ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Spring 2026	Assignment date: June 23, 2026, 09:15
Learning Theory	Due date: June 23, 2026, 12:15

Final Exam – CS 526 – Room INM 10

Use scratch paper if needed to figure out the solution. 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!

SCIPER No.: _______________

Problem 1	/ 10
Problem 2	/ 20
Problem 3	/ 20
Problem 4	/ 10
Total	/60

Problem 1 (10 pts). **Quick questions**

These questions have quick answers. You can directly use the relevant theorems or results seen in class without proof but you must clearly state your justifications.

1. **VC dimension I.** Let $\mathcal{H} = \{h_r(x) = \mathbf{1}[\|x\| \leq r] : r \geq 0\}$ be the class of classifiers given by Euclidean balls centered at the origin in $\mathbb{R}^d$. Determine the VC dimension of $\mathcal{H}$.

2. **VC dimension II.** Let $\mathcal{H}$ be a binary hypothesis class with exactly 8 hypotheses: $|\mathcal{H}| = 8$. Explain why $\mathcal{H}$ cannot shatter any set of 4 points. Give also an upper bound on the VC dimension of $\mathcal{H}$.

3. **Convexity.** Consider the function $f(x) = \frac{2x^2}{3} + \frac{|x|}{3}$ defined on the real line $\mathbb{R}$. Answer the following questions:

   - is f convex ?
   - is f differentiable everywhere ?
   - is f subdifferentiable everywhere ?
   - If you answered "yes" to the last question, then give its subdifferential set for each $x \in \mathbb{R}$.
   - Let $\omega \in 0, 1$ a Bernoulli random variable such that $\mathbb{P}(0) = \frac{2}{3}$ and $\mathbb{P}(1) = \frac{1}{3}$. Define the random function g_ω such that $g_0(x) = 2x$ and $g_1(x) = H(x) - H(-x)$ where $H(x) = 1$ for $x > 0$, $H(x) = 0$ for $x < 0$, and $H(0) = 0$.
     Does g_ω qualify for a stochastic gradient of f ?

4. **Tensors.** Let

$$T = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix}.$$

   Is this decomposition unique? What is the rank of T? Justify your answers.

Solution to Problem 1.

1. **VC dimension I. 2pt.** A single point $x \neq 0$ is shattered since choosing $r < \|x\|$ labels it 0 and choosing $r \geq \|x\|$ labels it 1.

   If x_1, x_2 satisfy $\|x_1\| < \|x_2\|$, then the labeling $(0, 1)$ cannot be realized, because every ball containing x_2 also contains x_1. Hence $\mathrm{VCdim}(\mathcal{H}) = 1$.

2. **VC dimension II. 2pt.** If $\mathcal{H}$ shatters a set of 4 points, it must realize all $2^4 = 16$ possible binary labelings of these points. A class with only 8 hypotheses can realize at most 8 labelings. Therefore no set of 4 points can be shattered. The argument also implies that $\mathrm{VCdim}(\mathcal{H}) \leq 3$.

3. **Convexity. 3pt.** The function is the sum of two convex functions so its convex.

   It is not differentiable at $x = 0$.

   But it has a subdifferential everywhere.

   At $x = 0$ this is the set $\{-1/3, +1/3\}$; for $x > 0$ it is $\{4x/3 + 1/3\}$; for for $x < 0$ it is $\{4x/3 - 1/3\}$.

   Finally yes g_ω is a stochastic gradient since its average is $\frac{1}{3}g_0 + \frac{2}{3}g_1$ which is equal to a subgradient (the one where we use 0 for the subgradient at 0.

