CS 512: Advanced Machine Learning **Due: 12 noon, Feb 27, 2020**

Assignment 1: Graphical Models (Programming Questions)

Student: *Email:*

1 Conditional Random Fields

The Conditional Random Field (CRF) model for a word/label pair $(X, \mathbf{y})$ can be written as

$$p(\mathbf{y}|X) = \frac{1}{Z_X} \exp\left(\sum_{s=1}^{m} \langle \mathbf{w}_{y_s}, \mathbf{x}_s \rangle + \sum_{s=1}^{m-1} T_{y_s, y_{s+1}} \right) \tag{1}$$

$$\text{where} \quad Z_X = \sum_{\hat{\mathbf{y}} \in \mathcal{Y}^m} \exp\left(\sum_{s=1}^{m} \langle \mathbf{w}_{\hat{y}_s}, \mathbf{x}_s \rangle + \sum_{s=1}^{m-1} T_{\hat{y}_s, \hat{y}_{s+1}} \right). \tag{2}$$

$\langle \cdot, \cdot \rangle$ denotes inner product between vectors. Two groups of parameters are used here:

- **Node weight:** Letter-wise discriminant weight vector $\mathbf{w}_k \in \mathbb{R}^{128}$ for each possible letter label $k \in \mathcal{Y}$;
- **Edge weight:** Transition weight matrix T which is sized 26-by-26. T_{ij} is the weight associated with the letter pair of the i-th and j-th letter in the alphabet. For example $T_{1,9}$ is the weight for pair ('a', 'i'), and $T_{24,2}$ is for the pair ('x', 'b'). In general T is not symmetric, *i.e.* $T_{ij} \neq T_{ji}$, or written as $T' \neq T$ where T' is the transpose of T.

Given these parameters (*e.g.* by learning from data), the model (1) can be used to predict the sequence label (*i.e.* word) for a new word image $X^* := (\mathbf{x}_1^*, \ldots, \mathbf{x}_m^*)'$ via the so-called maximum a-posteriori (MAP) inference:

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}^m}{\operatorname{argmax}}\, p(\mathbf{y}|X^*) = \underset{\mathbf{y} \in \mathcal{Y}^m}{\operatorname{argmax}} \left\{ \sum_{j=1}^{m} \left\langle \mathbf{w}_{y_j}, \mathbf{x}_j^* \right\rangle + \sum_{j=1}^{m-1} T_{y_j, y_{j+1}} \right\}. \tag{3}$$

(1a) **[5 Marks]** Show that $\nabla_{\mathbf{w}_y} \log p(\mathbf{y}|X)$—the gradient of $\log p(\mathbf{y}|X)$ with respect to $\mathbf{w}_y$—can be written as:

$$\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t) = \sum_{s=1}^{m} ([\![y_s^t = y]\!] - p(y_s = y|X^t))\mathbf{x}_s^t, \tag{4}$$

where $[\![\cdot]\!] = 1$ if $\cdot$ is true, and 0 otherwise. Show your derivation step by step.

Now derive the similar expression for $\nabla_{T_{ij}} \log p(\mathbf{y}|X)$.

[Answer:] (i) $\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t)$

$$\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t) = \nabla_{\mathbf{w}_y} \left(-logZ_{X^t} + \sum_{s=1}^{m} \left\langle \mathbf{w}_{{y_s}^t}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{{y_s}^t, {y_{s+1}}^t} \right) \tag{5}$$

$$= \nabla_{\mathbf{w}_y} \left(-logZ_{X^t} + \sum_{s=1}^{m} \left\langle \mathbf{w}_{{y_s}^t}, \mathbf{x}_s^t \right\rangle \right) \tag{6}$$

First, we take gradient of the second term:

$$\nabla_{\mathbf{w}_y} \sum_{s=1}^{m} \left\langle \mathbf{w}_{{y_s}^t}, \mathbf{x}_s^t \right\rangle = \sum_{s=1}^{m} \nabla_{\mathbf{w}_y} (\mathbf{w}_{y_s^t}^T {\mathbf{x}_s}^t) \tag{7}$$

$$= \sum_{s=1}^{m} [\![y_s^t = y]\!] \mathbf{x}_s^t \tag{8}$$

Now, we take the gradient of the first term:

