Variational Segmentation Algorithms with Label Frequency Constraints[1]

D. Kropotov[a], D. Laptev[b], A. Osokin[b], and D. Vetrov[b]

[a] *Dorodnicyn Computing Centre of the Russian Academy of Sciences, Vavilov str., 40, Moscow, 119333 Russia*

[b] *Lomonosov Moscow State University, Leninskie Gory, 1, 2nd ed. bldg., CMC Department, Moscow, 119992 Russia*

e-mail: dmitry.kropotov@gmail.com, laptev.d.a@gmail.com, anton.osokin@gmail.com, vetrovd@yandex.ru

Abstract—We consider image and signal segmentation problems within the Markov random field (MRF) approach and try to take into account label frequency constraints. Incorporating these constraints into MRF leads to an NP-hard optimization problem. For solving this problem we present a two-step approximation scheme that allows one to use hard, interval and soft constraints on label frequencies. On the first step a factorized approximation of the joint distribution is made (only local terms are included) and then, on the second step, the labeling is found by conditional maximization of the factorized joint distribution. The latter task is reduced to an easy-to-solve transportation problem. Basing on the proposed two-step approximation scheme we derive the ELM algorithm for tuning MRF parameters. We show the efficiency of our approach on toy signals and on the task of automated segmentation of Google Maps.

Key words: image segmentation, signal segmentation, Markov random fields, hidden Markov models, discrete optimization, area prior, variational inference, linear programming.

DOI: 10.1134/S1054661810030089

1. INTRODUCTION

Image and signal segmentation problems arise in many domains, e.g. in digital signal processing, computer vision, and image analysis. These problems can be treated as the generalizations of the well-studied classification problem, which is of great importance in the machine-learning field. As opposed to the latter we take into account not only the dependencies between features (observed variables) and class labels (hidden variables), but also the dependencies between the labels of the neighboring objects. The local neighborhood is generally given by an undirected graph known as the Markov network which is a particular case of graphical models, one of the most rapidly growing area of machine learning research.

It is a known fact that when the local neighborhood graph has no cycles the segmentation problem can be solved exactly in polynomial time by using dynamic programming approach (max-sum algorithm). In case of cyclic graphs the segmentation problem is generally NP-hard but, many efficient, approximate algorithms do exist (see Section 4).

In the paper we address the problem of conditional segmentation with the constraints on class sizes or equivalently on label frequencies. Denote $x_i \in \{1, \ldots, K\}$, $i = 1, \ldots, N$ the label of the ith object. Let m_k be the number of objects which are assigned label k. We consider the problems of the following form

$$\Phi(x_1, \ldots, x_N) = \sum_{i=1}^{N} \log \varphi_i(x_i) + \sum_{(i,j) \in \varepsilon} \log \psi_{ij}(x_i, x_j) + \sum_{k=1}^{K} f_k(m_k) \longrightarrow \min_{(x_1, \ldots, x_N)}, \quad (1)$$

where ε is a set of the neighboring objects and f_k is a function of the size of class k. Problem (1) is known to be NP-hard, so we focus on approximate methods. We suggest a two-stage scheme for solving problem (1). On the first stage we approximate Φ with a simpler function without the pairwise terms ψ_{ij} using variational methods. On the second stage we consider several types of functions f_k that correspond to the three kinds of constraints (hard, interval, and L1-penalty) and reduce each problem to the transportation problem that can be solved efficiently [17].

The paper is organized as follows. In Section 2 we briefly describe the formalism of Markov random fields (MRF). Section 3 defines the problem. Section 4 contains an overview of relevant literature on approximate methods of inference in MRF. In Section 5 we consider two ways of approximation of the joint distribution of hidden variables X in the family of factorized distributions. This is done by minimizing Kullback-Leibler divergence either direct or reverse. The three types of label frequency constraints and their reduction to the three types of the transportation problem are considered in Section 6. Section 7

[1] The article is published in the original.

Received April 7, 2010

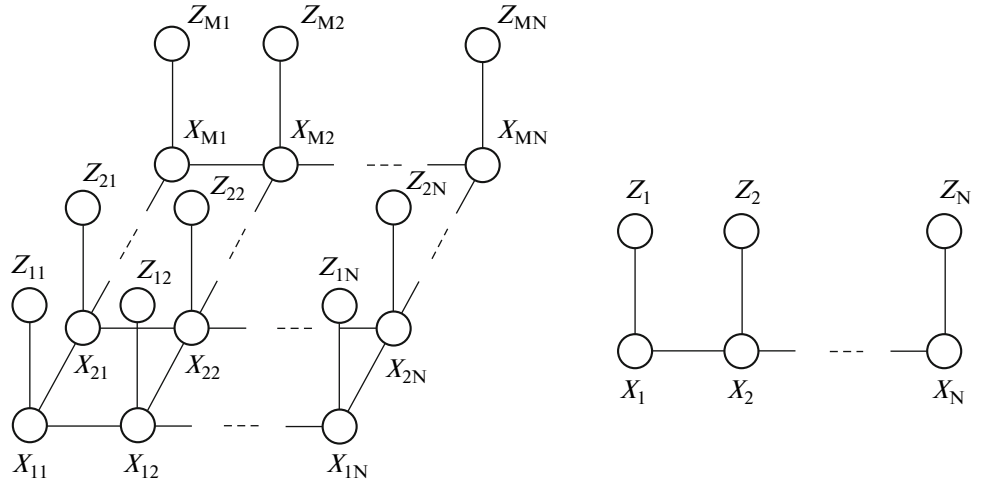


Fig. 1. Standard grid (left) and chain (right) neighborhood systems that are used for image and signal segmentation, respectively.

derives ELM-algorithm for tuning parameters of the MRF with label frequency constraints. This algorithm is a generalization of the well-known EM-algorithm. The results of the experiments both for signals and images are given in Section 8 and are discussed in Section 9.

2. MARKOV RANDOM FIELD

Suppose we have a probability model $P(X, Z)$ where Z is a set of *observed* variables and X is a set of discrete *hidden* variables. The *maximum a posteriori* (MAP) approach [18] is a standard way to find the hidden variables:

$$p(X|Z) = \frac{p(X, Z)}{p(Z)} = \frac{p(Z|X)p(X)}{p(Z)} \propto p(Z|X)p(X) \longrightarrow \max_X. \quad (2)$$

Here $p(X)$ is a prior distribution on the hidden variables and $p(Z|X)$ is a distribution of the observed variables given the hidden variables.

