ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Learning Theory
Spring 2021

Assignment date: June 24th, 2021, 08:15
Due date: June 24th, 2021, 11:15

Final Exam – CS 526 –

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	/ 15
Problem 2	/ 12
Problem 3	/ 15
Problem 4	/ 13
Total	/55

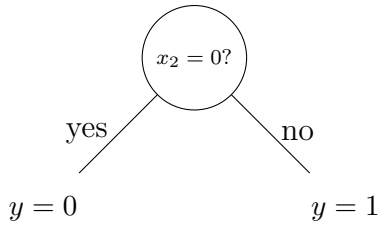

Figure 1: Example of single-node decision tree

Problem 1. *(VC dimension of decision trees with binary features)* (15pts)

In this problem, we consider the class $\mathcal{H}_{btree}$ of decision trees with binary features and binary labels. We have a set of samples $x^{(1)}, \ldots, x^{(m)}$, where $x^{(i)} \in \{0,1\}^d$. A decision tree is a classifier that returns the binary label y for a sample x after performing a series of tests of the type "$x_i = 0$?" for $0 \leq i < d$, which are organized in a binary tree-like manner. Nodes of this tree correspond to the tests and leaves to the returned label values. Note that it is allowed to return the same label value from both branches.

1. (5pts) Consider the subclass $\mathcal{H}_1$ of trees with a single decision node (see Fig. 1). Show that
$$\text{VCdim}\,\mathcal{H}_1 \leq \lfloor \log_2(d+1) \rfloor + 1.$$

2. (5pts) Show that
$$\text{VCdim}\,\mathcal{H}_1 \geq \lfloor \log_2(d+1) \rfloor + 1.$$

3. (5pts) Consider the subclass $\mathcal{H}_{deg,N}$ of degenerate trees. Now the tree has N decision nodes but each node except the bottom one has a single child node (see Fig. 2). Prove that
$$\text{VCdim}\,\mathcal{H}_{deg,N} \geq \lfloor \log_2(d-N+2) \rfloor + N.$$
Hint: Start from the case $N = 1$. What changes when we add another node to the tree?

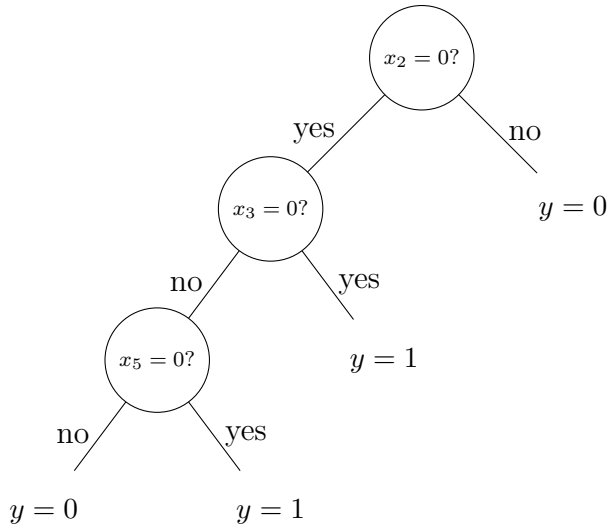

Figure 2: Example of degenerate tree with $N = 3$ nodes.

Solution:

1. For each feature i, there exist two trivial decision trees (that both return zero or both return one) and two non-trivial ones (the one that returns 0 if $x_i = 1$ and 1 otherwise and the one that returns 1 if $x_i = 1$ and 0 otherwise). Therefore, with d features we can have at most $2d + 2$ distinct labelings. In order to shatter m samples, we need to obtain all 2^m possible labelings, hence we have the bound

$$2d + 2 \geq 2^m.$$

   Resolving for m we get the stated upper bound.

2. To prove the lower bound, we need to construct the set of $m = \lfloor \log_2(d+1) \rfloor + 1$ samples that is shattered. To do this, take the set of all possible labelings except all-zero and all-one and for each labeling $(y_1, \ldots, y_m)$ remove its complement from the set. This leaves $2^{m-1} - 1$ distinct labelings $y^{(i)}$. Now we create the samples $x^{(1)}, \ldots, x^{(m)}$ s.t. $x_i^{(j)} = y_j^{(i)}$ for $1 \leq j \leq m, 1 \leq i \leq 2^{m-1} - 1 = d$. It remains to notice that a tree with node $x_i = 0?$ gives either the labeling $y^{(i)}$ or its complement (if we reverse the labels on branches) and in addition all-one and all-zero labelings if both branches return the same label, which completes the proof.

