Signal segmentation with label frequency constraints using dual decomposition approach for hidden Markov models*

Kropotov D., Laptev D., Osokin A., Vetrov D.
dmitry.kropotov@gmail.com, laptev.d.a@gmail.com, anton.osokin@gmail.com, vetrovd@yandex.ru
Moscow, Dorodnicyn Computing Centre of RAS, Moscow State University

We consider a signal segmentation problem within the hidden Markov model (HMM) approach and try to take into account label frequency constraints. Following the dual decomposition approach we maximize an energy lower bound via subgradient ascent method, where subgradient is found on each iteration by solving two subproblems. The first subproblem can be effectively solved by Viterbi algorithm and the other one can be reduced to an easy-to-solve transportation problem. We show the efficiency of our approach on toy signals and on the task of automated segmentation of mouse behavior.

Signal segmentation using hidden Markov models

Hidden Markov model (HMM) is a probabilistic model of a sequence that consists of a set of observed variables $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ (in arbitrary space) and discrete hidden variables $T = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N\}$, where $\boldsymbol{t}_n$ can take K different values. The joint distribution of all variables (see fig. 1) is given by

$$p(X, T) = p(\boldsymbol{t}_1) \prod_{n=1}^{N} p(\boldsymbol{x}_n \,|\, \boldsymbol{t}_n) \prod_{n=2}^{N} p(\boldsymbol{t}_n \,|\, \boldsymbol{t}_{n-1}). \quad (1)$$

Using 1-of-K coding scheme a discrete variable with K possible values can be represented as $\boldsymbol{t}_n = (t_{n1}, \ldots, t_{nK})$, where

$$t_{nj} = \begin{cases} 1, & \text{if at the moment } n \text{ the model is} \\ & \text{in the } j^{th} \text{ state;} \\ 0, & \text{otherwise.} \end{cases}$$

Hereinafter we suppose a homogeneous HMM, i. e. the probability $p(\boldsymbol{t}_n \,|\, \boldsymbol{t}_{n-1})$ and emission probability $p(\boldsymbol{x}_n \,|\, \boldsymbol{t}_n)$ do not depend on n. Hence, $p(\boldsymbol{t}_n \,|\, \boldsymbol{t}_{n-1})$ can be fully characterized by a transition matrix A of size $K \times K$, where $A_{ij} = p(t_{nj} = 1 \,|\, t_{n-1,i} = 1)$, $\sum_{j=1}^{K} A_{ij} = 1$. Equivalently,

$$p(\boldsymbol{t}_n \,|\, \boldsymbol{t}_{n-1}) = \prod_{i=1}^{K} \prod_{j=1}^{K} A_{ij}^{t_{n-1,i}\, t_{nj}}.$$

The prior probability $p(\boldsymbol{t}_1)$ at the first moment is given by $\boldsymbol{\pi}\colon p(t_{1j} = 1) = \pi_j$, $\sum_{j=1}^{K} \pi_j = 1$ and $p(\boldsymbol{t}_1) = \prod_{j=1}^{K} \pi_j^{t_{1j}}$.

We assume that emission probability $p(\boldsymbol{x}_n \,|\, \boldsymbol{t}_n)$ for state j is given by parametric distribution $p(\boldsymbol{x}_n \,|\, \boldsymbol{\varphi}_j)$, where $\boldsymbol{\varphi}_j$ is a set of parameters. Hence,

$$p(\boldsymbol{x}_n \,|\, \boldsymbol{t}_n) = \prod_{j=1}^{K} \big(p(\boldsymbol{x}_n \,|\, \boldsymbol{\varphi}_j)\big)^{t_{nj}}.$$

The work was supported by the Russian Foundation for Basic Research (projects 08-01-00405, 10-01-90419), the Russian President Grant MK-3827.2010.9 and Federal Target Program "Scientific and scientific-pedagogical personnel of innovative Russia in 2009–2013" (contract no. P1265).

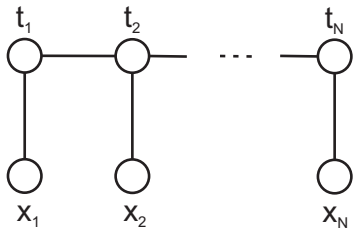


Fig. 1. Chain dependency of variables in HMM joint distribution (1). Variables in one factor are connected by edges.