A prior that can be factorized into a product of the pairwise functions is called a *Markov Random Field (MRF)* prior:

$$p(X) \propto \prod_{(i,j) \in \varepsilon} \psi_{ij}(x_i, x_j). \quad (3)$$

Functions ψ_{ij} are usually called *potentials*. A neighborhood system ε shows which variables are connected directly via the corresponding potential.

For simplicity assume that each $x \in X$ has a one-to-one correspondent $z \in Z$. Also assume that all observed variables Z are conditionally independent given X and each z depends only on the corresponding x. Such assumptions imply

$$p(Z|X) = \prod_{i=1}^{N} p(z_i|x_i).$$

Let us denote $p(z_i|x_i)$ by $\varphi_i(x_i, z_i)$. Now we can reformulate (2) in the following way:

$$p(X|Z) \propto \prod_{i=1}^{N} \varphi_i(x_i, z_i) \prod_{(i,j) \in \varepsilon} \psi_{ij}(x_i, x_j) \longrightarrow \max_X. \quad (4)$$

Functions $\varphi_i(x_i, z_i)$ are called *unary potentials* (with respect to variables X) and functions $\psi_{ij}(x_i, x_j)$ are correspondingly called *pairwise potentials*.

Taking the negative-log we get the following optimization problem:

$$\sum_{i=1}^{N} \log \varphi_i(x_i, z_i) + \sum_{(i,j) \in \varepsilon} \log \psi_{ij}(x_i, x_j) \longrightarrow \min_X. \quad (5)$$

Optimizing (5) w.r.t. hidden variables X is a standard, well-studied problem. A review of the methods for this problem is presented in Section 4.

Problem (5) most often arises in computer vision and signal processing. An image segmentation problem can be reduced to (5) in the following way. Each pixel has a corresponding observed and hidden variable. Observed variables can represent e.g. intensities. Hidden variables usually correspond to the pixels' class labels. Hidden variables are connected according to either the 4-connected or the 8-connected neighborhood grid. An example of the 4-connected grid is shown in Fig. 1 (left).

Signal segmentation problem can also be reduced to problem (5). Observed variables represent observed value of a signal. Hidden variables represent class labels. The hidden variables are organized as a chain. Such graph is shown in Fig. 1 (right). This model is

known as *hidden Markov model (HMM)*. The HMM is called *homogenious* if all unary and pairwise terms do not depend on i, j, i.e. $\varphi_i(z_i, x_i) = \varphi(z_i, x_i)$, $\psi_{ij}(x_i, x_j) = \psi(x_i, x_j)$. In this case $\psi(x_i, x_j)$ defines a *transition matrix*.

3. PROBLEM FORMULATION

The MRF prior model (3) has a major drawback. The marginal statistics of the most probable solution according to this prior do not correspond to the target marginal statistics. The MRF prior has a strong bias towards delta function marginal statistics. For example, in a binary image segmentation problem the most popular MRF prior is the one that penalizes the neighboring pixels to have different labels (i.e. Potts model). Suppose that the true prior learned from a series of images gives each pixel an independent probability of 60% of being white and 40% of being black. However, according to the MRF prior the most likely image will have either 100% white pixels or 100% black pixels, which does not correspond to the true marginal statistics. This bias away from the true marginal statistics does not make a big problem if the unary potentials are strong enough or the marginal statistics of the solution are not very important but it becomes crucial otherwise (e.g. in several important application, such as image segmentation and image denoising).

The problem of incorporating information about desirable marginal statistics into MRF prior is very challenging. One approach is to have a prior over the marginal statistics of the output solution. Since computing marginal statistics involves every output variable, this model generates a single clique over all variables—what is called *a Marginal Probability Field (MPF)* [16]. The powerful optimization techniques developed for solving MRFs with their small cliques are not suitable for MPFs.

Let us substitute the MRF prior by

$$p(X) \propto \prod_{(i,j)\in\varepsilon} \psi_{ij}(x_i, x_j) \prod_{k=1}^{K} p(m_k), \quad m_k = \sum_{i=1}^{N} [x_i = k]^1.$$

Thus problem (2) can be formulated in the following way:

$$\underbrace{\prod_{i=1}^{N} \varphi_i(x_i, z_i)}_{\text{Unary potentials}} \underbrace{\prod_{(i,j)\in\varepsilon} \psi_{ij}(x_i, x_j)}_{\text{Pairwise potentials}} \underbrace{\prod_{k=1}^{K} p(m_k)}_{\text{Global prior}} \longrightarrow \max_X \quad (6)$$

Note that this criterion function consists of the three groups of terms: unary potentials, pairwise potentials, and global terms that correspond to the prior on the marginal statistics on the hidden variables X.

Problem (6) is NP-hard so we propose the following approximate scheme:

(1) Unary and pairwise terms are nonnegative and therefore their product represents the distribution up to multiplicative constant. Thus we can compute a factorized approximation of the first two terms of expression (6) by minimizing the Kullback-Leibler divergence [18]:

$$\prod_{i=1}^{N} \varphi_i(x_i, z_i) \prod_{(i,j)\in\varepsilon} \psi_{ij}(x_i, x_j) \approx \prod_{i=1}^{N} q_i(x_i) = q(X). \quad (7)$$

Section 5 reviews this procedure and gives details for grid and chain models[2].

(2) Instead of solving original optimization problem (6) we minimize a factorized approximation of unary and pairwise terms $q(X)$ multiplied by marginal statistics prior $p(m_k)$:

$$\prod_{i=1}^{N} q_i(x_i) \prod_{k=1}^{K} p(m_k) \longrightarrow \max_X. \quad (8)$$

In general this optimization problem is NP-hard but for some specific global priors it can be solved in polynomial time. In this paper we focus on the three types of priors:

(a) Hard constraints on marginal statistics correspond to the delta-function prior:

$$p(m_k) = [m_k = b_k]. \quad (9)$$

(b) Interval hard constraints correspond to uniform prior over an interval:

$$p(m_k) \propto [m_k \in [b_k', b_k'']]. \quad (10)$$

(c) L1-penalty for the difference from the desirable value of marginal statistics corresponds to:

$$p(m_k) \propto \exp(-\lambda_k |m_k - b_k|). \quad (11)$$

For priors (9), (10) and (11) problem (8) is equivalent to a specific form of transportation problem. The reduction is presented in Section 6.