3. We need to construct the set of $m = \lfloor \log_2(d - N + 2) \rfloor + N$ samples on which we get all 2^m possible labels. We start from the case of one bottom node, with $d = 2^{m-1} - 1$ features for m samples. Now assume we get an extra feature x_{d+1} and an extra sample s.t. $x_{d+1}^{(m+1)} = 1$ and $x_i^{(m+1)} = 0$ for $i \neq d + 1$ ($x_{d+1}^{(i)} = 0$ for $i < m + 1$). We create a parent node that contains the existing node and our new sample as children and

the splitting rule is the new feature. The new splitting rule allows to label $x^{(m+1)}$ independently of other $x^{(i)}$, so we get all possible labelings on $m+1$ samples. This procedure can be performed $N-1$ times since we have N decision nodes in the tree. Therefore, for m samples we have $d = 2^{m-1-(N-1)} - 1 + (N-1) = 2^{m-N} + N - 2$ features that generate all 2^m possible labelings.

Problem 2. *A Conservation Law For Neural Networks* (12pts)

Consider a neural network (NN). To keep things as simple as possible assume that the activation functions have weights but no bias terms. Assume that we train the NN using gradient descent (GD). Let $\boldsymbol{w}$ denote the vector of weights. Let $L(\boldsymbol{w})$ denote our cost function (which depends of course on the given samples; but we suppress this dependence in our notation). Then, starting with an initial value $\boldsymbol{w}_0$, we proceed by computing the sequence $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta \nabla L(\boldsymbol{w}_{t-1})$ for a given step size η for a certain number of steps. This is our GD algorithm. As we already explored in the class, it is often easier to look at the continous-time version of this algorithm. This is called gradient flow (GF). The corresponding continuous-time version is the differential equation $\dot{\boldsymbol{w}}(t) = -\nabla L(\boldsymbol{w})$. GD (GF) is not the only possible algorithm. There are several variants that are used in practice, e.g., Nesterov's algorithm. For our purpose it is easiest to consider the so-called Newton dynamics (ND). The corresponding continuous-time version reads $\ddot{\boldsymbol{w}}(t) = -\nabla L(\boldsymbol{w}(t))$. For various reasons it is not used in practice but it is mathematically easier.

1. (6pts) Show that when we apply the ND to our system then $\frac{1}{2}\|\dot{\boldsymbol{w}}(t)\|^2 + L(\boldsymbol{w}(t))$ stays constant during the learning process, i.e., along the trajectory of the differential equation given by the ND. Why is this important? This says that the sum of the squares of the weights can grow by at most $L(\boldsymbol{w}(t=0))$, the initial loss. In particular, the weights cannot grow to infinity. This observation is important for the analysis.

2. (6pts) Assume further that $L(\boldsymbol{w}) = f(\|\boldsymbol{w}\|)$. Show that the (order-two antisymmetric) tensor $A = \boldsymbol{w}\dot{\boldsymbol{w}}^T - \dot{\boldsymbol{w}}\boldsymbol{w}^T$ (with $\boldsymbol{w}$ viewed as a column vector) stays constant during the learning process.

HINT: The proof is VERY easy. You have a function of time and want to show that it is constant.

P.S.: If we do not think of NNs but mechanics then $\|\dot{\boldsymbol{w}}(t)\|^2$ is the kinetic energy and $L(\boldsymbol{w}(t))$ is the potential energy. The statement is then the usual conservation law of energy. The second case corresponds to the conservation of the rotational momentum in case the potential is radially symmetric.

P.P.S.: We derived this conservation law for the ND. But similar expressions can be derived for other dynamics, such as GF.

P.P.P.S: In NN there are many other symmetries that stem e.g. from symmetries of the data or the activiation functions. It can be shown that each such symmetry leads to a conserved quantity.

