* possible label $y_i \in \mathcal{L}$ and store the results in a lookup-table (u^{t+1}, B^{t+1}).

Precisely, for each $1 \leq i \leq |\mathcal{V}|$ and each $y_i \in \mathcal{L}$ we create an entry $(u_{i,y_i}^{t+1}, B_{i,y_i}^{t+1})$ according to

$$\begin{aligned} B_{i,y_i}^{t+1} &:= \arg\min_{\beta_i} \psi_i(\beta_i; y_i, \alpha^t, \lambda_i^t), \\ u_{i,y_i}^{t+1} &:= \min_{\beta_i} \psi_i(\beta_i; y_i, \alpha^t, \lambda_i^t). \end{aligned} \tag{13}$$

In a second step, we determine the discrete variable update y^{t+1} as the (possibly suboptimal) solution of the MRF

$$y^{t+1} = \arg\min_{y \in \mathcal{Y}} \sum_{i \in \mathcal{V}} u_{i,y_i}^{t+1} + \sum_{C \in \mathcal{C}} E_C(y). \tag{14}$$

Afterwards, the continuous variable updates β_i^{t+1} can be read off from the solution of (13) via

$$\beta_i^{t+1} = B_{i,y_i^{t+1}}^{t+1}. \tag{15}$$

Note that there is an abundance of algorithms available to tackle problems of the form (14) such as graph cuts for binary submodular MRFs [28], move making and message passing algorithms [12, 27], primal-dual algorithms [30, 14] and more. For an overview, see also [23].

The matrix u specifies the unary energies in problem (14) that pushes the MRF to attain a labeling which corresponds to a more suitable classifier. The latter is determined by a tradeoff between minimizing the distance of β_i to the current consensus parameters $K_i\alpha^t + \lambda_i^t/\rho$ and minimizing the loss term corresponding to sample i.

In case of suboptimality of y^{t+1} we require that y^{t+1}, for some $\delta \geq 0$, satisfies a (sufficient) descent condition

$$\mathfrak{L}_\rho(\alpha^t, \beta^{t+1}, \lambda^t, y^{t+1}) - \mathfrak{L}_\rho(\alpha^t, B_{:,y^t}^{t+1}, \lambda^t, y^t) \leq -\delta. \tag{16}$$

If this condition is violated, then we keep the previous iterate $y^{t+1} = y^t$. Under condition (16), the overall convergence of our algorithm is guaranteed (cf. Prop. 1 and Prop. 2 in Sec. 3). We summarize our method in Alg. 1.

Note that if the discrete subproblem (14) is solved to global optimality, our method specializes to classical nonconvex ADMM applied to a purely continuous problem $\min_\alpha E(K\alpha) + f(\alpha)$. The function $E(\beta) = \min_{y \in \mathcal{Y}} \sum_{i \in \mathcal{V}} \ell(y_i; \beta) + \sum_{C \in \mathcal{C}} E_C(y)$ encapsulates the minimization over the discrete labelings y. This results in a pointwise minimum over exponentially many functions.

Distributed optimization. Distributed optimization is considered one of the main advantages of ADMM in supervised learning [15, 5]. In our method, the (β, y) update

Algorithm 1 Discrete-Continuous ADMM

Require: initialize $\alpha^0, \lambda^0, \rho_0 > 0, \tau > 1$, $\rho_{\max}$ as in (18)
while (not converged) **do**
  Compute lookup-table (u^{t+1}, B^{t+1}):
  for all $i \in \{1, \dots, |\mathcal{V}|\}$ **and** $y_j \in \mathcal{L}$ **do**
    In parallel update $(u^{t+1}_{i,y_j}, B^{t+1}_{i,y_j})$ as in (13).
  end for
  Update y^{t+1} as in (14).
  if y^{t+1} violates condition (16) **then**
    $y^{t+1} \leftarrow y^t$.
  end if
  $\beta_i^{t+1} \leftarrow B^{t+1}_{i,y_i^{t+1}}$
  Perform updates, as in (10),(11).
  if ρ violates condition (18) **then**
    $\rho_{t+1} \leftarrow \min\{\rho_{\max}, \tau\rho_t\}$.
  end if
  $t \leftarrow t + 1$
end while

