ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Learning Theory	Assignment date: July 2nd, 2022, 15:15
Spring 2022	Due date: July 2nd, 2022, 18:15

CS 526 – Final Exam – room INM 200

There are 4 problems: 3 “regular” problems and one that consists of 6 short questions. Use scratch paper if needed to figure out the solution. Write your final answer in the indicated space. This exam is open-book (lecture notes, exercises, course materials) but no electronic devices allowed. Good luck!

Name: ______________________________

Section: _______________

Sciper No.: _______________

Problem 1	/ 16
Problem 2	/ 18
Problem 3	/ 16
Problem 4	/ 12
Total	/62

The following properties of matrices might be useful:

- For an $n \times n$ matrix A, the trace is defined as: $\mathrm{Tr}(A) = \sum_{i=1}^{n} A_{ii}$.
- The trace of an outer product of two n-dimensional vectors is equal to their inner product: $\mathrm{Tr}\, vw^T = w^T v$.
- The inner product in the space of $n \times n$ real matrices is defined as $\langle M, N \rangle = \mathrm{Tr} M^T N$.
- If an $n \times n$ matrix B is real and symmetric, we have the eigen-decomposition $B = \sum_{j=1}^{n} \lambda_j u_j u_j^T$ where $\lambda_j \in \mathbb{R}$ and $\{u_j\}_{i=1}^{n}$ forms an orthonormal basis. If furthermore, the matrix is positive definite, then $\lambda_j > 0$ for all j.
- The operator norm of an $n \times n$ matrix C is $\|C\| = \max_{\|u\|=1} u^T C u$, $u \in \mathbb{R}^n$. And, we have the property that for two $n \times n$ matrices C, D : $\|CD\| \leq \|C\| \|D\|$.

Problem 1. *(Expectation Learnability)* (16 pts)

Assume that the realizability assumption holds throughout the problem.

A hypothesis class $\mathcal{H}$ is Expectation learnable (E learnable) if there exists a function $m_{\mathcal{H}}^{(\mathrm{E})} : (0,1) \to \mathbb{N}$ and a learning algorithm with the following property: For every $\gamma \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, when running the learning algorithm on a set S of $m \geq m_{\mathcal{H}}^{(\mathrm{E})}(\gamma)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h (which depends on S) such that $\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big] \leq \gamma$ (where the expectation is taken over the training set S). Recall that the error of a prediction is defined to be

$$L_{(\mathcal{D},f)}(h) := \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)].$$

1. (6 pts) Show that if a hypothesis class $\mathcal{H}$ is E learnable, then it is PAC learnable.

2. (6 pts) Show that if a hypothesis class $\mathcal{H}$ is PAC learnable, then it is E learnable.

3. (4 pts) Show that every finite hypothesis class $\mathcal{H}$ is E learnable with sample complexity

$$m_{\mathcal{H}}^{(\mathrm{E})}(\gamma) \leq \left\lceil \frac{2 \log\left(\frac{2|\mathcal{H}|}{\gamma}\right)}{\gamma} \right\rceil.$$

   Hint: You can use results proved in the course, and the relation between sample complexity of PAC learning and E learning derived in previous parts.

Solution to Problem 1:

1. Set $\gamma = \epsilon\delta$. By the E learnability, the algorithm running on $m \geq m_{\mathcal{H}}^{(\mathrm{E})}(\epsilon\delta)$ samples returns a hypothesis h so that $\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big] \leq \epsilon\delta$. Using the Markov inequality, we have:
$$\mathbb{P}\big[L_{(\mathcal{D},f)}(h) \geq \epsilon\big] \leq \frac{\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big]}{\epsilon} \leq \frac{\epsilon\delta}{\epsilon} = \delta.$$
Moreover, the number of samples needed to generate h is bounded by a function in $\epsilon\delta$, which is a function in ϵ, δ. Therefore, the requirements of the PAC learnability are satisfied.

