Machine Learning and Advanced Topics (Lecture Notes)

University of Copenhagen

Yijun Bian

SID: ??

Courses: ML, ATML

Department of Computer Science
University of Copenhagen (DIKU)

Postdoc Training
Prof. Yevgeny Seldin

23. November 2021

Contents

1 Preliminaries

1.1 Self-Assessment Assignment for the Machine Learning course

Needed:

- basic linear algebra
- calculus
- probability theory

1.1.1 Vectors and matrices

Problem 1 (Vectors and matrices). *Consider the two vectors* $\mathbf{a} = (1,2,2)^\mathsf{T}$ *and* $\mathbf{b} = (3,2,1)^\mathsf{T}$ *and the matrix*

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{1.1}$$

A 1–3: No, the inner product $\mathbf{a}^\mathsf{T}\mathbf{b}$ (i.e., scalar product, dot product) and the outer product $\mathbf{a}\mathbf{b}^\mathsf{T}$, they are not equal.

$$\langle \mathbf{a}, \mathbf{b} \rangle \overset{\text{def}}{=} \mathbf{a}^\mathsf{T}\mathbf{b} \overset{\text{def}}{=} \mathbf{a} \cdot \mathbf{b} = 1 \times 3 + 2 \times 2 + 2 \times 1 = 3 + 2 + 2 = 7\,, \tag{1.2a}$$

$$\|\mathbf{a}\| = \sqrt{1^2 + 2^2 + 2^2} = \sqrt{1 + 4 + 4} = \sqrt{9} = 3\,, \tag{1.2b}$$

$$\|\mathbf{b}\| = \sqrt{3^2 + 2^2 + 1^2} = \sqrt{9 + 4 + 1} = \sqrt{14}\,, \tag{1.2c}$$

$$\mathbf{a}\mathbf{b}^\mathsf{T} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 4 & 2 \\ 6 & 4 & 2 \end{pmatrix}. \tag{1.2d}$$

$$\tag{1.2e}$$

A4: The inner product is symmetric ($\mathbf{a}^\mathsf{T}\mathbf{b} = \mathbf{b}^\mathsf{T}\mathbf{a}$), but the outer product is not ($\mathbf{a}\mathbf{b}^\mathsf{T} \neq \mathbf{b}\mathbf{a}^\mathsf{T}$).

$$\mathbf{a}\mathbf{b}^\mathsf{T} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 4 & 2 \\ 6 & 4 & 2 \end{pmatrix}, \tag{1.3a}$$

$$\mathbf{a}^\mathsf{T}\mathbf{b} = \begin{pmatrix} 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} = 1 \times 3 + 2 \times 2 + 2 \times 1 = 3 + 4 + 2 = 9\,, \tag{1.3b}$$

$$\mathbf{b}\mathbf{a}^\mathsf{T} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 6 \\ 2 & 4 & 4 \\ 1 & 2 & 2 \end{pmatrix}, \tag{1.3c}$$

$$\mathbf{b}^\mathsf{T}\mathbf{a} = \begin{pmatrix} 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} = 3 \times 1 + 2 \times 2 + 1 \times 2 = 3 + 4 + 2 = 9\,. \tag{1.3d}$$

Q5: How to find inverse matrix?:
(1) $(A\ I)$, (2) elementary row transformation $(I\ B)$, (3) $B = A^{-1}$. e.g.,

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{1.4}$$

A5:

$$M^{-1}M = MM^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} M^{-1} = I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{1.5}$$

$$(M\,I) = \left(\begin{array}{ccc:ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc:ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{2} \end{array}\right) = (I\,M^{-1}), \tag{1.6a}$$

$$M^{-1}M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I, \tag{1.6b}$$

$$MM^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I. \tag{1.6c}$$

A6:

$$M\mathbf{a} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \\ 4 \end{pmatrix} \tag{1.7}$$

A7: No, the matrix is not symmetric ($A \neq A^\mathsf{T}$).

$$A = \mathbf{a}\mathbf{b}^\mathsf{T} = \begin{pmatrix} 3 & 2 & 1 \\ 6 & 4 & 2 \\ 6 & 4 & 2 \end{pmatrix}, \tag{1.8a}$$

$$A^\mathsf{T} = \begin{pmatrix} 3 & 6 & 6 \\ 2 & 4 & 4 \\ 1 & 2 & 2 \end{pmatrix} = \mathbf{b}\mathbf{a}^\mathsf{T} = (\mathbf{a}\mathbf{b}^\mathsf{T})^\mathsf{T}. \tag{1.8b}$$

A8: The rank of A is 1.

$$(A\,I) = \left(\begin{array}{ccc:ccc} 3 & 2 & 1 & 1 & 0 & 0 \\ 6 & 4 & 2 & 0 & 1 & 0 \\ 6 & 4 & 2 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 3 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{array}\right), \tag{1.9a}$$

$$\begin{pmatrix} A \\ I \end{pmatrix} = \left(\begin{array}{ccc} 3 & 2 & 1 \\ 6 & 4 & 2 \\ 6 & 4 & 2 \\ \hdots & \hdots & \hdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \\ \hdots & \hdots & \hdots \\ \frac{1}{3} & -\frac{2}{3} & -\frac{1}{3} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right). \tag{1.9b}$$

A9: A square matrix is invertible if and only if its rank is equal to the number of its columns/rows. $A = \mathbf{a}\mathbf{b}^\mathsf{T}$ is not invertible.

A10: The projection of vector **a** onto vector **b** is

$$\|\vec{a}\| = \sqrt{1^2+2^2+2^2} = \sqrt{1+4+4} = \sqrt{9} = 3\,, \tag{1.10a}$$

$$\|\vec{b}\| = \sqrt{3^2+2^2+1^2} = \sqrt{9+4+1} = \sqrt{14}\,, \tag{1.10b}$$

$$\cos(\theta) = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\cdot\|\vec{b}\|} = \frac{(1,2,2)\cdot(3,2,1)}{\sqrt{1+4+4}\sqrt{9+4+1}} = \frac{3+4+2}{\sqrt{9}\sqrt{14}} = \frac{9}{3\sqrt{14}} = \frac{3}{\sqrt{14}}\,, \tag{1.10c}$$

$$\Pi_{\vec{b}}(\vec{a}) = \|\vec{a}\|\cos(\theta)\,\vec{e}_b = 3\frac{3}{\sqrt{14}}\frac{(3,2,1)}{\sqrt{14}} = \frac{9}{14}(3,2,1) \tag{1.10d}$$

$$= \frac{\vec{a}\cdot\vec{b}}{\vec{b}\cdot\vec{b}}\vec{b} = \frac{3+4+2}{9+4+1}(3,2,1) = \frac{9}{14}(3,2,1) \tag{1.10e}$$

$$= \left(\vec{a}\cdot\frac{\vec{b}}{\|\vec{b}\|}\right)\frac{\vec{b}}{\|\vec{b}\|} = \|\vec{a}\|\cos(\theta)\,\frac{\vec{b}}{\|\vec{b}\|} = \frac{\vec{a}\cdot\vec{b}}{\vec{b}\cdot\vec{b}}\vec{b} \tag{1.10f}$$