Solution:

1.

$$\begin{aligned}\frac{d}{dt}(\frac{1}{2}\|\dot{\boldsymbol{w}}(t)\|^2 + L(\boldsymbol{w}(t))) &= \langle \dot{\boldsymbol{w}}(t), \ddot{\boldsymbol{w}}(t)\rangle + \langle \nabla L(\boldsymbol{w}(t)), \dot{\boldsymbol{w}}(t)\rangle \\ &= \langle \underbrace{\ddot{\boldsymbol{w}}(t) + \nabla L(\boldsymbol{w}(t))}_{=0}, \dot{\boldsymbol{w}}(t)\rangle = 0\end{aligned}$$

2.

$$\begin{aligned}\frac{d}{dt}A &= \dot{\boldsymbol{w}}\dot{\boldsymbol{w}}^T + \boldsymbol{w}\ddot{\boldsymbol{w}}^T - \ddot{\boldsymbol{w}}\boldsymbol{w}^T - \dot{\boldsymbol{w}}\dot{\boldsymbol{w}}^T \\ &= -\boldsymbol{w}(\nabla f(\|\boldsymbol{w}\|))^T + \nabla f(\|\boldsymbol{w}\|)\boldsymbol{w}^T \\ &= -\frac{\boldsymbol{w}\boldsymbol{w}^T}{\|w\|}f'(\|\boldsymbol{w}\|) + \frac{\boldsymbol{w}\boldsymbol{w}^T}{\|w\|}f'(\|\boldsymbol{w}\|) = 0,\end{aligned}$$

where in the last line we use $\nabla\|\boldsymbol{w}\| = \frac{\boldsymbol{w}}{\|w\|}$ and so $\nabla f(\|\boldsymbol{w}\|) = \frac{\boldsymbol{w}}{\|w\|}f'(\|\boldsymbol{w}\|)$.

Problem 3. *Mixture of linear regressions* (15pts)

In this problem we guide you through the mixture of linear regressions model. Assume we are given data points $(x_i, y_i)_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. To each data point i is associated a hidden label $z_i \in \{1, \cdots, K\}$ with iid distribution $\mathbb{P}(z_i = t) = p_t$, $t = 1, \cdots, K$. We assume that the data points can be explained by a mixture of linear regressions:

$$y_i = \sum_{t=1}^{K} w_t^T \cdot x_i \, \mathbb{I}(z_i = t)$$

and $x_i \sim \mathcal{N}(0, I_d)$, with I_d the $d \times d$ identity matrix, and $w_t \in \mathbb{R}^d$ are linear regression slope vectors. This identity says that if data point i has label $z_i = t$ then the pair (x_i, y_i) satisfies $y_i = w_t^T x_i$, but note that the label z_i is a hidden random variable.

The model is noiseless for simplicity here (but one could also have additive gaussian noise). We assume throughout that $K \leq d$ and that the vectors w_t are linearly independent.

The goal is to learn the parameters of the model: $\{p_t, w_t\}_{t=1}^K$. We are *not* interested in learning the labels. We will guide you through a sequence of questions leading to an algorithm (proposed in the recent literature). Note that the first question requires a bit of algebra but if you want you can proceed to the second and third questions directly.

1. (5pts) We define the "moments" $m_0 = \frac{1}{n}\sum_{i=1}^n y_i^2$ and $M_2 = \frac{1}{2n}\sum_{i=1}^n y_i^2 x_i \otimes x_i - \frac{1}{2} m_0 I_d$. Prove:
$$\mathbb{E}[M_2] = \sum_{t=1}^{K} p_t w_t \otimes w_t$$
Hint: It is best to work with components of vectors denoted by upper-script indices, e.g., x_i^α, w_t^α with $\alpha = 1, \cdots, d$. We recall that for $x_i \sim \mathcal{N}(0, I_d)$ we have $\mathbb{E}(x_i^\alpha x_i^\beta) = \delta_{\alpha\beta}$ and $\mathbb{E}(x_i^\alpha x_i^\beta x_i^\gamma x_i^\delta) = \delta_{\alpha\beta}\delta_{\gamma\delta} + \delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma}$.

2. (5pts) Assume first that w_t's are orthonormal vectors.
   - (2.5pts) What is the condition on p_t's so that the vectors w_t can be uniquely determined, and how do you determine them ? (answer with a few words - no calculation).
   - (2.5pts) Further explain how to get an estimate of p_t and w_t from *real* observations $(x_i, y_i)_{i=1}^n$ when n becomes very large. You can assume that for n very large M_2 concentrates.