requires the solution of only $|\mathcal{L}| \cdot |\mathcal{V}|$ many (instead of $|\mathcal{L}|^{|\mathcal{V}|}$ for the naive approach) independent and small-scale continuous minimization problems of the form (13) and one additional discrete problem (14). This suggests the distributed solution of the subproblems (13), for instance on a GPU. Subsequently, the optimization of the MRF (14) and the update of the consensus variable is carried out after gathering the solutions of the subproblems. Since (14) need not be solved to optimality, the MRF solver may be stopped early to speed up computation. This is particularly useful if a primal-dual algorithm for solving the LP-relaxation is used [30, 14].

Exploit duality. If the loss terms $\ell(y_i; \cdot)$ are convex and lower semicontinuous, then the independent subproblems (13) can be solved efficiently via duality as follows. For all loss functions we consider, it is convenient to solve the dual problem as it scales linearly with the number of training samples (which is equal to one in our case). For the Crammer and Singer multiclass SVM loss [11], for instance, there exists an efficient variable fixing algorithm [25] for solving the dual problem. For the softmax loss the dual problem reduces to a one-dimensional nonlinear equation via the Lambert-W function [35] and may be solved by performing a few iterations of Newton's or Halley's method. For the special case of the one-vs.-all hinge loss, (13) can be solved in closed form. In any case, each subproblem involves only a small number of instructions, which is important for a GPU-based implementation.

2.3. Consensus update

For a quadratic regularizer $h(x) = \nu x^2$, where ν is the regularization parameter, the update step (10) is equivalent to

$$\alpha^{t+1} = \arg\min_{\alpha} \; \nu \langle \alpha, K\alpha \rangle + \frac{\rho}{2} \|K\alpha - \beta^{t+1} + {}^{\lambda^t}\!/_{\rho}\|_F^2. \tag{17}$$

This is a quadratic problem that can be solved via a normal equation using either a cached eigenvalue decomposition of the kernel matrix, or an iterative algorithm such as conjugate gradient (CG). The latter is preferred for large scale applications, as each CG iteration involves a kernel-matrix-vector multiplication Kv. For the linear kernel $K = X^\top X$, this guarantees efficiency of our method, since K does not have to be stored explicitly. For general kernels such as the RBF kernel, a low rank approximation to the kernel matrix $K \approx GG^\top$, for some $G \in \mathbb{R}^{|\mathcal{V}| \times l}$ with $l \ll |\mathcal{V}|$ can be obtained for instance via the Nyström method [41, 58] or random features [45]. Furthermore, in practice, often only a small number of conjugate gradient iterations are necessary.

3. Convergence analysis

In this section, we provide a complete convergence analysis of the proposed algorithm. To this end, we make the following assumptions:

- The function f is L-smooth, m-semiconvex and lower-bounded, i.e. f is differentiable and ∇f is Lipschitz-continuous with modulus L and there exists $m > 0$ sufficiently large so that $f + \frac{m}{2}\|\cdot\|_F^2$ is convex.
- For all $y_i \in \mathcal{L}$, $\ell(y_i; \cdot)$ is lower-bounded.
- The kernel matrix $K \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ is surjective, i.e. the smallest eigenvalue $\sigma_{\min}(K^\top K) > 0$ is positive.
- After finitely many iterations t the penalty parameter ρ is sufficiently large and kept fixed such that

$$\frac{L^2}{\rho\sigma_{\min}(K^\top K)} + \frac{m - \rho\sigma_{\min}(K^\top K)}{2} < 0. \tag{18}$$

When the MRF subproblem is solved to global optimality, convergence can be guaranteed by considering a pointwise minimum over exponentially many augmented Lagrangians and applying existing theory [36, 19]. For the general case, however, the theory needs to be extended. Our convergence proof borrows arguments from [36, 19], where the convergence of ADMM in the nonconvex setting is shown via a monotonic decrease of the augmented Lagrangian. In our case, for a sufficiently large penalty parameter ρ, we can achieve a monotonic decrease of the "discrete-continuous" augmented Lagrangian (8), even if the MRF subproblem (14) is not solved globally optimal. This allows us to stop exact MRF solvers early or to apply heuristic solvers if computing global optima is intractable.

