Submodular Relaxation for MRFs with High-Order Potentials

Anton Osokin and Dmitry Vetrov

Moscow State University, Russia
anton.osokin@gmail.com,vetrovd@yandex.ru

Abstract. In the paper we propose a novel dual decomposition scheme for approximate MAP-inference in Markov Random Fields with sparse high-order potentials, i.e. potentials encouraging relatively a small number of variable configurations. We construct a Lagrangian dual of the problem in such a way that it can be efficiently evaluated by minimizing a submodular function with a min-cut/max-flow algorithm. We show the equivalence of this relaxation to a specific type of linear program and derive the conditions under which it is equivalent to generally tighter LP-relaxation solved in [1]. Unlike the latter our relaxation has significantly less dual variables and hence is much easier to solve. We demonstrate its faster convergence on several synthetic and real problems.

Key words: Markov random fields, energy minimization, MAP-inference, dual decomposition, high-order potentials

1 Introduction

Markov random fields (MRFs) are a popular approach for analyzing interdependent data. Its main advantage is the ability to express the joint distribution over the whole set of hidden variables in terms of a product of potentials usually depending on small subsets of the variables. Small-order MRFs became very wide-spread in many tasks that arise in computer vision mainly for regularization purposes since they allowed to take the context (or neighborhood) into account. However, recently it has been shown that the use of high-order terms can improve the accuracy in many vision problems (see, e.g. [2]).

One of the most important tasks in any MRF is the search of most probable configuration, also known as MAP-inference. In this paper we will describe this task in terms of minimizing the negative log-likelihood of MRF (energy).

Recently, the minimization of MRF energies with high-order potentials has attracted a lot of attention. One line of research is to develop reductions of general high-order energies to equivalent pairwise ones. The drawback of this approach is that the resulting minimization problem can be exponentially large [3]. However, for some classes of high-order potentials this transformation can be performed efficiently: Rother et al. in [4] show that sparse or pattern-based potentials (potentials that encourage small number of predefined node configurations) can be efficiently transformed into pairwise ones.

A popular way to minimize energies with high-order potentials without a reduction is applying a specialized version of α-expansion algorithm [5] that takes those potentials into account on each expansion-step. This approach can be applied only for specific type of high-order potentials: $\mathcal{P}^n$ Potts potentials [6] and their robust version [2] encourage all the nodes in a specific area to take the same label, label costs [7] penalize the total number of labels used in the solution. Although sometimes being efficient in practice this approach usually does not give any guarantees on the solution and can get stuck in local minima.

Another approach to minimizing energies with high-order potentials is to generalize the dual decomposition scheme proposed in [8]. The most obvious generalization adds a subproblem per each high-order potential (in what follows we refer to this approach by clique-wise decomposition, CWD)[1]. As any dual decomposition method CWD ends up with a convex non-smooth dual problem and is guaranteed to solve a specific LP-relaxation of the initial problem. The main drawback of CWD is its high computational complexity induced by two factors: the dual problem is non-smooth and its dimensionality is high. Komodakis and Paragios in [1] show that in case of pattern-based potentials it is possible to cleverly combine multiple high-order potentials in one subproblem and thus improve both the convergence speed and tightness of the relaxation.

In all dual decomposition methods the final optimization problem is convex but non-smooth and thus quite difficult to solve. The most standard way to tackle it is to use different subgradient-based schemes (see [8] for review). This methods are intuitive, produce almost zero computational overhead, but are not robust and are very sensitive to parameter choice. Finding better ways to optimize the dual is an area of active research (see e.g. [9, 10]).

In this paper we try to reduce the dimensionality of the dual problem and thus make it easier to solve. We propose a quite general framework that is based on a Lagrangian decomposition that extends recently proposed Submodular Decomposition method (SMD) [11]. Our approach in theory deals with arbitrary high-order potentials but is practical only for pattern-based potentials and their robust versions (including (robust) $\mathcal{P}^n$ Potts and label costs). We provide the theoretical analysis of our approach and its empirical evaluation.