4. RELATED WORK

Although segmentation problem is an extensions of the classification problem even the process of decision-making (inference) is computationally hard and not always tractable. The exact algorithms for (unconstrained) segmentation do exist for the case of acyclic MRFs (max-sum algorithm [18]), but the segmentation problem for arbitrary cyclic models is NP-hard in general. This fact motivates the development of numerous algorithms for approximate inference in graphical models with cycles. The most wide-spread ones are loopy belief propagation [3], which exploits message passing interface similar to max-sum, tree-

[1] Here we used Iverson brackets notation, i.e. [*statement*] = 1 if the statement is true and 0 otherwise.

[2] Figure 1 shows examples of such models' neighborhood systems.

reweighted message passing [4], where cyclic model is decomposed into a number of trees, variational inference [1], in which the joint distribution of variables within the model is factorized, expectation propagation [5], where each factor in the joint distribution is approximated by a simpler distribution, Monte Carlo Markov Chain methods of stochastic optimization [6], linear programming (LP) relaxation of the discrete problem [7, 8], etc.

One important subclass of the segmentation problems on Markov random fields with cycles is so-called submodular problems where hidden variables are binary, energy function consists of only unary and pairwise terms and the pairwise terms satisfy the submodularity condition. Then the segmentation problem can be solved exactly via max-flow/min-cut algorithms [9]. If the number of labels exceeds two, only approximate algorithms are available. When the pairwise terms are metrics the problem can approximately be solved e.g. by iterative application of max-flow algorithms [10].

The segmentation problem with global constraints is even harder since it requires to take into account global dependencies between variables. The theory of graphical models is based on the notion of conditional independence that allows to substitute global dependencies with local ones and hence to transform a clique where all variables are interconnected into Bayesian or Markov network with significantly smaller number of edges. The inevitable consequence of this reduction is the impossibility to consider global constraints at least in an exact manner. The attempts to take into account the global statistics have been made a number of times. In all cases the approximate schemes have been used and special priors on these statistics have been established. In [11] the authors established priors on pairs and triplets of variables. Non-parametric Bayesian methods and segment size priors based on Pitman-Yor processes have been suggested in [12]. The recent paper [14] proposes a generalization of LP relaxation approach to n-ary problems for soft constraint optimization MAP-MRF. In all three papers the constraints are soft and in general cannot guarantee desired label frequencies. Hard upper bounds on label frequencies with consequent LP-relaxation have been considered in [13]. Parametric max-flow [15] solves the binary problem with hard constraints for some specific values of constraint parameters.

One of the most recent attempts to establish constraints on global statistics in MRFs has been done in [16], where the problem was formulated in a very general form introducing the notion of Marginal Probability Fields (MPF). It was shown that a standard MAP-approach in MRF can be treated as a special linear case of the more general framework. Unfortunately the practical methods for inference in MPF are yet to be discovered.

5. FACTORIZATION OF THE JOINT DISTRIBUTION

Consider the joint distribution (4) in MRF for hidden and observed variables Z, X:

$$p(Z, X) = \frac{1}{C}\prod_{i=1}^{N}\varphi(x_i, z_i)\prod_{(i,j)\in\varepsilon}\psi_{ij}(x_i, x_j).$$

Here C is a normalization constant providing $\int p(Z, X)dZdX = 1$.

Suppose we would like to find a factorized approximation of the conditional distribution

$$p(X|Z) \approx q(X) = \prod_{i=1}^{N} q_i(x_i), \tag{12}$$

by minimizing KL($q\|p$)—Kullback-Leibler divergence between distributions $q(X)$ and $p(X|Z)$. KL divergence shows the difference between two probabilistic distributions and is defined as

$$\mathrm{KL}(q\|p) = -\int q(X)\log\frac{p(X|Z)}{q(X)}dX.$$

Note that KL divergence is not symmetric, and thus minimization of KL($q\|p$) and KL($p\|q$) w.r.t. $q(X)$ generally leads to different approximations.

First consider the minimization of KL($q\|p$). This approach is widely exploited in the variational inference and mean field theory [2, 1]. It can be shown that the optimal approximation satisfies the following equations

$$q_i(x_i) = \frac{1}{C_i}\exp\left(\int\log p(Z, X)\prod_{j\neq i} q_j(x_j)dx_j\right), \quad \forall i = 1, \ldots, N. \tag{13}$$

Here C_i is a normalization constant ensuring $\int q_i(x_i)dx_i = 1$. Note that expression (13) for $q_i(x_i)$ depends on all other $q_j(x_j)$. In practice we start from some initial distributions $q_i(x_i)$, $\forall i$ and iteratively update each of the $q_i(x_i)$ using (13). It can be shown [19] that this iterative process always converges.

Now consider the application of scheme (13) to HMM and MRF. In case of HMM the distribution $p(Z, X)$ is given by

$$p(Z, X) = \frac{1}{C}\varphi_1(z_1, x_1) \times \prod_{i=2}^{N}\varphi_i(z_i, x_i)\psi_{i,i-1}(x_i, x_{i-1}) \tag{14}$$

and we would like to find the approximation of this distribution in the factorized family $q(X) = \prod_i q_i(x_i)$.

Denote $q_{ij} = q_i(x_i = j)$. Then it is easy to show that scheme (13) transforms to

$$q_{ij} = \frac{\exp\left(\log\varphi_i(z_i,j) + \sum_k (\log\psi_{i,i-1}(j,k)q_{i-1,k} + \log\psi_{i+1,i}(k,j)q_{i+1,k})\right)}{\sum_m \exp\left(\log\varphi_i(z_i,m) + \sum_k (\log\psi_{i,i-1}(m,k)q_{i-1,k} + \log\psi_{i+1,i}(k,m)q_{i+1,k})\right)}, \quad i = 2, \ldots, N-1,$$

$$q_{1j} = \frac{\exp\left(\log\varphi_1(z_1,j) + \sum_k (\log\psi_{2,1}(k,j)q_{2,k})\right)}{\sum_m \exp\left(\log\varphi_1(z_1,m) + \sum_k (\log\psi_{2,1}(k,m)q_{2,k})\right)},$$

$$q_{Nj} = \frac{\exp\left(\log\varphi_N(z_N,j) + \sum_k (\log\psi_{N,N-1}(j,k)q_{N-1,k})\right)}{\sum_m \exp\left(\log\varphi_N(z_N,m) + \sum_k (\log\psi_{N,N-1}(m,k)q_{N-1,k})\right)}.$$