For the complete proofs of all the theoretical results, presented in this section, cf. the supplementary material.

Lemma 1. *Let $K \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ be surjective and $\delta \geq 0$. For ρ meeting condition (18) we have that*

1. *The "discrete-continuous" augmented Lagrangian (8) decreases monotonically with the iterates $(\alpha^t, \beta^t, \lambda^t, y^t)$:*

$$\begin{aligned}&\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1}) - \mathfrak{L}_\rho(\alpha^t, \beta^t, \lambda^t, y^t)\\ &\leq \left(\tfrac{L^2}{\rho\sigma_{\min}(K^\top K)} + \tfrac{m-\rho\sigma_{\min}(K^\top K)}{2}\right)\|\alpha^{t+1} - \alpha^t\|_F^2\\ &\qquad - \delta[\![y^{t+1} \neq y^t]\!], \quad (19)\end{aligned}$$

   where $[\![\cdot]\!]$ denotes the Iverson bracket.

2. *$\{\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1})\}_{t\in\mathbb{N}}$ is lower bounded.*

3. *$\{\mathfrak{L}_\rho(\alpha^{t+1}, \beta^{t+1}, \lambda^{t+1}, y^{t+1})\}_{t\in\mathbb{N}}$ converges.*

We are now able to guarantee that feasibility is achieved in the limit. This is in contrast to a dual-decomposition approach [29, 60, 51] or a Gauss-Seidel quadratic penalty method (with finite penalty parameter ρ), used for instance in [49, 24], where a violation of the consensus constraint remains in the limit. Moreover, if $\delta > 0$ is chosen strictly positive, then the discrete variable is guaranteed to converge, i.e. for T sufficiently large, we have $y^{t+1} = y^t$ for all $t > T$.

Lemma 2. *Let $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ be the iterates produced by Alg. 1. Then $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ is a bounded sequence. Furthermore, for $t \to \infty$ the distance of two consecutive continuous iterates vanishes, and feasibility is achieved in the limit:*

$$\|\alpha^{t+1} - \alpha^t\|_F \to 0, \quad (20)$$
$$\|\beta^{t+1} - \beta^t\|_F \to 0, \quad (21)$$
$$\|\lambda^{t+1} - \lambda^t\|_F \to 0, \quad (22)$$
$$\|K\alpha^{t+1} - \beta^{t+1}\|_F \to 0. \quad (23)$$

Finally, if $\delta > 0$ is chosen strictly positive, then there exists some $T \in \mathbb{N}$ such $y^{t+1} = y^t$ for all $t > T$.

The limit points of our algorithm correspond to "discrete-continuous" critical points of the augmented Lagrangian.

Definition 1 ("Discrete-continuous" critical point)**.** *We call $(\alpha^*, \beta^*, \lambda^*, y^*)$ a* "discrete-continuous" critical point *of the "discrete-continuous" augmented Lagrangian (8) if it satisfies*

$$0 \in \partial(\ell(y_i^*; \cdot))(\beta_i^*) - \lambda_i^*, \qquad \forall\, i \in \mathcal{V} \quad (24)$$
$$0 \in \partial g(\alpha^*) + K^\top\lambda^* \quad (25)$$
$$K\alpha^* = \beta^*, \quad (26)$$

for y^* with $E_C(y^*) < \infty$ for all $C \in \mathcal{C}$. Here, $\partial f(x)$ denotes the "limiting" subdifferential [46, Definition 8.3] of the function f at x with $f(x) < \infty$.

