ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Learning Theory	Assignment date: June 19th, 2023, 15:15
Spring 2023	Due date: June 19th, 2023, 18:15

CS 526 – Final Exam – room INJ 218

There are 4 problems: 3 “regular” problems and one that consists of 4 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	/ 16
Problem 3	/ 16
Problem 4	/ 12
Total	/ 60

Problem 1. *Consistent Learning* (16 pts)

Let $\mathcal{X}$ be a domain set and $\mathcal{Y}$ be a set of labels. Let $\mathcal{F}$ be a set of possible labelling functions, $\mathcal{F} \subset \{f | f : \mathcal{X} \to \mathcal{Y}\}$.

Definition: We say that A is a *consistent learner* for $\mathcal{F}$ using the hypothesis class $\mathcal{H}$, if for any labeling function $f \in \mathcal{F}$ and for all $m \geq 1$, when given as input the set of samples $S = \{(x_1, f(x_1)), \cdots, (x_m, f(x_m))\}$ where $x_i \in \mathcal{X}$, A outputs $h_S \in \mathcal{H}$ such that $h_S(x_i) = f(x_i)$ for $1 \leq i \leq m$.

Remark: Question 1 is about the proof of a statement and question 2 is an application. You can answer question 2 even if you do not prove the statement question 1.

1. (8 pts) Let $\mathcal{F}$ be a labelling class and $\mathcal{H}$ a finite hypothesis class which are not necessarily equal. We suppose there exists a consistent learner A for $\mathcal{F}$ using $\mathcal{H}$. Prove the following statement:

   For all $f \in \mathcal{F}$ and all distributions $\mathcal{D}$ over $\mathcal{X}$ and all $\epsilon, \delta \in (0, 1)$, if A is given a set of samples $S = \{(x_i, f(x_i))\}_{i=1}^m$ with $x_i \sim \mathcal{D}$ and size m such that

   $$m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right),$$

   then with probability at least $1 - \delta$ the learner A outputs a hypothesis $h_S \in \mathcal{H}$ that satisfies

   $$P_{x \sim \mathcal{D}}\big[h_S(x) \neq f(x)\big] \leq \epsilon$$

   <u>Hint</u>: *Fix the labeling function. Then, define a notion of "bad" hypotheses, and use union bound.*

Now, we consider the problem of learning conjunctions. Let $\mathcal{X} = \{0, 1\}^n$. Let $\mathcal{F} = \text{CONJUNCTIONS}_n$ denote the class of conjunctions over the n boolean variables $z_1, \ldots, z_n$. A *literal* is either a boolean variable z_i or its negation $\bar{z}_i$. A conjunction is simply an "and" ($\wedge$) of literals. An example conjunction φ with $n = 10$ is

$$\varphi(z_1, \ldots, z_{10}) = z_1 \wedge \bar{z}_3 \wedge \bar{z}_8 \wedge z_9$$

We want to learn a target conjunction $\phi^* \in \text{CONJUNCTIONS}_n$ from a sampling set $S = \{(x_i, \phi^*(x_i))\}_{i=1}^m$, and the hypothesis class is $\mathcal{H} = \text{CONJUNCTIONS}_n$. So here each sample x_i is a binary vector $(x_{i,1}, \cdots, x_{i,10})$ assigned to $(z_1, \ldots, z_{10})$. The corresponding label $\phi^*(x_i)$ equals 0 or 1.

2. (8 pts) Consider the following algorithm for learning conjunctions:

   1. Set $h = z_1 \wedge \bar{z}_1 \wedge z_2 \wedge \bar{z}_2 \wedge \cdots \wedge z_n \wedge \bar{z}_n$.

For $i = 1, \ldots, m$:
    If $\phi^*(x_i) == 1$: (Ignore samples with 0 label)
        For $j = 1, \ldots, n$:
            If $x_{i,j} == 0$: (j-th bit of x_i)
                Drop z_j from h.
            Else:
                Drop $\bar{z}_j$ from h.
Output h.

(a) Apply the algorithm to the sample set $S = \{(0001, 0), (0111, 0), (1001, 1), (1011, 0)\}$, and determine the output. Check that the algorithm has outputed a *consistent* hypothesis.

(b) Suppose now that the algorithm is indeed a consistent learner. Given (ϵ, δ) how many samples are needed to have:

$$P_{x \sim \mathcal{D}}\big[h_S(x) \neq f(x)\big] \leq \epsilon \qquad \text{with probability at least } 1 - \delta$$

for any distribution $\mathcal{D}$, and set S ?