The rest of the paper is organized as follows: we present submodular relaxation algorithm in section 2; section 3 explores its theoretical properties including convergence points, persistency property and considers an important special case of permuted $\mathcal{P}^n$ Potts potentials; in section 4 we provide experiments and and finish with a conclusion in section 5.

2 Submodular relaxation

Consider hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{C})$ where $\mathcal{V}$ is a set of nodes and $\mathcal{C}$ is a set of hyperedges. Let $x_i \in \{1, \dots, K\} = \mathcal{K}$, $i \in \mathcal{V}$ be a discrete variable (label) associated with each node. Consider a problem of minimizing the following energy

[1] In [1] this approach was referred to as "generic optimization".

(negative log-likelihood of an MRF up to a constant):

$$E(X) = \sum_{i \in \mathcal{V}} \phi_i(x_i) + \sum_{c \in \mathcal{C}} \phi_c(\boldsymbol{x}_c), \tag{1}$$

where $\phi_i(x_i)$, $\phi_c(\boldsymbol{x}_c)$ are unary and high-order potentials, $\boldsymbol{x}_c = (x_{i_1^c}, \dots, x_{i_{L(c)}^c})$ is a labeling of nodes incident to hyperedge c; $L(c)$ stands for the order of hyperedge c. Energy (1) can be rewritten in terms of indicator variables $y_{ik} \in \{0,1\}$ ($y_{ik} = 1 \Longleftrightarrow x_i = k$):

$$E(Y) = \sum_{i \in \mathcal{V}} \sum_{k=1}^{K} \phi_i(k) y_{ik} + \sum_{c \in \mathcal{C}} \sum_{\boldsymbol{k} \in \mathcal{K}^{L(c)}} \phi_c(\boldsymbol{k}) \prod_{l=1}^{L(c)} y_{i_l^c k_l}. \tag{2}$$

Unconstrained minimization of energy (1) over multi-label variables X is equivalent to minimization of energy (2) over binary variables Y under consistency constraints:

$$\sum_{k=1}^{K} y_{ik} = 1, \ \forall i \in \mathcal{V}. \tag{3}$$

Note that by adding constant term we may always ensure $\phi_c(\boldsymbol{k}) \leq 0$ for all c. In this case we can use identity $\left(-\prod_{l=1}^{L} y_{i_l}\right) = \min_{z \in \{0,1\}} \left((L-1)z - \sum_{l=1}^{L} y_{i_l} z\right)$ to transform high-order energy (2) into pairwise energy in such a way that $\min_{Y \in (3)} E(Y) = \min_{Z, Y : Y \in (3)} E(Y, Z)$[2].

Note that for general high-order potentials $\phi_c(\boldsymbol{x}_c)$ function $E(Y, Z)$ depends on exponentially many variables Z. In what follows we assume the pattern-based (or sparse) form of $\phi_c(\boldsymbol{x}_c)$, i.e. most of values are equal to zero and others are negative. Specifically, denote labelings of potential c that we encourage by $\mathcal{D}^c = \{\boldsymbol{d}^c\} = \{(d_1^c, \dots, d_{L(c)}^c)\}$[3]. Using this notation $E(Y, Z)$ can be written as follows:

$$E(Y,Z) = \sum_{i \in \mathcal{V}} \sum_{k=1}^{K} \phi_i(k) y_{ik} - \sum_{c \in \mathcal{C}} \sum_{\boldsymbol{d}^c \in \mathcal{D}^c} \phi_c(\boldsymbol{d}^c) \left((L(c)-1) z_{c,\boldsymbol{d}^c} - \sum_{l=1}^{L(c)} y_{i_l^c d_l^c} z_{c,\boldsymbol{d}^c} \right). \tag{4}$$

Energy $E(Y, Z)$ is submodular w.r.t. variables Y and Z and thus in absence of additional constraints can be efficiently minimized by min-cut/max-flow algorithms [12]. Adding and relaxing constraints (3) to the minimization of (4) gives us the following Lagrangian dual:

$$D(\Lambda) = \min_{Y,Z \in \{0,1\}} L(Y, Z, \Lambda) = \min_{Y,Z \in \{0,1\}} \left(E(Y,Z) + \sum_{i \in \mathcal{V}} \lambda_i \left(\sum_{k=1}^{K} y_{ik} - 1 \right) \right). \tag{5}$$

[2] This transformation of binary high-order function (2) is in fact equivalent to a special case Type-II binary transformation of [4], but in this form it was proposed much earlier (see e.g. [3] for review).