3. (5pts) From now on we do not assume anymore that the w_t's are orthogonal (we still assume they have unit norm).
   - (1 pt) Explain why we cannot uniquely determine w_t's if we only use M_2 ?

- (4 pt) We define $m_1 = \frac{1}{6n}\sum_{i=1}^n y_i^3 x_i$ and also

$$M_3 = \frac{1}{6n}\sum_{i=1}^n y_i^3 x_i \otimes x_i \otimes x_i - \sum_{\ell=1}^d (m_1 \otimes e_\ell \otimes e_\ell + e_\ell \otimes m_1 \otimes e_\ell + e_\ell \otimes e_\ell \otimes m_1)$$

where e_ℓ is the ℓ-th canonical basis vector of $\mathbb{R}^d$; in components $e_\ell^\alpha = \delta_{\alpha\ell}$. One can check with some algebra that (you are *not* asked to do so)

$$\mathbb{E}[M_3] = \sum_{t=1}^K p_t \, w_t \otimes w_t \otimes w_t$$

Given the two tensors $\mathbb{E}[M_2]$ and $\mathbb{E}[M_3]$ we have seen in class the theory combining the whitening procedure and tensor power method to determine the set of p_t and w_t's.

Write down a pseudocode (with a few comments) implementing an algorithm based on whitening and tensor power method that outputs the parameters of the mixture of regression model given real observations $(x_i, y_i)_{i=1}^n$.

Solution:

1. First we compute:

$$\mathbb{E}[m_0] = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{K}\mathbb{E}[y_i^2|z_i=t]\mathbb{P}(z_i=t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{K}\mathbb{E}[(w_t^T\cdot x_i)^2]p_t$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{K}p_t\sum_{\alpha,\beta=1}^{t}w_t^\alpha w_t^\beta\mathbb{E}[x_i^\alpha x_i^\beta] = \sum_{t=1}^{K}p_t\sum_{\alpha=1}^{t}(w_t^\alpha)^2 = \sum_{t=1}^{K}\|w_t\|^2 p_t$$

   Similarly:

$$\mathbb{E}[\frac{1}{2n}\sum_{i=1}^{n}y_i^2(x_i\otimes x_i)^{\gamma\delta}] = \frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{K}\mathbb{E}[y_i^2x_i^\gamma x_i^\delta|z_i=t]\mathbb{P}(z_i=t)$$
$$= \frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{K}\mathbb{E}[(w_t^T\cdot x_i)^2x_i^\gamma x_i^\delta]p_t = \frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{K}p_t\sum_{\alpha,\beta}\mathbb{E}[w_t^\alpha x_i^\alpha w_t^\beta x_i^\beta x_i^\gamma x_i^\delta]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{K}p_t\sum_{\alpha,\beta}w_t^\alpha w_t^\beta\mathbb{E}[w_t^\alpha x_i^\alpha w_t^\beta x_i^\beta x_i^\gamma x_i^\delta] = \frac{1}{2}\sum_{t=1}^{K}p_t\sum_{\alpha,\beta}w_t^\alpha w_t^\beta(\delta_{\alpha\beta}\delta_{\gamma\delta}+\delta_{\alpha\gamma}\delta_{\beta\delta}+\delta_{\alpha\delta}\delta_{\beta\gamma})$$
$$= \frac{1}{2}\sum_{t=1}^{K}p_t\sum_{\alpha}(w_t^\alpha)^2\delta_{\gamma\delta}+\frac{1}{2}\sum_{t=1}^{K}p_t(w_t^\gamma w_t^\delta+w_t^\delta w_t^\gamma) = \frac{1}{2}\mathbb{E}[m_0]I_d^{\gamma\delta}+\sum_{t=1}^{K}p_t(w_t\otimes w_t)^{\gamma\delta}$$

   These two results imply $\mathbb{E}[M_2] = \sum_t p_t w_t\otimes w_t = \sum_t p_t w_t w_t^T$

2. The $d\times d$ matrix $\mathbb{E}[M_2]$ is real symmetric. So if we know beforehand that w_t are orthonormal, they must be the eigenvectors, and p_t must be the eigenvalues. They are found by an SVD or diagonalization of the matrix. The condition of unicity is that all p_t's are distinct.