Solution to Problem 1:

1. Fix the labeling function f and a distribution $\mathcal{D}$ on $\mathcal{X}$. Call a hypothesis $h \in \mathcal{H}$ "bad" if $P_{x \sim \mathcal{D}}\big[h(x) \neq f(x)\big] > \epsilon$. Let E_h be the event that m independent samples in S drawn from $\mathcal{D}$ are all consistent with h, i.e. $h(x_i) = f(x_i)$, for $1 \leq i \leq m$. Then, if h is bad, $P[E_h] \leq (1-\epsilon)^m \leq e^{-\epsilon m}$.

   Consider the event
   $$E = \bigcup_{\text{bad } h \in \mathcal{H}} E_h$$
   Then, by union bound, we have:
   $$P[E] \leq \sum_{\text{bad } h \in \mathcal{H}} P[E_h] \leq |\mathcal{H}| e^{-\epsilon m}$$
   If $m \geq \frac{1}{\epsilon}\big(\log|\mathcal{H}| + \log\frac{1}{\delta}\big)$, then this probability is upper bounded by δ.

   Thus, whenever m is larger than the bound, the probability that a consistent learner returns a bad hypothesis $h_S \in E$ is at most δ. Which means that the event $P(h_S(x) \neq f(x)) > \epsilon$ has probability at most δ. Thus the event $P(h_S(x) \neq f(x)) > \epsilon$ has probability at least $1 - \delta$.

2. (a) The output is $h = z_1 \wedge \bar{z}_2 \wedge \bar{z}_3 \wedge z_4$. Consistency is checked by plugging all four $x_i \in S$ and checking that $h(x_i) = \phi^*(x_i)$.

   (b) We have that $|\mathcal{H}| = 3^n$, because any variable can appear as z_i or $\bar{z}_i$, or do not appear in a conjunction. Then using part 1, we should have
   $$m \geq \frac{1}{\epsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta}) = \frac{1}{\epsilon}(n\log 3 + \log\frac{1}{\delta})$$

Problem 2. Gradient descent(16 pts)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex Lipshitz continuous differentiable function with Lipshitz constant $\rho > 0$. Let S be a real symmetric strictly positive-definite $d \times d$ matrix with smallest eigenvalue $\lambda_{\min} > 0$. We consider a gradient descent iteration for $t \geq 1$ and step size $\eta > 0$:

$$x^{t+1} = x^t - \eta S^{-1} \nabla f(x^t) \tag{1}$$

with initial condition $x^1 = 0$. Further, define $x^* = \text{argmin}_{\|x\| \in B(0,R)} f(x)$, where $B(0, R)$ is the ball of radius R.

1. (4 pts) The update equation (1) is in the form of an Euler *forward* scheme. Write down the associated *backward* Euler scheme.

2. (6 pts) Consider the following iterations (assume the argmin exists and is unique)

$$x^{t+1} = \text{argmin}_x \left\{ f(x) + \frac{1}{2\eta}(x - x^t)^T S(x - x^t) \right\}$$

Is the quantity in the bracket simply convex or strictly convex ? Show that this iteration is equivalent to one of the two Euler schemes.

3. (6 pts) Show that if we choose the step size $\eta = \frac{R\sqrt{\lambda_{\max}\lambda_{\min}}}{\rho\sqrt{T}}$ after T iterations we have

$$f\Big(\frac{1}{T}\sum_{t=1}^{T} x^t\Big) - f(x^*) \leq \frac{\rho R}{\sqrt{T}}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

Hint: recall that in class we proved this statement when $S = I$ the identity matrix. Here you can use an eigenvalue decomposition $S^{-1} = U^T \Lambda^{-1} U$. The following is also useful:

$$\left\langle \underline{\nabla} f\left(x^t\right), x^t - x^*\right\rangle = \left\langle U\nabla f\left(x^t\right), Ux^t - Ux^*\right\rangle = \sum_{k=1}^{d} (U\nabla f)_k(x^t)\left(Ux^t - Ux^*\right)_k$$

Justify why these steps can be used.

Solution to Problem 2:

1. The backward Euler scheme is

$$x^{t+1} = x^t + S^{-1}\nabla f(x^{t+1}).$$

2. The first term f is convex. The second term is strictly convex because S is positive definite (with $\lambda_{min} > 0$). Thus the sum is strictly convex.