[3] All the remaining derivations in this section can be generalized to a robust version of pattern based potentials using robust Type-II transformation of [4].

Function $D(\Lambda)$ is a lower bound on energy (1) and is concave but non-smooth. Thus, it can be maximized e.g. by subgradient algorithms. Note that $L(Y, Z, \Lambda)$ remains submodular w.r.t. (Y, Z) for all Λ allowing us to efficiently evaluate D at any point. We refer to this approach as submodular relaxation (SMR).

It is worth emphasizing that in contrast to CWD and PatB methods of [1] number of dual variables in our approach does not depend on the number of high-order potentials (always equals to $|\mathcal{V}|$). In CWD dual there are $K \sum_{c \in \mathcal{C}} L(c)$ variables. PatB reduces this number by combining multiple high-order potentials in one subproblem but will still have in at least in a factor of K more variables than SMR for the case when high-order potentials densely cover all variables. SMR can be naturally combined with SMD [11] and include pairwise potentials without the increase of the number of dual variables but for both CWD and PatB this modification will require additional subproblems thus increasing the dimensionality of the dual even further.

3 Theoretical properties

In this section we explore some properties of the relaxation that is solved by maximizing lower bound $D(\Lambda)$ and address some practical issues of its application to inference problems.

3.1 General case

Theorem 1. *Maximum of the SMR dual function $D(\Lambda)$ (5) is equal to the solution of the following linear program:*

$$\min_{Y,Z} \quad Q(Y, Z) \tag{6}$$

$$s.t. \quad y_{ik}, z_{c,\boldsymbol{d}^c} \in [0, 1], \quad \forall i \in \mathcal{V}, k \in \mathcal{K}, c \in \mathcal{C}, \boldsymbol{d}^c \in \mathcal{D}^c \tag{7}$$

$$z_l^{c,\boldsymbol{d}^c} \leq z_{c,\boldsymbol{d}^c}, \quad z_l^{c,\boldsymbol{d}^c} \leq y_{i_l^c d_l^c}, \quad \forall c \in \mathcal{C}, \boldsymbol{d}^c \in \mathcal{D}^c, \ \forall l = 1, \ldots, L(c) \tag{8}$$

$$\sum_{k=1}^{K} y_{ik} = 1, \quad \forall i \in \mathcal{V} \tag{9}$$

where the target function $Q(Y, Z)$ is defined as follows:

$$Q(Y, Z) = \sum_{i \in \mathcal{V}} \sum_{k=1}^{K} \phi_i(k) y_{ik} - \sum_{c \in \mathcal{C}} \sum_{\boldsymbol{d}^c \in \mathcal{D}^c} \phi_c(\boldsymbol{d}^c) \left((L(c) - 1) z_{c,\boldsymbol{d}^c} - \sum_{l=1}^{L(c)} z_l^{c,\boldsymbol{d}^c} \right).$$

Proof. Denote

$$R(\Lambda) = \min_{Y,Z \in (7),(8)} \left(Q(Y, Z) + \sum_{i \in \mathcal{V}} \lambda_i \left(\sum_{k=1}^{K} y_{ik} - 1 \right) \right).$$