2. Set $\epsilon = \frac{\gamma}{2}, \delta = \frac{\gamma}{2}$, then by PAC learnability, we have an algorithm that running on $m \geq m_{\mathcal{H}}^{(\mathrm{PAC})}\big(\frac{\gamma}{2}, \frac{\gamma}{2}\big)$ samples returns a hypothesis h so that $\mathbb{P}\big[L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big] \leq \frac{\gamma}{2}$. We have
$$\begin{aligned}\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big] &= \mathbb{E}\big[L_{(\mathcal{D},f)}(h) | L_{(\mathcal{D},f)}(h) \leq \frac{\gamma}{2}\big]\mathbb{P}\big[L_{(\mathcal{D},f)}(h) \leq \frac{\gamma}{2}\big] \\ &\qquad + \mathbb{E}\big[L_{(\mathcal{D},f)}(h) | L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big]\mathbb{P}\big[L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big] \\ &\leq \frac{\gamma}{2}\mathbb{P}\big[L_{(\mathcal{D},f)}(h) \leq \frac{\gamma}{2}\big] + \mathbb{E}\big[L_{(\mathcal{D},f)}(h) | L_{(\mathcal{D},f)}(h) > \frac{\gamma}{2}\big]\frac{\gamma}{2} \\ &\leq \frac{\gamma}{2} + \frac{\gamma}{2} = \gamma\end{aligned}$$
where the last inequality is due to the boundedness of $L_{(\mathcal{D},f)}(h)$, since probability is bounded by 1.

   Moreover, the number of samples needed to generate h is bounded by a function in $\epsilon = \frac{\gamma}{2}, \delta = \frac{\gamma}{2}$ which is a function in γ. Therefore, the requirements of the E learnability are satisfied.

3. From the course, we know that every finite hypothesis class is PAC learnable with sample complexity $m_{\mathcal{H}}^{(\mathrm{PAC})}(\epsilon, \delta) \leq \left\lceil \frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon} \right\rceil$. Setting $\epsilon = \frac{\gamma}{2}, \delta = \frac{\gamma}{2}$, we get the result.

Problem 2. *Rayleigh Quotient*(18pts)

Consider a real symmetric $n \times n$ matrix M. Recall that in this case all the eigenvalues are real, call them $\lambda_{max} = \lambda_1 \geq \lambda_2 \cdots \geq \lambda_n = \lambda_{min}$. For a unit norm vector $x \in \mathbb{R}^n$, let $R(M, x) = x^T M x$ be the Rayleigh quotient for M and the vector x.

1. (3pts) Show that $\lambda_{min} \leq R(M, x) \leq \lambda_{max}$. When do you get equality?
2. (3pts) Suppose that you want to minimize or maximize the Rayleigh quotient with respect to the choice of x. Write down the condition of optimality.
3. (3pts) Now assume that the strict inequalities hold: $\lambda_1 > \lambda_2 \cdots > \lambda_n$. Propose an algorithm discussed in class for finding the maximum and minimum Rayleigh quotient.

   Hint: The optimization problem in the second question is the constrained optimization problem $\max_{\|x\|^2=1} x^T M x$. This can be converted into an unconstrained optimization problem using a Lagrange multiplier, call it γ. This leads to the unconstrained optimization of $\left\{x^T M x - \gamma \|x\|^2\right\}$.

Consider now a tensor S of order 3 and dimensions $n \times n \times n$. Assume that S is symmetric under permutation of indices. Let $x \in \mathbb{R}^n$. Recall that $S(x, x, x)$ is defined as $S(x, x, x) = \sum_{\alpha,\beta,\gamma} S^{\alpha,\beta,\gamma} x^\alpha x^\beta x^\gamma$. We follow here the notational convention from the course where the components are indexed by α, β, and γ. For a unit norm vector x define the "Rayleigh quotient" $R(S, x) = S(x, x, x)$.

4. (3pts) Assume that S has a unique decomposition, $S = \sum_{i=1}^n \mu_i \, u_i \otimes u_i \otimes u_i$ where the u_i's form an orthonormal basis, and $\mu_{max} = \mu_1 \geq \mu_2 \cdots \geq \mu_n = \mu_{min} \geq 0$ (note that here we assume all μ_i's non-negative). Show that $-\mu_{max} \leq R(S, x) \leq \mu_{max}$ (observe the difference with the matrix case!). When do you get equalities?
5. (3pts) Write down the optimality condition for maximizing the Rayleigh quotient with respect to x. Show that all vectors u_i, $i = 1, \cdots, n$ of the decomposition of S satisfy this condition.
6. (3pts) Now we further assume that there exists a vector x_0 such that

   $$\mu_1 |u_1^T x_0| > \mu_2 |u_2^T x_0| \geq \cdots \geq \mu_n |u_n^T x_0|$$

   where the first inequality is strict. Propose an algorithm to find the maximum Rayleigh quotient.

