**CS 726: Sample Questions on structured learning**

1. Consider a $n \times n$ grid graph $G = (V, E)$ where $V$ are vertices and $E$ are edges of $G$. Each node $k \in V$ is a binary random variable $y_k$ which takes value 1 or 0 depending on whether it is part of foreground or background. Each node is attached with a $x_k$ that is a real-value denoting its propensity to be foreground. There are only three features in this UGM

$$\begin{aligned} f_1((y_k),(k),\mathbf{x}) &= x_k y_k \\ f_2((y_k),(k),\mathbf{x}) &= y_k \\ f_3((y_k,y_j),(k,j),\mathbf{x}) &= y_k y_j + (1-y_k)(1-y_j) \text{ if } (k,j) \in E, 0 \text{ otherwise.} \end{aligned} \tag{1}$$

Let $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ denote the corresponding weights of these three features $\mathbf{f} = [f_1, f_2, f_3]$. Also, consider an instance $(\mathbf{x}^i, \mathbf{y}^i)$ for a $3 \times 3$ grid for which the value of features $x_k^i$ and correct label $y_k^i$ are as given as:

| | | |
|---|---|---|
| $x_1 = 0.0, y_1 = 0$ | $x_2 = 1.5, y_2 = 1$ | $x_3 = 1.0, y_3 = 0$ |
| $x_4 = 1.4, y_4 = 1$ | $x_5 = 2.6, y_5 = 1$ | $x_6 = 1.0, y_6 = 1$ |
| $x_7 = 0.5, y_7 = 0$ | $x_8 = 2.0, y_8 = 1$ | $x_9 = 0.0, y_9 = 0$ |

We will apply the techniques of end-to-end SPEN training on this problem. We will relax each $y_i$ to be continuous value between 0 and 1 using $y_k = \sigma(l_k) = \frac{1}{1+e^{-l_k}}$. The error function is $\text{err}(\mathbf{y}, \mathbf{y}^i) = \sum_{j \in V} |y_j - y_j^i|$ and $\mathbf{f}()$ will work for continuous $y_j$-s. And the energy function is: $E(\mathbf{x}, \mathbf{y}) = -\boldsymbol{\theta}\mathbf{f}(\mathbf{x}, \mathbf{y})$. Rewrite this as a energy over the continuos, unconstrained $l_k$, $x_k$, $\theta_1, \theta_2, \theta_3$ below.

(a) $E(\mathbf{x}, \mathbf{l}) =$ ..2

$-E(\mathbf{x}, \mathbf{l}) = \theta_1 \sum_k x_k \sigma(l_k) + \theta_2 \sum_k \sigma(l_k) + \theta_3 \sum_{k,j} [\sigma(l_k)\sigma(l_k) + (1 - \sigma(l_k))(1 - \sigma(l_k))]$

(b) At each step, we find initial $\mathbf{y}_0$ by ignoring edge potentials. Write the optimal solution for $\mathbf{y}_0$ for instance $(\mathbf{x}^i, \mathbf{y}^i)$ in the above table at $[\theta_1^t, \theta_2^t, \theta_3^t] = [1, 0, 0]$. [Note: this involves finding optimal solution for $l_k$ for each $k \in V$ ..2

We need to find the $\mathbf{y}$ for which $E(\mathbf{x}, \mathbf{y}) + \text{err}(\mathbf{y}, \mathbf{y}^i)$ is maximum. Since edge potentials is zero, this gives that:

$\max_{l_k} (\theta_1 x_k + \theta_2)\sigma(l_k) + |y_k^i - \sigma(l_k)|$

When $y_k^i = 1$, this simplifies to $(x_k - 1)\sigma(l_k) + 1$. This implies that for positions where $x_k \geq 1$ and $y_k^i = 1$ we pick $y_k = 1$ using $l_k = \infty$. The case where $x_k < 1$ and $y_k^i = 1$ does not exist in this example.

Similarly we can work out for $y_k^i = 0$ and get that we need to minimize $(x_k + 1)\sigma(l_k)$ This wants that $y_k = 1$ which mean $l_k$ should be a large positive value.

(c) Thereafter we compute the next $\mathbf{l}_1$ by gradient descent wrt continuous $\mathbf{l}$ on the loss function as discussed in the paper. Write the gradient of the loss function wrt $l_k$ for the same instance and $\theta$ as above. ..3 We need to compute gradient of this term wrt $l_k$ $E(\mathbf{x}, \mathbf{y}) + \text{err}(\mathbf{y}, \mathbf{y}^i)$. As per previous question this works out as:

compute gradient of $(x_k \sigma(l_k) + |y_k^i - \sigma(l_k)|)$ wrt $l_k$. We can do it separately for each case

i. $y_k^i = 0$ : The gradient of $(x_k + 1)\sigma(l_k)$ can be computed using standard techniques.

ii. $y_k^i = 1$: The gradient of $(x_k - 1)\sigma(l_k)$ can be computed using standard techniques.