Proposition 1. *Let $\delta \geq 0$. Then any limit point $(\alpha^*, \beta^*, \lambda^*, y^*)$ of the sequence $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ is a "discrete-continuous" critical point.*

Finally, under convexity of f and $\ell(y_i; \cdot)$, for all $y_i \in \mathcal{L}$ and strictly positive $\delta > 0$, we can guarantee that the sequence of iterates produced by Alg. 1 globally converges to a point $(\alpha^*, \beta^*, \lambda^*, y^*)$ which has the following property: α^* is the global optimum of the supervised learning problem w.r.t. the estimated training labels y^*:

$$\alpha^* = \arg\min_\alpha \sum_{i\in\mathcal{V}} \ell(y_i^*; K_i\alpha) + f(\alpha). \quad (27)$$

Proposition 2. *Let $\ell(y_i; \cdot)$ and g be proper, convex and lower-semicontinuous and let $\delta > 0$. Then the sequence $\{(\alpha^t, \beta^t, \lambda^t, y^t)\}_{t\in\mathbb{N}}$ produced by Alg. 1 converges to a "discrete-continuous" critical point $(\alpha^*, \beta^*, \lambda^*, y^*)$ of (8) and α^* solves the problem (27) to global optimality.*

Discussion of the assumptions. Note that in general the kernel matrix K is not surjective. However, for the strictly positive definite RBF kernel, K is strictly positive definite so that convergence can be achieved for finite ρ. In order to enforce theoretical convergence for general kernels, we may add a small constant to the diagonal of the kernel matrix $K := K + \gamma I$ that alters the model only slightly. In fact, for the binary SVM, this change is equivalent to replacing the hinge loss with its square.

4. Experiments

In this section, we present the experimental results of our method on several transductive learning tasks. First, we compare our method to an SDP relaxation method for transductive multinomial logistic regression by [22]. Second, we use our model and solver for the tasks of object video segmentation as well as image segmentation with user interaction, showing improvements on the false positive rate of object pixels.

4.1. Comparison with SDP relaxation for transductive learning

In this experiment, we consider the standard SSL benchmark [8] for a comparison with the SDP relaxation method for transductive multinomial logistic regression by [22]. The benchmark is a collection of several datasets, with varying feature dimensions and number of classes. Each dataset is provided with 12 splits into $l = 10$ or $l = 100$ labeled and $N - l$ unlabeled samples. We introduce additional unary energies E_C with $|C| = 1$ for all the labeled examples, to constrain their label to be fixed during optimization. While [22] incorporates an entropy prior on the labeling which favors an equal balance distribution, we introduce a higher

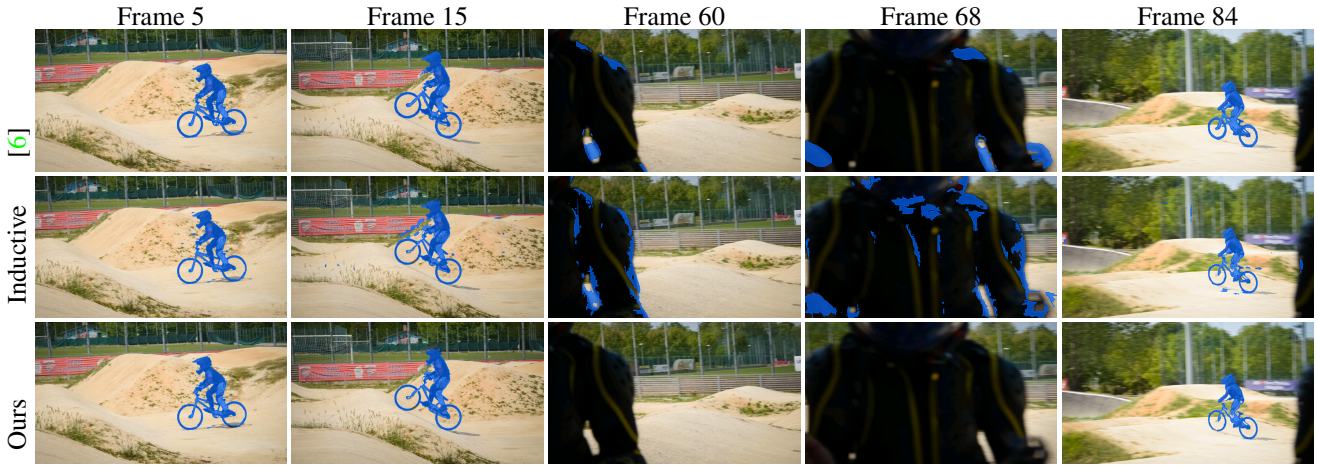

Figure 2: Exemplary results for video object segmentation on the DAVIS benchmark [44]. It can be seen that both, the inductive MRF inference approach and OSVOS produce a large number of false positive object pixels for the frames where the object is occluded.

