ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Learning Theory — Assignment date: June 25th, 2024, 15:15

Spring 2024 — Due date: June 25th, 2024, 18:15

CS 526 – Final Exam – room INF 119

Use scratch paper if needed to figure out the solution. Write your final solution and answers in the indicated space. This exam is open-book (lecture notes, exercises, course materials). You can use the uploaded material on your computer **but switch off the wifi**. Good luck!

Name: _____

Section: _____

Sciper No.: _____

Problem 1	/ 10
Problem 2	/ 14
Problem 3	/ 14
Problem 4	/ 12
Total	/ 50

Some useful facts:

1. **Taylor's theorem with remainder:** Let $f : \mathbb{R} \to \mathbb{R}$ be a twice differentiable function. Then for any $x, y \in \mathbb{R}^d$ we have

$$f(y) = f(x) + (y - x)^T \nabla f(x) + \frac{1}{2}(y - x)^T \text{Hessian}(f)(\xi x + (1 - \xi)y)(y - x),$$

   for some $\xi \in [0, 1]$.

2. **AM-GM inequality:** For $a_i > 0$, we have

$$\left[\prod_{i=1}^{m} a_i\right]^{1/m} \leq \frac{1}{m} \sum_{i=1}^{m} a_i.$$

Problem 1. *PAC learning* (10 pts)

Let $\mathcal{D}_i$, $i = 1, \cdots, m$ be a collection of *possibly different* distributions over $(x, y) \in \mathcal{X} \times \{0, 1\}$. Let $\mathcal{H}$ be a *finite* class of binary classifiers h. In other words for $|\mathcal{H}| < +\infty$ and $h : x \in \mathcal{X} \mapsto h(x) = y \in \{0, 1\}$. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $(x_i, y_i) \sim \mathcal{D}_i$ be a set of *i.i.d* samples. Consider the average distribution given by the convex combination

$$\mathcal{D} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{D}_i$$

We recall that the true (or population) risk for a distribution $\mathcal{P}$ is

$$L_{\mathcal{P}}(h) = \mathbb{E}_{\mathcal{P}}[\mathbb{1}(h(x) \neq y)]$$

(in this exercise $\mathcal{P} = \mathcal{D}, \mathcal{D}_i$) and the empirical risk is

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(h(x_i) \neq y_i).$$

1. Show that
$$\mathbb{P}[L_S(h) = 0] \leq (1 - L_{\mathcal{D}}(h))^m.$$

2. Prove that for any $\epsilon > 0$ we have
$$\mathbb{P}[\exists h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon \,\text{and}\, L_S(h) = 0] \leq |\mathcal{H}| e^{-\epsilon m}.$$

Solution to Problem 1:

1. (5 pts) Since the empirical risk is a sum of non-negative i.i.d terms distributed with $\mathcal{D}_i$:

$$\begin{aligned}
\mathbb{P}[L_S(h) = 0] &= \prod_{i=1}^{m} \mathbb{P}_{\mathcal{D}_i}[\mathbb{1}(h(x) \neq y) = 0] \\
&= \prod_{i=1}^{m} (1 - \mathbb{P}_{\mathcal{D}_i}[\mathbb{1}(h(x) \neq y) = 1]) \\
&= \prod_{i=1}^{m} (1 - \mathbb{P}_{\mathcal{D}_i}[\mathbb{1}(h(x) \neq y) = 1]) \\
&= \prod_{i=1}^{m} (1 - \mathbb{E}_{\mathcal{D}_i}[\mathbb{1}(h(x) \neq y)]) \\
&= \prod_{i=1}^{m} (1 - L_{\mathcal{D}_i}(h))
\end{aligned}$$

   Note that $\sum_{i=1}^{m} L_{\mathcal{D}_i}(h) = L_{\mathcal{D}}(h)$. Thus using the AGM inequality

$$\mathbb{P}_{\mathcal{D}}[L_S(h) = 0] \leq \left[\frac{1}{m} \sum_{i=1}^{m} (1 - L_{\mathcal{D}_i}(h)) \right]^m = (1 - L_{\mathcal{D}}(h))^m$$

2. (5 pts) We have

$$\{\exists h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon \text{ and } L_S(h) = 0\} = \cup_{h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon} \{L_S(h) = 0\}$$

   Thus by the union bound

$$\mathbb{P}[\exists h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon \text{ and } L_S(h) = 0] \leq \sum_{h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon} \mathbb{P}[L_S(h) = 0]$$

