ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Learning Theory	Assignment date: June 24th, 2019, 12:15
Spring 2019	Due date: June 24th, 2019, 15:15

Final Exam – CS 526 – CE4

There are 4 general problems and 4 multiple choice questions. Good luck!

Name: ______________________________

Section: _______________

Sciper No.: _______________

Problem 1	/ 20
Problem 2	/ 20
Problem 3	/ 20
Problem 4	/ 20
Problem 5: MCQ	/ 20
Total	/100

Problem 1. *VC Dimension* (20 pts)

In this problem we consider hypothesis functions from $\mathbb{R}^2$ to $\{0, 1\}$. We have seen in the homework that $\text{VCdim}(\mathcal{H}_{\text{rec}}) = 4$, where $\mathcal{H}_{\text{rec}}$ is the class of all rectangles in $\mathbb{R}^2$. Let us see some other examples.

1. (10 pts) (Circles) Let $\mathcal{H}_1 = \{h_C(\mathbf{x})\}$ with $h_C(\mathbf{x}) = \mathbb{I}(\mathbf{x} \text{ is inside the circle } C)$, where a circle C is determined by a center and a radius.

   (a) (3 pts) What is $\text{VCdim}(\mathcal{H}_1)$? Call your answer d_1.

   (b) (3 pts) Show that $\text{VCdim}(\mathcal{H}_1) \geq d_1$.
   (Hint: You can propose an instance of d_1 points and for each labelling draw the valid circle.)

   (c) (4 pts) Show that $\text{VCdim}(\mathcal{H}_1) \leq d_1$.
   Hint: You should consider two cases:

   - one of the points $\mathbf{x}$ is in the convex hull of the other points; OR
   - none of the points is in the convex hull of the other points.

   A formal proof might be difficult. It will suffice if you give us a "convincing" argument.

2. (10 pts)(Unbiased neurons) Let $\mathcal{H}_2 = \{h_{\alpha_1,\alpha_2}(\mathbf{x}) : \alpha_1, \alpha_2 \in \mathbb{R}\}$ with

$$h_{\alpha_1,\alpha_2}(\mathbf{x}) = \mathbb{I}\big(\tanh(\alpha_2 x_2 + \alpha_1 x_1) > 0\big).$$

   (a) (3 pts) What is $\text{VCdim}(\mathcal{H}_2)$? Call your answer d_2.

   (b) (3 pts) Show that $\text{VCdim}(\mathcal{H}_2) \geq d_2$.

   (c) (4 pts) Show that $\text{VCdim}(\mathcal{H}_2) \leq d_2$.

Solution:

1. (a) $\text{VCdim}(\mathcal{H}_1) = 3$

   (b) Take three points in $\mathbb{R}^2$ located at the corners of an equilateral triangle. It is then clear that a circle can select any single one of these points, but also any pair of points and of course also all three points together.

   (c) Take 4 points. Assume first that one of the points $\mathbf{x}$ is in the convex hull of the other 3 points. It is then impossible to label the 3 points with '1' and label the point $\mathbf{x}$ with '0'.

   If this is not the case, then the convex hull of the 4 points is a convex quadrilateral. Let $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(3)}$ be a pair of points along a diagonal, and let $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(4)}$ the other pair (along the second diagonal). The two diagonal line segments, called L_1 and

L_2, must intersect each other. Now we claim that it is impossible to have circles such that the corresponding functions implement both $(y_1, y_2, y_3, y_4) = (0, 1, 0, 1)$ and $(y_1, y_2, y_3, y_4) = (1, 0, 1, 0)$. This is true since it is impossible to have two circles C_1 and C_2 such that

- C_1 contains only $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(3)}$, C_2 contains only $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(4)}$, and
- L_1 cuts L_2.

If such C_1 and C_2 existed, it would imply that $(C_1 \cup C_2) \setminus (C_1 \cap C_2)$ has 4 disjoint parts.

2. Note that tanh does not change the sign of $\alpha_2 x_2 + \alpha_1 x_1$, so we don't need to bother with the tanh in analysis.