Denote all HMM parameters as $\Theta = \{\boldsymbol{\pi}, A, \boldsymbol{\varphi}\}$. The maximum a posteriori (MAP) approach [5] is a standard way to find the hidden variables T:

$$T^* = \arg\max_T p(T \,|\, X, \Theta) = \arg\max_T p(X, T \,|\, \Theta) = \arg\max_T \log p(X, T \,|\, \Theta). \quad (2)$$

Taking negative log and changing maximum to minimum the problem (2) can be equivalently rewritten as a min-energy problem with binary variables:

$$E_{\text{local}}(T) = -\log p(X, T \,|\, \Theta) = -\left(\sum_{j=1}^{K} t_{1j} \log \pi_j\right) - \left(\sum_{n=2}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} t_{n-1,i} t_{nj} \log A_{ij}\right) - \left(\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log p(\boldsymbol{x}_n \,|\, \boldsymbol{\varphi}_k)\right) \to \min_T. \quad (3)$$

This energy function is pairwise separable [1] and thus can be effectively minimized by Viterbi algorithm [6] in linear time w. r. t. signal length N.

Label frequency constraints

In this paper we consider the problem of signal segmentation with label frequency constraints. Denote $m_k = \sum_{n=1}^{N} t_{nk}$ — the total occurrence of label k in segmented signal. Suppose we have a function $f_k(m_k)$ that penalizes the deviation of frequency m_k from the desired one. Then signal segmentation problem with label frequency constraints can

be written as

$$E(T) = E_{\text{local}}(T) + E_{\text{global}}(T) = E_{\text{local}}(T) + \sum_{k=1}^{K} f_k(m_k) \to \min_T. \quad (4)$$

This corresponds to adding a global prior over T of the form $p(T) = \exp\bigl(-\sum_{k=1}^{K} f_k(m_k)\bigr)$ into the joint distribution (1).

In general case the problem (4) is NP-hard. From perspectives of graphical models of kind from Fig. 1 the problem (4) corresponds to the graph where all variables $\boldsymbol{t}_n$ are connected with each other (the full graph). Thus, the graph contains a lot of cycles, and therefore effective precise algorithms like Viterbi can not be applied. In this paper we consider three types of penalty functions $f_k(m_k)$ and for them derive approximate schemes for solving the problem (4):

Hard constraints:
$$f_k(m_k) = \begin{cases} 0, & \text{if } m_k = b_k; \\ +\infty, & \text{otherwise.} \end{cases} \quad (5)$$

Interval constraints:
$$f_k(m_k) = \begin{cases} 0, & \text{if } m_k \in [b'_k, b''_k]; \\ +\infty, & \text{otherwise.} \end{cases} \quad (6)$$

Soft constraints:
$$f_k(m_k) = \alpha_k |m_k - b_k|,\ \alpha_k > 0. \quad (7)$$

Dual Decomposition Approach

The energy (4) cannot be minimized effectively, but both summands $E_{\text{local}}(T)$ and $E_{\text{global}}(T)$ can be minimized separately in efficient manner. This leads to the idea of using the dual decomposition approach [2].

Following this approach we rewrite the problem (4) in the following way:

$$\min_{\substack{T_1,T_2:\\ T_1=T_2}} E_{\text{local}}(T_1) + E_{\text{global}}(T_2) = \min_{\substack{T_1,T_2:\\ T_1=T_2}} E_{\text{local}}(T_1) + E_{\text{global}}(T_2) + \boldsymbol{\lambda}^{\intercal}(T_1 - T_2).$$

Here T_1, T_2 are supposed to be vectors of length NK of the form $T = [t_{11}, \ldots, t_{1K}, t_{21}, \ldots, t_{NK}]$. Denoting

$$F_1(\boldsymbol{\lambda}) = \min_{T_1}\bigl(E_{\text{local}}(T_1) + \boldsymbol{\lambda}^{\intercal} T_1\bigr), \quad (8)$$

$$F_2(\boldsymbol{\lambda}) = \min_{T_2}\bigl(E_{\text{global}}(T_2) - \boldsymbol{\lambda}^{\intercal} T_2\bigr), \quad (9)$$

we get the dual problem:

$$\max_{\boldsymbol{\lambda}}\bigl(F_1(\boldsymbol{\lambda}) + F_2(\boldsymbol{\lambda})\bigr). \quad (10)$$

From (8), (9) it can be seen that $F_1(\boldsymbol{\lambda}) + F_2(\boldsymbol{\lambda})$ is the lower bound for $E_{\text{local}}(T_1) + E_{\text{global}}(T_2) + \boldsymbol{\lambda}^{\intercal}(T_1 - T_2)$, and taking $T_1 = T_2 = T$ we get that $F_1(\boldsymbol{\lambda}) + F_2(\boldsymbol{\lambda})$ is the lower bound for $E_{\text{local}}(T) + E_{\text{global}}(T)$. Solving the problem (10) we would approximately fit the min-energy in the original problem (4).