   Replacing the inequality obtained in the previous question in this union bound:

$$\begin{aligned}
\mathbb{P}[\exists h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon \text{ and } L_S(h) = 0] &\leq \sum_{h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon} (1 - L_{\mathcal{D}}(h))^m \\
&\leq \sum_{h \in \mathcal{H} : L_{\mathcal{D}}(h) > \epsilon} (1 - \epsilon)^m \\
&\leq |\mathcal{H}|(1 - \epsilon)^m \\
&\leq |\mathcal{H}| e^{-\epsilon m}
\end{aligned}$$

Problem 2. Gradient descent(14 pts)

Consider a dataset given by $S = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ satisfies $\|x_i\| = 1$, and $y_i \in \mathbb{R}$ for all $1 \leq i \leq n$. Let X be the matrix with x_i's as its rows. Assume that the smallest eigenvalue of the matrix $X^T X$ is $\mu > 0$. We consider the 'linear noiseless setting', where we assume that there exists a $\beta^* \in \mathbb{R}^d$ such that $y_i = x_i^T \beta^*$ for all $i \leq i \leq n$. We want to find β^* by minimizing the loss function

$$L(\beta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\beta, x_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2.$$

1. Show that for any $\beta, \beta' \in \mathbb{R}^d$,

$$L(\beta') - L(\beta) \geq (\beta' - \beta)^T \nabla L(\beta) + \frac{\mu}{n} \|\beta' - \beta\|^2.$$

2. Consider the following stochastic gradient descent for minimizing the loss function L: At each step k, we sample i_k uniformly at random from $\{1, 2, \cdots, n\}$ independent of the previous steps and do the SGD step given by

$$\beta_{k+1} = \beta_k - \eta \nabla \ell(\beta_k, x_{i_k}, y_{i_k}).$$

   Show that for sufficiently small η, we have

$$\mathbb{E}\|\beta_k - \beta^*\|^2 \leq \left(1 - \frac{2\eta\mu}{n}\right)^k \|\beta_0 - \beta^*\|^2.$$

   Find the values of η for which the above convergence rate is satisfied.
   Hint: First estimate the conditional expectation of $\|\beta_k - \beta^*\|^2$ given β_{k-1}.

3. Discuss the differences between the convergence result in Problem 2.2 and the convergence result for SGD discussed in class for convex functions with bounded stochastic gradients.

Solution to Problem 2:

1. (4 pts) From the Taylor's theorem, we have

$$L(\beta') = L(\beta) + (\beta' - \beta)^T \nabla L(\beta) + \frac{1}{2}(\beta' - \beta)^T \text{Hessian}(L)(\xi\beta + (1-\xi)\beta)(\beta' - \beta),$$

for some $\xi \in [0, 1]$. We can rewrite L as

$$L(\beta) = \frac{1}{n}\|X\beta - y\|^2,$$

which gives $\text{Hessian}(L)(\beta) = \frac{2}{n}X^T X$. Hence, we have

$$(\beta' - \beta)^T \text{Hessian}(L)(\xi\beta + (1-\xi)\beta)(\beta' - \beta) \geq \frac{2\mu}{n}\|\beta' - \beta\|^2,$$

giving the desired result.

2. (8 pts) Expanding $\|\beta_{k+1} - \beta^*\|^2$, we have

$$\begin{aligned}\|\beta_{k+1} - \beta^*\|^2 &= \|\beta_k - \eta\nabla\ell(\beta_k, x_{i_k}, y_{i_k}) - \beta^*\|^2 \\ &= \|\beta_k - \beta^*\|^2 - 2\eta\langle\beta_k - \beta^*, \nabla\ell(\beta_k, x_{i_k}, y_{i_k})\rangle + \eta^2\|\nabla\ell(\beta_k, x_{i_k}, y_{i_k})\|^2\end{aligned}$$

Let $\mathbb{E}_k$ denote expectation conditioned on the randomness till step k. We have

$$\begin{aligned}\mathbb{E}_k\|\beta_{k+1} - \beta^*\|^2 &= \|\beta_k - \beta^*\|^2 - 2\eta\langle\beta_k - \beta^*, \mathbb{E}_k\nabla\ell(\beta_k, x_{i_k}, y_{i_k})\rangle + \eta^2\mathbb{E}_k\|\nabla\ell(\beta_k, x_{i_k}, y_{i_k})\|^2 \\ &= \|\beta_k - \beta^*\|^2 - 2\eta\langle\beta_k - \beta^*, \nabla L(\beta_k)\rangle + 4\eta^2\mathbb{E}_k[(x_{i_k}^T\beta_k - y_{i_k})^2\|x_{i_k}\|^2]\end{aligned}$$