Table 1: Comparison with the method of [22] on the SSL benchmark [8]. Reported are the average label-accuracy (in %) and variance over the splits. Our evaluation suggests that our method performs better for a standard hyperparameter setting except for three out of 20 settings. Moreover, it produces more consistent results, i.e. lower variances over the splits.

	Linear Kernel		RBF Kernel	
Dataset	SDP	Ours	SDP	Ours
Digit1,10l	69.27±27.56	**82.20±4.54**	53.93±9.43	**78.18±8.43**
USPS,10l	57.72±13.73	**64.58±3.37**	40.10±11.67	**48.19±5.84**
BCI,10l	50.44±3.16	**50.62±2.08**	50.00±3.06	**51.67±0.44**
g241c,10l	49.88±38.92	**55.42±3.95**	62.33±36.84	**89.98±0.32**
g241n,10l	52.77±34.37	**57.61±4.44**	50.13±0.53	**51.13±0.13**
Digit1,100l	75.74±29.73	**85.60±2.91**	**88.65±0.49**	87.61±3.44
USPS,100l	63.44±9.97	**72.14±0.84**	39.83±12.63	**56.54±3.31**
BCI,100l	60.58±6.87	**65.23±1.25**	**64.19±1.23**	62.62±1.00
g241c,100l	64.92±17.47	**86.31±0.91**	85.63±0.76	**89.34±1.07**
g241n,100l	**54.14±17.13**	54.11±0.64	52.23±1.61	**53.98**±0.38

order potential E_C, with $C = \mathcal{V}$, that restricts the solution to deviate at most 10 percent from the equal balance distribution. We solve the LP-relaxation of the higher-order MRF subproblem (14) with the dual-simplex method and round the solution. The baseline results are computed with a MATLAB implementation that is provided by the authors. For these experiments, we use the softmax loss and set the regularization parameter $\nu = 0.05$ for the linear kernel. For the RBF kernel we manually chose the variance parameter $\sigma = 0.5477$ and the regularization parameter $\nu = 0.0025$. We chose the initial penalty parameter $\rho_0 = 0.001$ and $\tau = 1.003$. All values are averaged over 12 different splits. The evaluation in Tab. 1 suggests, that our method performs better for a standard hyperparameter setting except for three out of 20 settings. Moreover, it produces more consistent results, i.e. lower variances over the splits, which suggests that our method is more robust towards noise and poorly labeled data.

4.2. Video object segmentation

In this experiment, we evaluated our method on video object segmentation. Here, the task is to segment an object throughout a video, given its mask in the first frame. This problem has been successfully approached by [6], using end-to-end deep learning with fully convolutional neural networks. At test time, their classifier is fine-tuned on the appearance of the object and the background in the first frame and predicts the object pixels of individual later frames. However, this method struggles with drastic appearance changes of the object, which have not been learned in advance. These include pose changes, sharp lighting and background changes or severe occlusions as shown in Fig. 2.