$\underline{\mathrm{VCdim}(\mathcal{H}_2) \geq 2}$: given any two samples $(\mathbf{x}^{(1)}, y^{(1)})$ and $(\mathbf{x}^{(2)}, y^{(2)})$ with $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ linearly independent, we can find valid α_1, α_2 by solving

$$\begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} b^{(1)} \\ b^{(2)} \end{bmatrix}$$

where $b^{(i)}$ is any real numbers that has the same sign with $(-1)^{1+y^{(i)}}$.

$\underline{\mathrm{VCdim}(\mathcal{H}_2) \leq 2}$: For any three points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$, one can propose $y^{(1)}, y^{(2)}, y^{(3)}$ such that $\mathcal{H}_2$ does not shatter the 3 points. This amounts to showing that there exists $y^{(1)}, y^{(2)}, y^{(3)}$ such that

$$\begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \mathbf{x}^{(3)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} b^{(1)} \\ b^{(2)} \\ b^{(3)} \end{bmatrix} \tag{1}$$

has no solutions, with $b^{(i)}$ as defined above. In $\mathbb{R}^2$ any three points are linearly dependent. So (1) is degenerated. We can assume $\mathbf{x}^{(3)} = w_1 \mathbf{x}^{(1)} + w_2 \mathbf{x}^{(2)}$ for some $w_1, w_2 \in \mathbb{R}$. Suppose $y^{(1)}, y^{(2)}$ allows a solution of α_1, α_2 for the first two equations of (1). However, if one chooses $y^{(3)}$ such that $\sum_{i=1}^{2} \sum_{j=1}^{2} w_i \alpha_j x_j^{(i)}$ has a different sign from $(-1)^{1+y^{(3)}}$, then (1) has no solution.

Problem 2. *GD and SGD* (20 pts)

1. (15 pts) Consider the Least Squares optimization problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}),$$

where $f(\mathbf{x}) = \frac{1}{2}||A\mathbf{x}-\mathbf{b}||_2^2$, $b \in \mathbb{R}^m$. We assume that A is a full column rank matrix in $\mathbb{R}^{m\times n}, n \leq m$, and that there exists a solution to the linear system $A\mathbf{x} = \mathbf{b}$. Let $\sigma_{\max}$ and $\sigma_{\min}$ be the largest and the smallest singular values of A and consider the gradient descent method

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha\nabla f(\mathbf{x}^t)$$

with a fixed step size $\alpha = 1/\sigma_{\max}(A)^2$.

(a) (5 pts) Show that $\sigma_{\max}(I - \alpha A^T A) = 1 - \alpha\sigma_{\min}(A)^2 = 1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2}$.

(b) (5 pts) Calculate the gradient $\nabla f(\mathbf{x})$ and rewrite the GD using this gradient.

(c) (5 pts) Show that the procedure converges as

$$||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 \leq (1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2})||\mathbf{x}^t - \mathbf{x}^*||_2.$$

2. (5 pts) Let us now consider the SGD. In this case one can show a convergence of the form

$$\mathbb{E}[||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2] \leq (1 - \frac{\sigma_{\min}(A)^2}{||A||_F^2})\mathbb{E}[||\mathbf{x}^t - \mathbf{x}^*||_2^2]$$

where $\|A\|_F$ is the Frobenius norm. How does this compare to GD? Which is better?

Solution:

1. (a) Assume that A has the singular value decomposition UDV^T. Plugging this into the expression $I - \alpha A^T A$ we see that $I - \alpha A^T A$ has the singular value decomposition $VD'V^T$, where D' is of dimension $n \times n$ and has the singular values $1 - \alpha\sigma_i^2$. For the given choice of α all these singular values are non-negative and the largest is $1 - \alpha\sigma_{\min}^2(A) = 1 - \frac{\sigma_{\min}^2(A)}{\sigma_{\max}^2(A)}$.