$$-\nabla_{\mathbf{w}_y} logZ_{X^t} = -\frac{1}{Z_{X^t}} \sum_{\mathbf{y} \in \mathcal{Y}^m} \exp\left(\sum_{s=1}^{m} \left\langle \mathbf{w}_{y_s}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{y_s, y_{s+1}} \right) \nabla_{\mathbf{w}_y} \sum_{s=1}^{m} \left\langle \mathbf{w}_{y_s}, \mathbf{x}_s^t \right\rangle \tag{9}$$

$$= -\sum_{\mathbf{y} \in \mathcal{Y}^m} p(\mathbf{y}|X^t) \sum_{s=1}^{m} [\![y_s = y]\!] \mathbf{x}_s^t \tag{10}$$

$$= -\sum_{s=1}^{m} p(y_s = y|X^t) \mathbf{x}_s^t \tag{11}$$

Therefore, we get:

$$\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t) = \sum_{s=1}^{m} ([\![y_s^t = y]\!] - p(y_s = y|X^t)) \mathbf{x}_s^t \tag{12}$$

(ii) $\nabla_{T_{ij}} \log p(\mathbf{y}^t|X^t)$

$$\nabla_{T_{ij}} \log p(\mathbf{y}^t|X^t) = \nabla_{T_{ij}} \left(-logZ_{X^t} + \sum_{s=1}^{m} \left\langle \mathbf{w}_{{y_s}^t}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{{y_s}^t, {y_{s+1}}^t} \right) \tag{13}$$

$$= \nabla_{T_{ij}} \left(-logZ_{X^t} + \sum_{s=1}^{m} T_{{y_s}^t, {y_{s+1}}^t} \right) \tag{14}$$

First, we take gradient of the second term:

$$\nabla_{T_{ij}} \sum_{s=1}^{m-1} T_{{y_s}^t, {y_{s+1}}^t} = \sum_{s=1}^{m-1} \nabla_{T_{ij}} T_{{y_s}^t, {y_{s+1}}^t} \tag{15}$$

$$= \sum_{s=1}^{m-1} [\![y_s^t = i, y_{s+1}^t = j]\!] \tag{16}$$

Now, we take the gradient of the first term:

$$-\nabla_{T_{ij}} logZ_{X^t} = -\frac{1}{Z_{X^t}} \sum_{\mathbf{y}\in\mathcal{Y}^m} \exp\left(\sum_{s=1}^{m} \left\langle \mathbf{w}_{y_s}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{y_s,y_{s+1}}\right) \nabla_{T_{ij}} \sum_{s=1}^{m-1} T_{y_s,y_{s+1}} \tag{17}$$

$$= -\sum_{\mathbf{y}\in\mathcal{Y}^m} p(\mathbf{y}|X^t) \sum_{s=1}^{m-1} [\![y_s = i, y_{s+1} = j]\!] \tag{18}$$

$$= -\sum_{s=1}^{m-1} p(y_s = i, y_{s+1} = j | X^t) \tag{19}$$

Therefore, we get:

$$\nabla_{T_{ij}} \log p(\mathbf{y}^t|X^t) = \sum_{s=1}^{m-1} ([\![y_s^t = i, y_{s+1}^t = j]\!] - p(y_s = i, y_{s+1} = j | X^t)) \tag{20}$$

Note that in the above notations, y_s^t are known labels that are given, while y_s is random variable.

(1b) **[5 Marks]** A feature is a function that depends on X and $\mathbf{y}$, but not $p(X|\mathbf{y})$. Show that the gradient of $\log Z_X$ with respect to $\mathbf{w}_y$ and T is exactly the expectation of some features with respect to $p(\mathbf{y}|X)$, and what are the features? Include your derivation.

[Answer:]

(1c) **[20 Marks]** Implement the decoder (3) with computational cost $O(m|\mathcal{Y}|^2)$.

In your submission, create a folder `result` and store the result of decoding (the optimal $\mathbf{y}^* \in \mathcal{Y}^{100}$ of (3)) in `result/decode_output.txt`. It should have 100 lines, where the i-th line contains one integer in $\{1, \ldots, 26\}$ representing y_i^*. In your report, provide the maximum objective value $\sum_{j=1}^{m} \left\langle \mathbf{w}_{y_j}, \mathbf{x}_j \right\rangle + \sum_{j=1}^{m-1} T_{y_j,y_{j+1}}$ for this test case. If you are using your own dynamic programming algorithm (*i.e.* not max-sum), give a brief description especially the formula of recursion.

[Answer:]