   In practice with *real* data one considers the empirical matrix $M_2^{\text{emp}} = \frac{1}{2n}\sum_{i=1}^{n}y_i^2(x_i\otimes x_i - I_d)$. This is a real symmetric $d\times d$ matrix and one can do an SVD (or diagonalization). We have $K\leq d$ by assumption. One will keep the top K eigenvalues and and eigenvectors as estimates of p_t and w_t's. For n large the other $d-K$ eigenvalues are close to zero (and we discard them) as one expects that M_2 concentrates on $\mathbb{E}[M_2]$.

3. If the w_t's are not orthogonal they do not constitute the set of eigenvectors. The decomposition of the matrix into rank one matrices is not unique because of the "rotation problem". More specifically for any $K\times K$ orthogonal matrix R we have that

$$\sum_{t=1}^{K}p_t w_t\otimes w_t = [\sqrt{p_1}w_1,\cdots,\sqrt{p_K}w_K][\sqrt{p_1}w_1,\cdots,\sqrt{p_K}w_K]^T$$
$$= [\sqrt{p_1}w_1,\cdots,\sqrt{p_K}w_K]RR^T[\sqrt{p_1}w_1,\cdots,\sqrt{p_K}w_K]^T$$
$$= [u_1,\cdots,u_K][u_1,\cdots,u_K]^T = \sum_t u_t\otimes u_t$$

where the new vectors are $u_t = \sum_{s=1}^K \sqrt{p_s} w_s R_{st}$.

4. The pseudocode takes input data $(x_i, y_i)_{i=1}^n$ and outputs $\{w_t, p_t\}_{t=1}^K$:

   (a) *Compute empirical moments*:
   $m_0 \leftarrow \frac{1}{n}\sum_{i=1}^n y_i^2$
   $m_1 \leftarrow \frac{1}{6n}\sum_{i=1}^n y_i^3 x_i$
   $M_2 \leftarrow \frac{1}{2n}\sum_{i=1}^n y_i^2 x_i \otimes x_i - \frac{1}{2} m_0 I_d$
   $M_3 \leftarrow \frac{1}{6n}\sum_{i=1}^n y_i^3 x_i \otimes x_i \otimes x_i - \sum_{\ell=1}^d (m_1 \otimes e_\ell \otimes e_\ell + e_\ell \otimes m_1 \otimes e_\ell + e_\ell \otimes e_\ell \otimes m_1)$.

   (b) *Compute the SVD of M_2*:
   $M_2 = UDU^T$ (as M_2 is real symmetric $D = \mathrm{diag}(d_1, \cdots, d_K)$ are the K non-zero eigenvalues and U is a $d \times K$ matrix with K orthonormal columns) and do $W \leftarrow UD^{-1/2} = U\mathrm{diag}(d_1^{-1/2}, \cdot, d_K^{-1/2})$.

   (c) *Whitening of M_3*:
   $T_3^{\alpha'\beta'\gamma'} \leftarrow \sum_{\alpha,\beta,\gamma} M_3^{\alpha\beta\gamma} W^{\alpha\alpha'} W^{\beta\beta'} W^{\gamma\gamma'}$.

   (d) *Tensor power method applied to T_3*:
   Input T_3. Output set $\{\lambda_t, v_t\}_{t=1,\cdots,K}$ such that $T_3 = \sum_{t=1}^K \lambda_t v_t \otimes v_t \otimes v_t$. The theory shows that v_t is orthonormal which is necessary for the tensor power iterations.
   For $s = 1, \cdots, K-1$ do:

      i. Initialize the iterations with a random vector $v^{(0)}$ and for $n = 0, \cdots, T$ (T large enough) do $v^{(n+1)} \leftarrow \frac{T_3(I, v^{(n)}, v^{(n)})}{\|T_3(I, v^{(n)}, v^{(n)})\|}$. Recall the notation: $T_3(I, v, v)^\alpha = \sum_{\beta,\gamma} T_3^{\alpha\beta\gamma} v^\beta v^\gamma$
      ii. Do $v_s \leftarrow w^{(T)}$ and $\lambda_s \leftarrow T_3(v^{(T)}, v^{(T)}, v^{(T)}) = \sum_{\alpha,\beta,\gamma} T_3^{\alpha\beta\gamma} v^{(T)\alpha} v^{(T)\beta} v^{(T)\gamma}$
      iii. Deflation step $T_3 \leftarrow T_3 - p_s v_s \otimes v_s \otimes v_s$.