1.1.2 Derivatives

Problem 2 (Derivatives). *We denote the derivative of a univariate function $f(x)$ with respect to the variable x by $\frac{df(x)}{dx}$. We denote the partial derivative of a multivariate function $f(x_1,...,x_n)$ with respect to the variable x_i, where $1 \leqslant i \leqslant n$, by $\frac{\partial f(x_1,...,x_n)}{\partial x_i}$. The partial derivative $\frac{\partial f(x_1,...,x_n)}{\partial x_i}$ is the derivative of $f(x_1,...,x_n)$ with respect to x_i when we treat all other variables x_j for $j \neq i$ in f as constants.*
Please recall the basic rules for derivatives, namely the sum rule, the chain rule, and the product rule.

Q1: What is the derivative of $f(x) = \frac{1}{1+\exp(-x)}$ with respect to x?

Q2: What is the partial derivative of $f(w,x) = 2(wx+5)^2$ with respect to w?

A 1–2:

$$\frac{df(x)}{dx} = (1+e^{-x})^{-1} = -(1+e^{-x})^{-2}e^{-x}(-1) = \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{x}}\,. \tag{1.11a}$$

$$\frac{\partial f(w,x)}{\partial x} = 2\cdot 2(wx+5)^1 w = 4w(wx+5)\,, \tag{1.11b}$$

$$\frac{\partial f(w,x)}{\partial w} = 2\cdot 2(wx+5)^1 x = 4x(wx+5)\,. \tag{1.11c}$$

1.1.3 Probability Theory: Sample Space

Problem 3 (Probability theory: sample space). *An urn contains five red, three orange, and one blue ball. Two balls are randomly selected (without replacement).*

A1: The sample space is balls in the urn, that is, 5 red, 3 orange, and 1 blue ball.

A2:

$$\mathbf{P}_{red}(x) = \frac{5}{5+3+1} = \frac{5}{9}\,, \tag{1.12a}$$

$$\mathbf{P}_{orange}(x) = \frac{3}{5+3+1} = \frac{3}{9} = \frac{1}{3}\,, \tag{1.12b}$$

$$\mathbf{P}_{blue}(x) = \frac{1}{5+3+1} = \frac{1}{9}\,. \tag{1.12c}$$

A 3–4: The possible values of X are 0, 1, 2, and 3. ($X = 3$ will never be true.)

$$\mathbf{P}[X=3] = 0\,, \tag{1.13a}$$

$$\mathbf{P}[X=2] = \frac{C_3^2}{C_9^2} = \frac{3}{1}\cdot\frac{2\times 1}{9\times 8} = \frac{1}{12}\,, \tag{1.13b}$$

$$\begin{aligned}\mathbf{P}[X=1] &= \frac{C_3^1C_5^1 + C_3^1C_1^1}{C_9^2} = \frac{3\times 5+3\times 1}{1}\cdot\frac{2\times 1}{9\times 8} = \frac{6}{3\times 4} = \frac{1}{2}\\ &= \frac{C_3^1C_{5+1}^1}{C_9^2} = \frac{3\times 6}{1}\cdot\frac{2\times 1}{9\times 8} = \frac{1}{2}\,,\end{aligned} \tag{1.13c}$$

$$\begin{aligned}\mathbf{P}[X=0] &= 1-0-\frac{1}{12}-\frac{1}{2} = 1-\frac{7}{12} = \frac{5}{12}\\ &= \frac{C_1^0C_5^2 + C_1^1C_5^1}{C_9^2} = \frac{1\times\frac{5\times 4}{2\times 1}+1\times 5}{1}\cdot\frac{2\times 1}{9\times 8} = \frac{10+5}{1}\frac{1}{9\times 4} = \frac{15}{36} = \frac{5}{12}\,.\end{aligned} \tag{1.13d}$$

Table 1: Probability of orange balls. Let Y represent the number of selected red balls. There are $C_9^2 = \frac{9\times 8}{2\times 1} = 36$ probabilities.

(a) No blue ball

X \ Y	0	1	2	3	4	5	
0	0	0	10	0	0	0	10
1	0	15	0	0	0	0	15
2	3	0	0	0	0	0	3
3	0	0	0	0	0	0	0
	3	15	10	0	0	0	28

(b) One blue ball

X \ Y	0	1	2	3	4	5	
0	0	5	0	0	0	0	5
1	3	0	0	0	0	0	3
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
	3	5	0	0	0	0	8

A 5–6:

$$\mathbf{E}[X] = \sum_{x=0}^{3} x\mathbf{P}[X=x] = 3\times 0+2\times\frac{1}{12}+1\times\frac{1}{2}+0\times\frac{5}{12} = 0+\frac{1}{6}+\frac{1}{2}+0 = \frac{4}{6} = \frac{2}{3}\,, \tag{1.14a}$$

$$\begin{aligned}\mathbf{V}[X] = \mathbf{E}[(X-\mu)^2] &= \sum_{x=0}^{3}(x-\mu)^2\mathbf{P}(X=x)\\ &= \left(-\frac{2}{3}\right)^2\frac{5}{12}+\left(\frac{1}{3}\right)^2\frac{1}{2}+\left(\frac{4}{3}\right)^2\frac{1}{12}+\left(\frac{7}{3}\right)^2\cdot 0\\ &= \frac{5}{9\times 3}+\frac{1}{9\times 2}+\frac{4}{27}+0 = \frac{9}{27}+\frac{1}{18} = \frac{1}{3}+\frac{1}{18} = \frac{7}{18}\,.\end{aligned} \tag{1.14b}$$

1.1.4 Probability Theory: Properties of Expectation

Basic Theorems ref

Remark 1 (*Linear Function*). *If $Y = aX + b$, where a and b are finite constants, then*

$$\mathbf{E}(Y) = a\mathbf{E}(X) + b\,. \tag{1.15}$$

Proof.

$$\begin{aligned}\mathbf{E}(Y) = \mathbf{E}(aX+b) &= \int_{-\infty}^{\infty}(ax+b)f(x)dx \\ &= a\int_{-\infty}^{\infty}xf(x)dx + b\int_{-\infty}^{\infty}f(x)dx \;= a\mathbf{E}(X)+b\,.\end{aligned} \tag{1.16}$$