4. **Tensors. 3pt.** We have

$$T = \vec{a}_1 \otimes \vec{b}_1 \otimes \vec{c}_1 + \vec{a}_2 \otimes \vec{b}_2 \otimes \vec{c}_2 + \vec{a}_3 \otimes \vec{b}_3 \otimes \vec{c}_3,$$

   where

$$\begin{bmatrix}\vec{a}_1 \ \vec{a}_2 \ \vec{a}_3\end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 3 \\ 1 & 2 & 5 \end{bmatrix}$$

   has pairwise independent columns,

$$\begin{bmatrix}\vec{b}_1 \ \vec{b}_2 \ \vec{b}_3\end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

   has linearly independent columns, and

$$\begin{bmatrix}\vec{c}_1 \ \vec{c}_2 \ \vec{c}_3\end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 3 \\ 0 & 1 & 0 \end{bmatrix}$$

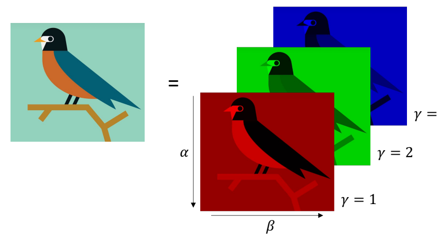
has linearly independent columns.

Hence by Jenrich's theorem the decomposition is unique and

$$\operatorname{rank}(T) = 3.$$

Problem 2 (20 pts)**.** *Tucker Decomposition for RGB Image Compression*

Let $\mathcal{X} \in \mathbb{R}^{m\times n\times 3}$ represent an RGB image admitting the Tucker representation



$$\mathcal{X}^{\alpha\beta\gamma} = \sum_{p=1}^{r_1}\sum_{q=1}^{r_2}\sum_{r=1}^{r_3} \mathcal{G}^{pqr}\, A^{\alpha p}\, B^{\beta q}\, C^{\gamma r},$$

where

$$A \in \mathbb{R}^{m\times r_1}, \qquad B \in \mathbb{R}^{n\times r_2}, \qquad C \in \mathbb{R}^{3\times r_3},$$

have **orthonormal columns** (i.e. $A^T A = I_{r_1}$, $B^T B = I_{r_2}$, $C^T C = I_{r_3}$).

We introduce the following notation. For a tensor $\mathcal{T}$ and matrices M_1, M_2, M_3 of compatible sizes, define the *multilinear multiplication* $\mathcal{T}(M_1, M_2, M_3)$ as the tensor with entries

$$\mathcal{T}(M_1, M_2, M_3)^{\delta\varepsilon\eta} = \sum_{\alpha,\beta,\gamma} \mathcal{T}^{\alpha\beta\gamma}\, M_1^{\delta\alpha}\, M_2^{\varepsilon\beta}\, M_3^{\eta\gamma}.$$

For any core $\mathcal{Y} \in \mathbb{R}^{r_1\times r_2\times r_3}$, we also write

$$\mathcal{T}_{\mathcal{Y}}^{\alpha\beta\gamma} = \sum_{p=1}^{r_1}\sum_{q=1}^{r_2}\sum_{r=1}^{r_3} \mathcal{Y}^{pqr}\, A^{\alpha p}\, B^{\beta q}\, C^{\gamma r},$$

so that in particular $\mathcal{X} = \mathcal{T}_{\mathcal{G}}$.

1. Consider an image of size $1024 \times 1024 \times 3$ and a Tucker approximation of multilinear rank $(r_1, r_2, r_3) = (20, 20, 3)$.

a) Count the number of real scalars needed to store the original image and the Tucker factors A, B, C together with the core $\mathcal{G}$.

b) Deduce the compression ratio.

2. Suppose $r_3 = 3$. Show that one can always reparametrize the decomposition so that $C = I_3$, without changing $\mathcal{X}$.

Hint: absorb C into the core by defining a new core $\widetilde{\mathcal{G}}$.

3. From now on, assume that $C = I_3$ (and hence $r_3 = 3$). Define the linear subspace of $\mathbb{R}^{m\times n\times 3}$

$$S = \{\, \mathcal{T}_{\mathcal{Y}} : \mathcal{Y} \in \mathbb{R}^{r_1\times r_2\times 3} \,\},$$

and set $\widehat{\mathcal{X}} = \mathcal{X}\big(A^T, B^T, I_3\big) \in \mathbb{R}^{r_1\times r_2\times 3}$.