Solution to Problem 2:

1. Consider the eigen-decomposition of M, $M = \sum_{i=1}^{n} \lambda_i u_i u_i^T$, where the u_i's form an orthonormal basis. The Rayleigh quotient for the vector x is

$$R(M,x) = x^T \big(\sum_{i=1}^{n} \lambda_i u_i u_i^T\big) x = \sum_{i=1}^{n} \lambda_i (x^T u_i)^2 \leq \lambda_{max} \sum_{i=1}^{n} (x^T u_i)^2 = \lambda_{max},$$

   where in the last equality, we used $\|x\| = 1$ and the fact that the u_i's form an orthonormal basis. Similarly, we can get the lower bound, $R(M,x) \geq \lambda_{min}$.

   Since eigenvectors are orthogonal, the upper bound is attained for u_1, and the lower bound is attained for u_n.

2. Differentiating the Lagrangian $L(x) = x^T M x - \gamma x^T x$, we have $\frac{dL(x)}{dx} = 2Mx - 2\gamma x$. Setting the derivative to zero, we find the optimality condition $Mx = \gamma x$. (This is just the eigenvalue equation and is satisfied by the eigenvalue-eigenvector pairs of M.)

3. When all eigenvalues are distinct the power method allows to find them all. To simplify the exposition assume that all eigenvalues are non-negative (other-wise we need to consider the absolute value). We take an initial vector not orthogonal to the eigenvectors (typically we choose it at random). First we find the largest eigenvalue $\lambda_1 = \lambda_{\max}$ by power iterations, and then by deflating the matrix we find λ_2, and so on till we get $\lambda_n = \lambda_{\min}$.

4. With the decomposition $S = \sum_{i=1}^{n} \mu_i u_i \otimes u_i \otimes u_i$, the Rayleigh quotient is

$$\begin{aligned} R(S,x) = \sum_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \mu_i \, u_i^\alpha \, u_i^\beta \, u_i^\gamma \, x^\alpha \, x^\beta \, x^\gamma &= \sum_{i=1}^{n} \mu_i \big(\sum_\alpha u_i^\alpha x^\alpha\big)\big(\sum_\beta u_i^\beta x^\beta\big)\big(\sum_\gamma u_i^\gamma x^\gamma\big) \\ &= \sum_{i=1}^{n} \mu_i \big(x^T u_i\big)^3 \\ &\leq \sum_{i=1}^{n} \mu_i \big(x^T u_i\big)^2 \\ &\leq \mu_{max} \sum_{i=1}^{n} \big(x^T u_i\big)^2 = \mu_{max} \end{aligned}$$

   where in the first inequality, we used $\mu_i \geq 0$ and $x^T u_i \leq 1$ since $\|x\| = 1$. Similarly, we get the lower bound, $R(S,x) \geq -\mu_{max}$ using $-1 \leq x^T u_i$.

   The upper bound is attained for $x = u_1$, and the lower bound is attained for $x = -u_1$.

5. The derivative of the objective function with respect to a component x^α is

$$\frac{d}{dx^\alpha}\big(S(x,x,x) - l\|x\|^2\big) = 3 \sum_{\beta,\gamma} S^{\alpha,\beta,\gamma} x^\beta x^\gamma - 2 l x^\alpha$$

$$\to \nabla_x \big(S(x,x,x) - l\|x\|^2 \big) = 3S(I,x,x) - 2l\, x$$

Setting the gradient to zero, we find the optimality condition $S(I,x,x) = \frac{2l}{3}\, x$. The vectors u_i satisfy this equation with $l = \frac{3\mu_i}{2}$. Indeed

$$S(I,u_i,u_i) = \sum_k \mu_k\, u_i\, (u_k^T u_i)^2 = \mu_i u_i$$

since $u_k^T u_i = \delta_{ki}$.

6. By the tensor power method, under the assumption, iterating from the initial vector x_0 we converge towards $x_t \to u_1$ and $S(x_t,x_t,x_t) \to \mu_1 = \mu_{max}$.

Problem 3. *Gradient Descent* (16 pts)

Let $X, Y \in \mathbb{R}^{n\times n}$ be $n \times n$ real matrices and $A, B \in \mathbb{R}^{n\times n}$ be $n \times n$ real symmetric and positive definite matrices. Let $F : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$ the function $F(X) = \frac{1}{2}\mathrm{Tr}X^T BX$.