We propose to use a transductive approach instead. More precisely, we use the pre-trained (not fine-tuned) OSVOS parent network [6] as a deep feature extractor and a MRF model with a variable classifier in the form of (2). We use a simple linear kernel SVM in our model, as the extracted deep features are almost linearly separable. Further, we introduce unary indicator energies E_i to fix the labels of the user-annotated pixels in the first frame and pairwise energies E_{ij} for adjacent pixels in any frame to favor spatially smooth solutions. Similar to [6], we do not use any temporal consistency terms. To reduce the number of examples, we apply our method on a superpixel level and extract 6000 super-pixels [1] for each frame and apply average pooling

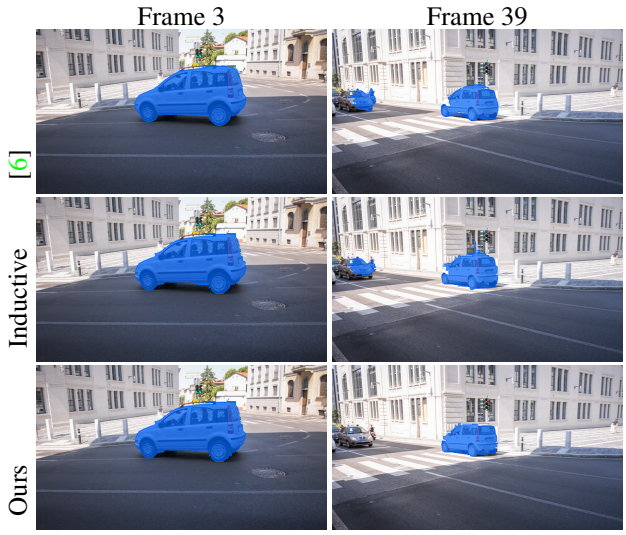

Figure 3: Further results for video object segmentation on the car-shadow sequence from the DAVIS benchmark [44].

over the superpixels. We compare the proposed transductive approach to OSVOS [6] and the classical (inductive) MRF inference approach (where the classifier is learned with the first frame only) on the DAVIS benchmark [44]. For both the inductive and the transductive approach the used linear kernel SVM model, the higher order energies E_C and the extracted superpixels are the same. The results are shown in Fig. 2. It can be seen that [6] works well as long as the appearance of the object and background are sufficiently similar to the first frame (first column). In frames 60 to 68, where the object is occluded, both the inductive MRF inference approach and OSVOS produce a large number of false positive object pixels. In this experiment the intersection-over-union scores (the higher the better) are 0.7087 for our method, 0.6452 for OSVOS and 0.5063 for the inductive approach. Similarly in the car-shadow sequence, OSVOS and the inductive approach mask additionally the other car and the motorbike in frame 39 (cf. Fig. 3). In contrast our method masks the correct car only. Here, the intersection-over-union scores are 0.9262 for OSVOS, 0.9196 for our method and 0.8844 for the inductive approach.

4.3. Image segmentation with user interaction

We evaluated our method on the task of interactive foreground-background segmentation with deep features. Like in the previous experiment we used OSVOS as a deep feature extractor. On this task we compare our method to the Chan-Vese kernel k-means approach proposed in [53, 54] as a baseline method. Since the features are almost linearly separable, we use a simple linear kernel for both our model and the baseline model. As it is shown in Fig. 4, the k-means approach often fails to find a good cluster-center

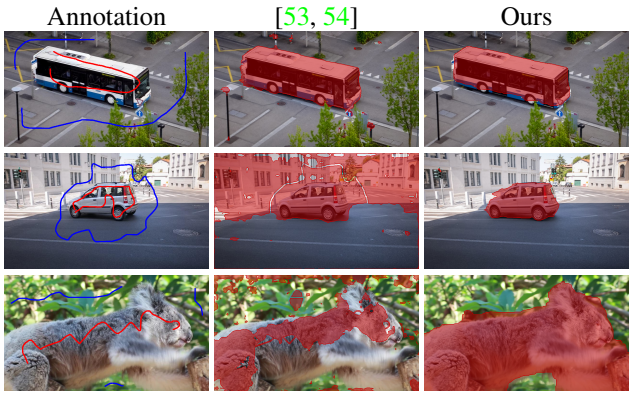

Figure 4: Exemplary results for interactive binary image segmentation with deep features. **Left**: Input images along with user scribbles in red for foreground and blue for background. **Middle**: Segmentation results (red masks) obtained from k-means. **Right**: Segmentation results (red masks) obtained with the proposed method.

assignment, despite of strong supervision (provided in the form of user-scribbles) and richness of the features. This is due to the fact that deep high dimensional features are in general not k-means friendly [59], which means further pre-processing or a k-means suited kernel would be required. In contrast, our method provides a reasonable result for all cases, without the need for feature-pre-processing or kernel-parameter tuning.