The similar approach was used in [3] for image segmentation problem, but there it was assumed that E_{local} can be minimized only approximately by algorithms providing a lower bound for local energy, e. g. by tree-reweighted message passing [7]. In this paper we consider the case when local energy can be minimized in an exact way by Viterbi algorithm.

It's easy to see that $F(\boldsymbol{\lambda}) = F_1(\boldsymbol{\lambda}) + F_2(\boldsymbol{\lambda})$ is a concave piecewise-linear function and thus it can be effectively maximized by subgradient ascent method. A particular subgradient of $F_1(\boldsymbol{\lambda}) + F_2(\boldsymbol{\lambda})$ can be found analytically and is equal to $T_1(\boldsymbol{\lambda}) - T_2(\boldsymbol{\lambda})$, where $T_1(\boldsymbol{\lambda})$ and $T_2(\boldsymbol{\lambda})$ are argmins of the problems (8) and (9) respectively. Hence, iteration i of the subgradient ascent method is the following:

$$\boldsymbol{\lambda}_{i+1} = \boldsymbol{\lambda}_i + \delta_i(T_1 - T_2), \quad (11)$$

where δ_i is a step value.

Note that if $T_1(\boldsymbol{\lambda}) = T_2(\boldsymbol{\lambda})$ for some $\boldsymbol{\lambda}$, than we obtain the global optimum of (4). However, in general case, since energy function (4) is not convex, the maximum of (10) doesn't have to coincide with the minimum of (4). Besides, $T_1(\boldsymbol{\lambda})$ and $T_2(\boldsymbol{\lambda})$ do not coincide during subgradient ascent iterations and hence we need a special procedure for solution harmonization. Here we simply keep track of energy value (4) for both $T_1(\boldsymbol{\lambda})$ and $T_2(\boldsymbol{\lambda})$ for all iterations and return the one with the minimum value of (4).

Next we show how to solve the problems (8) and (9) on each step of the subgradient ascent method.

Viterbi algorithm

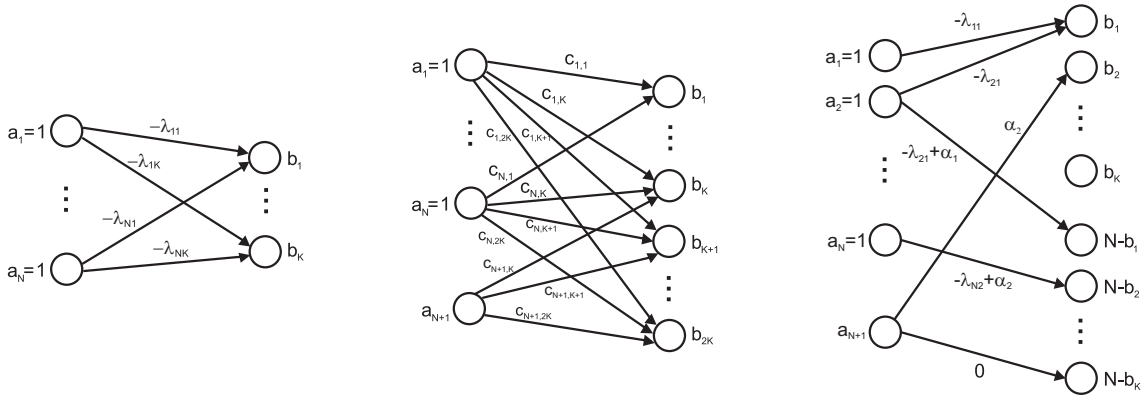
The criterion function in the optimization problem (8) can be written as

$$E_{\text{local}}(T_1) + \boldsymbol{\lambda}^T T_1 = -\left(\sum_{j=1}^{K} t_{1j} \log \pi_j\right) - \left(\sum_{n=2}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K} t_{n-1,i}\, t_{nj} \log A_{ij}\right) - \left(\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\bigl(\log p(\boldsymbol{x}_n \mid \boldsymbol{\varphi}_k) - \lambda_{nk}\bigr)\right).$$

Note that this function coincides with the pairwise-separable local energy (3) with a simple modification of unary terms. Hence, this energy can be minimized efficiently by Viterbi algorithm.