□

Corollary 1.1. *If $X = c$ with probability 1, then $\mathbf{E}(X) = c$.*

Proof.

$$\mathbf{E}(X) = \int_{-\infty}^{\infty}cf(x)dx = c\int_{-\infty}^{\infty}f(x)dx = c\,. \tag{1.17}$$

□

Theorem 1.2. *If there exists a constant such that $\mathbf{Pr}(X \geqslant a) = 1$, then $\mathbf{E}(X) \geqslant a$. If there exists a constant b such that $\mathbf{Pr}(X \leqslant b) = 1$, then $\mathbf{E}(X) \leqslant b$.*

Proof.

$$\mathbf{E}(X) = \int_{-\infty}^{\infty}xf(x)dx = 0 + \int_{a}^{\infty}xf(x)dx \geqslant \int_{a}^{\infty}af(x)dx = a\,\mathbf{Pr}(X \geqslant a) = a\,, \tag{1.18}$$

$$\mathbf{E}(X) = \int_{-\infty}^{\infty}xf(x)dx = \int_{-\infty}^{b}xf(x)dx + 0 \leqslant \int_{-\infty}^{b}bf(x)dx = b\,\mathbf{Pr}(X \leqslant b) = b\,. \tag{1.19}$$

□

Theorem 1.3. *Suppose that $\mathbf{E}(x) = a$ and that either $\mathbf{Pr}(X \geqslant a) = 1$ or $\mathbf{Pr}(X \leqslant a) = 1$. Then $\mathbf{Pr}(X = a) = 1$.*

Proof. I don't know why? Discrete situations here. (1) When $\mathbf{Pr}(X \geqslant a) = 1$, suppose $x_1, x_2, ...$ are all situations of $x > a$, then $\mathbf{Pr}(X = x) > 0$. Let $p_0 = \mathbf{Pr}(X = a)$, then

$$\mathbf{E}(X) = p_0 a + \sum_{j=1}^{\infty}x_j\mathbf{Pr}(X = x_j) \geqslant p_0 a + \sum_{j=1}^{\infty}a\mathbf{Pr}(X = x_j) = a \tag{1.20}$$

(2) When $\mathbf{Pr}(X \leqslant a) = a$, suppose $..., x_{-2}, x_{-1}$ are all situations of $x < a$, then $\mathbf{Pr}(X = x) > 0$. Let $p_0 = \mathbf{Pr}(X = a)$, then

$$\mathbf{E}(X) = p_0 a + \sum_{j=-\infty}^{-1}x_j\mathbf{Pr}(X = x_j) \leqslant p_0 a + \sum_{j=-\infty}^{-1}a\mathbf{Pr}(X = x_j) = a \tag{1.21}$$

□

Theorem 1.4. *If $X_1, ..., X_n$ are n random variables such that each expectation $E(X_i)$ is finite $(i = 0, ..., n)$, then*

$$\mathbf{E}(X_1 + \cdots + X_n) = \mathbf{E}(X_1) + \cdots + \mathbf{E}(X_n)\,. \tag{1.22}$$

Proof. Two continuous variables

$$\begin{aligned}
\mathbf{E}(X_1+X_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1+x_2)f(x_1,x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1f(x_1,x_2)dx_1dx_2+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_2f(x_1,x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1f(x_1,x_2)dx_2dx_1+\infty\int_{-\infty}^{\infty}x_2f(x_1,x_2)dx_1dx_2 \qquad (1.23a)\\
&= \int_{-\infty}^{\infty}x_1\left[\int_{-\infty}^{\infty}f(x_1,x_2)dx_2\right]dx_1+\int_{-\infty}^{\infty}x_2\left[\int_{-\infty}^{\infty}f(x_1,x_2)dx_1\right]dx_2 \qquad (1.23b)\\
&= \int_{-\infty}^{\infty}x_1f_1(x_1)dx_1+\int_{-\infty}^{\infty}x_2f_2(x_2)dx_2 \\
&= \mathbf{E}(X_1)+\mathbf{E}(X_2)\,. \qquad (1.23c)
\end{aligned}$$

□

Corollary 1.5. *Assume that* $\mathbf{E}(x_i)$ *is finite for* $i=1,...,n$*. For all constants* $a_1,...,a_n$ *and* b*,*

$$\mathbf{E}(a_1X_1+\cdots+a_nX_n+b)=a_1\mathbf{E}(X_1)+\cdots a_n\mathbf{E}(X_n)+b\,. \qquad (1.24)$$

Definition 1.6 (Convex Functions)**.** *A function* g *of a vector argument is convex if, for every* $\alpha\in(0,1)$*, and every* x *and* y*,*

$$g[\alpha x+(1-\alpha)y]\geqslant\alpha g(x)+(1-\alpha)g(y)\,. \qquad (1.25)$$

Theorem 1.7 (Jensen's Inequality)**.** *Let* g *be a convex function, and let* X *be a random vector with finite mean. Then* $\mathbf{E}[g(X)]\geqslant g(\mathbf{E}[X])$*. The equality is satisfied/met if and only if* g *is a linear function.*

Remark 2. *If* $X_1,...,X_n$ *are* n *independent random variables such that each expectation* $\mathbf{E}(X_i)$ *is finite* $(i=1,...,n)$*, then*

$$\mathbf{E}(\Pi_{i=1}^nX_i)=\Pi_{i=1}^n\mathbf{E}(X_i)\,. \qquad (1.26)$$

Proof. The p.d.f. are f_i for the variable X_i in this group, then

$$f(x_1,...,x_i)=\Pi_{i=1}^nf_i(x_i)\,, \qquad (1.27)$$

therefore,

$$\begin{aligned}
\mathbf{E}(\Pi_{i=1}^nX_i) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\Pi_{i=1}^nx_i)f(x_1,...,x_n)dx_1,...,x_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Pi_{i=1}^n[x_if_i(x_i)]dx_1,...,x_n \qquad (1.28a)\\
&= \mathbf{E}(X_1)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Pi_{i=2}^n[x_if_i(x_i)]dx_2,...,x_n \\
&= \mathbf{E}(X_1)\mathbf{E}(X_2)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Pi_{i=3}^n[x_if_i(x_i)]dx_3,...,x_n \\
&= \Pi_{i=1}^n\int_{-\infty}^{\infty}x_if_i(x_i)dx_i \quad =\Pi_{i=1}^n\mathbf{E}(X_i)\,. \qquad (1.28b)
\end{aligned}$$

□

Problem 4 (Probability theory: properties of expectation)**.** *Let* X *and* Y *be two discrete random variables taking values in* $\mathcal{X}$ *and* $\mathcal{Y}$*, respectively. Starting from the definitions, prove the following identities:*