The result from the previous question with $\beta = \beta_k, \beta' = \beta^*$ gives

$$\langle\beta_k - \beta^*, \nabla L(\beta_k)\rangle \geq L(\beta_k) + \frac{\mu}{n}\|\beta_k - \beta^*\|^2.$$

Hence,

$$\mathbb{E}_k\|\beta_{k+1} - \beta^*\|^2 \leq \|\beta_k - \beta^*\|^2 - 2\eta L(\beta_k) - \frac{2\eta\mu}{n}\|\beta_k - \beta^*\|^2 + 4\eta^2\mathbb{E}_k[(x_{i_k}^T\beta_k - y_{i_k})^2\|x_{i_k}\|^2]$$

Using the fact that $\|x_{i_k}\|^2 = 1$, we get

$$\begin{aligned}\mathbb{E}_k\|\beta_{k+1} - \beta^*\|^2 &\leq \|\beta_k - \beta^*\|^2 - 2\eta L(\beta_k) - \frac{2\eta\mu}{n}\|\beta_k - \beta^*\|^2 + 4\eta^2 L(\beta_k) \\ &= \left(1 - \frac{2\eta\mu}{n}\right)\|\beta_k - \beta^*\|^2 - 2(\eta - 2\eta^2)L(\beta_k), \\ &\leq \left(1 - \frac{2\eta\mu}{n}\right)\|\beta_k - \beta^*\|^2, \quad \text{for } \eta \leq \frac{1}{2}.\end{aligned}$$

Taking expectation, we get

$$\mathbb{E}\|\beta_{k+1} - \beta^*\|^2 \leq \left(1 - \frac{2\eta\mu}{n}\right) \mathbb{E}\|\beta_k - \beta^*\|^2.$$

Now, recursively applying the above result, we get

$$\mathbb{E}\|\beta_k - \beta^*\|^2 \leq \left(1 - \frac{2\eta\mu}{n}\right)^k \mathbb{E}\|\beta_0 - \beta^*\|^2.$$

3. (2 pts) In Problem 2.2, the convergence is guaranteed for the iterates (β_k) itself without averaging, and is exponentially fast. The convergence result derived in class is for function value at the average of iterates, and the convergence is polynomial in the number of iterations.

Problem 3. *Tensor decomposition* (14 pts)

Consider a collection of R of d-dimensional vectors $\vec{w}_i \in \mathbb{R}^d$, $\|\vec{w}_i\|^2 = d$, $i = 1, \dots, R$. We assume that their barycenter is at the origin, that is $\sum_{i=1}^R \vec{w}_i = 0$. Consider the following 'model'

$$y = \frac{1}{6} \sum_{i=1}^{R} \Big(\frac{\vec{w}_i^T \vec{x}}{\sqrt{d}} \Big)^3 + \xi$$

where the observation $y \in \mathbb{R}$, the signal $\vec{x} \in \mathbb{R}^d$ is random distributed as $\mathcal{N}(0, I_d)$ (with I_d the $d \times d$ identity matrix), and the additive noise $\xi \in \mathbb{R}^d$ is gaussian distributed as $\mathcal{N}(0, I_d)$ (independent of $\vec{x}$).

In this problem the goal is to construct an algorithm (using tensor methods) to estimate the vectors $\vec{w}_i$, $i = 1, \dots, R$ given n training data samples $(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$.

1. What is the mean, variance, and covariance, of the random variables $Z_i = \frac{\vec{w}_i^T \vec{x}}{\sqrt{d}}$?

2. Form the tensor

$$T = d^{\frac{3}{2}} \mathbb{E}[y \, \vec{x} \otimes \vec{x} \otimes \vec{x}]$$

where the expectation is with respect to $\vec{x}$ and ξ. Compute and find an expression for this expectation which involves only the $\vec{w}_i$'s.