Show that $\mathcal{T}_{\widehat{\mathcal{X}}}$ is the *orthogonal projection* of $\mathcal{X}$ onto S; that is, verify that

$$\langle \mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}},\, \mathcal{Z}\rangle_F = 0 \qquad \forall\, \mathcal{Z} \in S,$$

where the Frobenius inner product is $\langle \mathcal{T}, \mathcal{Z} \rangle_F = \sum_{\alpha,\beta,\gamma} \mathcal{T}^{\alpha\beta\gamma} \mathcal{Z}^{\alpha\beta\gamma}$.

Hint : Show that it suffices that for every p, q, γ, $\quad \sum_{\alpha,\beta} \big(\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\big)^{\alpha\beta\gamma} A^{\alpha p}\, B^{\beta q} = 0.$

4. Consider the least-squares problem

$$\min_{\mathcal{Y} \in \mathbb{R}^{r_1 \times r_2 \times 3}} \big\| \mathcal{X} - \mathcal{T}_{\mathcal{Y}} \big\|_F^2.$$

a) Show that the minimum is achieved for $\mathcal{Y} = \widehat{\mathcal{X}}$.

b) Deduce that $\min_{\mathcal{Y}} \big\| \mathcal{X} - \mathcal{T}_{\mathcal{Y}} \big\|_F^2 = \| \mathcal{X} \|_F^2 - \| \widehat{\mathcal{X}} \|_F^2$

c) Interpret this result in the context of image compression.

Hint : Justify that you can use the Pythagorean identity $\| \mathcal{X} \|_F^2 \; = \; \| \mathcal{T}_{\widehat{\mathcal{X}}} \|_F^2 \; + \; \| \mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}} \|_F^2$.

Solution to Problem 2.

1. **(a) Parameter counts. 3pt.**

   Original image. The image $\mathcal{X} \in \mathbb{R}^{1024\times1024\times3}$ requires

   $$1024 \times 1024 \times 3 = 3{,}145{,}728 \text{ scalars.}$$

   Tucker decomposition. The four objects to store are:

   - Core $\mathcal{G} \in \mathbb{R}^{20\times20\times3}$: $20 \times 20 \times 3 = 1{,}200$ scalars.
   - Factor $A \in \mathbb{R}^{1024\times20}$: $1024 \times 20 = 20{,}480$ scalars.
   - Factor $B \in \mathbb{R}^{1024\times20}$: $1024 \times 20 = 20{,}480$ scalars.
   - Factor $C \in \mathbb{R}^{3\times3}$: $3 \times 3 = 9$ scalars.

   Total: $1{,}200 + 20{,}480 + 20{,}480 + 9 = 42{,}169$ scalars.

   (b) Compression ratio. 1pt.

   $$\text{Compression ratio} = \frac{3{,}145{,}728}{42{,}169} \approx 74.6.$$

   The Tucker decomposition requires roughly 75 times fewer parameters than the original image.

2. **4pt.**

   Since $r_3 = 3$ and $C \in \mathbb{R}^{3\times3}$ has orthonormal columns, C is a square orthogonal matrix satisfying $CC^T = C^TC = I_3$. Define the new core $\widetilde{\mathcal{G}} \in \mathbb{R}^{r_1\times r_2\times3}$ by

   $$\widetilde{\mathcal{G}}^{pq\gamma} = \sum_{r=1}^{3} \mathcal{G}^{pqr}\, C^{\gamma r}, \qquad \text{i.e.} \quad \widetilde{\mathcal{G}} = \mathcal{G}(I_{r_1},\, I_{r_2},\, C).$$

   Then

   $$\mathcal{X}^{\alpha\beta\gamma} = \sum_{p,q,r} \mathcal{G}^{pqr}\, A^{\alpha p}\, B^{\beta q}\, C^{\gamma r} = \sum_{p,q} \widetilde{\mathcal{G}}^{pq\gamma}\, A^{\alpha p}\, B^{\beta q},$$

   which is exactly the Tucker representation with color factor I_3. Hence $\mathcal{X}$ is unchanged and we may always set $C = I_3$.