1. $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$
2. *If X and Y are independent then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. (Mark the step where you are using the independence assumption. Note that this assumption was not required in point 1.)*
3. *Provide an example of two random variables X and Y for which $\mathbf{E}[XY] \neq \mathbf{E}[X]\mathbf{E}[Y]$. (Describe how you define the random variables, provide a joint probability distribution table [see comment below], and calculate $\mathbf{E}[XY]$ and $\mathbf{E}[X]\mathbf{E}[Y]$.)*
4. $\mathbf{E}[\mathbf{E}[X]] = \mathbf{E}[X]$.
5. *Variance of a random variable is defined as $\mathbf{V}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$. Show that $\mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.*

A0: We have already known that

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} x_k p_k \,, \qquad \mathbf{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \,, \tag{1.29a}$$

$$\mathbf{E}[Y] = \sum_{k=1}^{\infty} y_k p_k \,, \qquad \mathbf{E}[Y] = \int_{-\infty}^{\infty} y f(y) dy \,, \tag{1.29b}$$

$$\mathbf{E}[X+Y] = \,. \tag{1.29c}$$

A1: therefore,

$$\begin{aligned} \mathbf{E}[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f(x+y) dx dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x+y) dy dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x+y) dx dy && \text{(1.30a)} \\ &= \int_{-\infty}^{\infty} x f_x(x) dx + \int_{-\infty}^{\infty} y f_y(y) dy && \text{(1.30b)} \\ &= \mathbf{E}[X] + \mathbf{E}[Y] \,. && \text{(1.30c)} \end{aligned}$$

A2: therefore,

$$\begin{aligned} \mathbf{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy) f(xy) dx dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_x(x) f_y(y) dx dy && \text{(1.31a)} \\ &= \int_{-\infty}^{\infty} x f_x(x) dx \int_{-\infty}^{\infty} y f_y(y) dy && \text{(1.31b)} \\ &= \mathbf{E}[X]\mathbf{E}[Y] \,. && \text{(1.31c)} \end{aligned}$$

A4:

$$\mathbf{E}[X] = \textbf{ a constant} \tag{1.32a}$$

$$\mathbf{E}[\mathbf{E}[X]] = \int_{-\infty}^{\infty} \mathbf{E}[X]\mathbf{Pr}(x) dx = \mathbf{E}[X] \int_{-\infty}^{\infty} \mathbf{Pr}(x) dx = \mathbf{E}[X] \tag{1.32b}$$

A5:

$$\begin{aligned} \mathbf{V}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + (\mathbf{E}[X])^2] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + (\mathbf{E}[X])^2 = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \,. \end{aligned} \tag{1.33}$$

1.1.5 Probability Theory: Complements of Events

Problem 5 (1). *The complement of event A is denoted by $\bar{A}$ and defined by $\bar{A} = \Omega \setminus A$. Starting from probabilities axioms prove that $\mathbf{Pr}\{A\} = 1 - \mathbf{Pr}\{\bar{A}\}$.*

Problem 6 (2). *In many cases it is easier to calculate the probability of a complement of an event than to calculate the probability of the event itself. Use this to solve the following question. We flip a fair coin 10 times.*

- *What is the probability to observe at least one tail?*
- *What is the probability to observe at least two tails?*

A2:

$$\mathbf{Pr}[x \geqslant 1] = 1 - \mathbf{Pr}[x = 0] = 1 - \frac{1^1 0}{2^{10}} = \frac{1023}{1024} \tag{1.34}$$

$$\mathbf{Pr}[x \geqslant 2] = 1 - \mathbf{Pr}[x = 0] - \mathbf{Pr}[x = 1] = 1 - \frac{1^{10}}{2^{10}} - A^1_{10}\frac{1^9}{2^{10}} = 1 - \frac{1}{1024} - \frac{10}{1024} = \frac{1024 - 11}{1024} = \frac{1013}{1024} \tag{1.35}$$

1.1.6 Induction

Problem 7. *Prove by induction that for all integer n and d, such that $n \geqslant d \geqslant 1$:*

$$\sum_{i=0}^{d} \binom{n}{i} \leqslant n^d + 1\,. \tag{1.36}$$

Proof. I don't know why? (a1) When $d = 1$,

$$Eq.^{\text{left}} = \sum_{i=0}^{1} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} = \frac{n!}{n!0!} + \frac{n!}{(n-1)!1!} = 1 + n = n^d + 1 = Eq.^{\text{right}} \tag{1.37}$$

(a2) When $d = j$, the inequality is met, then for $d = j + 1$,

$$Eq.^{\text{left}} = \sum_{i=0}^{j+1} \binom{n}{i} = \sum_{i=0}^{j} \binom{n}{i} + \binom{n}{j+1} \leqslant n^j + 1 + \frac{n!}{(n-j-1)!(j+1)!}\,, \tag{1.38a}$$

$$Eq.^{\text{right}} = n^{j+1} + 1 = \tag{1.38b}$$

(a3) In conclusion.

(b1) When $n = 1$, then $d = 1$ as well,

$$Eq.^{\text{left}} = \sum_{i=0}^{1} \binom{1}{i} = \binom{1}{0} + \binom{1}{1} = \frac{1!}{1!0!} + \frac{1!}{0!1!} = 1 + 1 = n^d + 1 = Eq.^{\text{right}} \tag{1.39}$$

(b2) When $n = t$, the inequality is met, then for $n = t + 1$,

$$Eq.^{\text{left}} = \sum_{i=0}^{d} \binom{t+1}{i} = \sum_{i=0}^{d} \frac{(t+1)!}{(t+1-i)!i!} = \sum_{i=0}^{d} \frac{t+1}{t+1-i}\frac{t!}{(t-i)!i!} = \sum_{i=0}^{d} \binom{t}{i}\frac{t+1}{t+1-i}$$

$$\leqslant (t^d + 1)\frac{t+1}{t+1-i} \tag{1.40a}$$

$$Eq.^{\text{right}} = (t+1)^d + 1 = \tag{1.40b}$$

(b3) In conclusion. □

1.1.7 Programming

1.2 ATML Self-Preparation Assignment

1.2.1 Illustration of Hoeffding's Inequality

Q1:

1. Make $1,000,000$ repetitions of the experiment of drawing 20 i.i.d. Bernoulli random variables $X_1, ..., X_{20}$ (20 coins) with bias $\frac{1}{2}$.
2. Plot the empirical frequency of observing $\frac{1}{20}\sum_{i=1}^{20} X_i \geqslant \alpha$ for $\alpha \in \{0.5, 0.55, 0.6, ..., 0.95, 1\}$.
3. Explain why the above granularity of α is sufficient. I.e., why, e.g., taking $\alpha = 0.51$ will not provide any extra information about the experiment.
4. In the same figure plot the Hoeffding's bound[1] on $\mathbf{Pr}(\frac{1}{20}\sum_{i=1}^{20} X_i \geqslant \alpha)$ for the same values of α.
5. In the same figure plot the Markov's bound[2] on $\mathbf{Pr}(\frac{1}{20}\sum_{i=1}^{20} X_i \geqslant \alpha)$.
6. Compare the three plots.
7. For $\alpha = 1$ and $\alpha = 0.95$ calculate the exact probability $\mathbf{Pr}(\frac{1}{20}\sum_{i=}^{20} X_i \geqslant \alpha)$ and compare it with the Hoeffding's bound. (No need to add this one to the plot.)

