Deep Learning and Practice — Final Exam

Date: Tuesday, June 23, 2020
Time: 18:20pm – 21:20pm (180 minutes)
Format: Open book
Instructions:
1) You may give your answers in Chinese or English.
2) Please give your answers in succinct phrases or point form.
3) Please write your answers clearly (with explicit denotation of labels and symbols used).

1. (15 pts) Consider an energy-based model with the following probability distribution

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right)$$

where $\boldsymbol{v} = (v_1, v_2, \ldots, v_m)$ are *binary visible units*; $\boldsymbol{h} = (h_1, h_2, \ldots, h_n)$ are *binary hidden units*; $Z = \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right)$ is the partition function; and $E(\boldsymbol{v}, \boldsymbol{h}))$ is the energy function defined as

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{b}^T \boldsymbol{v} - \boldsymbol{c}^T \boldsymbol{h} - \boldsymbol{v}^T \boldsymbol{W} \boldsymbol{h},$$

with the vectors $\boldsymbol{b}, \boldsymbol{c}$ and the matrix $\boldsymbol{W}$ denoting the model parameters.

(a) (5 pts) Show that $p(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{n} p(h_j|\boldsymbol{v})$ is factorial and $p(h_j = 1|\boldsymbol{v}) = \sigma(c_j + \boldsymbol{v}^T \boldsymbol{W}_{:,j})$, where $\boldsymbol{W}_{:,j}$ is the j-th column vector of $\boldsymbol{W}$.

(b) (5 pts) Show that $p(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{m} p(v_i|\boldsymbol{h})$ is factorial and $p(v_i = 1|\boldsymbol{h}) = \sigma(b_i + \boldsymbol{W}_{i,:} \boldsymbol{h})$, where $\boldsymbol{W}_{i;:}$ is the i-th row vector of $\boldsymbol{W}$.

(c) (5 pts) Assuming the model parameters are known, how can the $p(\boldsymbol{h}|\boldsymbol{v})$ and $p(\boldsymbol{v}|\boldsymbol{h})$ be utilized to draw samples of $\boldsymbol{v}$ (and/or $\boldsymbol{h}$)? How would you draw independent samples of $\boldsymbol{v}$?

2. (15 pts) In the **linear regression** problem with **Bayesian** statistics, the settings are as follows:

$$\begin{aligned} &\text{Visible variables: } y_i = \boldsymbol{\phi}(x_i)^T \boldsymbol{w} + \varepsilon_i, i = 1, 2, \ldots, N \\ &\text{Latent variables: } \boldsymbol{w} = (w_1, w_2, \ldots, w_M) \end{aligned}$$

where ε_i are independently and identically distributed Gaussian noises, and independent of $\boldsymbol{w}$ with[1]

$$\begin{aligned} p(\varepsilon_i) &= \mathcal{N}(\varepsilon_i; 0, \beta^{-1}), i = 1, 2, \ldots, N \\ p(\boldsymbol{w}) &= \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \lambda^{-1} \boldsymbol{I}) \end{aligned}$$

(a) (5 pts) Show that the posterior $p(\boldsymbol{w}|\boldsymbol{y}), \boldsymbol{y} = (y_1, y_2, ..., y_N)$ is given by

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{u}_N, \boldsymbol{\Lambda}_N^{-1}),$$

---

[1] n-dimensional Gaussian: $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \triangleq \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Lambda}^{-1}|} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{x} - \boldsymbol{\mu}))$

with

$$
\begin{aligned}
\boldsymbol{\Lambda}_N &= \lambda \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \\
\boldsymbol{u}_N &= \boldsymbol{\Lambda}_N^{-1}(\lambda \boldsymbol{\mu} + \beta \boldsymbol{\Phi}^T \boldsymbol{y}) \\
\boldsymbol{\Phi} &= \begin{bmatrix} \boldsymbol{\phi}(x_1)^T \\ \boldsymbol{\phi}(x_2)^T \\ \vdots \\ \boldsymbol{\phi}(x_N)^T \end{bmatrix}
\end{aligned}
$$