Reduction to transportation problem

The problem (9) for penalty functions (5), (6) and (7) can be reduced to a transportation problem which is a particular case of linear programming problem that can be solved efficiently.



(a) (b) (c)

Fig. 2. Transportation problems for the hard constraints (a), the interval constraints (b) and the soft constraints (c).

First consider the case of hard constraints (5). Then the problem (9) transforms to the following one:

$$\begin{cases} -\sum\limits_{n=1}^{N}\sum\limits_{k=1}^{K}\lambda_{nk}t_{nk} \to \min\limits_{T}; \\ \sum\limits_{k=1}^{K} t_{nk} = 1; \quad \sum\limits_{n=1}^{N} t_{nk} = b_k; \\ t_{nk} \in \{0,1\}. \end{cases} \tag{12}$$

Note that the problem (12) is the binary transportation problem with N sources and K targets. Source nodes have capacity 1 and correspond to time moments. Target nodes correspond to labels and their capacities equal the corresponding values b_k (see fig. 2a). Now consider LP-relaxation of (12) by allowing t_{nk} to be continuous: $t_{nk} \in [0,1]$. It is a well-known fact [4] that in case of integer b_k the optimal values of t_{nk} are also integer. Hence the optimal solutions of (12) and its LP-relaxation are the same and we may use efficient continuous methods, e. g. simplex-method, in order to find the best labeling.

Now consider the case of interval constraints (6). The corresponding problem (9) can be solved by a straightforward generalization of the transportation problem (12). We add an extra source node (see fig. 2b) to the graph with capacity

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

and K extra target nodes with capacities

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

Also we define the capacity of the remaining target nodes as

$$b_k = b'_k, \quad k = 1,\dots,K.$$

The cost terms are defined in the following way

$$\begin{aligned} &c_{nk} = c_{n,K+k} = -\lambda_{nk}, \quad n = 1,\dots,N,\ k = 1,\dots,K, \\ &c_{N+1,k} = +\infty, \quad k = 1,\dots,K, \\ &c_{N+1,K+k} = 0, \quad k = 1,\dots,K. \end{aligned}$$

The final transportation problem has the form

$$\begin{cases} \sum\limits_{n=1}^{N+1}\sum\limits_{k=1}^{2K} c_{nk}t_{nk} \to \min\limits_{T}; \\ \sum\limits_{k=1}^{K} t_{nk} = 1, \quad n = 1,\dots,N; \\ \sum\limits_{k=1}^{K} t_{N+1,k} = \sum\limits_{k}^{K} b''_k - N; \\ \sum\limits_{n=1}^{N+1} t_{nk} = b_k, \quad k = 1,\dots,2K; \\ t_{nk} \geqslant 0. \end{cases} \tag{13}$$

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.

Consider the soft constraints (7). This case corresponds to the following optimization problem:

$$\begin{cases} -\sum\limits_{n=1}^{N}\sum\limits_{k=1}^{K}\lambda_{nk}t_{nk} + \sum\limits_{k=1}^{K}\alpha_k\left|\sum\limits_{n=1}^{N} t_{nk} - b_k\right| \to \min\limits_{T}; \\ \sum\limits_{k=1}^{K} t_{nk} = 1; \quad t_{nk} \in \{0,1\}. \end{cases}$$

This problem 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. 2c):

$$\begin{aligned} &c_{nk} = -\lambda_{nk}, \quad n = 1,\dots,N,\ k = 1,\dots,K, \\ &c_{N+1,k} = \alpha_k, \quad k = 1,\dots,K, \\ &c_{n,K+k} = -\lambda_{nk} + \alpha_k, \quad n = 1,\dots,N,\ k = 1,\dots,K, \\ &c_{N+1,K+k} = 0, \quad k = 1,\dots,K. \end{aligned}$$

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 that with such transportation costs the situation when some real units are transported to the target $K+k$ while some

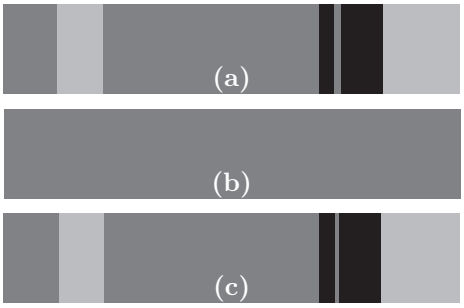


Fig. 3. Segmentation results for a toy signal. a: true segmentation, b: Viterbi segmentation, c: DD HMM segmentation.

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 $K + k$.