In a similar way the recalculation scheme for MRF is given by

$$q_{ij} = \frac{\exp\left(\log\varphi_i(z_i,j) + \sum_{k:(k,i)\in\varepsilon} \sum_n \log\psi_{ki}(n,j)q_{kn}\right)}{\sum_m \exp\left(\log\varphi_i(z_i,m) + \sum_{k:(k,i)\in\varepsilon} \sum_n \log\psi_{ki}(n,m)q_{kn}\right)}.$$

Together with variational inference we consider also an alternative factorized approximation $q(X) = \prod_i q_i(x_i)$ of conditional distribution $p(X|Z)$ which is obtained by minimizing $\mathrm{KL}(p\|q)$. It can be shown [5] that optimal factors $q_i(x_i)$ correspond to the marginals in this case, i.e.

$$q_i(x_i) = \int p(X|Z) \prod_{j\neq i} dx_j.$$

The computation of the marginal distributions is NP-hard in cyclic graphs but can be done efficiently by sum-product algorithm for trees. In case of HMM sum-product algorithm transforms to classical Baum-Welch algorithm [21].

6. REDUCTION TO THE TRANSPORTATION PROBLEM

In this section we show how the problem $q(X) = \prod_i q_i(x_i) \longrightarrow \max_X$ with global constraints (9), (10) and (11) can be reduced to the easy-to-solve transportation problem.

Recall that m_k denotes the size of each class

$$m_k = \sum_{i=1}^{N} [x_i = k].$$

For each discrete variable x_i we associate a binary vector $\vec{y}_i$ in the way that $y_{ik} = 1 \Leftrightarrow x_i = k$ and 0 otherwise.

6.1. Hard Constraints

Consider the following global constraints on m_k:

$$m_k = b_k, \quad k = 1, \ldots, K.$$

Here b_k are some predefined constants such that $\sum_{k=1}^{K} b_k = N$. These constraints are hard since we assign the particular values b_k to the sizes m_k of all classes. The optimization problem then takes the following form:

$$\begin{cases} q(X) \longrightarrow \max_X, \\ m_k = b_k, \quad \forall k = 1, \ldots, K. \end{cases} \tag{15}$$

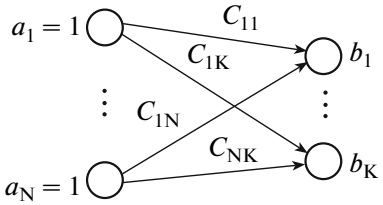


Fig. 2. The transportation problem for the hard constraints.

By applying negative logarithm this optimization problem can be equivalently rewritten as

$$\begin{cases} \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik}y_{ik} \longrightarrow \min\limits_{y}, \\ \sum_{k=1}^{K} y_{ik} = 1, \\ \sum_{i=1}^{N} y_{ik} = b_k, \\ y_{ik} \in \{0, 1\}, \end{cases} \tag{16}$$

where $c_{ik} = -\log q_{ik}$. The optimal $\hat{y}$ gives us all the labels x_i that agree with the constraints $\sum_i [x_i = k] = b_k$. Note that the problem (16) is the binary transportation problem with N sources and K targets. Source nodes have capacity 1 and correspond to variables x_i. Target nodes correspond to classes and their capacities equal the corresponding values b_k (see Fig. 2). Now consider LP-relaxation of (16) by allowing y_{ik} to be continuous:

$$\begin{cases} \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik}y_{ik} \longrightarrow \min\limits_{y}, \\ \sum_{k=1}^{K} y_{ik} = 1, \\ \sum_{i=1}^{N} y_{ik} = b_k, \\ y_{ik} \geq 0. \end{cases} \tag{17}$$

It is a well-known fact [17] that in case of integer b_k the optimal values of y_{ik} are also integer. Hence the optimal solutions of (16) and (17) are the same and we may use efficient continuous methods, e.g. simplex-method, in order to find the best labeling.

6.2. Interval Constraints

Now consider the case of interval constraints (10):

$$\begin{cases} q(X) \longrightarrow \max\limits_{X}, \\ m_k \in [b'_k, b''_k], \quad \forall k = 1, \ldots, K. \end{cases} \tag{18}$$

Here $\sum_k b'_k \leq N \leq \sum_k b''_k$. This case can be solved by straightforward generalization of the transportation problem (17). We add an extra source node (see Fig. 3 (left)) to the graph with capacity

$$a_{N+1} = \sum_k b''_k - N$$

and K extra target nodes with capacities

$$b_{K+k} = b''_k - b'_k, \quad \forall k \in \{1, \ldots, K\}.$$

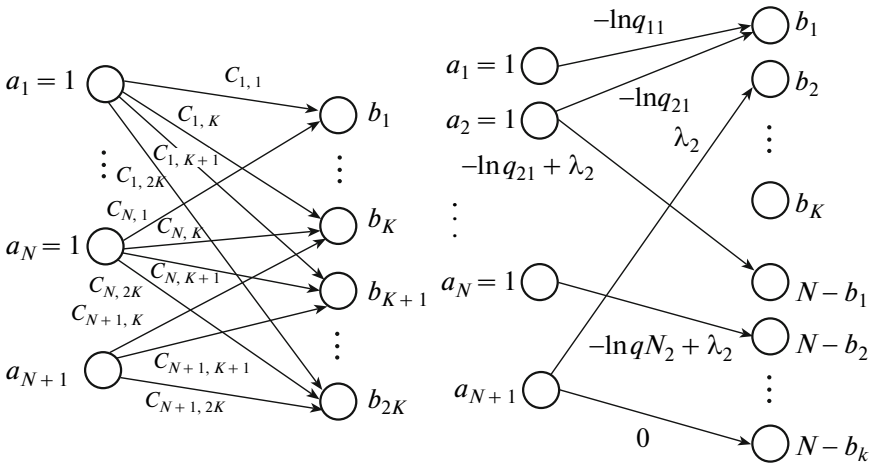


Fig. 3. Illustration of the transportation problem for the interval constraints (left) and for L1-penalty (right).