3. **6pt.**

   Take any $\mathcal{Z} \in S$, so $\mathcal{Z} = \mathcal{T}_{\mathcal{Y}}$ for some $\mathcal{Y} \in \mathbb{R}^{r_1\times r_2\times3}$, i.e.

   $$\mathcal{Z}^{\alpha\beta\gamma} = \sum_{p,q} \mathcal{Y}^{pq\gamma}\, A^{\alpha p}\, B^{\beta q}.$$

Following the hint, we first show that for every p_0, q_0, γ_0,

$$\sum_{\alpha,\beta} \big(\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\big)^{\alpha\beta\gamma_0} A^{\alpha p_0}\, B^{\beta q_0} = 0. \tag{1}$$

By definition of $\widehat{\mathcal{X}} = \mathcal{X}(A^T, B^T, I_3)$, the first term gives $\sum_{\alpha,\beta} \mathcal{X}^{\alpha\beta\gamma_0} A^{\alpha p_0} B^{\beta q_0} = \widehat{\mathcal{X}}^{p_0 q_0 \gamma_0}$. For the term involving $\mathcal{T}_{\widehat{\mathcal{X}}}$,

$$\begin{aligned}
\sum_{\alpha,\beta} \widehat{\mathcal{T}}^{\alpha\beta\gamma_0}\, A^{\alpha p_0}\, B^{\beta q_0} &= \sum_{\alpha,\beta} \sum_{p,q} \widehat{\mathcal{X}}^{pq\gamma_0}\, A^{\alpha p}\, B^{\beta q}\, A^{\alpha p_0}\, B^{\beta q_0} \\
&= \sum_{p,q} \widehat{\mathcal{X}}^{pq\gamma_0} \Big(\sum_{\alpha} A^{\alpha p}\, A^{\alpha p_0}\Big)\Big(\sum_{\beta} B^{\beta q}\, B^{\beta q_0}\Big) \\
&= \sum_{p,q} \widehat{\mathcal{X}}^{pq\gamma_0}\, (A^T A)_{pp_0}\, (B^T B)_{qq_0} = \widehat{\mathcal{X}}^{p_0 q_0 \gamma_0},
\end{aligned}$$

using $A^T A = I_{r_1}$ and $B^T B = I_{r_2}$. Subtracting confirms (1). Substituting the definition of $\mathcal{Z}$,

$$\begin{aligned}
\langle \mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}},\, \mathcal{Z} \rangle_F &= \sum_{\alpha,\beta,\gamma} \big(\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\big)^{\alpha\beta\gamma} \sum_{p,q} \mathcal{Y}^{pq\gamma}\, A^{\alpha p}\, B^{\beta q} \\
&= \sum_{p,q,\gamma} \mathcal{Y}^{pq\gamma} \underbrace{\sum_{\alpha,\beta} \big(\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\big)^{\alpha\beta\gamma} A^{\alpha p}\, B^{\beta q}}_{=\,0 \text{ by } (1)} \;=\; 0.
\end{aligned}$$

Hence $\mathcal{T}_{\widehat{\mathcal{X}}}$ is the orthogonal projection of $\mathcal{X}$ onto S.