Recall that $\phi_c(\boldsymbol{d}^c) \leq 0$. In this case problem (6), (7), (8) is equivalent to the standard (Schlesinger's) LP-relaxation of binary submodular energy (4) which

is known to be tight (see, e.g. [13]) and hence $R(\Lambda) = D(\Lambda)$. Consider $\Lambda^* = \arg\max R(\Lambda)$. Due to the strong duality in LP problems there exist primal and dual variables such that

$$(\Lambda^*, Y^*, Z^*) = \arg\max_{\Lambda} \min_{Y,Z\in(7),(8)} \left(Q(Y,Z) + \sum_{i\in\mathcal{V}} \lambda_i^* \left(\sum_{k=1}^{K} y_{ik} - 1 \right) \right)$$
$$= \arg \min_{Y,Z\in(7),(8)} \max_{\Lambda} \left(Q(Y,Z) + \sum_{i\in\mathcal{V}} \lambda_i^* \left(\sum_{k=1}^{K} y_{ik} - 1 \right) \right).$$

This implies that (Y^*, Z^*) satisfy (9), i.e. $R(\Lambda^*)$ is equal to the solution of the problem (6)—(9). Finally, equality $R(\Lambda^*) = D(\Lambda^*)$ completes the proof.

3.2 Permuted $\mathcal{P}^n$ Potts

Relaxation (6)-(9) is not the best possible LP-relaxation of initial problem. In [1] the authors formulated a tighter relaxation with additional constraint responsible for marginalization of higher order potential w.r.t all but one variable:

$$\sum_{\boldsymbol{k}\in\mathcal{K}^{L(c)}:\ k_l=k_0} z_l^{c,\boldsymbol{k}} = y_{i_l^c k_0}, \quad \forall c \in \mathcal{C},\ k_0 \in \mathcal{K},\ l = 1,\dots,L(c). \tag{10}$$

Definition 1. *A higher order potential is called permuted $\mathcal{P}^n$ Potts iff*

$$\forall c \in \mathcal{C}\ \forall i \in c\ \forall \boldsymbol{d}', \boldsymbol{d}'' \in \mathcal{D}^c: \quad \boldsymbol{d}' \neq \boldsymbol{d}'' \Rightarrow d_i' \neq d_i''.$$

In permuted $\mathcal{P}^n$ Potts potentials all preferable configurations differ from each other in all variables. $\mathcal{P}^n$ Potts potential described in [6] is a special case.

Theorem 2. *If all higher-order potentials are permuted $\mathcal{P}^n$ Potts, then the maximum of dual function* (5) *is equal to the solution of LP* (6)-(10).

Proof. The proof follows from the fact that in the case of permuted $\mathcal{P}^n$ potential for each i_l^c and for each $k \in \mathcal{K}$ there exists no more than one $\boldsymbol{d}^c \in \mathcal{D}^c$ such that $d_l^c = k$. But each single $z_l^{c,\boldsymbol{d}^c} = y_{i_l^c k_l}$ due to (8) and negative sign of $\phi_c(\boldsymbol{d}^c)$.

3.3 Consistency

Now examine the properties of point $\Lambda^* = \arg\max D(\Lambda)$. Denote

$$Z_{ik}(\Lambda) = \operatorname{Arg}\min_{y_{ik}} \left[\min_{Y\setminus\{y_{ik}\},Z} L(Y,Z,\Lambda) \right].$$

Definition 2. *Point Λ is a weak agreement point if*

$$\forall i \in \mathcal{V}\ \exists k: \{1\} \subseteq Z_{ik}(\Lambda),\ \forall k' \neq k, \{0\} \subseteq Z_{ik'}(\Lambda).$$

Informally this definition means that at weak agreement point Λ means for each node i of MRF there is at least one label k such that there exists unconstrained minimum of $L(Y, Z, \Lambda)$ w.r.t. $Y, Z \in \{0, 1\}$ such that $y_{ik} = 1$ and there exist optimal configurations where all other $y_{ik'} = 0$.

Theorem 3. *Point Λ^* satisfies weak agreement.*

The proof is performed by contradiction and is almost identical to the proof of theorem 4 in [11].