   Since f is differentiable we can differentiate the gradient of the quantity in the bracket in order to find the argmin:

$$\nabla f(x) + \eta^{-1}S(x - x^t) = 0$$

   which implies the backward Euler scheme:

$$x^{t+1} = x^t - \eta S^{-1}\nabla f(x^{t+1})$$

3. Let $S^{-1} = U^T\Lambda^{-1}U$ with U an orthogonal matrix, and $\Lambda = \operatorname{Diag}(\lambda_1 \cdots \lambda_d)$. With $\bar{x} = \frac{1}{T}\sum_{t=1}^T x^t$, we have

$$\begin{aligned}
f(\bar{x}) - f(x^*) &\leqslant \frac{1}{T}\sum_{t=1}^T \left(f\left(x^t\right) - f\left(x^*\right)\right) \qquad \text{convexity} \\
&\leqslant \frac{1}{T}\sum_{t=1}^T \left\langle \nabla f\left(x^t\right), x^t - x^* \right\rangle \qquad \text{convexity} \\
&= \frac{1}{T}\sum_{t=1}^T \left\langle U\nabla f\left(x^t\right), Ux^t - Ux^* \right\rangle \\
&= \sum_{k=1}^d \frac{1}{T}\sum_{t=1}^T (U\nabla f)_k(x^t)\left(U\left(x^t - x^*\right)\right)_k \\
&= \sum_{k=1}^d \frac{\lambda_k}{\eta T}\sum_{t=1}^T \left(\frac{\eta}{\lambda_k}\right)(U\nabla f)_k(x^t)\left(U\left(x^t - x^*\right)\right)_k
\end{aligned}$$

$$= \sum_{k=1}^d \frac{\lambda_k}{2\eta T}\sum_{t=1}^T \left\{ -\left(\left(U\left(x^t - x^*\right)\right)_k - \frac{\eta}{\lambda_k}(U\nabla f)_k(x^t)\right)^2 + \left(U\left(x^t - x^*\right)\right)_k^2 + \frac{\eta^2}{\lambda_k^2}\left(U\nabla_f\right)_k(x^t)^2 \right\}$$

   Now, from the backward equation we have:

$$\begin{aligned}
&x^{t+1} = x^t - \eta U^T\Lambda^{-1}U\nabla(x^t) \\
&\Rightarrow Ux^{t+1} = Ux^t - \eta\Lambda^{-1}U\nabla f(x^t) \\
&\left(Ux^{t+1}\right)_k = \left(Ux^t\right)_k - \frac{\eta}{\lambda_k}(U\nabla f)_k(x^t)
\end{aligned}$$

From which we get

$$
\begin{aligned}
f(\bar{x})-f\left(x^*\right) &\leq \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \sum_{t=1}^{T}\left\{-\left(U\left(x^{t+1}-x^*\right)\right)_k^2+\left(U\left(x^t-x^*\right)\right)_k^2+\frac{\eta^2}{\lambda_k^2}\left(U\nabla_f\right)_k\left(x^t\right)^2\right\} \\
&= \sum_{k=1}^{d} \frac{\lambda_k}{2\eta T}\left[\left(U\left(x^1-x^*\right)\right)_k^2-\left(U\left(x^{T+1}-x^*\right)\right)_k^2\right]+\sum_{k=1}^{d} \frac{\lambda_k}{2\eta T} \sum_{t=1}^{T} \frac{\eta^2}{\lambda_k^2}(U\nabla f)_k(x^t)^2 \\
&\leq \frac{\lambda_{\max}}{2\eta T} \sum_{k=1}^{d}\left(U\left(x^1-x^*\right)\right)_k^2+\frac{\eta}{2T\lambda_{\min}} \sum_{t=1}^{T}\|U\nabla f\|^2 \\
&= \frac{\lambda_{\max}}{2\eta T}\left\|U\left(x^1-x^*\right)\right\|^2+\frac{\eta}{2\lambda_{\min}}\|\nabla f\|^2 \\
&\leq \frac{\lambda_{\max}}{2\eta T} R^2+\frac{\eta}{2\lambda_{\min}} \rho^2
\end{aligned}
$$

where we used that $x^1=0$ and $\|x^*\|^2 \leq R^2$ (by assumption) in the last inequality.
Set

$$
\eta^2=\frac{\lambda_{\max}\lambda_{\min}R^2}{\rho^2 T}
$$

Then, we find:

$$
\begin{aligned}
f(\bar{x})-f\left(x^*\right) &\leq \frac{\lambda_{\max}R^2\rho\sqrt{T}}{2\sqrt{\lambda_{\max}\lambda_{\min}}RT}+\frac{\sqrt{\lambda_{\max}\lambda_{\min}}R}{\rho\sqrt{T}} \frac{\rho^2}{2\lambda_{\min}} \\
&= \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{2\sqrt{T}}+\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{2\sqrt{T}} \\
&= \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\rho R}{\sqrt{T}}
\end{aligned}
$$