(b) We get

$$\nabla f(\mathbf{x}) = A^T(A\mathbf{x} - \mathbf{b}) = A^T A(\mathbf{x} - \mathbf{x}^*),$$

where we used the fact that A has full column rank so that $A\mathbf{x}^* = b$. Hence GD can be rewritten as

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha A^T A(\mathbf{x}^t - \mathbf{x}^*). \tag{2}$$

(c) Subtracting $\mathbf{x}^*$ from both sides of (2) gives

$$\mathbf{x}^{t+1} - \mathbf{x}^* = \mathbf{x}^t - \mathbf{x}^* - \alpha A^T A(\mathbf{x}^t - \mathbf{x}^*) = (I - \alpha A^T A)(\mathbf{x}^t - \mathbf{x}^*).$$

By taking norms we obtain

$$\begin{aligned} ||\mathbf{x}^{t+1} - \mathbf{x}^*||_2 &\leq \sigma_{\max}(I - \alpha A^T A)||\mathbf{x}^t - \mathbf{x}^*||_2 \\ &= (1 - \alpha\sigma_{min}(A)^2)||\mathbf{x}^t - \mathbf{x}^*||_2. \end{aligned}$$

2. Recall that $||A||_F^2 = \sum_i \sigma_i(A)^2$, where $\sigma_i(A)$ is the i-th singular value of A. Therefore for GD we have a factor $1 - \frac{\sigma_{min}(A)^2}{\sigma_{max}(A)^2}$ and for SGD a factor $\sqrt{1 - \frac{\sigma_{min}(A)^2}{\sum_i \sigma_i(A)^2}}$ (because of the squared norm). The second expression is closer to 1, so GD converges faster.

Problem 3. *Probabilistic graphical models* (20 pts)

Let X_t, $t = 0, 1, 2$ a random walk on the state space $\mathbb{Z}$ (Markov chain) with initial distribution $\mathbb{P}(X_0)$ and transition probability $\mathbb{P}(X_{t+1} = i+1|X_t = i) = p$, $\mathbb{P}(X_{t+1} = i-1|X_t = i) = 1-p$, and zero otherwise (here $0 < p < 1$). We suppose that we have "observations" Y_t of the state at time t given by the output of an additive Gaussian noise channel:

$$Y_t = X_t + \sigma\xi_t, \qquad t = 0, 1, 2$$

where $\xi_t \sim \mathcal{N}(0, 1)$ is Gaussian of mean zero and variance 1. The setting corresponds to the belief network of a Hidden Markov Model (Figure 1).

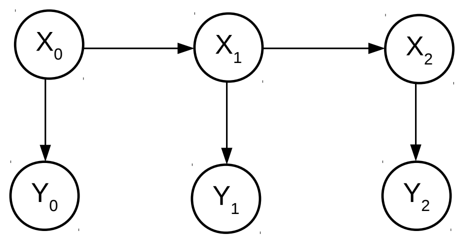

Figure 1: Belief Network

1. (4 pts) Write down the joint probability distribution of the whole belief network.
2. (4 pts) Are Y_0 and Y_2 independent random variables when conditioned on X_1 ? Are they independent when we do not condition ? (no calculation but justification required).
3. (2 pts) Convert the belief network to a Markov Random Field and identify the maximal cliques, the corresponding factors, and the normalization factor Z.
4. (2 pts) *From now on we initialize the Markov chain at time* $t = 0$ *with* $X_0 = 0$. What is the initial distribution $\mathbb{P}(X_0)$? And what is the effective alphabet (or possible values) of the random variables X_1, X_2, Y_1, Y_2, Y_3 ?
5. For this question the initialization is again $X_0 = 0$. We consider the Factor Graph representation of Figure 2.
   a) (6 pts) Set up the message passing equations and compute the marginal $\mu(Y_2)$ from those (see the recap of message passing equations below if needed). Express the result explicitly in terms of p and σ.
   b) (2 pts) Do you think this calculation gives the exact marginal ? Say why.

RECAP: Message passing equations for a general factor graph model $p(\mathbf{x}) \propto \prod_a f_a(\{x_j : j \in \partial a\})$:

$$\mu_{i\to a}(x_i) = \prod_{b\in\partial i\setminus a} \mu_{b\to i}(x_i), \qquad \mu_{a\to i}(x_i) = \sum_{x_j: j\in\partial a\setminus i} f_a(\{x_j, j \in \partial a\}) \prod_{j\in\partial a\setminus i} \mu_{j\to a}(x_j)$$

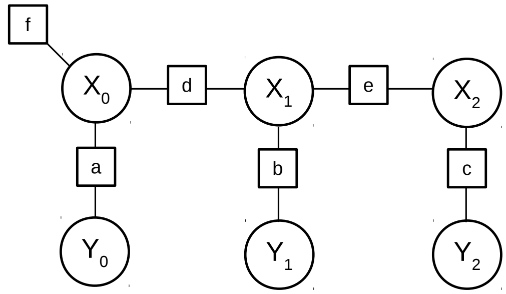

Figure 2: Factor Graph

A leaf node is initialized with $\mu_{i\to a}(x_i) = 1$ and marginals are given by $\mu_i(x_i) \propto \prod_{a\in\partial i} \mu_{a\to i}(x_i)$.