1. (4 pts) Show that $F(X) \geq 0$ for any X.
2. (4 pts) Compute the second derivative of

$$f(s) = \mathrm{Tr}(sX^T + (1-s)Y^T)B(sX + (1-s)Y)$$

   for $s \in [0, 1]$ and deduce that F is a convex function.
3. (4 pts) Deduce the inequality $F(Y) - F(X) \geq \mathrm{Tr}X^T B(Y - X)$. Is F Lipschitz ?
4. (4 pts) Consider now the function $G : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$ with $G(X) = \frac{1}{2}\mathrm{Tr}(X - I)^T A(X - I)$ where I is the identity matrix. Define $L(X) = F(X) + G(X)$.
   (a) (2 pts) Write down the gradient descent algorithm for L. Call X_t the updated matrix at time t.
   (b) (2 pts) Assume that the operator norm $\|X_t\| \leq M$ stays bounded uniformly in n. Show that

$$\|\frac{1}{T}\sum_{t=1}^{T} X_t - (B + A)^{-1}A\| \leq \frac{2M}{\eta T}\|(B + A)^{-1}\|$$

Solution:

1. Use the spectral decomposition $B = \sum_{j=1}^n \lambda_j u_j u_j^T$ and since B is positive definite all $\lambda_j > 0$ (and we can take eigenvectors with real components). Then

$$\begin{aligned}F(X) &= \sum_{j=1}^n \lambda_j \mathrm{Tr} X^T u_j u_j^T X = \sum_{j=1}^n \lambda_j \mathrm{Tr}(X^T u_j)(X^T u_j)^T \\ &= \sum_{j=1}^n \lambda_j (X^T u_j)^T (X^T u_j) = \sum_{j=1}^n \lambda_j \|X^T u_j\|^2 \geq 0\end{aligned}$$

   since $\lambda_j > 0$ for all j.

2. We find

$$\begin{aligned}f''(s) &= 2\mathrm{Tr} X^T B X + 2\mathrm{Tr} Y^T B Y - \mathrm{Tr} X^T B Y - \mathrm{Tr} Y^T B X \\ &= 2\mathrm{Tr}(X-Y)^T B (X-Y) \geq 0\end{aligned}$$

   Thus f is convex. Since $f(s) = f((1-s).0 + s.1)$ we have $f(s) \leq (1-s)f(0) + sf(1)$. This inequality reads

$$F((sX + (1-s)Y) \leq sF(X) + (1-s)F(Y)$$

3. The gradient of $F(X)$ is the matrix

$$\nabla_X F(X) = BX$$

   This can be computed using components $\frac{\partial}{\partial X_{ij}} F(X)$. Since F is convex it is above its tangent and this shows (see class)

$$F(Y) - F(X) \geq \langle \nabla_X F(X), Y - X \rangle = \mathrm{Tr}(BX)^T (Y - X)$$

   Note the last result can also be found working with components.

   The function is not Lipschitz because the gradient BX is not bounded (locally it is Lipschitz but we did not talk about this in class).

4. For L the gradient is $\nabla L(X) = BX + AX - A$. The gradient descent algorithm is as follows: initialize with X_1 and for $t = 1, \cdots, T$ do

$$X_{t+1} = X_t - \eta(BX_t + AX_t - A)$$

   Summing over $t = 1, \cdots, T$ we get

$$\frac{1}{T}(X_{T+1} - X_1) = -\eta((B + A)\frac{1}{T}\sum_{t=1}^T X_t - A)$$

   Since we assume $\|X_t\| \leq M$ uniformly in t, we can use $\|X_1\| \leq M$ and $\|X_{T+1}\| \leq M$ to get

$$\|\frac{1}{T}\sum_{t=1}^T X_t - (B + A)^{-1} A\| \leq \frac{2M}{\eta T} \|(B + A)^{-1}\|$$

Problem 4. *This problem consists of 6 short questions. Answer each point with a short justification or calculation.* [12 pts]

1. (2 pt) Let $\mathcal{H}$ be the class of indicator functions defined by the intervals over $\mathbb{R}$, $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where $h_{a,b}(x) = \mathbb{1}_{[x \notin (a,b)]}$. What is the VC dimension of $\mathcal{H}$?

