STRUCTURED MODEL LEARNING
EARLY TRACK EXAM SS2020 (24P)

Assignment 1. (4p)
Let $\mathcal{X}$ be a set of inputs and $\mathcal{Y} = \mathcal{A}^n$ a set of sequences of length n defined over a finite alphabet $\mathcal{A}$. Let $h\colon \mathcal{X} \to \mathcal{Y}$ be a prediction rule that for each $x \in \mathcal{X}$ returns a sequence $h(x) = (h_1(x), \dots, h_n(x))$. Assume that we want to measure the prediction accuracy of $h(x)$ by the expected Hamming distance $R(h) = \mathbb{E}_{(x,y_1,\dots,y_n)\sim p}(\sum_{i=1}^n [\![h_i(x) \neq y_i]\!])$ where $p(x, y_1, \dots, y_n)$ is a p.d.f. defined over $\mathcal{X} \times \mathcal{Y}$. As the distribution $p(x, y_1, \dots, y_n)$ is unknown we estimate $R(h)$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{j=1}^{l} \sum_{i=1}^{n} [\![y_i^j \neq h_i(x^j)]\!]$$

where $\mathcal{S}^l = \{(x^i, y_1^i, \dots, y_n^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, l\}$ is a set of examples drawn from i.i.d. random variables with the distribution $p(x, y_1, \dots, y_n)$.

a) Assume that the sequence length is $n = 10$ and that we compute the test error from $l = 1000$ examples. What is the minimal probability that $R(h)$ will be in the interval $(R_{\mathcal{S}^l}(h) - 1, R_{\mathcal{S}^l}(h) + 1)$?

b) What is the minimal number of the test examples l which we need to collect in order to guarantee that $R(h)$ is in the interval $(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon)$ with probability δ at least? Write l as a function of ε, n and δ.

Assignment 2. (8p)
Let $\mathcal{X}$ be a set of inputs and $\mathcal{Y} = \mathcal{A}^n$ a set of hidden sequences of length n defined over finite alphabet $\mathcal{A}$. Let $h\colon \mathcal{X} \to \mathcal{Y}$ be a prediction rule that for each $x \in \mathcal{X}$ returns a sequence $h(x) = (h_1(x), \dots, h_n(x))$ obtained solving

$$h(x) = \operatorname*{argmax}_{(y_1,\dots,y_n)\in\mathcal{A}^n} \left(\sum_{i=1}^{n} q(x, y_i) + \sum_{i=2}^{n} g(y_{i-1}, y_i) \right) \tag{1}$$

where $q\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $g\colon \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ are quality functions describing compatibility between input and hidden states. Let $\mathcal{T}^m = \{(x^j, y_1^j, \dots, y_n^j) \in \mathcal{X} \times \mathcal{A}^n \mid j = 1, \dots, m\}$ be a training set of examples drawn from i.i.d. random variables with a distribution $p(x, y_1, \dots, y_n)$. The goal is to learn q and g such that the predictor (1) has a small expected Hamming distance $R(h) = \mathbb{E}_{(x,,y_1,\dots,y_n)\sim p}(\sum_{i=1}^n [\![h_i(x) \neq y_i]\!])$. To this end, we employ the SO-SVM algorithm learning parameters $\boldsymbol{w} \in \mathbb{R}^d$ of a linear classifier

$$h'(x) \in \operatorname*{argmax}_{(y_1,\dots,y_n)\in\mathcal{A}^n} \langle \boldsymbol{w}, \boldsymbol{\phi}(x, y_1, \dots, y_n) \rangle \tag{2}$$

by solving a convex problem $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \big(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R(\boldsymbol{w})\big)$ where $\lambda > 0$ is a regularization constant and

$$R(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{(y_1,\ldots,y_n)\in\mathcal{A}^n} \left(\sum_{j=1}^{n} [\![y_j^i \neq y_j]\!] + \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y_1, \ldots, y_n) - \boldsymbol{\phi}(x^i, y_1^i, \ldots, y_n^i)\rangle \right).$$

a) Define $\boldsymbol{w}$ and $\boldsymbol{\phi}$ such that (2) and (1) are equivalent.

b) Write a formula for the sub-gradient of $R(\boldsymbol{w})$.

c) Describe a polynomial time algorithm which evaluates the risk $R(\boldsymbol{w})$ and its subgradient $R'(\boldsymbol{w})$. How does the time complexity of the algorithm scale with the number of examples m, sequence length n and the alphabet size $|\mathcal{A}|$?

Assignment 3. (4p)
Consider the following two definitions of a stochastic binary neuron with output $y = \pm 1$ and input $x \in \mathbb{R}^n$

(1) $p_w(y \mid x) = 1/(1 + e^{-y\langle w,x\rangle})$
(2) $y = \operatorname{sign}\big[\langle w, x\rangle - z\big]$ where z is a random variable with standard logistic distribution.

Prove that the definitions are equivalent. Use the fact that the cumulative distribution function of the logistic distribution is $F_z(u) = 1/(1 + e^{-u})$.

Assignment 4. (8p)
Consider the task of semantic segmentation of images $x\colon V \to \mathbb{R}^3$. Let us denote the segmentations by $y\colon V \to K$, where K is the set of labels. We want to use a discriminative model combining a convolutional neural network and a Markov random field

$$p_w(y \mid x) = \frac{1}{Z(x,w)} \exp\Big[-\alpha \sum_{ij\in E} |y_i - y_j| + \sum_{i\in V} u_i(y_i, x_{C_i}, w)\Big],$$

where $u_i(y_i, x_{C_i}, w)$ is the output of the CNN in pixel $i \in V$ and $C_i \subset V$ denotes its transitive receptive field. The network parameters are denoted by w. The MRF parameter α is known. We are given a training set of pairs (x, y) and want to learn the network parameters using the maximum conditional likelihood estimate.

a) Show that learning the network parameters w by gradient descent requires computing marginal probabilities $p_w(y_i \mid x)$.

b) Propose a suitable approximation algorithm for computing the required marginal probabilities and explain it.