Solution:

1. We have

$$\mathbb{P}(X_0, X_1, X_2, Y_0, Y_1, Y_2) = \mathbb{P}(X_0)\mathbb{P}(X_1|X_0)\mathbb{P}(X_2|X_1)\frac{e^{-\frac{1}{2\sigma^2}(Y_0-X_0)^2}}{\sqrt{2\pi\sigma^2}}\frac{e^{-\frac{1}{2\sigma^2}(Y_1-X_1)^2}}{\sqrt{2\pi\sigma^2}}\frac{e^{-\frac{1}{2\sigma^2}(Y_2-X_2)^2}}{\sqrt{2\pi\sigma^2}}$$

2. The path connecting Y_0 and Y_2 hits X_1 in a head to tail configuration. Therefore (as seen in class), Y_0, Y_2 are independent conditioned on X_1. They are not independent without conditioning: indeed we have

$$\mathbb{P}(Y_0, Y_2) = \sum_{X_0\in\mathbb{Z}} \mathbb{P}(X_0)\frac{e^{-\frac{1}{2\sigma^2}(Y_0-X_0)^2}}{\sqrt{2\pi\sigma^2}}\mathbb{P}(Y_2|X_0)$$

   Since $\mathbb{P}(Y_2|X_0)$ non-trivialy depends on X_0, $\mathbb{P}(Y_0, Y_2) \neq \mathbb{P}(Y_0)\mathbb{P}(Y_2)$.

3. The MRF graph is the same graph but undirected. Maximal cliques are all edges. The MRF is

$$\frac{1}{Z}\psi_1(X_0, X_1)\psi_2(X_1, X_2)\psi_3(X_0, Y_0)\psi_4(X_1, Y_1)\psi_5(X_2, Y_2)$$

   with $Z = 1$ and

$$\begin{aligned}
\psi_1(X_0, X_1) &= \mathbb{P}(X_0)\mathbb{P}(X_1|X_0) \\
\psi_2(X_1, X_2) &= \mathbb{P}(X_2|X_1) \\
\psi_3(X_0, Y_0) &= \frac{e^{-\frac{1}{2\sigma^2}(Y_0-X_0)^2}}{\sqrt{2\pi\sigma^2}} \\
\psi_4(X_1, Y_1) &= \frac{e^{-\frac{1}{2\sigma^2}(Y_1-X_1)^2}}{\sqrt{2\pi\sigma^2}} \\
\psi_5(X_2, Y_2) &= \frac{e^{-\frac{1}{2\sigma^2}(Y_2-X_2)^2}}{\sqrt{2\pi\sigma^2}}
\end{aligned}$$

4. Because of the initialization we have $\mathbb{P}(X_0) = \delta_{X_0,0}$. And the possible values for the r.v's are $X_1 \in \{\pm 1\}$, $X_2 \in \{0, \pm 2\}$, $Y_0, Y_1, Y_2 \in \mathbb{R}$.

5. a) First we identify carefully the factors. Then

$$\mu_{Y_0 \to a}(Y_0) = 1, \qquad \mu_{a \to X_0}(X_0) = \int dY_0 \frac{e^{-\frac{1}{2\sigma^2}(Y_0 - X_0)^2}}{\sqrt{2\pi\sigma^2}} \times 1 = 1$$

$$\mu_{X_0 \to d}(X_0) = \delta_{X_0,0} \times 1, \qquad \mu_{d \to X_1}(X_1) = \sum_{X_0} \mathbb{P}(X_1|X_0)\delta_{X_0,0} = \mathbb{P}(X_1|X_0 = 0)$$

$$\mu_{b \to X_1}(X_1) = 1, \qquad \mu_{X_1 \to e}(X_1) = \mathbb{P}(X_1|X_0 = 0) \times 1$$