2. (2 pt) Let $\mathcal{H}$ be the class of indicator functions defined by the intervals over $\mathbb{R}$, $\mathcal{H} = \{h_{a,b,c,d} : a, b, c, d \in \mathbb{R}, a < b, c < d\}$ where $h_{a,b,c,d}(x) = \mathbb{1}_{[x \in (a,b) \text{ OR } x \in (c,d)]}$. What is the VC dimension of $\mathcal{H}$?

3. (2 pt) Let $\mathcal{H}$ be the class of triangles in $\mathbb{R}^2$, $\mathcal{H} = \{h_{a,b,c} : a, b, c \in \mathbb{R}^2, a, b, c \text{ form a triangles}\}$ where $h_{a,b,c}(x) = \mathbb{1}_{[x \in \triangle abc]}$. What is the VC dimension of $\mathcal{H}$?

4. (2 pts) Let T be a $3 \times 3 \times 3$ tensor, all of its entries are 1 except one, that is 2. What is the minimum and what is the maximum multi-linear rank of such a tensor?

5. (2 pts) Let $T = \sum_{r=1}^{4} a_r \otimes b_r \otimes c_r$, where the a_r, b_r, and c_r form the columns of the matrices A, B, and C. Is this decomposition unique? If yes, give the smallest change you can think of to make it potentially non-unique. If no, give the smallest change you can think of to make it unique. The matrices $A = [a_1, \cdots, a_4]$, $B = [b_1, \cdots, b_4]$, $C = [c_1, \cdots, c_4]$ are:

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 3 & 0 & 0 & 2 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 2 & 0 & 0 & 3 \end{pmatrix} \quad C = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

6. (2 pts) Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a differentiable Lipschitz function with constant ρ. Define $h_\alpha : \mathbb{R}^d \mapsto \mathbb{R}$, with $h_\alpha(x) = g(\|x\|^\alpha)$ where $\alpha > 0$. For which values of $\alpha > 0$ can we conclude that h_α a Lipschitz function without further information on g ? Give a Lipschitz constant when this is the case.

Solution:

1. The VC dimension is 2: A set of size 2 can be shattered by $\mathcal{H}$, but for a set of size 3 with elements $x_1 < x_2 < x_3$ the labeling $(0, 1, 0)$ cannot be obtained by any $h_{a,b} \in \mathcal{H}$. Therefore, the VC dimension is 2.

2. The VC dimension is 4: A set of size 4 can be shattered, but a set of size 5 with elements $x_1 < \ldots < x_5$ with labels $(1, 0, 1, 0, 1)$ cannot be obtained by any $h_{a,b,c,d} \in \mathcal{H}$. Therefore, the VC dimension is 4.

3. The VC dimension is 7: A set of size 7 which form convex hull can be shattered by class of triangles. Consider a set of size 8, if one point is in the convex hull of the others, it cannot be shattered. If the set form a convex hull, then the alternating labeling of the points cannot be obtained by any triangle.

4. The multilinear-rank of any such tensor is $(2, 2, 2)$ since regardless where we place the entry 2, each T_x, T_y and T_z will be of size 3×9 and will have two rows that are all ones, and one row of all-ones except a single 2.

5. The determinants of A and B are non-zero (easily computed since we have block matrices). Thus these two matrices are full column rank. For C we easily see that all column pairs are independent vectors. Thus Jennrich's theorem applies so the decomposition is unique. There are infinite ways to make it potentially non-unique: for example change $C_{14} \to 2$ or change $A_{22} \to 1$, etc.

6. We have $\nabla \|x\|^\alpha = \alpha \|x\|^{(\alpha-1)} \frac{x}{\|x\|}$. Therefore $\nabla h_\alpha(x) = \alpha \|x\|^{(\alpha-1)} \frac{x}{\|x\|} g'(\|x\|^\alpha)$ and

   $$\|\nabla h_\alpha(x)\| = \alpha \|x\|^{(\alpha-1)} |g'(\|x\|^\alpha)| \leq \alpha\rho \|x\|^{(\alpha-1)}$$

   So $h_{\alpha=1}$ is a Lipschitz function with constant ρ. For $\alpha > 1$ the equality shows that $\|\nabla h_\alpha(x)\|$ is not bounded so we dont have a Lipschitz function. For $\alpha < 1$ $\|\nabla h_\alpha(x)\|$ is unbounded when $x \to 0$ unless we assume that g vanishes fast enough at the origin so we dont have a Lipschitz constant.