5. Conclusion

We considered the joint solution of MAP-inference in MRFs and parameter learning, which can be viewed as a transductive inference problem. To solve this task, we proposed a novel algorithm that jointly optimizes over the discrete label variables and the continuous model parameters. The proposed method is related to classical ADMM from continuous optimization and admits a convergence proof under suitable assumptions even though the objective function is discrete-continuous and nonconvex. Our algorithm makes use of a decoupling of the problem into purely discrete and purely continuous subproblems and can be implemented in a distributed fashion. We evaluated our approach in several experiments including video object segmentation and interactive image segmentation. Our results suggest that the proposed optimization method performs favorable compared to alternating optimization (as in k-means) and convex relaxations. In particular, this indicates that the presented method also serves as an alternative approach to optimization problems arising in semi-supervised or transductive learning, e.g., in the case of SVMs. Furthermore, the visual results show that the transductive inference model is able to reduce the hallucination of false object pixels in image and video segmentation tasks.

References

[1] R. Achanta and S. Susstrunk. Superpixels and polygons using simple non-iterative clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017. 7

[2] S. Basu, I. Davidson, and K. Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 1 edition, 2008. 2, 3

[3] K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems, NIPS*, pages 368–374, 1999. 3

[4] A. Blake and A. Zisserman. *Visual reconstruction*. MIT press, 1987. 2

[5] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 3, 4

[6] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017. 7, 8

[7] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001. 2

[8] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006. 6, 7

[9] R. Chartrand and B. Wohlberg. A nonconvex ADMM algorithm for group sparsity with sparse groups. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6009–6013, May 2013. 3

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014. 1, 3

[11] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. 5

[12] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2173–2180, June 2010. 1, 4

[13] J. Eckstein and D. P. Bertsekas. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992. 3

[14] A. Fix, C. Wang, and R. Zabih. A primal-dual algorithm for higher-order multilabel markov random fields. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1138–1145, 2014. 4, 5

[15] P. A. Forero, A. Cano, and G. B. Giannakis. Consensus-based distributed support vector machines. *Journal of Machine Learning Research*, 11:1663–1707, 2010. 4

[16] D. Gabay. Applications of the method of multipliers to variational inequalities. *Studies in Mathematics and Its Applications*, 15:299–331, 1983. 3

[17] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976. 3

[18] R. Glowinski and A. Marrocco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue Française d'Automatique, Informatique, Recherche Opérationnelle. Analyse Numérique*, 9(2):41–76, 1975. 3

[19] M. Hong, Z.-Q. Luo, and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016. 3, 5

[20] K. Huang and N. D. Sidiropoulos. Consensus-ADMM for general quadratically constrained quadratic programming. *IEEE Transactions on Signal Processing*, 64(20):5297–5310, 2016. 3

[21] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning ICML*, pages 200–209, 1999. 3

[22] A. Joulin and F. R. Bach. A convex relaxation for weakly supervised classifiers. In *Proceedings of the 29th International Conference on Machine Learning, ICML*, 2012. 3, 6, 7

[23] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, J. Lellmann, N. Komodakis, and C. Rother. A comparative study of modern inference techniques for discrete energy minimization problem. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2013. 4

[24] S. Kim, D. Min, S. Lin, and K. Sohn. Dctm: Discrete-continuous transformation matching for semantic flow. In *IEEE International Conference on Computer Vision, ICCV*, Oct 2017. 6

[25] K. C. Kiwiel. Variable fixing algorithms for the continuous quadratic knapsack problem. *Journal of Optimization Theory and Applications*, 136(3):445–458, 2008. 5

[26] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. 1

[27] V. Kolmogorov. A new look at reweighted message passing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):919–930, 2015. 4

[28] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. 1, 4

[29] N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *IEEE 11th International Conference on Computer Vision, ICCV*, pages 1–8. IEEE Computer Society, 2007. 3, 6