Do not forget to put axis labels and a legend in your plot!

1.2.2 The effect of scale (range) and normalization of random variables in Hoeffding's inequality

1.2.3 Probability in Practice

1.2.4 Occam's Razor

1.2.5 Train-Validation Split Trade-off

1.2.6 Optional Assignment Covering the VC-Analysis

1.2.7 The growth function

1.2.8 VC-dimension

1.2.9 Airline Revisited

[1] Hoeffding's bound is the right hand side of Hoeffding's inequality.
[2] Markov's bound is the right hand side of Markov's inequality.

2 Chapter 2. Concentration of Measure Inequalities

2.1 Markov's Inequality

Theorem 2.1 (Markov's Inequality). *For any non-negative random variable X and $\varepsilon > 0$:*

$$\mathbf{Pr}(X \geqslant \varepsilon) \leqslant \frac{\mathbf{E}[X]}{\varepsilon}. \tag{2.1}$$

Proof. Define a random variable $Y = \mathbb{I}(X \geqslant \varepsilon)$ to be the indicator function of whether X exceeds ε. Then $Y \leqslant \frac{X}{\varepsilon}$. Since Y is a Bernoulli random variable, $\mathbf{E}[Y] = \mathbf{Pr}(Y = 1)$. We have:

$$\mathbf{Pr}(X \geqslant \varepsilon) = \mathbf{Pr}(Y = 1) = \mathbf{E}[Y] \leqslant \mathbf{E}\left[\frac{X}{\varepsilon}\right] = \frac{\mathbf{E}[X]}{\varepsilon}. \tag{2.2}$$

□

A Question: Because $Y = \mathbb{I}[\frac{X}{\varepsilon} \geqslant 1] \in [0, 1]$, and $\frac{X}{\varepsilon} \in [0, +\infty)$ due to the non-negativity of X.

1. If $\frac{X}{\varepsilon} \in [0, 1)$, then $Y = 0$, and $Y \leqslant \frac{X}{\varepsilon}$;
2. If $\frac{X}{\varepsilon} = 1$, then $Y = 1$, and $Y = \frac{X}{\varepsilon}$;
3. If $\frac{X}{\varepsilon} \in (1, +\infty)$, then $Y = 1$, and $Y < \frac{X}{\varepsilon}$;

Overall, $Y \leqslant \frac{X}{\varepsilon}$. Then $\mathbf{E}[Y] = \mathbf{Pr}(Y = 1)$.

$$\mathbf{Pr}(X \geqslant \varepsilon) = \mathbf{Pr}(Y = 1) = \mathbf{E}[Y] \leqslant \mathbf{E}[\tfrac{X}{\varepsilon}] = \frac{\mathbf{E}[X]}{\varepsilon} \tag{2.3a}$$

$$\mathbf{Pr}(X \geqslant \tfrac{1}{\delta}\mathbf{E}[X]) \leqslant \frac{\mathbf{E}[X]}{\frac{1}{\delta}\mathbf{E}[X]} = \delta \tag{2.3b}$$

By denoting the right hand side of Markov's inequality by δ we obtain the following equivalent statement. For any non-negative random variable X:

$$\mathbf{Pr}\left(X \geqslant \tfrac{1}{\delta}\mathbf{E}[X]\right) \leqslant \delta. \tag{2.4}$$

2.2 Chebyshev's Inequality

This one exploits variance to obtain tighter concentration.

Theorem 2.2 (Chebyshev's inequality). *For any $\varepsilon > 0$,*

$$\mathbf{Pr}(|X - \mathbf{E}[X]| \geqslant \varepsilon) \leqslant \frac{\mathbf{V}[X]}{\varepsilon^2}. \tag{2.5}$$

Proof. We use a transformation of a random variable. We have that $\mathbf{Pr}(|X - \mathbf{E}[X]| \geqslant \varepsilon) = \mathbf{Pr}((X - \mathbf{E}[X])^2 \geqslant \varepsilon^2)$, because the first statement holds if and only if the second holds. In addition, using Markov's inequality and the fact that $(X - \mathbf{E}[X])^2$ is a non-negative random variable we have

$$\mathbf{Pr}(|X - \mathbf{E}[X]| \geqslant \varepsilon) = \mathbf{Pr}\left((X - \mathbf{E}[X])^2 \geqslant \varepsilon^2\right) \leqslant \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}{\varepsilon^2} = \frac{\mathbf{V}[X]}{\varepsilon^2}. \tag{2.6}$$

□

In order to illustrate the relative advantage of Chebyshev's inequality compared to Markov's consider the following example. Let $X_1, ..., X_n$ be n independent identically distributed Bernoulli random variables and let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be their average. We would like to bound the probability that $\hat{\mu}_n$ deviates from $\mathbf{E}[\hat{\mu}_n]$ by more than ε (this is the central question in machine learning). We have $\mathbf{E}[\hat{\mu}_n] = \mathbf{E}[X_1] = \mu$ and by independence of X_i-s and Theorem B.26[3] we have $\mathbf{V}[\hat{\mu}_n] = \frac{1}{n^2}\mathbf{V}[n\hat{\mu}_n] = \frac{1}{n^2}\sum_{i=1}^n \mathbf{V}[X_i] = \frac{1}{n}\mathbf{V}[X_1]$. By Markov's inequality,

$$\mathbf{Pr}(\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n] \geqslant \varepsilon) = \mathbf{Pr}(\hat{\mu}_n \geqslant \mathbf{E}[\hat{\mu}_n] + \varepsilon) \leqslant \frac{\mathbf{E}[\hat{\mu}_n]}{\mathbf{E}[\hat{\mu}_n] + \varepsilon} = \frac{\mathbf{E}[X_1]}{\mathbf{E}[X_1] + \varepsilon}\,. \tag{2.9}$$

Note that as n grows the inequality stays the same. By Chebyshev's inequality we have

$$\mathbf{Pr}(\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n] \geqslant \varepsilon) \leqslant \mathbf{Pr}(|\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n]| \geqslant \varepsilon) \leqslant \frac{\mathbf{V}[\hat{\mu}_n]}{\varepsilon^2} = \frac{\mathbf{V}[X_1]}{n\varepsilon^2}\,. \tag{2.10}$$

Note that as n grows the right hand side of the inequality decreases at the rate of $\frac{1}{n}$. Thus, in this case, Chebyshev's inequality is much tighter than Markov's and it illustrates that as the number of random variables grows the probability that their average significantly deviates from the expectation decreases. In the next section we show that this probability actually decreases at an exponential rate.