4. **a) $\widehat{\mathcal{X}}$ is the minimizer. 2pt.** For any $\mathcal{Y} \in \mathbb{R}^{r_1 \times r_2 \times 3}$, decompose

$$\mathcal{X} - \mathcal{T}_{\mathcal{Y}} = \underbrace{\big(\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\big)}_{\perp S} \;-\; \underbrace{\big(\mathcal{T}_{\mathcal{Y}} - \mathcal{T}_{\widehat{\mathcal{X}}}\big)}_{\in S}.$$

Since $\mathcal{T}_{\mathcal{Y}} - \widehat{\mathcal{T}} \in S$, it is orthogonal to $\mathcal{X} - \widehat{\mathcal{T}}$ by Question 3, so the Pythagorean theorem gives

$$\|\mathcal{X} - \mathcal{T}_{\mathcal{Y}}\|_F^2 = \|\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2 + \|\mathcal{T}_{\mathcal{Y}} - \mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2 \;\geq\; \|\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2,$$

with equality if and only if $\mathcal{T}_{\mathcal{Y}} = \mathcal{T}_{\widehat{\mathcal{X}}}$, i.e. $\mathcal{Y} = \widehat{\mathcal{X}}$. Hence $\min_{\mathcal{Y}} \|\mathcal{X} - \mathcal{T}_{\mathcal{Y}}\|_F^2 = \|\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2$.

b) Minimal loss. 3pt. Like in a) we can deduce the Pythagorean identity. Since

$$\mathcal{X} = \underbrace{\big(\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\big)}_{\perp S} \;+\; \underbrace{\mathcal{T}_{\widehat{\mathcal{X}}}}_{\in S}.$$

we obtain directly

$$\|\mathcal{X} - \mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2 = \|\mathcal{X}\|_F^2 - \|\mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2.$$

Finally, $\|\mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2 = \|\widehat{\mathcal{X}}\|_F^2$. Since A and B have orthonormal columns, the map $\mathcal{Y} \mapsto \mathcal{T}_{\mathcal{Y}}$ is an isometry. Indeed, expanding the square,

$$\begin{aligned}
\|\mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2 &= \sum_{\alpha,\beta,\gamma}\Big(\sum_{p,q}\widehat{\mathcal{X}}^{pq\gamma}\,A^{\alpha p}\,B^{\beta q}\Big)^2 = \sum_{p,p',q,q',\gamma}\widehat{\mathcal{X}}^{pq\gamma}\,\widehat{\mathcal{X}}^{p'q'\gamma}\Big(\sum_{\alpha}A^{\alpha p}\,A^{\alpha p'}\Big)\Big(\sum_{\beta}B^{\beta q}\,B^{\beta q'}\Big)\\
&= \sum_{p,p',q,q',\gamma}\widehat{\mathcal{X}}^{pq\gamma}\,\widehat{\mathcal{X}}^{p'q'\gamma}\,\delta_{pp'}\,\delta_{qq'} = \sum_{p,q,\gamma}\big(\widehat{\mathcal{X}}^{pq\gamma}\big)^2 = \|\widehat{\mathcal{X}}\|_F^2,
\end{aligned}$$

using $(A^T A)_{pp'} = \delta_{pp'}$ and $(B^T B)_{qq'} = \delta_{qq'}$.

Combining:

$$\min_{\mathcal{Y}}\|\mathcal{X} - \mathcal{T}_{\mathcal{Y}}\|_F^2 = \|\mathcal{X}\|_F^2 - \|\mathcal{T}_{\widehat{\mathcal{X}}}\|_F^2 = \|\mathcal{X}\|_F^2 - \|\widehat{\mathcal{X}}\|_F^2.$$

c) Interpretation. 1pt. For fixed A and B, the core $\widehat{\mathcal{X}}$ give the best approximation of the original picture. Choosing A and B to maximize $\|\widehat{\mathcal{X}}\|_F^2$ therefore minimizes the distortion, which is the goal of optimal image compression.

Problem 3 (20 pts)**.** Stochastic Gradient Descent

Consider a dataset consisting of n given pairs $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}$ represent the features and $y_i \in \mathbb{R}$ represent the labels. We fit a linear model without intercept: $y = wx$.

We define the mean squared error objective function $f(w)$ as:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(wx_i - y_i)^2$$

1. Compute w^* the unique global minimum of $f(w)$. Justify why it is unique. Write the standard Gradient Descent (GD) update rule with step size $\eta > 0$.