Experimental results

First consider a toy signal generated from HMM with 3 hardly distinguishable states and the following parameters:

$$\boldsymbol{\pi} = [0.2, 0.2, 0.6]^\intercal, \quad A = \begin{bmatrix} 0.98 & 0.01 & 0.01 \\ 0.01 & 0.98 & 0.01 \\ 0.01 & 0.01 & 0.98 \end{bmatrix}.$$

The observed values $\boldsymbol{x}_n \in \mathbb{R}^2$, $n = 1, \ldots, 500$ were generated from Gaussian emission probabilities $\mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \Sigma_k)$ with parameters:

$$\boldsymbol{\mu}_1 = [-1, 0]^\intercal, \quad \boldsymbol{\mu}_2 = [0, 1]^\intercal, \quad \boldsymbol{\mu}_3 = [0, 0]^\intercal,$$

$$\Sigma_1 = \begin{bmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 0.6 & 0.2 \\ 0.2 & 0.6 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 1.2 \end{bmatrix}.$$

We compared the true segmentation (see Fig. 3a) with the segmentation of Viterbi algorithm without constraints (see Fig. 3b) and DD HMM with hard constraints obtained from the true segmentation (see Fig. 3c). As can be seen, Viterbi failed to distinguish different states while DD HMM gave the result similar to the true one due to considering of label frequency constraints. The behavior of energy function (4) and its lower bound (10) during DD HMM iterations are shown in Fig. 4.

The next experiment is mouse video tracking segmentation into four behavior acts: sitting in one place,

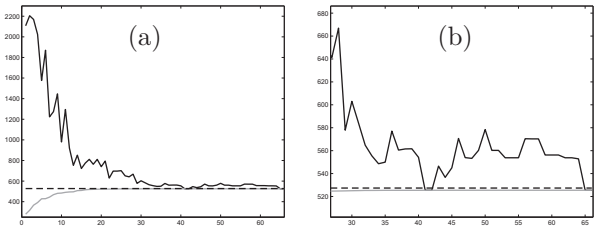


Fig. 4. Behavior of energy function (black) and its lower bound (grey) during DD HMM iterations for a toy signal segmentation problem. The energy value for the true segmentation is shown by dotted line. Case b is an enlarged part of a.

Table 1. Confusion matrix for DD HMM with hard constraints.

	sitting	grooming	walking	running
sitting	4223	126	93	48
grooming	126	237	0	0
walking	83	0	807	67
running	58	0	57	75

Table 2. Confusion matrix for standard HMM.

	sitting	grooming	walking	running
sitting	4226	0	47	217
grooming	363	0	0	0
walking	24	0	685	248
running	16	0	10	164

grooming, walking and running. The video tracking system measures a set of characteristics for each time moment: mouse contour and three points – gravity centre, nose point and tail point. Using these characteristics we calculate for each time moment a set of features like speed, acceleration, different angles, etc. Then using these features and a set of manually segmented mouse trajectories we learn emission probabilities $p(\boldsymbol{x}_n \,|\, \boldsymbol{t}_n)$ for each state (behavior act) by means of mixture of Gaussians as well as transition matrix A and prior probabilities $\boldsymbol{\pi}$. Finally we make segmentation using HMM without constraints (Viterbi algorithm) and DD HMM with hard constraints obtained from the true segmentation. Tables 1 and 2 show confusion matrices for both cases. As can be seen DD HMM shows much better performance resulting in accuracy 89.03% comparing to accuracy 84.58% for standard HMM.

References

[1] *Perner P.* Machine learning and data mining in pattern recognition. — New York: Springer. — 2009.

[2] *Komodakis N., Paragios N., Tziritas G.* MRF optimization via dual decomposition: message-passing revisited // ICCV. — 2007.

[3] *Woodford O., Rother C., Kolmogorov V.* A global perspective on MAP inference for low-level vision // ICCV. — 2009.

[4] *Sigal I., Ivanova A.* An introduction to discrete programming: models and computational algorithms, 2nd ed. — Moscow: Fizmatlit. — 2007. (in Russian).

[5] *Bishop C.* Pattern recognition and machine learning. — New York: Springer. — 2006.

[6] *Viterbi A.* Error bounds for convolutional codes and an asymptotically optimum decoding algorithm // IEEE Transactions on Information Theory. — 1967. — Vol. 13. — Pp. 260–267.

[7] *Kolmogorov V.* Convergent tree-reweighted message passing for energy minimization // IEEE TPAMI. — 2006. — Pp. 1568–1583.