STRUCTURED MODEL LEARNING
EXAM SS2020 (25P)

Assignment 1. **(14p)** Consider a linear classifier $h\colon \mathcal{X} \times \mathcal{X} \to \mathcal{Y} \times \mathcal{Y}$ predicting a pair of labels $(y_1, y_2) \in \mathcal{Y} \times \mathcal{Y}$ from a pair of inputs $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$ based on the rule

$$h(x_1, x_2; \boldsymbol{\theta}) = \operatorname*{argmax}_{y_1 \in \mathcal{Y}, y_2 \in \mathcal{Y}} \left(\langle \boldsymbol{\phi}(x_1), \boldsymbol{w}_{y_1} \rangle + \langle \boldsymbol{\phi}(x_2), \boldsymbol{w}_{y_2} \rangle + g(y_1, y_2) \right) \tag{1}$$

where $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^n$ is a feature map, $\boldsymbol{w}_y \in \mathbb{R}^n$, $y \in \mathcal{Y}$, are vectors and $g\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a function. The vector $\boldsymbol{\theta} \in \mathbb{R}^{n|\mathcal{Y}|+|\mathcal{Y}|^2}$ encapsulates all parameters of the classifier, i.e. vectors $\{\boldsymbol{w}_y \in \mathbb{R}^n \mid y \in \mathcal{Y}\}$ and function values $\{g(y, y') \in \mathbb{R} \mid y \in \mathcal{Y}, y' \in \mathcal{Y}\}$. Let $\mathcal{T}^m = \{(x_1^j, x_2^j, y_1^j, y_2^j) \in (\mathcal{X}^2 \times \mathcal{Y}^2) \mid j = 1, \ldots, m\}$ and $\mathcal{S}^l = \{(x_1^j, x_2^j, y_1^j, y_2^j) \in (\mathcal{X}^2 \times \mathcal{Y}^2) \mid j = 1, \ldots, l\}$ be a set of training and testing examples, respectively, both being drawn from i.i.d. random variables with a distribution $p(x_1, x_2, y_1, y_2)$. The goal is to use $\mathcal{T}^m$ to learn a predictor h with small expected risk $R(h) = \mathbb{E}_{(x_1,x_2,y_1,y_2)\sim p} \ell\big(y_1, y_2, h_1(x_1), h_2(x_2)\big)$, where the loss $\ell(y_1, y_2, \hat{y}_1, \hat{y}_2) = |y_1 + y_2 - \hat{y}_1 - \hat{y}_2|$ measures the absolute deviation between the sum of the correct and the predicted labels.

The Structured Output SVM can learn parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ of a linear classifier

$$h'(x_1, x_2) \in \operatorname*{argmax}_{(y_1, y_2) \in \mathcal{Y}^2} \langle \boldsymbol{\theta}, \boldsymbol{\psi}(x_1, x_2, y_1, y_2) \rangle\,, \tag{2}$$

by solving a convex problem $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \left(\frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 + \hat{R}(\boldsymbol{\theta}) \right)$ where $\lambda > 0$ is a regularization constant, $\boldsymbol{\psi}\colon \mathcal{X}^2 \times \mathcal{Y}^2 \to \mathbb{R}^n$ is an input-output feature map, and

$$\hat{R}(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \max_{(y_1, y_2) \in \mathcal{Y}^2} \Big(\ell(y_1^i, y_2^i, y_1, y_2) + \langle \boldsymbol{\theta}, \boldsymbol{\psi}(x_1^i, x_2^i, y_1, y_2) - \boldsymbol{\psi}(x_1^i, x_2^i, y_1^i, y_2^i) \rangle \Big)\,.$$

a) (3p) Define $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ such that (2) and (1) are equivalent.

b) (3p) Write a formula for computing the sub-gradient of $\hat{R}(\boldsymbol{\theta})$.

c) (4p) Describe a variant of the Perceptron algorithm learning parameters $\boldsymbol{\theta}$ such that the classifier (1) predicts all examples from $\mathcal{T}^m$ correctly provided such parameters exist.

d) (4p) Assume that we want to estimate the expected risk $R(h)$ of the learned predictor h by computing the test risk $R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{j=1}^{l} \ell(y_1^j, y_2^j, h_1(x_1^j), h_2(x_2^j))$. What is the minimal number of the test examples l 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 and δ.

Assignment 2. (**3p**) Consider a binary valued Gibbs random field on a bipartite graph with $n + m$ nodes. Its distribution is given by

$$p(x, y) = \frac{1}{Z} \exp\left[a^{\mathsf{T}} x + x^{\mathsf{T}} W y + b^{\mathsf{T}} y\right],$$

where $x \in \mathcal{B}^n$ and $y \in \mathcal{B}^m$ are the label vectors of the two node sets and $\mathcal{B} = \{\pm 1\}$. The $n \times m$ matrix W and the vectors a, b are model parameters. Explain how to learn them from a training sample of pairs (x, y).

Assignment 3. (**8p**) Let $x \in \mathcal{B}^n$ denote n-dimensional binary vectors, where $\mathcal{B} = \{\pm 1\}$. Let W be a symmetric, real valued $n \times n$ matrix. Consider the following parallel sampler on $\mathcal{B}^n$.

$$T(x_{t+1} \mid x_t) = \frac{1}{Z(x_t)} \exp[x_{t+1}^{\mathsf{T}} W x_t]$$

a) Find a close form expression for the normalising factor $Z(x)$. *Hint:* You may want to use the $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$ function for this.

b) Prove that the sampler has a unique limiting distribution $p^*(x)$.

c) Show that the limiting distribution is

$$p^*(x) = \alpha \prod_{i=1}^{n} \cosh(w_i^{\mathsf{T}} x),$$

where w_i, $i = 1, \dots, n$ denote the row vectors of the matrix W.

d) Check whether the sampler has the detailed balance property.