Definition 3. *For a weak agreement point Λ define a strong agreement set:*

$$S(\Lambda) = \left\{ i \in \mathcal{V} \;\middle|\; |Z_{ik}(\Lambda)| = 1 \;\forall k \right\}.$$

For any node from the strong agreement set the consistent labeling (i.e. labeling that satisfies (3)) can be restored from $Z_{ik}(\Lambda)$. It is easy to show that if $S(\Lambda) = \mathcal{V}$ then $D(\Lambda)$ is equal to the optimal solution of (2) and the optimal labeling can be extracted uniquely from $Z_{ik}(\Lambda)$. A more interesting question is whether there exists optimal labeling such that for each node from $S(\Lambda^*)$ its labeling equals the one obtained from $Z_{ik}(\Lambda)$ (so-called, persistency property). Unfortunately, the answer is generally negative even for maximal cliques of size two (see the appendix for a counter-example).

Note that this result is similar to the analogous result for tree-reweighted message passing (TRW) for pairwise MRFs. In [13] it was proven that for binary problems ($K = 2$) persistency holds for nodes strongly labeled with TRW, but for $K > 2$ it does not hold. If $K = 2$ in pairwise MRF SMR is equivalent to QPBO [14] method for which persistency is well-known to hold.

3.4 Getting a primal solution

Due to the lack of persistency property the question of a getting primal solution after solving the dual is not trivial. We suggest the following post-processing procedure. First assign $y_{ik} = 1$ if $\{1\} \subseteq Z_{ik}(\Lambda)$ and $\nexists k' < k : \; \{1\} \subseteq Z_{ik'}(\Lambda)$ and $y_{ik} = 0$ otherwise. Define

$$\hat{Y}_c = \arg \min_{\sum_{k=1}^K y_{ik}=1} \left(\sum_{k=1}^{K} \sum_{i \in c} (\phi_i(k) + \lambda_i^*) y_{ik} + \sum_{\boldsymbol{d}^c \in \mathcal{D}^c} \phi_c(\boldsymbol{d}^c) \prod_{l=1}^{L(c)} y_{i_l^c d_l^c} \right).$$

For each $c \in \mathcal{C}$ try to improve the current labeling Y by setting $Y_c = \hat{Y}_c$. If $E(Y)$ is reduced we change the current labeling. Then we switch to next c until all hyperedges have been considered. The process is repeated several times. This procedure is similar to Iterated Conditional Modes and therefore we refer to it as ICM.

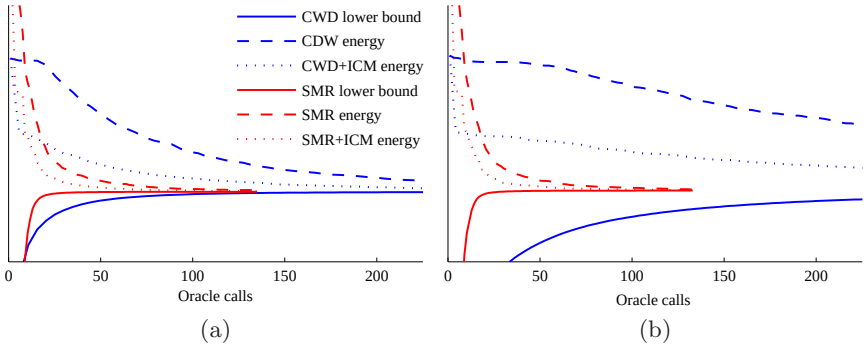


Fig. 1. Energies/lower bounds produced by SMR and CWD on synthetic datasets. All blue curves correspond to CWD, red curves – to SMR; solid lines show the lower bounds, dashed lines — energies obtained via random agreement of the subproblems, dotted lines – energies after 1 cycle of greedy improving via ICM (see sec. 3.4). (a) corresponds to the dataset with high-order potentials of size 50, (b) – of size 500.

4 Experiments

In this section we present an empirical evaluation of SMR compare it against CWD. We also evaluate the effect ICM-based postprocessing scheme on both methods. We perform the comparison on two synthetic sets of energies and one energy acquired from an image segmentation problem.