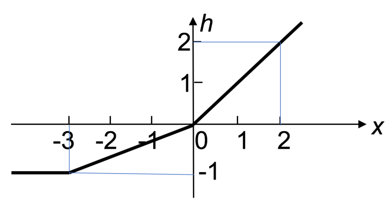
(b) Approximate the posterior $p(\boldsymbol{w}|\boldsymbol{y})$ with the variational mean-field inference $p(\boldsymbol{w}|\boldsymbol{y}) \approx \prod_{i=1}^{M} q(w_i|\boldsymbol{y})$.

(b1) (5 pts) Find the functional form for $q(w_i|\boldsymbol{y})$.

(b2) (5 pts) Provide fixed-point update equations for their parameters.

3. (20 pts) Maxout units, pooling and CNN.

(a) (8 pts) Use the **maxout** unit to design the activation function below.



(b) (6 pts) Enumerate and describe the pooling functions, as many as you know.

(c) (6 pts) What are the major reasons that contribute to the success of convolutional neural networks?

4. (15 pts) Training the VAE.

(a) (3 pts) In training the VAE, we try to maximize a variational lower bound on the data log-likelihood. Explain the main idea and provide the exact objective function to be maximized.

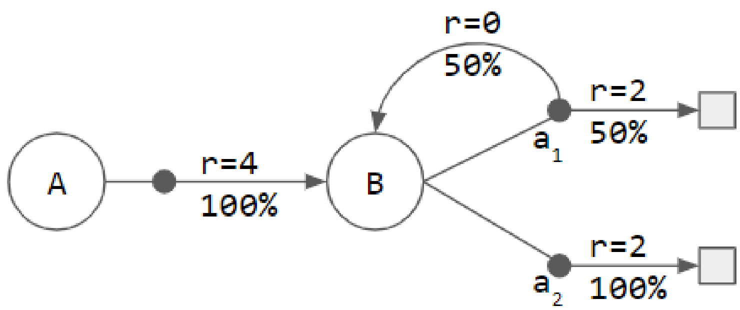
(b) (3 pts) What distribution does the approximate posterior $q(z|x)$ take for training VAE? Is this an assumption?

(c) (3 pts) Explain the notion of the re-parameterization trick.

(d) (3 pts) True or False: In maximizing the variational lower bound, the approximate posterior $q(z|x)$ should ideally be identical to the prior $p(z)$ when the variational lower bound is maximized. Explain your answer.

(e) (3 pts) How would you evaluate the KL divergence $KL(q(z|x)||p(z))$ if the prior $p(z)$ is replaced with a Gaussian Mixture distribution?

5. (5 pts) In evaluating the KL divergence between the ground-truth distribution $p(z)$ and the learned distribution $q(z)$, explain how $q(z)$ may turn out to be if the objective is to minimize $KL(p(z)||q(z))$ and $KL(q(z)||p(z))$. Here $q(z)$ is assumed to be an uni-modal distribution while $p(z)$ has two peaks.



6. (18 pts) Consider a MDP shown below.

   The non-terminal states are $S = \{A, B\}$, and the terminal states are the shaded squares in the figure. There are two actions, $\{a_1, a_2\}$, at state B.

   (a) (8 pts) Given $\pi(B, a_1) = 25\%$, $\pi(B, a_2) = 75\%$.

      i. (4 pts) What is $V_\pi(A)$ when $\gamma = 1$?

      ii. (4 pts) What is $V_\pi(A)$ when $\gamma = 0.5$?

   (b) (10 pts) Given $\gamma = 0.5$,

      i. (7 pts) What is the optimal value $V^*(A)$? (Hint: Bellman optimality equation)

      ii. (3 pts) Give an example of optimal policy and justify.

7. (12 pts) Answer the following questions related to DQN and DDPG.

   (a) (6 pts) What techniques are used for exploration in DQN and DDPG respectively?

   (b) (6 pts) Explain the importance of the target network in DQN.