   (e) *Undo the whitening*:
   For all $t = 1, \cdots, K$ do $\mu_t \leftarrow 1/\lambda_t^2$, $w_t = (1/\sqrt{p_t}) W v_t$.

Problem 4. *These are 4 short questions. Answer each point with a short justification or calculation.* [13pts]

1. (2 pts) Let $\vec{x}_i, \vec{y}_i \in \mathbb{R}^N$, $N \geq 2$, $i = 1, 2, 3$ be six N-dimensional real component vectors such that their components satisfy $x_i^\alpha = y_i^\alpha$ for $\alpha = 1, \cdots, N-1$ and $x_i^N \neq y_i^N$ (here the upper index labels the component of the vector). Consider the tensor $T = \vec{x}_1 \otimes \vec{x}_2 \otimes \vec{x}_3 + \vec{y}_1 \otimes \vec{y}_2 \otimes \vec{y}_3$.

   (a) (1pt) Does Jennrich's theorem apply ?

   (b) (1pt) The tensor-rank equals 2 ?

2. (4 pts) For two perpendicular (and non-zero) vectors $\vec{x}$, $\vec{y}$ in $\mathbb{R}^N$, $N \geq 2$, and for $t \geq 1$, let $S(t) = \vec{x} \otimes \vec{x} \otimes (\vec{x} - t\vec{y}) + (\vec{x} + \frac{1}{t}\vec{y}) \otimes (\vec{x} + \frac{1}{t}\vec{y}) \otimes ty$.

   (a) (1pt) The tensor-rank of $S(t)$ equals 2 for all fixed $t \geq 1$?

   (b) (3pt) What is the rank of the limiting tensor $\lim_{t \to +\infty} S(t)$?

3. (3 pts) Consider the function

$$f(x) = x^2 + 2.5 \cos x + |x|,$$

   defined on the real line $\mathbb{R}$. Which of the following statements is correct and why/why not? The function f is:

   (a) (1pt) convex

   (b) (1pt) differentiable everywhere

   (c) (1pt) subdifferentiable everywhere

4. (4pts) Consider some hypothesis class $\mathcal{H}$. Which of the following is true? Why or why not?

   (a) (1pt) If $|\mathcal{H}|$ is infinite, it is not PAC learnable.

   (b) (1pt) If $\mathcal{H}$ is PAC learnable, it has finite VC dimension.

   (c) (1pt) If $\mathcal{H}$ is specified by a finite number of parameters, it has finite VC dimension.

   (d) (1pt) If $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are some hypothesis classes that are PAC learnable, then $\mathcal{H}$ is also PAC learnable.

Solution:

1. (a) Yes, Jennrich's theorem applies because $[x_1, y_1]$, $[x_2, y_2]$, $[x_3, y_3]$ are $N \times 2$ full column rank matrices (i.e., with independent columns) as long as the first $N-1$ components do not vanish. Otherwise it doesnt apply.

   (b) Yes, according to Jennrich's theorem the tensor-rank is necessarily equal to 2 as long as the first $N-1$ components do not vanish. Otherwise the rank is not necessarily two.

2. (a) Yes the rank is 2 for all $t \geq 1$, because the matrices $[\vec{x}, \vec{x} + \frac{1}{t}\vec{y}]$, $[x - t\vec{y}, t\vec{y}]$ are full column rank as long as $t \neq 0$ and t fixed finite.

   (b) Developing $S(t) = \vec{x} \otimes \vec{x} \otimes \vec{x} + \vec{x} \otimes \vec{y} \otimes \vec{y} + \vec{y} \otimes \vec{x} \otimes \vec{y} + \frac{1}{t}\vec{y} \otimes \vec{y} \otimes \vec{y}$. Thus $\lim_{t \to +\infty} = \vec{x} \otimes \vec{x} \otimes x + \vec{x} \otimes \vec{y} \otimes \vec{y} + \vec{y} \otimes \vec{x} \otimes \vec{y}$. Note that Jennrich's theorem does not apply to the limiting tensor. Despite this its rank is 3. Indeed we show by contradiction that the rank cannot be 1 or 2.