$$\mu_{e \to X_2}(X_2) = \sum_{X_1} \mathbb{P}(X_2|X_1)\mu_{X_1 \to e}(X_1) = \sum_{X_1} \mathbb{P}(X_2|X_1)\mathbb{P}(X_1|X_0 = 0)$$

$$\mu_{X_2 \to c}(X_2) = \mu_{e \to X_2}(X_2) = \sum_{X_1} \mathbb{P}(X_2|X_1)\mathbb{P}(X_1|X_0 = 0)$$

$$\begin{aligned}
\mu_{e \to Y_2}(Y_2) &= \sum_{X_2} \frac{e^{-\frac{1}{2\sigma^2}(Y_2 - X_2)^2}}{\sqrt{2\pi\sigma^2}} \mu_{X_2 \to c}(X_2) \\
&= \sum_{X_2} \frac{e^{-\frac{1}{2\sigma^2}(Y_2 - X_2)^2}}{\sqrt{2\pi\sigma^2}} \sum_{X_1} \mathbb{P}(X_2|X_1)\mathbb{P}(X_1|X_0 = 0)
\end{aligned}$$

This last expression is also the marginal. Explicitly in terms of p and σ:

$$\mu(Y_2) = p^2 \frac{e^{-\frac{1}{2\sigma^2}(Y_2 - 2)^2}}{\sqrt{2\pi\sigma^2}} + (1-p)^2 \frac{e^{-\frac{1}{2\sigma^2}(Y_2 + 2)^2}}{\sqrt{2\pi\sigma^2}} + p(1-p) \frac{e^{-\frac{1}{2\sigma^2}(Y_2)^2}}{\sqrt{2\pi\sigma^2}} + p(1-p) \frac{e^{-\frac{1}{2\sigma^2}(Y_2)^2}}{\sqrt{2\pi\sigma^2}}$$

b) This marginal is exact because the factor graph is a tree.

Problem 4. *Tensor methods* (20 pts)

Let $[\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k]$ a set of k *linearly independent* column vectors of dimension n (with real components). We will assume throughout that these vectors have *unit norm.* Set

$$T_2 = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i, \qquad T_3 = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i$$

where $w_i, i = 1, \cdots, k$, are real nonzero values.

We are given the arrays of components $T_2^{\alpha\beta}$, $T_3^{\alpha\beta\gamma}$, $\alpha, \beta, \gamma \in \{1, \cdots, n\}$ and want to determine $w_1, \cdots, w_k$ and $[\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k]$. This problem guides you through a method that uses the simultaneous diagonalization of appropriate matrices.

The following multilinear transformation of T_3 will be used

$$T_3(I, I, \mathbf{u}) = \sum_{i=1}^{k} w_i (\boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i)(\mathbf{u}^T \boldsymbol{\mu}_i)$$

where I denotes the identity matrix and $\mathbf{u}$ an n-dimensional real column vector, $\mathbf{u}^T$ the transposed vector.

1. (7 pts) Let $V = [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k]$ a square matrix. Show that

   $$T_2 = V\, \mathrm{Diag}(w_1, \cdots, w_k) V^T, \quad T_3(I, I, \mathbf{u}) = V\, \mathrm{Diag}(w_1, \cdots, w_k) \mathrm{Diag}(\mathbf{u}^T \boldsymbol{\mu}_1, \cdots, \mathbf{u}^T \boldsymbol{\mu}_k) V^T$$

   where $\mathrm{Diag}(a_1, \cdots, a_k)$ is the diagonal matrix with a_i's on the diagonal.

2. (2 pts) Now we specialize to $n = k$. Why is T_2 an invertible matrix ?

3. We choose $\mathbf{u}$ from a continuous distribution over $\mathbb{R}^n$. Still in the case $n = k$.

   a) (7 pts) Explain how to uniquely recover almost surely the set of μ_i's from the matrix

   $$M = T_3(I, I, \mathbf{u}) T_2^{-1}$$

   using standard linear algebra methods.

   b) (4 pts) How do you then recover the w_i's ?