2. Now we consider Stochastic Gradient Descent. We initialize at w_1 (given) and at each step $k \geq 1$ we sample a fresh index $i_k \in \{1, \ldots, n\}$ uniformly at random (with replacement).

(a) Justify why the SGD update step can be written as:

$$w_{k+1} = w_k - \eta(w_k x_{i_k}^2 - x_{i_k} y_{i_k}), \qquad k \geq 1$$

(b) Define the error $e_k = w_k - w^*$, $k \geq 1$. Prove that the expected error evolves as $\mathbb{E}[e_{k+1}] = (1 - \eta\left(\frac{1}{n}\sum x_i^2\right))^k e_1$, where the expectation is over the past history $i_k, i_{k-1}, \cdots, i_1$.

Hint: Compute first the conditional expectation $\mathbb{E}_{i_k}[e_{k+1} \mid i_{k-1}, \cdots, i_1]$*. You can use* $\mathbb{E}_{i_k}[e_{k+1}]$ *as a short hand notation for this conditional expectation.*

(c) Give a condition on $\eta > 0$ to guarantee the convergence of $\mathbb{E}[e_k] \longrightarrow 0$ as $k \to +\infty$.

3. Prove that the expected squared error satisfies

$$\mathbb{E}_{i_k}[e_{k+1}^2] = \frac{1}{n}\sum_{i=1}^{n}\left(e_k(1 - \eta x_i^2) + \eta(x_i y_i - w^* x_i^2)\right)^2$$

4. From the previous result deduce the steady state behavior $\lim_{k\to+\infty} \mathbb{E}[e_k^2]$ where the expectation is over the whole (past) history.

Solution to Problem 3 (total 20 pts)**.**

1. Taking the derivative of $f(w)$ with respect to w:

$$f'(w) = \frac{1}{n}\sum_{i=1}^{n} x_i(wx_i - y_i) = w\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

   We get $f'(w) = w - w^*$ which vanishes at

$$w^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

   which is the global minimum. The second derivative is equal to $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ and is strictly positive. Thus the function is strictly convex, hence there exist only a unique global minimum.

   The GD update rule is:

$$w_{k+1} = w_k - \eta(w_k - w^*)\left(\frac{1}{n}\sum x_i^2\right) \implies e_{k+1} = \left(1 - \eta\left(\frac{1}{n}\sum x_i^2\right)\right)e_k$$

   where $e_k = w_k - w^*$and this converges to zero for $\left|1 - \eta\frac{1}{n}\sum x_i^2\right| < 1$, which yields the step size in the range $0 < \eta < \frac{2n}{\sum_i x_i^2}$.

2. The gradient of a single loss term $f_{i_k}(w)$ is $f'_{i_k}(w) = x_{i_k}(wx_{i_k} - y_{i_k}) = wx_{i_k}^2 - x_{i_k}y_{i_k}$. The average over a uniform distribution of indices i_k is exactly the full gradient found in the previous question. This justifies the SGD update equation.

   Substituting $w_k = e_k + w^*$ into the SGD equation:

$$e_{k+1} + w^* = e_k + w^* - \eta\left((e_k + w^*)x_{i_k}^2 - x_{i_k}y_{i_k}\right)$$

$$e_{k+1} = e_k(1 - \eta x_{i_k}^2) - \eta(w^* x_{i_k}^2 - x_{i_k}y_{i_k})$$

   We take the conditional expectation $\mathbb{E}_{i_k}[\cdot \mid e_k]$ and note that e_k depends only on the past $i_{k-1}, \cdots, i_1$. Since the index i_k is chosen uniformly at random, so we have $\mathbb{E}_{i_k}[x_{i_k}^2] = \frac{1}{n}\sum x_i^2$ and $\mathbb{E}_{i_k}[x_{i_k}y_{i_k}] = \frac{1}{n}\sum x_i y_i$. Thus $\mathbb{E}_{i_k}[w^* x_{i_k}^2 - x_{i_k}y_{i_k}] = 0$ (by the formula found above for w^*). Thus

$$\mathbb{E}_{i_k}[e_{k+1} \mid e_k] = e_k(1 - \eta\frac{1}{n}\sum x_i^2)$$

   and hence for the total expectation we get $\mathbb{E}[e_{k+1}] = (1 - \eta\frac{1}{n}\sum x_i^2)\mathbb{E}[e_k]$.