   If the rank was 1 then we should have $\vec{x} \otimes \vec{x} \otimes \vec{x} + \vec{x} \otimes \vec{y} \otimes \vec{y} + \vec{y} \otimes \vec{x} \otimes \vec{y} = \vec{a} \otimes \vec{b} \otimes \vec{c}$. But then by orthogonality of $\vec{x}$ and $\vec{y}$ we would have $\|x\|^2 \vec{x} \otimes x + \|x\|^2 \vec{y} \otimes y = (\vec{x}^T \cdot \vec{a})\vec{b} \otimes \vec{c}$ which is not possible because the l.h.s is rank 2 and the r.h.s rank 1 (or nul).

   If the rank was 2 then we would have $\vec{x} \otimes \vec{x} \otimes \vec{x} + \vec{x} \otimes \vec{y} \otimes \vec{y} + \vec{y} \otimes \vec{x} \otimes \vec{y} = \vec{a} \otimes \vec{b} \otimes \vec{c} + \vec{e} \otimes \vec{f} \otimes \vec{g}$ with $[\vec{a}, \vec{e}]$, $[\vec{b}, \vec{f}]$, $[\vec{c}, \vec{g}]$ with independent columns. Viewing both sides of the equation as multilinear transforms on $(\vec{y}, I, I)$ we find $\|\vec{y}\|^2 \vec{x} \otimes \vec{y} = (\vec{y}^T \cdot \vec{a})\vec{b} \otimes \vec{c} + (\vec{y}^T \cdot \vec{e})\vec{f} \otimes \vec{g}$ which means $(\vec{y}^T \cdot \vec{a}) = 0$ or $(\vec{y}^T \cdot \vec{e}) = 0$ (because otherwise the r.h.s is rank 2). Suppose w.l.o.g that $(\vec{y}^T \cdot \vec{e}) = 0$. Then $\vec{b} \sim \vec{x}$, $\vec{c} \sim \vec{y}$. We get $\vec{x} \otimes \vec{x} \otimes \vec{x} + \vec{x} \otimes \vec{y} \otimes \vec{y} + \vec{y} \otimes \vec{x} \otimes \vec{y} = \lambda \vec{a} \otimes \vec{x} \otimes \vec{y} + \vec{e} \otimes \vec{f} \otimes \vec{g}$. Applying $(I, I, \vec{x})$ we get $\|\vec{x}\|^2 \vec{x} \otimes \vec{x} = \vec{e} \otimes \vec{f}(\vec{g}^T \cdot \vec{x})$ and thus $\vec{e} \sim \vec{x}$, $\vec{f} \sim \vec{x}$. We get $\vec{x} \otimes \vec{x} \otimes \vec{x} + \vec{x} \otimes \vec{y} \otimes \vec{y} + \vec{y} \otimes \vec{x} \otimes \vec{y} = \lambda \vec{a} \otimes \vec{x} \otimes \vec{y} + \mu \vec{x} \otimes \vec{x} \otimes \vec{g}$. Applying $(I, \vec{y}, I)$ we get $\|\vec{y}\|^2 \vec{x} \otimes \vec{y} = 0$ which is a contradiction.

3. (a) For $x \geq 0$, $f''(x) = 2 - 2.5\cos(x)$, which is negative for some values of $x \geq 0$. Hence the function is not convex.

   (b) f is not differentiable at $x = 0$ due to the term $|x|$.

   (c) This is a little tricky. The function has a derivative everywhere except at 0 where it has a subderivative. But it is not convex and hence not subdifferentiable everywhere. :-(We gave you a point either way. :-)

4. (a) False. If $\mathcal{H}$ has finite VC dimension then it is PAC learnable due to the Fundamental theorem of Statistical learning.

   (b) True. According to the Fundamental theorem of Statistical learning.

   (c) False. We saw in the homework that there are hypotheses classes with infinite VC dimension that are specified by a single parameter.

(d) True. If $\mathcal{H}_1, \mathcal{H}_2$ have finite VC dimension then the VC dimension of their union is also finite and therefore $\mathcal{H}$ is also PAC learnable.