Both synthetic datasets consist of 20 energies depending on 10-label variables that form 4-neighborhood grids graphs of 50x50 nodes. All unary potentials are generated from gaussian distribution $\mathcal{N}(0, 10)$. The pairwise potentials are all Potts: $0.1|c_{ij}|[x_i \neq x_j]$, where $c_{ij} \sim \mathcal{N}(0, 1)$. Afterwards we add 50 $\mathcal{P}^n$ Potts potentials. For the first experiment we add potentials connecting random 50 nodes each, for the second experiment – 500 random nodes each. The costs of the potentials are generated from the uniform distribution on the segment $[0, 100]$.

As an image segmentation problem we take segmentation of image (fig. 3a) into three classes: “fern”, “ground”, “grass”. For each class we select a small region (seeds) and fit a Gaussian mixture distribution with 5 components to colors of selected pixels in CIELUV color-space. Negative log-densities form unary terms, pairwise terms are Potts with weight 1, high-order potentials are posed on mean-shift superpixels computed with EDISON[4] system [15] with default parameters, the weight of high-order potentials is 100.

In SMR we minimize submodular functions using Boykov-Kolmogorov max-flow/min-cut algorithm [16][5]. To implement CWD we make the following decomposition of the energy: all pairwise potentials are separated into vertical and horizontal chains, each high-order potential forms a separate subproblem, unary

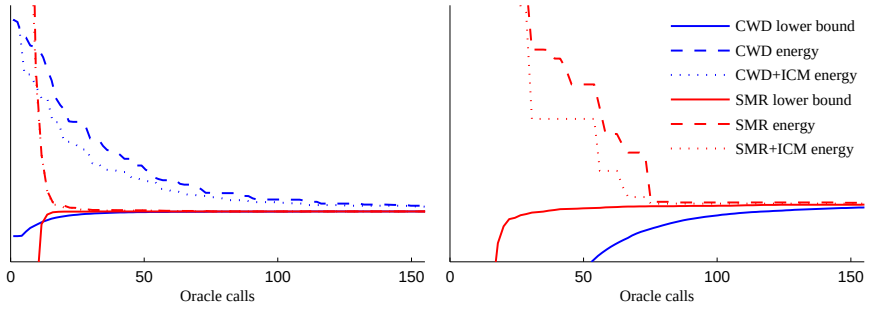
[4] `http://coewww.rutgers.edu/riul/research/code/EDISON/`
[5] `http://pub.ist.ac.at/~vnk/software/maxflow-v3.02.src.tar.gz`

potentials are evenly distributed between the horizontal and the vertical forests. In such decomposition the dual function depends on $|\mathcal{V}|K(1+h)$ variables. Here h is an average number of high-order potentials incident to individual nodes. The dual function in SMR depends on $|\mathcal{V}|$ variables. For synthetic experiments the CWD's dual has 50000 and 275000 variables respectively, the SMR's dual has 2500 variables. For image segmentation experiment the duals have 675000 (CWD) and 112500 (SMR) variables. As an oracle for SMR we use Boykov-Kolmogorov max-flow algorithm [16]; for CWD we use dynamic programming to perform MAP-inference in each row and column and follow [1] in inferring in high-order cliques. In our experiments oracles in CWD and SMR run roughly for the same time, and to exclude the issue of efficient implementation of the oracles we measure the algorithm complexity in number of oracle calls. Note, that more sophisticated version of CWD, PatB [1], could potentially improve the performance only when the high-order potentials intersect. For our first synthetic dataset high-order potentials almost do not intersect, for image segmentation example the do not intersect at all.