A Question: I don't know why $\mathbf{E}[\hat{\mu}_n] = \mathbf{E}[X_1] = \mu$?

2.3 Hoeffding's Inequality

Hoeffding's inequality is a much more powerful concentration result.

Theorem 2.5 (Hoeffding's Inequality). *Let $X_1, ..., X_n$ be independent real-valued random variables, such that for each $i \in \{1, ..., n\}$ there exist $a_i \leqslant b_i$, such that $X_i \in [a_i, b_i]$. Then for every $\varepsilon > 0$:*

$$\mathbf{Pr}\left(\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right] \geqslant \varepsilon\right) \leqslant e^{-2\varepsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}\,, \tag{2.11}$$

and

$$\mathbf{Pr}\left(\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right] \leqslant -\varepsilon\right) \leqslant e^{-2\varepsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}\,. \tag{2.12}$$

By taking a union bound of the events, we obtain the following corollary.

3

Theorem 2.3 (B.26). *If $X_1, ..., X_n$ are independent random variables then*

$$\mathbf{V}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{V}[X_i]\,. \tag{2.7}$$

Theorem 2.4 (B.23). *If X and Y are independent random variables, then*

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]\,. \tag{2.8}$$

We emphasize that in contrast with Theorem B.22, this property does not hold in the general case (if X and Y are not independent).

Corollary 2.6. *Under the assumptions of Theorem 2.5:*

$$\mathbf{Pr}\left(\left|\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right| \geqslant \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i-a_i)^2}. \tag{2.13}$$

Equations (2.11) and (2.12) are known as "one-sided Hoeffding's inequalities" and 2.13 is known as "two-sided Hoeffding's inequality". If we assume that X_i-s are identically distributed and belong to the $[0,1]$ interval we obtain the following corollary.

A Proof of Corollary 2.6: Let the event A be $\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]$, then

$$\mathbf{Pr}(A_+) = \mathbf{Pr}(A \geqslant \varepsilon) \tag{2.14a}$$
$$\mathbf{Pr}(A_-) = \mathbf{Pr}(A \leqslant -\varepsilon) \tag{2.14b}$$
$$\begin{aligned}\mathbf{Pr}(|A| \geqslant \varepsilon) &= \mathbf{Pr}\big((A \geqslant \varepsilon) \vee (A \leqslant -\varepsilon)\big) \\ &= \mathbf{Pr}(A \geqslant \varepsilon) + \mathbf{Pr}(A \leqslant -\varepsilon) \\ &\leqslant 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)\end{aligned} \tag{2.14c}$$

Corollary 2.7. *Let $X_1, ..., X_n$ be independent random variables, such that $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$:*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2} \tag{2.15}$$

and

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2}. \tag{2.16}$$

A Proof of Corollary 2.7: According to Theorem 2.5, we get two inequalities. Let $Y_i = \frac{1}{n}X_i \in [0,1]$, we can get

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}\frac{1}{n^2}(b_i-a_i)^2}\right) = \exp\left(\frac{-2\varepsilon^2}{\frac{1}{n}1^2}\right) = e^{-2n\varepsilon^2} \tag{2.17a}$$
$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \leqslant -\varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}\frac{1}{n^2}(b_i-a_i)^2}\right) = \exp\left(\frac{-2\varepsilon^2}{\frac{1}{n}1^2}\right) = e^{-2n\varepsilon^2} \tag{2.17b}$$
$$\mathbf{E}\left[\sum_{i=1}^{n}\frac{1}{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[X_i] = \mu \tag{2.17c}$$

Recall that by Chebyshev's inequality $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ at the rate of n^{-1}. Hoeffding's inequality demonstrates that the convergence is actually much faster, at least at the rate of e^{-n}. The proof of Hoeffding's inequality is based on Hoeffding's lemma.

Lemma 2.8 (Hoeffding's Lemma)**.** *Let X be a random variable, such that $X \in [a,b]$. Then for any $\lambda \in \mathbb{R}$:*

$$\mathbf{E}\left[e^{\lambda X}\right] \leqslant e^{\lambda\mathbf{E}[X] + \frac{\lambda^2(b-a)^2}{8}}. \tag{2.18}$$

The function $f(\lambda) = \mathbf{E}[e^{\lambda X}]$ is known as the *moment generating function* of X, since $f'(0) = \mathbf{E}[X]$, $f''(0) = \mathbf{E}[X^2]$, and, more generally, $f^{(k)}(0) = \mathbf{E}[X^k]$. We provide the proof of the lemma immediately after the proof of Theorem 2.5.

Proof of Theorem 2.5. We prove the first inequality in Theorem 2.5. The second inequality follows by applying the first inequlity to $-X_1, ..., -X_n$. The proof is based on Chernoff's bounding technique. For any $\lambda > 0$ the following holds:

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) = \mathbf{Pr}\left(e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}[\sum_{i=1}^{n} X_i]\right)} \geqslant e^{\lambda\varepsilon}\right) \leqslant \frac{\mathbf{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}[\sum_{i=1}^{n} X_i]\right)}\right]}{e^{\lambda\varepsilon}}, \tag{2.19}$$

where the first step holds since $e^{\lambda x}$ is a monotonously increasing function for $\lambda > 0$ and the second step holds by Markov's inequality. We now take a closer look at the nominator:

$$\begin{aligned}
\mathbf{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}[\sum_{i=1}^{n} X_i]\right)}\right] &= \mathbf{E}\left[e^{\sum_{i=1}^{n} \lambda(X_i - \mathbf{E}[X_i])}\right] \\
&= \mathbf{E}\left[\prod_{i=1}^{n} e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \\
&= \prod_{i=1}^{n} \mathbf{E}\left[e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] && \text{(2.20a)} \\
&\leqslant \prod_{i=1}^{n} e^{\lambda^2 (b_i - a_i)^2/8} && \text{(2.20b)} \\
&= e^{(\lambda^2/8)\sum_{i=1}^{n}(b_i - a_i)^2}, && \text{(2.20c)}
\end{aligned}$$

where (2.20a) holds since $X_1, ..., X_n$ are independent and (2.20b) holds by Hoeffding's lemma applied to a random variable $Z_i = X_i - \mathbf{E}[X_i]$ (note that $\mathbf{E}[Z_i] = 0$ and that $Z_i \in [a_i - \mu_i, b_i - \mu_i]$ for $\mu_i = \mathbf{E}[X_i]$). *Put attention to the crucial role that independence of $X_1, ..., X_n$ plays in the proof! Without independence we would not have been able to exchange the expectation with the product and the proof would break down!* To complete the proof we substitute the bound on the expectation into the previous calculation and obtain:

