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

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

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

(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:

$$\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t) = \sum_{s=1}^{m} (\llbracket y_s^t = y \rrbracket - p(y_s = y|X^t))\mathbf{x}_s^t, \tag{3}$$

where $\llbracket \cdot \rrbracket = 1$ 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) $\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t)$

$$\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t) = \nabla_{\mathbf{w}_y} \left(-logZ_{X^t} + \sum_{s=1}^{m} \left\langle \mathbf{w}_{{y_s}^t}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{{y_s}^t, {y_{s+1}}^t}\right) \tag{4}$$

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

First, we take gradient of the second term:

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

$$= \sum_{s=1}^{m} \llbracket y_s^t = y \rrbracket \mathbf{x}_s^t \tag{7}$$

Now, we take the gradient of the first term:

$$-\nabla_{\mathbf{w}_y} logZ_{X^t} = -\frac{1}{Z_{X^t}} \sum_{\mathbf{y}\in\mathcal{Y}^m} \exp\left(\sum_{s=1}^{m} \left\langle \mathbf{w}_{y_s}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{y_s,y_{s+1}}\right) \nabla_{\mathbf{w}_y} \sum_{s=1}^{m} \left\langle \mathbf{w}_{y_s}, \mathbf{x}_s^t \right\rangle \tag{8}$$

$$= -\sum_{\mathbf{y}\in\mathcal{Y}^m} P(y|X^t) \sum_{s=1}^{m} [\![y_s^t = y]\!] \mathbf{x}_s^t \tag{9}$$

$$= -\sum_{s=1}^{m} P(y_s = y|X^t)\mathbf{x}_s^t \tag{10}$$

Therefore, we get:

$$\nabla_{\mathbf{w}_y} \log p(\mathbf{y}^t|X^t) = \sum_{s=1}^{m} ([\![y_s^t = y]\!] - p(y_s = y|X^t))\mathbf{x}_s^t \tag{11}$$

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

$$\nabla_{T_{ij}} \log p(\mathbf{y}^t|X^t) = \nabla_{T_{ij}} \left(-logZ_{X^t} + \sum_{s=1}^{m} \left\langle \mathbf{w}_{y_s{}^t}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{y_s{}^t, y_{s+1}{}^t}\right) \tag{12}$$

$$= \nabla_{T_{ij}} \left(-logZ_{X^t} + \sum_{s=1}^{m} T_{y_s{}^t, y_{s+1}{}^t}\right) \tag{13}$$

First, we take gradient of the second term:

$$\nabla_{T_{ij}} \sum_{s=1}^{m-1} T_{y_s{}^t, y_{s+1}{}^t} = \sum_{s=1}^{m-1} \nabla_{T_{ij}} T_{y_s{}^t, y_{s+1}{}^t} \tag{14}$$

$$= \sum_{s=1}^{m-1} [\![y_s^t = i, y_{s+1}^t = j]\!] \tag{15}$$

Now, we take the gradient of the first term:

$$-\nabla_{T_{ij}} logZ_{X^t} = -\frac{1}{Z_{X^t}} \sum_{\mathbf{y}\in\mathcal{Y}^m} \exp\left(\sum_{s=1}^{m} \left\langle \mathbf{w}_{y_s}, \mathbf{x}_s^t \right\rangle + \sum_{s=1}^{m-1} T_{y_s,y_{s+1}}\right) \nabla_{T_{ij}} \sum_{s=1}^{m-1} T_{y_s{}^t, y_{s+1}{}^t} \tag{16}$$

$$= -\sum_{\mathbf{y}\in\mathcal{Y}^m} P(y|X^t) \sum_{s=1}^{m-1} [\![y_s^t = i, y_{s+1}^t = j]\!] \tag{17}$$

$$= -\sum_{s=1}^{m-1} P(y_s = i, y_{s+1} = j|X^t) \tag{18}$$

Therefore, we get:

$$\nabla_{T_{ij}} \log p(\mathbf{y}^t|X^t) = [\![y_s^t = i, y_{s+1}^t = j]\!] - \sum_{s=1}^{m-1} P(y_s = i, y_{s+1} = j|X^t) \tag{19}$$

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