Hint: you are advised to work with the components $T^{\alpha\beta\gamma}$, $\alpha, \beta, \gamma = 1, \cdots, d$. You can also use the following property for standard Gaussian variables (sometimes called Wick's theorem)

$$\begin{aligned} \mathbb{E}[x^k x^l x^m x^\alpha x^\beta x^\gamma] = & (\delta_{k\alpha}\delta_{l\beta}\delta_{m\gamma} + \text{permutations of} k, l, m) \\ & + \delta_{\alpha\beta}\big(\delta_{\gamma k}\delta_{lm} + \text{cyclic permutations of} k, l, m\big) \\ & + \delta_{\alpha\gamma}\big(\delta_{\beta k}\delta_{lm} + \text{cyclic permutations of} k, l, m\big) \\ & + \delta_{\beta\gamma}\big(\delta_{\alpha k}\delta_{lm} + \text{cyclic permutations of} k, l, m\big) \end{aligned}$$

3. Now suppose that we are given data $(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$ and that we assume it follows the above model. Suggest a tensor-based algorithm to estimate $\vec{w}_1, \dots, \vec{w}_R$. Justify your answer by specifying:

   (a) What tensor exactly do you suggest we should look at ?

   (b) Suggest an algorithm of your choice and discuss the "chances" of succeeding.

4. *Bonus question (3 pts):* Now we want to investigate how many samples we would need in practice, given a dimensionality d. How would you go about finding the minimum number of samples required for your algorithm to work ? Your answer may consist of a qualitative argument.

Solution to Problem 3:

1. (3 pts) Since $\vec{x} \sim \mathcal{N}(0, I_d)$ and $Z_i = \frac{1}{\sqrt{d}} \sum_{\alpha=1}^{d} w_i^\alpha x^\alpha$,

$$\mathbb{E}[Z_i] = 0, \quad \mathrm{Var}[Z_i] = \frac{1}{d} \sum_i \mathrm{Var}[w_i^\alpha x^\alpha] = \frac{1}{d} \sum_i (w_i^\alpha)^2 = 1$$

and

$$\mathrm{Cov}(Z_i; Z_j) = \mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j] = \frac{1}{d} \sum_{\alpha,\beta} w_i^\alpha w_j^\beta \mathbb{E}[x^\alpha x^\beta] = \frac{1}{d} \vec{w}_i \cdot \vec{w}_j$$

2. (7 pts) Working in components, using that ξ is zero mean and independent of $\vec{x}$, and using Wick's theorem (in the hint):

$$\begin{aligned}
T^{\alpha\beta\gamma} &= d^{3/2} \mathbb{E}[y x^\alpha x^\beta x^\gamma] \\
&= \frac{1}{6} \sum_{i=1}^{R} \mathbb{E}[\left(\vec{w}_i^T \cdot \vec{x}\right)^3 x^\alpha x^\beta x^\gamma] \\
&= \frac{1}{6} \sum_{i=1}^{R} \sum_{k,l,m=1}^{d} w_i^k w_i^l w_i^m \mathbb{E}[x^k x^l x^m x^\alpha x^\beta x^\gamma] \\
&= \sum_{i=1}^{R} w_i^\alpha w_i^\beta w_i^\gamma + \frac{1}{2} \delta_{\alpha\beta} \sum_{i=1}^{R} \sum_{l=1}^{d} w_i^\gamma (w_i^l)^2 + \frac{1}{2} \delta_{\alpha\gamma} \sum_{i=1}^{R} \sum_{l=1}^{d} w_i^\beta (w_i^l)^2 + \frac{1}{2} \delta_{\beta\gamma} \sum_{i=1}^{R} \sum_{l=1}^{d} w_i^\alpha (w_i^l)^2
\end{aligned}$$

Using $\sum_{l=1}^{d} (w_i^l)^2 = d$ and $\sum_{i=1}^{R} w_i^\alpha = 0$ (barycenter at origin) we find

$$T^{\alpha\beta\gamma} = \sum_{i=1}^{R} w_i^\alpha w_i^\beta w_i^\gamma$$

in other words $T = \sum_{i=1}^{R} \vec{w}_i \otimes \vec{w}_i \otimes \vec{w}_i$.

3. (4 pts) Given the data we have access to the empirical tensor

$$T_{\mathrm{emp}} = \frac{d^{3/2}}{n} \sum_{k=1}^{n} y_k \vec{x}_k \otimes \vec{x}_k \otimes \vec{x}_k$$

We expect that for n large enough this will concentrate on $T = \sum_{i=1}^{R} \vec{w}_i \otimes \vec{w}_i \otimes \vec{w}_i$.