Also we define the capacity of the remaining target nodes as

$$b_k = b'_k \ \forall k \in \{1, \ldots, K\}.$$

The cost terms are defined in the following way

$$c_{ik} = c_{i,K+k} = -\log q_{ik},$$
$$\forall i \in \{1, \ldots, N\}, \ \forall k \in \{1, \ldots, K\};$$
$$c_{N+1,k} = +\infty, \ \forall k \in \{1, \ldots, K\};$$
$$c_{N+1,K+k} = 0, \ \forall k \in \{1, \ldots, K\}.$$

The final transportation problem has the form

$$\begin{cases} \sum_{i=1}^{N+1}\sum_{k=1}^{2K} c_{ik}y_{ik} \longrightarrow \min\limits_{y}, \\ \sum_{k=1}^{K} y_{ik} = 1, \ \forall i \in \{1, \ldots, N\}, \\ \sum_{k=1}^{K} y_{N+1,k} = \sum_{k}^{K} b''_k - N, \\ \sum_{i=1}^{N+1} y_{ik} = b_k, \ \forall k \in \{1, \ldots, 2K\}, \\ y_{ik} \geq 0. \end{cases} \tag{19}$$

Hence we want to distribute $\sum_{k=1}^{K} b'_k$ units from the first N sources among the first K targets. The remaining units from all sources are distributed among the last K targets. The extra source node is required to deal with surplus of $\sum_{k=1}^{K} b''_k - N$ units. The final labeling is performed in the following way: $x_i \equiv k (\mathrm{mod}\ K) \Leftrightarrow y_{ik} = 1$, $\forall i \in \{1, \ldots, N\}$.

6.3. L1-Penalty

Consider the soft constraints (11) with L1-penalties for label frequency violation:

$$\begin{cases} \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik}y_{ik} + \sum_{k=1}^{K} \lambda_k |m_k - b_k| \longrightarrow \min\limits_{y}, \\ \sum_{k=1}^{K} y_{ik} = 1, \\ \sum_{i=1}^{N} y_{ik} = b_k, \\ y_{ik} \geq 0. \end{cases} \tag{20}$$

The problem (20) can be reduced to the transportation problem with $2K$ target nodes and $N+1$ source nodes. The first K target nodes have capacities b_k while the remaining ones have capacities $N - b_k$. The transportation costs are defined in the following way (see Fig. 3 (right)):

$$c_{ik} = -\log q_{ik}, \ \forall i \in \{1, \ldots, N\}, \ \forall k \in \{1, \ldots, K\};$$
$$c_{N+1,k} = \lambda_k, \ \forall k \in \{1, \ldots, K\};$$
$$c_{i,K+k} = -\log q_{ik} + \lambda_k,$$
$$\forall i \in \{1, \ldots, N\}, \ \forall k \in \{1, \ldots, K\};$$
$$c_{N+1,K+k} = 0, \ \forall k \in \{1, \ldots, K\}.$$

The capacity of the source node $N+1$ (virtual units) is defined so that the transportation problem is balanced, i.e., $a_{N+1} = (K-1)N$. Note target $K+k$ while some virtual units are transported to the target k is impossible since it is always better to assign those real units to the target k and virtual ones to the target $K+k$.

Algorithm 1. ELM-algorithm

Regiure: Undirected graph $G = (\mathcal{V}, \varepsilon)$, parametric model $\psi_{ij}(x_i, x_j, \vec{\theta})$, $\varphi_i(x_i, z_i, \vec{\theta})$, the values of observed variables Z, frequency constraints $p(m_k)$, $k = 1, \ldots, K$.

Ensure: Set of parameters $\vec{\theta}$.

1: Initialize $\vec{\theta}_{\text{old}} = \vec{\theta}_0$.

2: **repeat**

3: E-step. Compute factorized approximation $q(X) = \prod_{i=1}^{N} q_i(x_i)$ of the posterior distribution $p(X|Z, \vec{\theta}_{\text{old}})$.

4: L-step. Find the most probable values of X

$$X^* = \arg\max_{X} q(X) \prod_{k=1}^{K} p(m_k)$$

by solving either the problem (15), (18) or (20) depending on the type of the constraint.

5: M-step. Find new values of $\vec{\theta}$ as follows

$$\vec{\theta}_{\text{new}} = \arg\max_{\vec{\theta}} p(X^*, Z|\vec{\theta})$$

and set $\vec{\theta}_{\text{old}} = \vec{\theta}_{\text{new}}$.

6: until $\|\vec{\theta}_{\text{new}} - \vec{\theta}_{\text{old}}\| < \eta$.

7. CONDITIONAL LEARNING

Suppose that the joint distribution $p(X, Z|\vec{\theta})$ is known up to some parameters $\vec{\theta}$ which define the potential functions $\psi_{ij}(x_i, x_j, \vec{\theta})$ and $\varphi_i(x_i, z_i, \vec{\theta})$. Iter-

ative expectation-maximization (EM) algorithm is widely used for searching

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} p(Z|\vec{\theta})$$

$$= \arg\max_{\vec{\theta}} \sum_{x_1, \ldots, x_N} p(X, Z|\vec{\theta}) \quad (21)$$

$$= \arg\max_{\vec{\theta}} \sum_{x_1, \ldots, x_N} \prod_{i=1}^{N} \varphi_i(x_i, z_i, \vec{\theta}) \prod_{(i,j) \in \varepsilon} \psi_{ij}(x_i, x_j, \vec{\theta}).$$

The algorithm consists of two iterative steps usually referred to as E-step and M-step. During E-step on n-th iteration for all pairs $(i,j) \in \varepsilon$ the pairwise marginals $p(x_i, x_j|Z, \vec{\theta}_n)$ and for all i the unary marginals $p(x_i|Z, \vec{\theta}_n)$ are computed either exactly or approximately given current values of $\vec{\theta}_n$. On M-step parameters $\vec{\theta}$ are re-estimated as follows

$$\vec{\theta}_{n+1} = \arg\max_{\vec{\theta}} \mathbb{E}_{X|Z, \vec{\theta}_n} \log p(X, Z|\vec{\theta})$$

$$= \arg\max_{\vec{\theta}} \left[\sum_{(i,j) \in \varepsilon} p(x_i, x_j|Z, \vec{\theta}_n) \log \psi_{ij}(x_i, x_j, \vec{\theta}) \right.$$

$$\left. + \sum_{i=1}^{N} p(x_i|Z, \vec{\theta}) \log \varphi_i(x_i, \vec{\theta}) \right].$$

One of the possible modifications of EM-algorithm is "hard" version where on E-step the most probable values of X are found:

$$X^* = \arg\max_X p(X|Z, \vec{\theta}_n).$$

In some cases (e.g. in cyclic graphs with submodular energies (negative log-potentials)) the process of finding X^* is easier than the computation of the marginals.