(d) Write derivative of $E(\mathbf{x}^i, \mathbf{l}_1)$ wrt to $\theta_1$. You may assume $\mathbf{l}_0$ is a constant but not $\mathbf{l}_1$.? ..3

The relevant expression is: $\nabla_{\theta_1} E(\mathbf{x}^i, l_{1k}) - \eta_t H_{\theta_1,\mathbf{l}}(E(\mathbf{x}^i, \mathbf{l}_0))\nabla_{l_1}(E)$
$\nabla_{\theta_1} E(\mathbf{x}^i, l_{1k}) = -\sum_k x_k l_{1k}$.
The Hessian when computed by the difference method becomes: $\frac{1}{r}(-\sum_k x_k(l_{0k} + rv_k) + -\sum_k x_k(l_{0k}))$ where $\mathbf{v} = \nabla_{\mathbf{l}}(E)$ which is
$[(\theta_1 x_k + \theta_2)\sigma(l_k)(1 - \sigma(l_k))] + \theta_3$CumbersomeEdgeTerms.

2. A limitation of VAE is that often it is not easy to train a single set of encoder parameters $\phi$ such that $q_\phi(\mathbf{z}|\mathbf{x})$ is close to the true $P_\theta(\mathbf{z}|\mathbf{x})$ for all $\mathbf{x}$. Can you propose how the ideas in the end-2-end SPEN paper can be used to design an alternative method for estimating $P_\theta(\mathbf{z}|\mathbf{x})$? Assume we approximate $P_\theta(\mathbf{z}|\mathbf{x}^i)$ for each instance $\mathbf{x}^i$ by a Gaussian distribution $q(\mathbf{z}|\mathbf{x}^i)$ whose parameters $\mu_{\mathbf{z}|\mathbf{x}^i}, \Sigma_{\mathbf{z}|\mathbf{x}^i}$ are together denoted by the short form $\lambda^i$. Clearly, specify how you would train the parameters of the network you define. ..4

We identify that optimal $q(\mathbf{z}|\mathbf{x}^i)$ can be obtained by $\max_q \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}^i) \log P(\mathbf{x}^i|\mathbf{z}) - KL(q; p(\mathbf{z}))$

Following the ideas in SPEN paper, we will compute the parameters $\lambda^i$ of $q$ by using fixed number of coordinate update steps on the above function.

The equivalent of 'energy' in this case will be the above ELBO objective. Earlier in VAE, the ELBO objective was calculated using posteriors computed from the encoder. The same encoder can be used to get an initial value for the posterior (defined by $\lambda^i$). Thereafer $\lambda^i$ will be refined by computing gradient of ELBO objective unrolled up to $T$ time steps. When training $\theta$, the gradient will flow through these $T$ updates much like in SPENs.

3. Let's say we try to learn the parameters of a single dimensional Gaussian distribution using various RL loss functions discussed in class. The parameters are $\theta = [\mu, \sigma]$ of the Gaussian. Let reward for a sampled $y$ be $R(y) = 100 - (y - 10)^2$. Let initial values of $\theta^0 = [\mu^0, \sigma^0] = [0, 1]$. Now let us see we draw $M = 1000$ samples from each of the following three algorithms for RL training. Where do you expect the new mean to be after updating the $\theta$ using the corresponding objective for each.

   (a) Policy gradient The $M$ samples will be obtained in this case from $P_{\theta^0}(y)$ which is a $\mathcal{N}(0, 1)$. The reward values of all these samples will be very close to zero but the reward will be higher for positive values. So, we expect the $\mu^t$ to be shifted slightly in the positive direction.

   (b) Reward Augmented Maximum Likelihood with $\tau = 1$. The $M$ samples in this case will be obtained from $\exp^{R(y)/\tau}$ which is a Gaussian $\mathcal{N}(10, 1)$. Hence most of the samples will be close to 10. So, we expect the $\mu^t$ to shift to be around 10 in the next iteration.

   (c) Softmax policy gradient The $M$ samples in this case will be obtained from $\exp^{R(y)/\tau} P_{\theta^0}(y)$ which is a product of two Gaussians $\mathcal{N}(10, 1)$ and $\mathcal{N}(0, 1)$. Hence most of the samples will be in the middle of 10 and 0, that is around 5. So, we expect the $\mu^t$ to shift to be around 5 in the next iteration.

4. Let's say we are estimating a regression function $f(\mathbf{x})$ as a Guassian Process $GP(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$ where the covariance or kernel function $\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{x}.\mathbf{x}'$ and $m(x) = 0$. Each $y_i = f(x_i) + \epsilon_i$ where $\epsilon_i$ $\mathcal{N}(0, \sigma^2)$ where $\sigma^2$ denotes the variance of the noise. Let the training data be $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$. Express the mean $\mu*$ and variance $\sigma*^2$ of the posterior prediction $P(y^*|D, x^*)$ on a test $x^*$ for this Gaussian process in closed form [Hint: will match linear least square solution]

5. Let's say we observe events at the following times $D = \{1, 1, 5, 7, 7, 10, 20, 40\}$. Estimate the maximum likelihood values of the parameters of each of following models for estimating the intensity function at time $T = 100$ where the $t = 40$ is the last event.

   (a) Uniform intensity $f(t) = \lambda e^{-\lambda t}$

   (b) Non-uniform intensity with $\lambda(t) = \alpha_1 \frac{1}{(t-2)^2+1} + \alpha_2 \frac{1}{(t-15)^2+1}$ [Do not need to calculate exact value if the expression is correct.]