[30] N. Komodakis, G. Tziritas, and N. Paragios. Fast, approximately optimal solutions for single and dynamic mrfs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1–8. IEEE, 2007. 1, 4, 5

[31] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, NIPS*, 2012. 3

[32] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1056–1077, 2014. 1

[33] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning, ICML*, pages 282–289, 2001. 1

[34] R. Lai and S. Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449, 2014. 3

[35] M. Lapin, M. Hein, and B. Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. 5

[36] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015. 3, 5

[37] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. 2, 3

[38] A. F. T. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing. An augmented lagrangian approach to constrained MAP inference. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML*, pages 169–176, 2011. 3

[39] O. Miksik, V. Vineet, P. Pérez, and P. H. S. Torr. Distributed non-convex ADMM-based inference in large-scale random fields. In M. F. Valstar, A. P. French, and T. P. Pridmore, editors, *British Machine Vision Conference, BMVC*, 2014. 3

[40] S. Nowozin and C. H. Lampert. Structured prediction and learning in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3–4), 2010. 1

[41] E. J. Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54(1):185–204, 1930. 5

[42] V. Ozoliņš, R. Lai, R. Caflisch, and S. Osher. Compressed modes for variational problems in mathematics and physics. *Proceedings of the National Academy of Sciences*, 110(46):18368–18373, 2013. 3

[43] R. R. Paulsen, J. A. Bærentzen, and R. Larsen. Markov random field surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, pages 636–646, 2010. 1

[44] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. J. V. Gool, M. H. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 724–732, 2016. 7, 8

[45] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems, NIPS*, pages 1177–1184, 2008. 5

[46] R. Rockafellar and R.-B. Wets. *Variational Analysis*. Springer, 1998. 6

[47] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004. 2

[48] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *14th Annual Conference on Computational Learning Theory, COLT*, volume 2111, pages 416–426, 2001. 3

[49] F. Steinbrücker, T. Pock, and D. Cremers. Large displacement optical flow computation withoutwarping. In *IEEE 12th International Conference on Computer Vision, ICCV*, pages 1609–1614, 2009. 6

[50] M. Storath, A. Weinmann, and L. Demaret. Jump-sparse and sparse recovery using potts functionals. *IEEE Transactions on Signal Processing*, 62(14):3654–3666, 2014. 3

[51] P. Swoboda and B. Andres. A message passing algorithm for the minimum cost multicut problem. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, July 2017. 3, 6

[52] R. Takapoui, N. Moehle, S. Boyd, and A. Bemporad. A simple effective heuristic for embedded mixed-integer quadratic programming. *International Journal of Control*, pages 1–23, 2017. 3

[53] M. Tang, I. B. Ayed, D. Marin, and Y. Boykov. Secrets of grabcut and kernel k-means. In *IEEE International Conference on Computer Vision, ICCV*, pages 1555–1563, 2015. 2, 3, 8

[54] M. Tang, D. Marin, I. B. Ayed, and Y. Boykov. Normalized cut meets MRF. In *European Conference on Computer Vision, ECCV*, pages 748–765, 2016. 2, 3, 8

[55] V. Trajkovska, P. Swoboda, F. Åström, and S. Petra. Graphical model parameter learning by inverse linear programming. In *Scale Space and Variational Methods in Computer Vision - 6th International Conference, SSVM*, pages 323–334, 2017. 2

[56] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995. 2, 3

[57] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the 18th International Conference on Machine Learning, ICML*, pages 577–584. Morgan Kaufmann, 2001. 2, 3

[58] C. K. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems, NIPS*, pages 682–688, 2001. 5

[59] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, pages 3861–3870, 2017. 3, 8

[60] C. Zach. Dual decomposition for joint discrete-continuous optimization. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 31 of *Proceedings of Machine Learning Research*, pages 632–640, 2013. 3, 6

[61] K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. In Z. Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning, ICML*, volume 227, pages 1119–1126, 2007. 2, 3

[62] S. C. Zhu and A. L. Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image

segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996. 2