On M-step the new values of $\vec{\theta}$ are computed in the following manner:

$$\vec{\theta}_{n+1} = \arg\max_{\vec{\theta}} \log p(X^*, Z|\vec{\theta}).$$

It can be shown [18] that during E- and M-steps a lower bound on $\log p(Z|\vec{\theta})$ monotonically increases and thus the iterative process always converges. This hard EM-procedure can be easily adopted for the case of label frequency constraints. We establish intermediate L-step (L stands for labeling) on which the constrained segmentation is found using current values of the parameters $\vec{\theta}$. On E-step variational approximation is performed by minimizing KL divergence as described in Section 5. M-step remains the same. Note that like in EM-algorithm we may get stuck in a local extremum, hence in practice several runs from various initializations are desirable.

8. EXPERIMENTS

8.1. Signals

We made several illustrative experiments on toy signals. For segmenting signals we used homogenious hidden Markov models (14). It means that the joint distribution of variables X corresponds to tree graph structure and hence we have two ways for getting factorized approximation $q(X)$. The first is variational inference (13) and the second is taking marginals by applying Baum-Welch algorithm [21].

In the first experiment we modeled the signal as a realization of a two-dimensional stationary Gaussian process (see Fig. 4a). We used HMM with Gaussian distributions $\varphi(z_i, x_i) = \mathcal{N}(z_i|\vec{\mu}_{x_i}\Sigma_{x_i})$, where parameters $\vec{\mu}$, Σ for all states were obtained by k-means clustering algorithm in two-dimensional phase space. Then using this HMM we segmented the signal first by Viterbi algorithm [20] (the same results were obtained by maximizing both factorized approximations) and second by establishing label frequency constraints slightly changed from the ones obtained by Viterbi. Hereinafter for the signals we used factorized approximation obtained by the minimization of the reverse Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. Factorization obtained by the minimization of direct divergence $\mathrm{KL}(q\|p)$ led to very fragmented conditional segmentation and was considered to be unsuitable for signal segmentation problems. The segmentation results are shown in Fig. 4. As we can see, changing label frequencies does not lead to highly fragmented segmentation and thus varying constraints we can specify more accurate outcome.

Another toy example is a signal capable of an ambiguous segmentation. The signal and its correct labeling are shown in Fig. 5a. Here we again use HMM with Gaussian distributions and apply EM algorithm for tuning parameters in classical HMM without constraints [18] and ELM algorithm with hard constraints. After EM-algorithm one state collapses (see Fig. 5b) while using ELM with correct label frequency prior leads to the correct segmentation (see Fig. 5c).

8.2. Images

Another illustrative example is the segmentation of seashore satellite photos from the Google Maps (see Fig. 6a) into two segments—sea and land. We used the dataset of such images which contains 40 examples and is available on http://xorio.net/maps_dataset.rar. Google Maps service provides also the correct labeling (see Fig. 6b).

For image segmentation we used MRF (4), where unary potentials (color models) are Gaussian distribu-

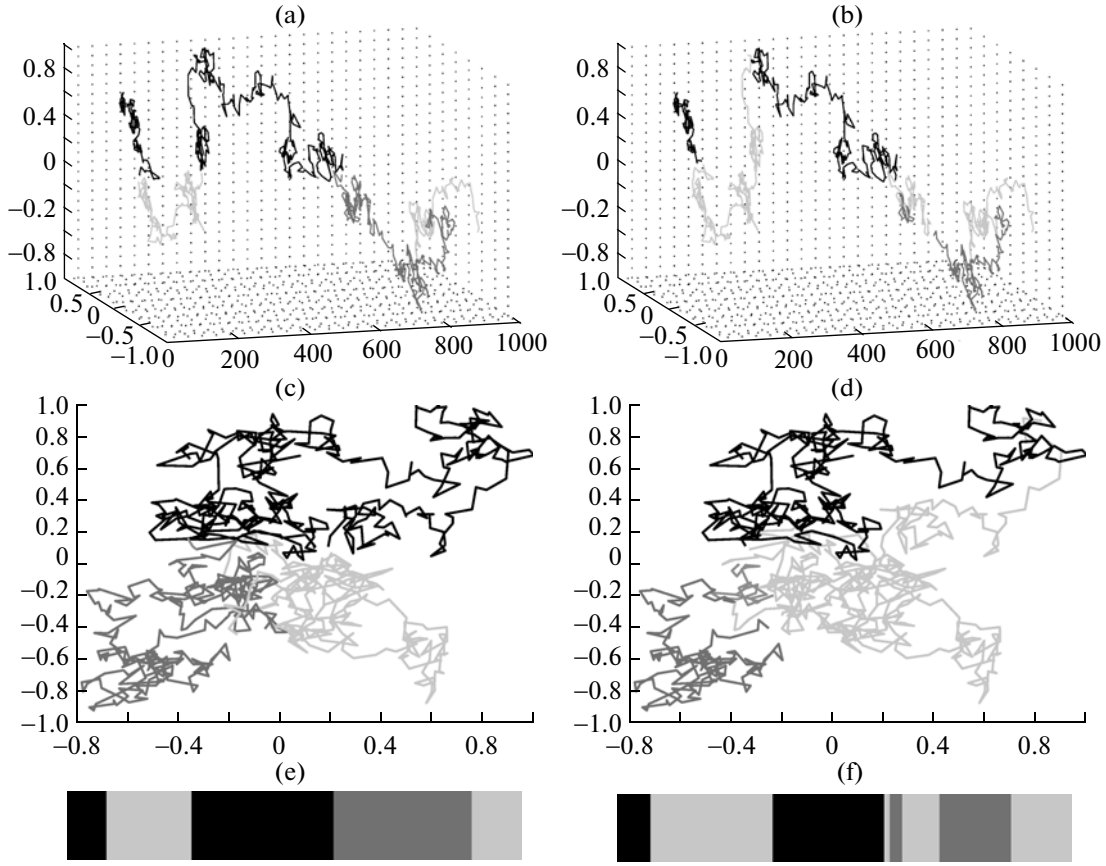


Fig. 4. Segmentation of 2D Gaussian process realization using standard HMM (a) and HMM with label frequency constraints changed from the ones obtained by Viterbi (+80, −200, −100) (b). Figures (c) and (d) show the corresponding segmentations in the phase space, while figures (e) and (f) show the segmentations themselves.