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant e^{(\lambda^2/8)\left(\sum_{i=1}^{n}(b_i - a_i)^2\right) - \lambda\varepsilon}. \tag{2.21}$$

This expression is minimized[4] by

$$\lambda^* = \operatorname*{argmin}_{\lambda} e^{(\lambda^2/8)\left(\sum_{i=1}^{n}(b_i - a_i)^2\right) - \lambda\varepsilon} = \operatorname*{argmin}_{\lambda}\left((\lambda^2/8)\left(\sum_{i=1}^{n}(b_i - a_i)^2\right) - \lambda\varepsilon\right) = \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}. \tag{2.23}$$

[4]

$$\begin{aligned}
\lambda^* &= \operatorname*{argmin}_{\lambda} \frac{1}{8}\left(\left(\sum_{i=1}^{n}(b_i - a_i)^2\right)\lambda^2 - 8\varepsilon\lambda\right) = \operatorname*{argmin}_{\lambda} \frac{1}{8}\sum_{i=1}^{n}(b_i - a_i)^2\left(\lambda^2 - \frac{8\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}\lambda\right) \\
&= \operatorname*{argmin}_{\lambda} \frac{1}{8}\sum_{i=1}^{n}(b_i - a_i)^2\left(\left(\lambda - \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)^2 - \frac{16\varepsilon^2}{\left(\sum_{i=1}^{n}(b_i - a_i)^2\right)^2}\right) \\
&= \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2} \qquad \text{s.t.} \quad \min_{\lambda} = -\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2} && \text{(2.22a)}
\end{aligned}$$

It is important to note that the best choice of λ does not depend on the sample. In particular, it allows to fix λ before observing the sample. By substituting λ^* into the calculation we obtain the result of the theorem. □

Proof of Lemma 2.8. Note that

$$\mathbf{E}\left[e^{\lambda X}\right] = \mathbf{E}\left[e^{\lambda(X-\mathbf{E}[X])+\lambda\mathbf{E}[X]}\right] = e^{\lambda\mathbf{E}[X]} \times \mathbf{E}\left[e^{\lambda(X-\mathbf{E}[X])}\right]. \tag{2.24}$$

Hence, it is sufficient to show that for any random variable Z with $\mathbf{E}[Z] = 0$ and $Z \in [a, b]$ we have

$$\mathbf{E}\left[e^{\lambda Z}\right] \leqslant e^{\lambda^2(b-a)^2/8}. \tag{2.25}$$

By convexity of the exponential function, for $z \in [a, b]$ we have:

$$e^{\lambda Z} \leqslant \tfrac{z-a}{b-a}e^{\lambda b} + \tfrac{b-z}{b-a}e^{\lambda a}. \tag{2.26}$$

Let $p = -a/(b-a)$. Then:

$$\begin{aligned}
\mathbf{E}\left[e^{\lambda Z}\right] &\leqslant \mathbf{E}\left[\tfrac{Z-a}{b-a}e^{\lambda b} + \tfrac{b-Z}{b-a}e^{\lambda a}\right] \\
&= \frac{\mathbf{E}[Z]-a}{b-a}e^{\lambda b} + \frac{b-\mathbf{E}[Z]}{b-a}e^{\lambda a} && \text{(2.27a)} \\
&\textit{nonumber} && \text{(2.27b)} \\
&= \tfrac{-a}{b-a}e^{\lambda b} + \tfrac{b}{b-a}e^{\lambda a} \\
&= \left(1-p+pe^{\lambda(b-a)}\right)e^{-p\lambda(b-a)} \\
&= e^{\phi(u)}, && \text{(2.27c)}
\end{aligned}$$

where $u = \lambda(b-a)$ and $\phi(u) = -pu + \ln(1-p+pe^u)$ and we used the fact that $\mathbf{E}[Z] = 0$. It is easy to verify that the derivative of ϕ is

$$\phi'(u) = -p + \tfrac{p}{p+(1-p)e^{-u}}, \tag{2.28}$$

and, therefore, $\phi(0) = \phi'(0) = 0$. Furthermore,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{\left(p+(1-p)e^{-u}\right)^2} \leqslant \tfrac{1}{4}. \tag{2.29}$$

By Taylor's theorem, $\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta)$ for some $\theta \in [0, u]$. Thus, we have:

$$\phi(u) = \phi(0) + u\phi'(0) + \tfrac{u^2}{2}\phi''(\theta) = \tfrac{u^2}{2}\phi''(\theta) \leqslant \tfrac{u^2}{8} = \tfrac{\lambda^2(b-a)^2}{8}. \tag{2.30}$$

□

2.3.1 Understanding Hoeffding's Inequality

Hoeffding's inequality involves three interconnected terms: n, ε, and $\delta = 2e^{-2n\varepsilon^2}$, which is the bound on the probability that the event under $\mathbf{Pr}(\cdot)$ holds (for the purpose of the discussion we consider two-sided Hoeffding's inequality for random variables bounded in $[0, 1]$). We can fix

any two of the three terms n, ε, and δ and then the relation $\delta = e^{-2n\varepsilon^2}$ provides the value of the third. Thus, we have

$$\delta = 2e^{-2n\varepsilon^2}\,, \tag{2.31a}$$

$$\varepsilon = \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\,, \tag{2.31b}$$

$$n = \frac{\ln\left(\frac{2}{\delta}\right)}{2\varepsilon^2}\,. \tag{2.31c}$$

Overall, Hoeffding's inequality tells by how much the empircial average $\frac{1}{n}\sum_{i=1}^{n} X_i$ can deviate from its expectation μ, but the interplay between the three parameters provides several ways of seeing and using Hoeffding's inequality. For example, if the number of samples n is fixed (we have made a fixed number of experiments and now analyze what we can get from them), there is an interplay between the precision ε and confidence δ. We can request higher precision ε, but then we have to compromise on the confidence δ that the desired bound $\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leqslant \varepsilon$ holds. And the other way around: we can request higher confidence δ, but then we have to compromise on precision ε, i.e., we have to increase the allowed range $\pm\varepsilon$ around μ, where we expect to find the empirical average $\frac{1}{n}\sum_{i=1}^{n} X_i$.