3. Let $A_k = 1 - \eta x_{i_k}^2$ and $B_k = x_{i_k}y_{i_k} - w^* x_{i_k}^2$. The update equation is $e_{k+1} = e_k A_k + \eta B_k$. Squaring both sides yields:

$$e_{k+1}^2 = e_k^2 A_k^2 + 2\eta e_k A_k B_k + \eta^2 B_k^2$$

   Taking the condition expectation first:

- e_k is independent of index i_k, so $\mathbb{E}_{i_k}[e_k^2 A_k^2] = e_k^2 \mathbb{E}_{i_k}[A_k^2] = e_k^2 \frac{1}{n} \sum_{i=1}^n (1 - \eta x_i^2)^2$.
- The cross-term satisfies $\mathbb{E}_{i_k}[e_k A_k B_k] = e_k \frac{1}{n} \sum_{i=1}^n (1 - \eta x_i)^2 (x_i y_i - w^* x_i^2)$.
- The last term is $\mathbb{E}_{i_k}[B_k^2] = \frac{1}{n} \sum_{i=1}^n (x_i y_i - w^* x_i^2)^2$.

This yields:

$$\begin{aligned}
\mathbb{E}_{i_k}[e_{k+1}^2] &= e_k^2 \frac{1}{n} \sum_{i=1}^n (1 - \eta x_i^2)^2 + 2\eta e_k \frac{1}{n} \sum_{i=1}^n (1 - \eta x_i)^2 (x_i y_i - w^* x_i^2) + \eta^2 \frac{1}{n} \sum_{i=1}^n (x_i y_i - w^* x_i^2)^2 \\
&= \frac{1}{n} \sum_{i=1}^n \big(e_k (1 - \eta x_i^2) + \eta (x_i y_i - w^* x_i^2) \big)^2
\end{aligned}$$

4. From the expression with the square expanded and noting that $\mathbb{E}[e_k] \to 0$ as $k \to +\infty$,

$$\lim_{k \to +\infty} \mathbb{E}[e_k^2] \big(1 - \frac{1}{n} \sum_{i=1}^n (1 - \eta x_i^2)^2 \big) = \eta^2 \frac{1}{n} \sum_{i=1}^n (x_i y_i - w^* x_i^2)^2$$

which leads to

$$\begin{aligned}
\lim_{k \to +\infty} \mathbb{E}[e_k^2] &= \frac{\eta^2 \frac{1}{n} \sum_{i=1}^n (x_i y_i - w^* x_i^2)^2}{\big(1 - \frac{1}{n} \sum_{i=1}^n (1 - \eta x_i^2)^2 \big)} \\
&= \frac{\eta}{2} \frac{\sum_{i=1}^n (x_i y_i - w^* x_i^2)^2}{\sum_{i=1}^n x_i^2 (1 - \eta x_i^2)}
\end{aligned}$$

Problem 4 (10 pts)**.** Multiclass Shattering

Definition 1 (Shattering Multiclass Version)**.** We say that a set $C \subseteq \mathcal{X}$ is *N-shattered* by $\mathcal{H}$ if there exist two functions $f_0, f_1 : C \to [k]$ such that

- For every $x \in C,\ f_0(x) \neq f_1(x)$.
- For every $B \subseteq C$, there exists a function $h \in \mathcal{H}$ such that

$$\forall x \in B,\ h(x) = f_0(x) \qquad \text{and} \qquad \forall x \in C \setminus B,\ h(x) = f_1(x).$$

Definition 2 (Natarajan Dimension)**.** The *Natarajan dimension* of $\mathcal{H}$, denoted $\mathrm{Ndim}(\mathcal{H})$, is the maximal size of a shattered set $C \subseteq \mathcal{X}$.