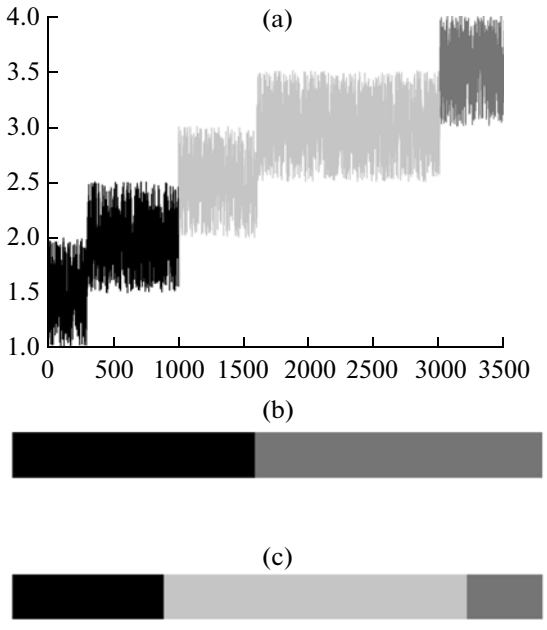


Fig. 5. Segmentation of toy signal (a) using EM algorithm for standard HMM (b) and ELM algorithm with true hard label frequency constraints (c).

tions and pairwise potentials are the generalized Potts model[3] of the following form:

$$\log \psi_{ij}(x_i, x_j, z_i, z_j) = \begin{cases} 0, & \text{if } x_i = x_j, \\ \rho \exp\left(-\frac{\gamma}{2}\|z_i - z_j\|^2\right), & \text{otherwise.} \end{cases}$$

Parameters of color models were obtained using our dataset while $\gamma = 1$, $\rho = 5$. All $0 \le z_i \le 1$ represented intensity values. Results of constrained and unconstrained segmentations are shown on Fig. 6. As we can see considering label frequency constraints leads to much more appropriate segmentation results. Also we measured the segmentation error rate for our dataset. The result is about 7 percent pixels wrong labeled for unconstrained segmentation and about 3.4 percent for constrained segmentation.

[3] To be strictly correct these potentials correspond to the special case of MRF—so-called Conditional Random Field (CRF) [22].

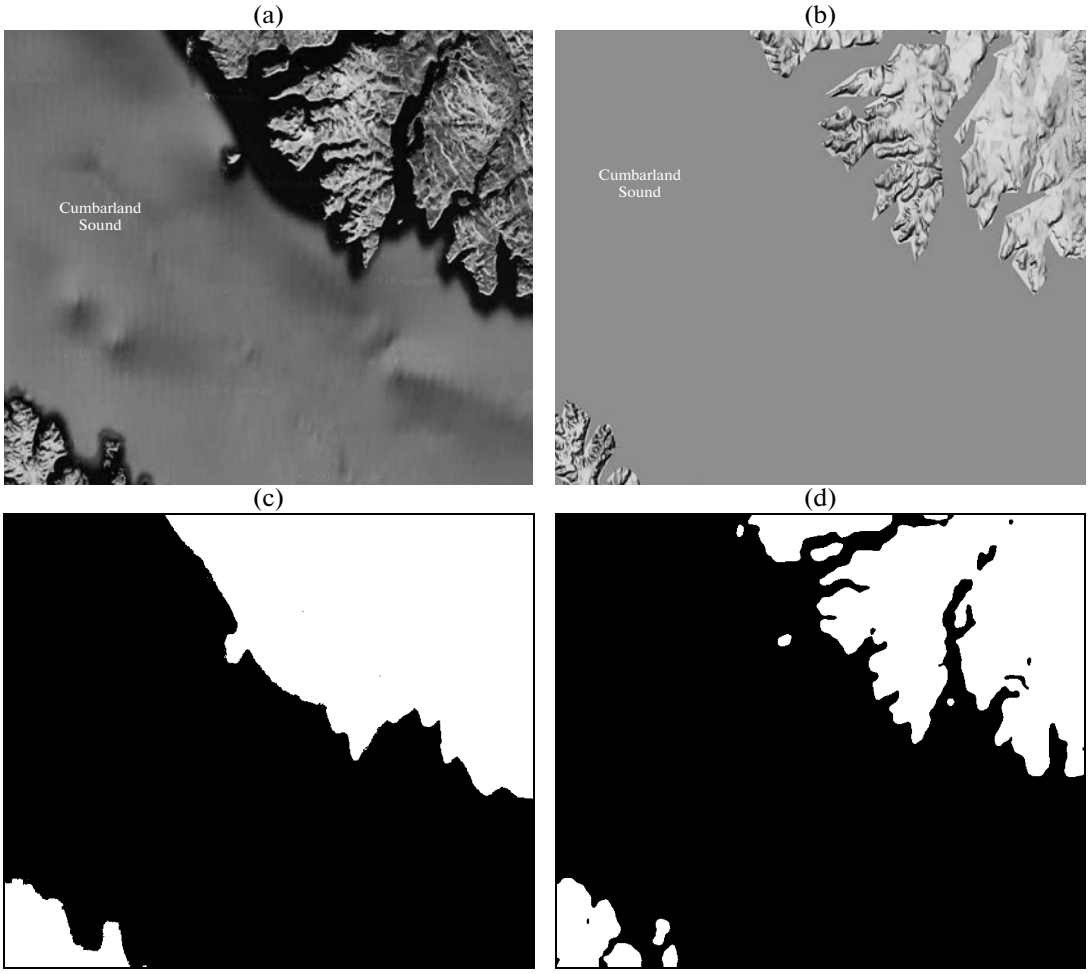


Fig. 6. initial satellite photo (a), true segmentation into land and sea (b), the result of unconstrained segmentation (c) and the result of constrained segmentation (d).