One could try Jennrich's algorithm as we have a three mode tensor. However the w_i's are not independent so the guarantees of success are not fulfilled. One could try the tensor power method but again this would require to whiten the Tensor first and again the vectors are not linearly independent. Finally the alternating minimisation method often works in practice but without any a priori guarantees.

4. (3 pts) *Bonus question*: We must ensure that T_{emp} concentrates on T. Since we have a sum of i.i.d terms the squared-fluctuations of T_{emp} are of the order of

$$\text{Var}[T^{\alpha\beta\gamma}_{\text{emp}}] = \frac{d^3}{n^2} n \text{Var}[y x^\alpha_k x^\beta_k x^\gamma_k]$$

With ξ independent additive noise we have two contributions (cross terms vanish). The first one is

$$\frac{d^3}{n}\mathbb{E}[\xi^2]\mathbb{E}[(x^\alpha x^\beta x^\gamma)^2] = O(\frac{d^3}{n})$$

The second one is

$$\frac{d^3}{n}\text{Var}[\sum_{i=1}^R Z_i^3 x^\alpha x^\beta x^\gamma] = \frac{d^3}{n}\sum_{i,j=1}^R \text{Cov}[Z_i^3 x^\alpha x^\beta x^\gamma; Z_j^3 x^\alpha x^\beta x^\gamma] = O(\frac{d^3}{n})$$

(R dependent). To see the last fact we check that Z_i's, x^k's are jointly gaussian with covariances independent of d. For example:

$$\mathbb{E}[Z_i Z_j] = \frac{1}{d}\sum_{k,l} w_i^k w_j^l \mathbb{E}[x^k x^l] = \frac{1}{d}\vec{w}_i \cdot \vec{w}_j \leq \frac{1}{d}\|\vec{w}_i\|\|w_j\| = 1$$

$$\mathbb{E}[Z_i x^k] = \frac{1}{\sqrt{d}}\sum_{l,k} w_i^l \mathbb{E}[x^l x^k] = \frac{1}{\sqrt{d}} w_i^k \leq O(1/\sqrt{d}) \text{ roughly speaking}$$

ect...

In conclusion to have small fluctuations we should have $n >> d^3$ samples.

Problem 4 (12 pts)**.** *This problem consists of 4 short questions. Answer each point with a short justification, picture, or calculation.*

1. (3 pts) Determine the VC-dimension of the following hypothesis class defined on $x \in \mathbb{R}$:

$$\mathcal{H} = \left\{ \operatorname{sgn} \prod_{i=1}^{2k} (x - a_i), \quad a_1 < a_2 < \dots a_{2k} \right\}$$

   where sgn means the sign of the product.

2. (3 pts) Let $f(x) = a|x|^3 + b|x| + c$ for $a, b \in \mathbb{R}_+$ and $c \in \mathbb{R}$. Is this function convex ? If yes what are the subgradient sets $\partial f(x)$?

3. (3 pts) Let $G(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$ and the convolution $f_G(x) = \int_{\mathbb{R}} dz G(z - x) f(z)$. Consider the standard Gaussian random variable $Z \sim \mathcal{N}(0, 1)$. Consider the random map $x \mapsto Z f(x + Z)$. Which is true ?

   (a) This random map is a stochastic gradient of f_G.

   (b) This random map cannot be a stochastic gradient since it does not contain any derivative.

4. (3 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 square matrices A, B, and C. Let $\det A = \det B \neq 0$ and $\det C = 0$. Which of the following is true ? Justify.

   (a) For all matrices A, B, C satisfying the assumptions above, the decomposition of T is not unique (up to trivial rescalings and permutation of terms).

   (b) There exist matrices A, B, C satisfying the assumptions above, such that the decomposition of T is unique (up to trivial rescalings and permutation of terms).

Solution:

1. The VC dimension is 2k+1. Pictures.

2. The function $|x|^3$ is convex (but not strictly convex) and this can be seen by computing the second derivative. Also $|x|$ is convex. The sum of convex fuunctions is convex therefore the whole function is convex. For $x \neq 0$ the subgradient is just the derivative $3ax^2 + b\text{sgn}(x)$. For $x = 0$ the subgradient is $[-b, +b]$.

3. Small calculation shows that (a) is true.

4. By Jenrich's theorem (b) is true.