An important part of experimental comparison is the choice of the optimization method used to solve the dual problems. Originally [1] discusses two strategies: subgradient ascent and averaging of min-marginals. Min-marginal averaging corresponds to coordinate ascent and can get stuck in arbitrary poor coordinate-wise optimum of the dual [17, 11]. Subgradient ascent technics are in theory guaranteed to converge to the optimum but our experiments (confirmed by the recent research [9]) show that the subgradient schemes are unstable and sensitive to parameter choice, and their convergence can be very slow. To exclude the aspect of parameter choice in all experiments we use an off-the-shelf HANSO optimization system v2.01[6]. HANSO system consists of a specific version of BFGS algorithm that is applicable to non-differentiable functions [18] and a robust gradient sampling method [19] that is used to adjust the solution at the end of the process. The computation overhead for HANSO system (time spent not on evaluating the function, but on inner manipulations) is non-zero and increases with the growth of the dimensionality of dual variables, but in all our experiments it spends less than 5% of total running time. We are aware that recently there've been much research in developing fast and robust methods to solve such duals (e.g. [9, 10]). We believe that at least some of these technics can be applied to improve SMR as well and leave this as future work.

Figures 1 and 2 present the evaluation on synthetic and segmentation datasets. Figure 3 shows the initial image (a),(b) and the resulting segmentation with (c) and without (d),(e) high-order terms. In all cases we observed that SMR converges faster than CWD in terms of the value of the dual and obtains better primal discrete solution. Our experiments show that ICM postprocessing greatly helps CWD but has very little effect on SMR, even slowing it down in terms of time.

[6] `http://www.cs.nyu.edu/overton/software/hanso/`



(a) (b)

Fig. 2. Comparison of SMR and CWD on the image segmentation experiment. Both plots show energies/lower bounds against number of oracle calls. (b) is a zoomed version of (a). Note that both SMR and CWD were able to achieve the global maximum of the lower bound, and SMR was able to retrieve the global minimum of the initial energy as well (the duality gap is zero).

5 Conclusion

In the paper we suggest a new type of approximate inference algorithm for the case of sparse higher-order potentials that encourage few configurations of variables. Our method (SMR) is based on relaxation of the energy by making it submodular. Such relaxation corresponds to the solution of specific LP-relaxation of the discrete problem. Although in general this relaxation is less tight then clique-wise decomposition (CWD) we have derived an important case when they are equivalent. In comparison to CWD and its improved version PatB our method (SMR) requires less Lagrangian multipliers, converges faster and the process of obtaining primal solutions is easier. These properties make SMR a promising tool for training and MAP-inference in MRFs with higher order potentials.

Acknowledgments. This work was supported by the Russian Foundation for Basic Research (projects 12-01-00938, 12-01-31254, 12-01-33085).

References

1. Komodakis, N., Paragios, N.: Beyond pairwise energies: Efficient optimization for higher-order MRFs. In: CVPR. (2009)
2. Kohli, P., Ladický, L., Torr, P.H.S.: Robust higher order potentials for enforcing label consistency. IJCV **82** (2009) 302–324
3. Ishikawa, H.: Transformation of general binary MRF minimization to the first order case. IEEE TPAMI **33** (2011) 1234–1249
4. Rother, C., Kohli, P., Feng, W., Jia, J.: Minimizing sparse higher order energy functions of discrete variables. In: CVPR. (2009)

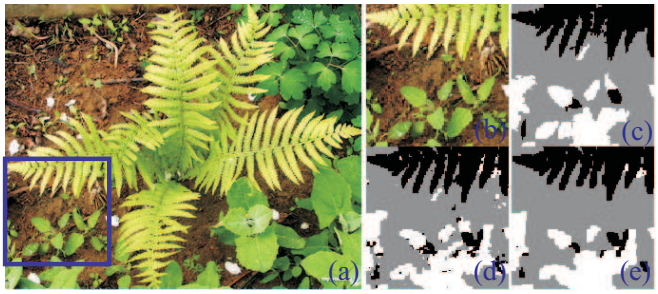


Fig. 3. (a) – the initial image; (b) – the zoomed part of the initial image; (c) – the result of SMR with high-order cliques, SMR found the global minimum; (d),(e) - the results of segmentation without high-order potentials, computed with α-expansion algorithm.