9. DISCUSSION

We present a general two-step approximate algorithm for solving segmentation problems using MRFs with three variants of label frequency constraints. On the first step we look for a factorized approximation of unary and pairwise terms, and then, on the second step, we solve the problem by its reduction to a variant of the transportation problem. We consider two approaches for factorized approximation: either variational inference or computing marginals. In the first case we approximate a small neighborhood of a local maximum of the true distribution $p(X, Z)$ [19]. Hence this approach can be appropriate only for relatively smooth distributions $p(X, Z)$. Otherwise maximization of the variational approximation s.t. the label frequency constraints leads to a very fragmented solution. This was demonstrated for the case of HMM for the signal segmentation task. On the contrary, computing marginals corresponds to the approximation of the true distribution for all values of X where $p(X|Z)$ is high. As a result the subsequent maximization s.t. the label frequency constraints significantly less suffers from the oversegmentation.

However, effective marginals' calculation is possible only for a very limited number of models, e.g. for HMMs.

Together with a smooth variational approximation the second requirement for the successful application of the proposed approach is establishing the adequate label frequency constraints. Our experiments on the signal segmentation problem show that choosing the label frequency prior far from the real frequencies enforces oversegmentation and thus leads to an inappropriate solution.

We illustrated the proposed approach only for signal and image segmentation problems, however it can be applied for an arbitrary segmentation problem which can be formulated within the MRF framework.

ACKNOWLEDGMENTS

We would like to thank Valery Vishnevskiy for the idea of considering soft constraints. The work was supported by the Russian Foundation for Basic Research

(projects 08-01-00405, 09-01-92474, 09-04-12215) and the Russian President Grant MK-3827.2010.9.

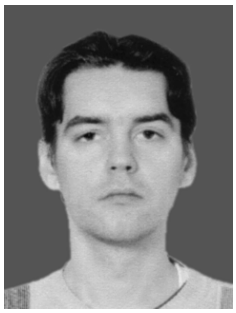
REFERENCES

1. C. Bishop, D. Spiegelhalter, and J. Winn, "VIBES: A Variational Inference Engine for Bayesian Networks," NIPS (2003).
2. G. Parisi, *Statistical Field Theory* (Addison-Wesley, 1988).
3. B. Frey, "A Revolution: Belief Propagation in Graphs with Cycles," NIPS (1998).
4. V. Kolmogorov and M. Wainwright, "On the Optimality of Tree-Reweighted Max-Product Message-Passing," UAI (2005).
5. T. Minka, "Expectation Propagation for Approximate Bayesian Inference," UAI (2001).
6. R.M. Neal, "Probabilistic Inference using Markov Chain Monte Carlo Methods," Tech. report CRG-TR-93-1, 1993.
7. J. Kleinberg and E. Tardos, "Approximation Algorithms for Classification Problems with Pairwise Relationships: Metric Labeling and Markov Random Fields," J. ACM, No. 5 (2002).
8. T. Werner, "A Linear Programming Approach to Max-Sum Problem: A Review," EEE TPAMI **29** (7), 1165–1179 (2007).
9. D. Greig, B. Porteous, and A. Seheult, "Exact Maximum a Posteriori Estimation for Binary Images," J. Royal Statistical Soc., No. 2 (1989).
10. Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," IEEE Trans. Pattern Analysis Machine Intelligence, No. 11 (2001).
11. D. Cremers and L. Grady, Statistical Priors for Efficient Combinatorial Optimization via Graph Cuts," ECCV (2006).
12. E. Sudderth and M. Jordan, "Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes," NIPS (2008).
13. J. Naor and R. Schwartz, "Balanced Metric Labeling," in *Proc. of Symposium on Theory of Computing (STOC), 2005.*
14. T. Werner, "High-Arity Interactions, Polyhedral Relaxations, and Cutting Plane Algorithm for Soft Constraint Optimisation (MAP-MRF)," CVPR (2008).
15. V. Kolmogorov, Y. Boykov, and C. Rother, "Applications of Parametric Maxflow in Computer Vision," ICCV (2007).
16. O. Woodford, C. Rother, and V. Kolmogorov, "A Global Perspective on MAP Inference for Low-Level Vision," ICCV (2009).
17. I. H. Sigal and A. P. Ivanova, *An Introduction to Discrete Programming: Models and Computational Algorithms*, 2nd ed. (Fizmatlit, Moscow, 2007) [in Russian].
18. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).
19. M. I. Jordan, Z. Gharamani, T. S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," in *Learning in Graphical Models,* Ed. by M. I. Jordan, 1998, pp. 105–162.
20. A.J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," IEEE Trans. Inform. Theory **13**, 260–267 (1967).
21. L. E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," Inequalities **3**, 1–8 (1972).
22. J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," ICML, 282–289 (2001).

Dmitry Kropotov, born in 1981, graduated from Moscow State University in 2003, currently works as junior researcher at Dorodnicyn Computing Centre of the Russian Academy of Sciences, areas of interests include machine learning, data mining, pattern recognition, artificial intelligence, image analysis, bioinformatics and language processing, has 19 articles including 1 monograph, a member of Russian Association for Pattern Recognition and Image Analysis, Russian Association for Artificial Intelligence, has best paper awards on different Russian and international conferences.

Dmitry Laptev, born in 1989, a student at Department of Computational Mathematics and Cybernetics in Moscow State University from 2006, areas of interests include machine learning, pattern recognition, data mining, image analysis, author of 1 article.

Anton Osokin, born in 1988, currently is a 5-th year student of Moscow State University, Computational Mathematics and Cybernetics department. Areas of interest include machine learning, computer vision, and cognitive science. Author of 2 journal and 7 conference papers, a student member of IEEE.



Dmitry Vetrov, PhD. Born in 1981, graduated from Moscow State University in 2003, got hit PhD degree there in 2006. Currently works as researcher at Department of Computational Mathematics and Cybernetics in Moscow State University. His area of interests includes Bayesian methods of machine learning, statistical relational learning, data mining, image analysis, bioinformatics, cognitive science. Author of 21 articles including 1 monograph, a member of Russian Association for Pattern Recognition and Image Analysis. During his work on PhD he received President scholarship for PhD students in 2005.