Problem 3. *Tensor decomposition* (16 pts)

Consider the tensor

$$T = \sum_{i=1}^{K} \lambda_i \, \vec{a}_i \otimes \vec{b}_i \otimes \vec{c}_i$$

where $\vec{a}_i \in \mathbb{R}^{d_1}$ are orthogonal and $\vec{b}_i \in \mathbb{R}^{d_2}$ are orthogonal, $\vec{c}_i \in \mathbb{R}^{d_3}$, and λ_i's are positive and distinct. The goal is to recover the factors $(\lambda_i, \vec{a}_i, \vec{b}_i, \vec{c}_i)$ up to rescaling. Therefore, without loss of generality, we assume that $\|\vec{a}_i\|_2 = \|\vec{b}_i\|_2 = \|\vec{c}_i\|_2 = 1$ for all $1 \leq i \leq K$.

Let $T_{(1)} \in \mathbb{R}^{d_1 \times d_2 d_3}$ be the mode-1 matrization (or unfolding) of T obtained from the vertical fibers of T. $T_{(1)}$ can be expressed in terms of $\lambda_i, \vec{a}_i, \vec{b}_i, \vec{c}_i$'s as:

$$T_{(1)} = \sum_{i=1}^{K} \lambda_i \, \vec{a}_i \big(\vec{c}_i \otimes_{\mathrm{Kro}} \vec{b}_i\big)^{\mathrm{T}}$$

with $\otimes_{\mathrm{Kro}}$ denoting the Kronecker product of two vectors:

$$x \otimes_{\mathrm{Kro}} y = [x_1 y^{\mathrm{T}}, x_2 y^{\mathrm{T}}, \cdots, x_n y^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{nm} \qquad \text{for } x \in \mathbb{R}^n, y \in \mathbb{R}^m$$

1. (5 pts) Let $X = T_{(1)} T_{(1)}^{\mathrm{T}}$. Express X in terms of $\lambda_i, \vec{a}_i, \vec{b}_i, \vec{c}_i$'s, and write its spectral decomposition. What is the rank of X? Explain how to recover the vectors $\vec{a}_i$'s and corresponding λ_i's.

2. (5 pts) Explain how to recover the vectors $\vec{b}_i$'s and how to pair them with the $\vec{a}_i$'s and λ_i's.

3. (6 pts) Now that we have found $(\lambda_i, \vec{a}_i, \vec{b}_i)$'s, describe a way to recover $\vec{c}_i$'s.

   Hint: Try multilinear transformations of T!

Solution:

1. $$X = \sum_{i,j=1}^{K} \lambda_i \, \lambda_j \, \vec{a}_i \big(\vec{c}_i \otimes_{\mathrm{Kro}} \vec{b}_i\big)^{\mathrm{T}} \big(\vec{c}_j \otimes_{\mathrm{Kro}} \vec{b}_j\big) \vec{a}_j^{\mathrm{T}}$$

   From the definition of the Kronecker product, we have that $\big(\vec{c}_i \otimes_{\mathrm{Kro}} \vec{b}_i\big)^{\mathrm{T}} \big(\vec{c}_j \otimes_{\mathrm{Kro}} \vec{b}_j\big) = \big(\vec{c}_i^{\mathrm{T}} \vec{c}_j\big)\big(\vec{b}_i^{\mathrm{T}} \vec{b}_j\big)$. Using the orthogonality of $\vec{b}_i$'s, and the assumption that the vectors are unit norm, we find:

   $$X = \sum_{i=1}^{K} \lambda_i^2 \, \vec{a}_i \vec{a}_i^{\mathrm{T}} \tag{2}$$

   Since $\vec{a}_i$'s are orthogonal, (**??**) is the spectral decomposition of X, thus X has rank K.

   From the tensor T, we can find the matrix X. Computing spectral decomposition of X, we find λ_i^2, and the vectors $\vec{a}_i$'s. Since, λ_i's are assumed to be positive, we can find λ_i's.