Consider the binary classification case $k = 2$:

1. Show that every VC-shattered set is N-shattered.

2. Show that every N-shattered set is VC-shattered.
Hint: Fix an arbitrary labeling of C and show that there exists an hypothesis $h \in \mathcal{H}$ that realize this labeling.

3. Deduce that, in the binary classification case,

$$\mathrm{Ndim}(\mathcal{H}) = \mathrm{VCdim}(\mathcal{H}).$$

Solution to Problem 4.

1. **Every VC-shattered set is N-shattered. 4pt.**

   Let $C \subseteq \mathcal{X}$ be VC-shattered by $\mathcal{H}$. Define two functions

   $$f_0(x) = 1, \qquad f_1(x) = 0, \qquad x \in C.$$

   Clearly,

   $$f_0(x) \neq f_1(x)$$

   for every $x \in C$.

   Let $B \subseteq C$ be arbitrary. Since C is VC-shattered, there exists a hypothesis $h \in \mathcal{H}$ realizing the labeling

   $$h(x) = 1 \quad \forall x \in B, \qquad h(x) = 0 \quad \forall x \in C \setminus B.$$

   Equivalently,

   $$h(x) = f_0(x) \quad \forall x \in B, \qquad h(x) = f_1(x) \quad \forall x \in C \setminus B.$$

   Therefore the conditions of N-shattering are satisfied, and hence C is N-shattered.

2. **Every N-shattered set is VC-shattered. 4pt.**

   Let $C \subseteq \mathcal{X}$ be N-shattered. Then there exist functions

   $$f_0, f_1 : C \to \{0, 1\}$$

   such that

   (a) $f_0(x) \neq f_1(x)$ for every $x \in C$;

   (b) for every $B \subseteq C$, there exists $h \in \mathcal{H}$ such that

   $$h(x) = f_0(x) \quad \forall x \in B, \qquad h(x) = f_1(x) \quad \forall x \in C \setminus B.$$

   Since there are only two labels, the condition $f_0(x) \neq f_1(x)$ implies that

   $$\{f_0(x), f_1(x)\} = \{0, 1\}$$

   for every $x \in C$.

   Let $g : C \to \{0, 1\}$ be an arbitrary binary labeling of C. Define

   $$B = \{x \in C : \ g(x) = f_0(x)\}.$$

   By N-shattering, there exists $h \in \mathcal{H}$ such that

   $$h(x) = f_0(x) \quad \forall x \in B, \qquad h(x) = f_1(x) \quad \forall x \in C \setminus B.$$

If $x \in B$, then $h(x) = f_0(x) = g(x)$. If $x \notin B$, then $g(x) \neq f_0(x)$; since the only labels are 0 and 1, it follows that

$$g(x) = f_1(x) = h(x).$$

Hence $h(x) = g(x)$ for every $x \in C$.

Since every binary labeling g of C is realized by some hypothesis in $\mathcal{H}$, the set C is VC-shattered.

3. **Equality of the dimensions. 2pt.**

   Part (1) shows that every VC-shattered set is N-shattered. Therefore

   $$\mathrm{VCdim}(\mathcal{H}) \leq \mathrm{Ndim}(\mathcal{H}).$$

   Part (2) shows that every N-shattered set is VC-shattered. Therefore

   $$\mathrm{Ndim}(\mathcal{H}) \leq \mathrm{VCdim}(\mathcal{H}).$$

   Combining the two inequalities yields

   $$\boxed{\mathrm{Ndim}(\mathcal{H}) = \mathrm{VCdim}(\mathcal{H}).}$$