5. Boykov, Y., Veksler, O., Zabih., R.: Fast approximate energy minimization via graph cuts. IEEE TPAMI **23** (2001) 1222–1239
6. Kohli, P., Kumar, M.P., Torr, P.: $\mathcal{P}^3$&Beyond: Move making algorithms for solving higher order functions. IEEE TPAMI **31** (2008) 1645–1656
7. Delong, A., Osokin, A., Isack, H., Boykov, Y.: Fast approximate energy minimization with label costs. IJCV **96** (2012) 1–27
8. Komodakis, N., Paragios, N., Tziritas, G.: MRF energy minimization and beyond via dual decomposition. IEEE TPAMI **33** (2010) 531–552
9. Kappes, J.H., Savchynskyy, B., Schnörr, C.: A bundle approach to efficient MAP-inference by lagrangian relaxation. In: CVPR. (2012)
10. Komodakis, N.: Towards more efficient and effective LP-based algorithms for MRF optimization. In: ECCV. (2010)
11. Osokin, A., Vetrov, D., Kolmogorov, V.: Submodular decomposition framework for inference in associative Markov networks with global constraints. In: CVPR. (2011)
12. Boros, E., Hammer, P.L.: Pseudo-boolean optimization. Discrete Appl. Math **123** (2002) 155–225
13. Kolmogorov, V., Wainwright, M.: On the optimality of tree-reweighted max-product message passing. In: UAI. (2005)
14. Kolmogorov, V., Rother, C.: Minimizing non-submodular functions with graph cuts – a review. IEEE TPAMI **29** (2007) 1274–1279
15. Christoudias, C., Georgescu, B., Meer, P.: Synergism in low-level vision. In: CVPR. (2002)
16. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE TPAMI **26** (2004) 1124–1137
17. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE TPAMI **28** (2006) 1568–1583
18. Lewis, A., Overton, M.: Nonsmooth optimization via BFGS. SIAM J. Optimization (2008)
19. Burke, J., Lewis, A., Overton, M.: A robust gradient sampling algorithm for non-smooth, nonconvex optimization. SIAM J. Optimization **15** (2005) 751–779

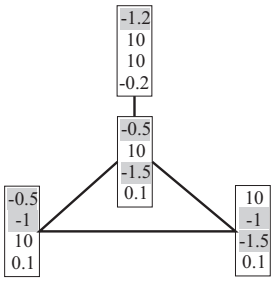


Fig. 4. A counter-example to persistency in SMR. Unary potentials which are set to maximize the lower bound $D(\Lambda)$; pairwise potentials are Potts models: $\phi_{ij}(x_i, x_j) = 2[x_i \neq x_j]$. Labels k such that $\{1\} \subseteq Z_{ik}(\Lambda)$ are shown in grey. The upper node forms a strong agreement set. The persistency property does not hold since the optimal labeling is to assign all nodes to the forth class, i.e. $x_i = 4,\ \forall i \in \mathcal{V}$.

A counter-example to persistency property in SMR

Consider the 4-class pairwise MRF (fig. 4). The SMR lower bound has its maximum with zero Lagrange multiplyers: $\mathbf{0} = \arg\max_\Lambda D(\Lambda)$. According to theorem 3 it is a weak agreement point. The upper node is consistently labeled by 1 and thus forms a strong agreement set. However, in the optimal labeling all nodes are labeled by 4 although all $Z_{i4}(\mathbf{0}) = \{0\}$. In other words, we cannot guarantee persistency property in submodular relaxation even for pairwise Potts models in case of more than two classes.

Consider the 4-class pairwise MRF with structure and the unary potentials shown in figure 4. Let all pairwise terms be Potts potentials $\phi_{ij}(x_i, x_j) = 2[x_i \neq x_j]$. It is easy to see that $\mathbf{0} = \arg\max_\Lambda D(\Lambda)$ and it is a weak agreement point. The upper node has consistent labeling and it is assigned to class 1. It forms strong agreement set. However in optimal labeling all nodes are assigned class 4 although all $Z_{i4}(\mathbf{0}) = \{0\}$. In other words we cannot guarantee persistency property in submodular relaxation even for pairwise Potts models in the case of more than two classes.