2. Form the mode-2 matrization of T, which can be expressed as:

   $$T_{(2)} = \sum_{i=1}^{K} \lambda_i \, \vec{b}_i \big(\vec{a}_i \otimes_{\mathrm{Kro}} \vec{c}_i\big)^{\mathrm{T}}$$

   Then, compute the matrix $Y = T_{(2)} T_{(2)}^{\mathrm{T}}$. Following the same steps as in the previous part, the spectral decomposition of Y is:

   $$Y = \sum_{i=1}^{K} \lambda_i^2 \, \vec{b}_i \vec{b}_i^{\mathrm{T}}$$

   Therefore, $\vec{b}_i$'s can be recovered as the eigenvectors of Y.

3. For each $1 \leq i \leq K$, we consider the following linear transformation of T:

   $$T(\vec{a}_i, \vec{b}_i, .) = \sum_{j=1}^{K} \lambda_j \, (\vec{a}_i^{\mathrm{T}} \vec{a}_j) \, (\vec{b}_i^{\mathrm{T}} \vec{b}_j) \vec{c}_j = \lambda_i \, \vec{c}_i$$

   where in the last equality we used the orthogonality assumption of $\vec{a}_i$'s (or $\vec{b}_i$'s). Since, we know λ_i, we can find $\vec{c}_i$ from the above transformation.

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

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

$$\mathcal{H} = \left\{ \operatorname{sgn}\left(ax^2 + bx + c\right); a, b, c, \in \mathbb{R} \right\}$$

where

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(ii) (3 pts) Let $f(\vec{x}) = \sum_{i=1}^{d} |x_i|^\alpha$, $\vec{x} \in \mathbb{R}^d$, $\alpha \geq 0$. State for which values of α the function is convex, and when this is the case give the subgradient set for each $\vec{x}$.

(iii) (3 pts) Let $\{\vec{u}_i, i = 1, \ldots, d\}$ be an orthogonal basis of column vectors in $\mathbb{R}^d$ where each vector has norm $\sqrt{d}$. We assign *some* probability distribution to the vectors of this basis. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and let $V_i = \vec{u}_i \vec{u}_i^T \nabla f(\vec{x})$, $i = 1 \ldots, d$. Answer by true or false and justify:

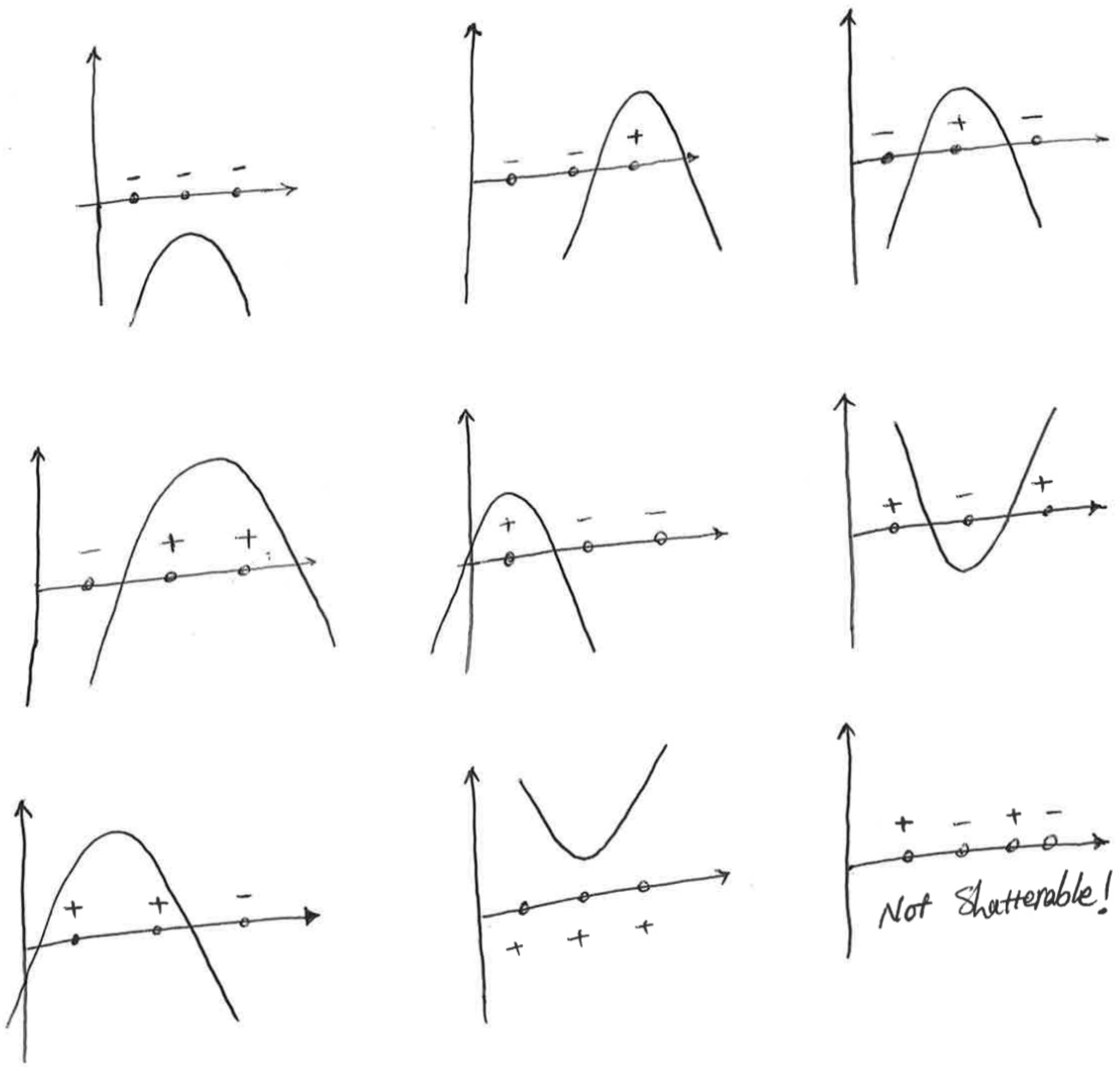
(a) V_i is always a stochastic gradient.