Solution:

1. Working with components we have on one hand

$$T_2^{\alpha\beta} = \sum_{i=1}^{k} w_i \mu_i^\alpha \mu_i^\beta$$

and on the other hand

$$
\begin{aligned}
(V\mathrm{Diag}(w_1, \cdots, w_k)V^T)^{\alpha\beta} &= \sum_{i,j=1}^{n} V^{\alpha i} w_i \delta_{ij} (V^T)^{j\beta} = \sum_{i,j=1}^{n} V^{\alpha i} w_i \delta_{ij} V^{\beta j} \\
&= \sum_{i=1}^{n} V^{\alpha i} w_i V^{\beta i} = \sum_{i=1}^{n} \boldsymbol{\mu}_i^\alpha w_i \boldsymbol{\mu}_i^\beta
\end{aligned}
$$

Exactly the same calculation applies to:

$$
T_3(I, I, \mathbf{u}) = \sum_{i=1}^{k} w_i (\boldsymbol{u}^T \boldsymbol{\mu}_i)(\boldsymbol{\mu}_i \otimes \boldsymbol{\mu}_i)
$$

with w_i replaced by $w_i(\mathbf{u}^T \boldsymbol{\mu}_i)$. It remains to notice that

$$
\mathrm{Diag}(w_1(\mathbf{u}^T \boldsymbol{\mu}_1), \cdots, w_k(\mathbf{u}^T \boldsymbol{\mu}_k)) = \mathrm{Diag}(w_1, \cdots, w_k)\mathrm{Diag}(\mathbf{u}^T \boldsymbol{\mu}_1, \cdots, \mathbf{u}^T \boldsymbol{\mu}_k)
$$

2. When $n = k$, since μ_i are linearly independent the matrix V is square and full rank, so invertible. This also holds for V^T. Thus since w_i's are non-zero T_2 is also invertible and
$$
T_2^{-1} = (V^T)^{-1} \mathrm{Diag}(\frac{1}{w_1}, \cdots, \frac{1}{w_k}) V^{-1}
$$

3. a) First note that

$$
\begin{aligned}
M &= T_3(I, I, \mathbf{u}) T_2^{-1} \\
&= V\mathrm{Diag}(w_1, \cdots, w_k)\mathrm{Diag}(\mathbf{u}^T \boldsymbol{\mu}_1, \cdots, \mathbf{u}^T \boldsymbol{\mu}_k) V^T (V^{-1})^T \mathrm{Diag}(\frac{1}{w_1}, \cdots, \frac{1}{w_k}) V^{-1} \\
&= V\mathrm{Diag}(\mathbf{u}^T \boldsymbol{\mu}_1, \cdots, \mathbf{u}^T \boldsymbol{\mu}_k) V^{-1}
\end{aligned}
$$

Thus
$$
MV = V\mathrm{Diag}(\mathbf{u}^T \boldsymbol{\mu}_1, \cdots, \mathbf{u}^T \boldsymbol{\mu}_k)
$$
which is equivalent to
$$
M\boldsymbol{\mu}_i = \lambda_i \boldsymbol{\mu}_i, \qquad \lambda_i = \mathbf{u}^T \boldsymbol{\mu}_i
$$
When $\mathbf{u}$ is taken at random from a continuous distribution the inner products $\boldsymbol{\mu}_i^T \mathbf{u}$ are all distinct and non-zero with probability one (indeed the set of $\mathbf{u}$'s satisfying equalities has measure zero). Therefore we uniquely find (normalized) eigenvectors $\boldsymbol{\mu}_i$'s simply by diagonalizing M.

b) Once we have recovered V we find the w_i's from $V^{-1} M_2 (V^{-1})^T$.

Problem 5. *Multiple choice questions* (20 pts)

<u>Circle</u> the correct answers. No justification required

1. (5 pts) [Several correct answers possible.] Let $\mathcal{H} = \{h_\theta\}_{\theta \in \Theta}$ be a hypothesis class such that $\mathrm{VCdim}(\mathcal{H}) = +\infty$. Then the set of parameters Θ:

   A. is finite.

   B. can be countable.

   C. can be uncountable.

   D. can be finite, countable or uncountable.

2. (5 pts) [Several correct answers possible.] Let $(x_i, y_i) \in \mathbb{R} \times \{0, 1\}$ for $i \in \{1, \ldots, n\}$. Let $\hat{y}_i(w) = 1/(1 + e^{-wx_i})$. Define

$$f : w \in \mathbb{R} \mapsto -\sum_{i=1}^{n} \big[y_i \log(\hat{y}_i(w)) + (1 - y_i) \log(1 - \hat{y}_i(w))\big] + \lambda |w| \ ,$$

   where $\lambda > 0$. The function f is:

   A. convex.

   B. differentiable everywhere.

   C. subdifferentiable everywhere.

   D. Lipschitzian.

3. (5 pts) [Single correct answer.] According to the Hammersley-Clifford theorem the MRF property for a probability distribution $p(\mathbf{x}) > 0$ implies

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{\text{maximal cliques } C} \psi_C(\{x_i, i \in C\})$$

   where $\psi_C(\{x_i, i \in C\}) > 0$ and Z is the normalization factor. This decomposition is unique (up to the absorption of Z into factors):

   A. always.

   B. never.

   C. only when the MRF comes from a Belief Network.

   D. only if the graph of the MRF is a tree.

4. (5 pts) [Single correct answer.] Let $w_i(\epsilon)$, $i = 1, \cdots, K$ be continuous functions of $\epsilon \in [0,1]$. Let also $[\mathbf{a}_1+\epsilon\mathbf{a}'_1, \cdots, \mathbf{a}_K+\epsilon\,\mathbf{a}'_K]$, $[\mathbf{b}_1+\epsilon\,\mathbf{b}'_1, \cdots, \mathbf{b}_K+\epsilon\,\mathbf{b}'_K]$, $[\mathbf{c}_1+\epsilon\,\mathbf{c}'_1, \cdots, \mathbf{c}_K+\epsilon\,\mathbf{c}'_K]$ be $N \times K$ rank-K matrices for all ϵ. Consider the tensor

$$T(\epsilon) = \sum_{i=1}^{K} w_i(\epsilon)\,(\mathbf{a}_i + \epsilon\mathbf{a}'_1) \otimes (\mathbf{b}_i + \epsilon\mathbf{b}'_1) \otimes (\mathbf{c}_i + \epsilon\mathbf{c}'_1)$$

A. The tensor rank always equals K for all $\epsilon \in [0,1]$.

B. The tensor rank equals K for all $\epsilon \in [0,1]$ such that $w_i(\epsilon) \neq 0$, $i = 1, \cdots, K$.

C. When we take a limit $\lim_{\epsilon\to 0} T(\epsilon)$ it may happen that the tensor rank of the limit is $K+1$.

D. If we replace the assumption that $[\mathbf{c}_1 + \epsilon\,\mathbf{c}'_1, \cdots, \mathbf{c}_K + \epsilon\,\mathbf{c}'_K]$ is rank K, by the assumption that these vectors are pairwise independent, then the tensor rank can never be K whatever we assume for $w_i(\epsilon)$, $i = 1, \cdots, K$.

Solutions:

1. **B and C.** The set Θ parametrizing the hypothesis class must be infinite: if $\mathcal{H}$ has finite cardinality then $\mathrm{VCdim}(\mathcal{H}) \leq \log|\mathcal{H}|$. In the second graded homework, we studied the hypothesis class $\mathcal{H} = \{\lceil \sin(\theta\pi\cdot)\rceil\}_{\theta\in\Theta}$ and proved that it has an infinite VC dimension if $\Theta = \{2^n\}_{n\in\mathbb{N}}$ (and by extension $\Theta = \mathbb{R}$). Therefore B and C are correct.

2. **A, C and D.** The first sum is the classical cross entropy loss in a logistic regression problem. We can check that this first sum is convex (nonnegative second derivative) and Lipschitzian (bounded first derivative). These properties remain when summing the regularization term.

3. **A.** Because the product is over *maximal* cliques.

4. **B.** When $w_i(\epsilon) \neq 0$ for all i and all $\epsilon \in [0,1]$, according to Jennrich's theorem, since the three arrays have rank K, and there are K terms in the tensor decomposition, this decomposition is unique and therefore the rank is K. A is not true when for some i and ϵ the $w_i(\epsilon)$ vanishes. C is not true because all functions of ϵ are continuous therefore $\lim_{\epsilon\to 0} T(\epsilon) = T(0)$ and by Jennrich's theorem the rank is K. D is not true because if $w_i(\epsilon) \neq 0$ for all i and $\epsilon \in [0,1]$ then the rank is K.