(b) V_i is a stochastic gradient only if the probability distribution is uniform.

(iv) (3 pts) Suppose that the order-3 tensor $T \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ has Tucker decomposition with core tensor $G \in \mathbb{R}^{R_1 \times R_2 \times R_3}$, and factor matrices $A \in \mathbb{R}^{I_1 \times R_1}, B \in \mathbb{R}^{I_2 \times R_2}, C \in \mathbb{R}^{I_3 \times R_3}$. Under what condition on G is the Tucker decomposition the same as the Canonical Polyadic Decomposition (CPD) in terms of rank one tensors with factor matrices A, B, C? Under what condition on A, B, C is the CPD unique?

Solution:

1. The VC dimension is 3 . To prove this, we need to show that there is one configuration of three points such that all its labelings can be shattered, and that no set of 4 points can be shattered. Note that, from the definition of H we are only dealing with points on the x axis (although the VC dimension is still 3 in two dimensions). The case of 3 can easily be verified by checking the 8 possible labelings. And, any alternating labeling of four points will result in a configuration that cannot be shattered because quadratic functions can change signs at most twice.



2. For $\alpha = 0$ the function is constant equal to 1. Therefore it is convex and the subgradient is always $\{0\}$. For $0 < \alpha < 1$ the function is not convex. For $\alpha = 1$ the function is

convex: the subgradient is constituted of vectors of the form $(v_1, \ldots, v_d)$ with $v_i = 1$ if $x_i > 0$, $v_i = -1$ if $x_i < 0$, and $-1 \leq v_i \leq 1$ if $x_i = 0$. For $\alpha > 1$ the function is convex and differentiable and the subgradient is constituted of vectors $(v_1, \ldots, v_d)$ with $v_i = \alpha |x_i|^{\alpha-1}$ for $x_i \geq 0$ and $v_i = -\alpha |x_i|^{\alpha-1}$ for $x_i \leq 0$.

3. To have a stochastic gradient one has to check that $\mathbb{E}[\vec{u}_i \vec{u}_i^T \nabla f(\vec{x})] = \nabla f(\vec{x})$. Since $\mathbb{E}[\vec{u}_i \vec{u}_i^T] = \sum_{i=1}^d p_i \vec{u}_i \vec{u}_i^T$ we get the identity matrix for $p_i = \frac{1}{d}$ (since $\|\vec{u}_i\| = \sqrt{d}$). Therefore we have (a) is false; (b) is true; and (c) is of course false.

4. If G is a super diagonal tensor $G_{i,j,k} = \lambda_i \delta(i,j)\delta(j,k)$. Then, the CPD of T is:

$$T = \sum_{i=1}^{R} \lambda_i a_i \otimes b_i \otimes c_i$$

with $R = \min\{R_1, R_2, R_3\}$, and is unique under conditions of Jenrich' theorem.