As another example, we may have target precision ε and confidence δ and then the inequality provides us the number of experiments n that we have to perform in order to achieve the target.

It is often convenient to write the inequalities (2.15) and (2.16) with a fixed confidence in mind, thus we have

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right) \leqslant \delta\,, \tag{2.32a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right) \leqslant \delta\,, \tag{2.32b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\right) \leqslant \delta\,. \tag{2.32c}$$

(Put attention that the $\ln 2$ factor in the last inequality comes from the union bound over the first two inequalities: if we want to keep the same confidence we have to compromise on precision.)

A1: I don't know why, that is, Corollary 2.6.

In many situations we are interested in the complimentary events. Thus, for example, we have

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right) \geqslant 1 - \delta\,. \tag{2.33}$$

Careful reader may point out that the inequalities above should be strict ("$<$" and "$>$"). This is true, but if it holds for strict inequalities it also holds for non-strict inequalities ("$\leqslant$" and "$\geqslant$"). Since strict inequalities provide no practical advantages we will use the non-strict inequalities to avoid the headache of remembering which inequalities should be strict and which should not.

The last inequality essentially says that with probability at least $(1-\delta)$ we have

$$\mu \leqslant \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\,, \tag{2.34}$$

and this is how we will occasionally use it. Note that the random variable is $\frac{1}{n}\sum_{i=1}^{n} X_i$ and the right way of interpreting the above inequality is actually that with probability at least $(1-\delta)$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \mu - \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\,, \tag{2.35}$$

that is, the probability is over $\frac{1}{n}\sum_{i=1}^{n} X_i$ and not over μ. However, many generalisation bounds that we study in Chapter 3 are written in the first form in the literature and we follow the tradition.

2.3.2 Example Appendix

Q Example: We would like to bound the probability that we flip a fair coin 10 times and obtain 8 or more heads. Let $X_1, \ldots, X_{10}$ be i.i.d. Bernoulli random variables with bias $\frac{1}{2}$. The question is equivalent to asking what is the probability that $\sum_{i=1}^{10} X_i \geqslant 8$. We have $\mathbf{E}\left[\sum_{i=1}^{10} X_i\right] = 5$ and by Markov's inequality,

$$\mathbf{Pr}\left(\sum_{i=1}^{10} X_i \geqslant 8\right) \leqslant \frac{\mathbf{E}\left[\sum_{i=1}^{10} X_i\right]}{8} = \frac{5}{8}\,. \tag{2.36}$$

A Example: Let the event A denote the value/number of $\sum_{i=1}^{10} X_i$.

$$\mathbf{Pr}(A \geqslant 8) \leqslant \tfrac{\mathbf{E}[A]}{8} = \tfrac{5}{8}\,. \tag{2.37a}$$

$$\mathbf{Pr}(A \geqslant 8) = \frac{A_{10}^{8} + A_{10}^{9} + A_{10}^{10}}{2^{10}} = \frac{\frac{10!}{2!} + \frac{10!}{1!} + \frac{10!}{0!}}{2^{10}} = \frac{}{2^{10}}\,, \tag{2.37b}$$

$$\mathbf{Pr}(A \geqslant 8) = \frac{C_{10}^{8} + C_{10}^{9} + C_{10}^{10}}{2^{10}} = \frac{\frac{10!}{8!2!} + \frac{10!}{9!1!} + \frac{10!}{10!0!}}{2^{10}} = \tfrac{45+10+1}{1024} = \tfrac{56}{1024} = \frac{7}{128}\,. \tag{2.37c}$$

According to the Markov's inequality, we have

$$C_{10}^{i} = C_{10}^{i} \quad \text{for } i \in \{0, 1, \ldots, 10\} \tag{2.38a}$$

$$= \tfrac{10!}{10!0!}\, \tfrac{10!}{9!1!}, \tfrac{10!}{8!2!}, \tfrac{10!}{7!3!}, \tfrac{10!}{6!4!}, \tfrac{10!}{5!5!}, \tfrac{10!}{4!6!}, \tfrac{10!}{3!7!}, \tfrac{10!}{2!8!}, \tfrac{10!}{1!9!}, \tfrac{10}{0!10!}, \tag{2.38b}$$

$$= 1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1, \tag{2.38c}$$

in summary, 1024 (2.38d)

$$\mathbf{Pr}(A \geqslant 0) \leqslant \tfrac{5}{0} \approx +\infty \tag{2.39a}$$
$$\mathbf{Pr}(A \geqslant 1) \leqslant \tfrac{5}{1} = 5 \tag{2.39b}$$
$$\mathbf{Pr}(A \geqslant 2) \leqslant \tfrac{5}{2} = 2.5 \tag{2.39c}$$
$$\mathbf{Pr}(A \geqslant 3) \leqslant \tfrac{5}{3} \approx 1.6667 \tag{2.39d}$$
$$\mathbf{Pr}(A \geqslant 4) \leqslant \tfrac{5}{4} = 1.25 \tag{2.39e}$$
$$\mathbf{Pr}(A \geqslant 5) \leqslant \tfrac{5}{5} = 1 \tag{2.39f}$$
$$\mathbf{Pr}(A \geqslant 6) \leqslant \tfrac{5}{6} \approx 0.8333 \tag{2.39g}$$
$$\mathbf{Pr}(A \geqslant 7) \leqslant \tfrac{5}{7} \approx 0.7143 \tag{2.39h}$$
$$\mathbf{Pr}(A \geqslant 8) \leqslant \tfrac{5}{8} = 0.625 \tag{2.39i}$$
$$\mathbf{Pr}(A \geqslant 9) \leqslant \tfrac{5}{9} \approx 0.5556 \tag{2.39j}$$
$$\mathbf{Pr}(A \geqslant 10) \leqslant \tfrac{5}{10} = 0.5 \tag{2.39k}$$

$$\mathbf{Pr}(A \geqslant 0) = 1 \tag{2.40a}$$
$$\mathbf{Pr}(A \geqslant 1) = \tag{2.40b}$$
$$\mathbf{Pr}(A \geqslant 2) = \tag{2.40c}$$
$$\mathbf{Pr}(A \geqslant 3) = \tag{2.40d}$$
$$\mathbf{Pr}(A \geqslant 4) = \tag{2.40e}$$
$$\mathbf{Pr}(A \geqslant 5) = \tag{2.40f}$$
$$\mathbf{Pr}(A \geqslant 6) = \tag{2.40g}$$
$$\mathbf{Pr}(A \geqslant 7) = \tag{2.40h}$$
$$\mathbf{Pr}(A \geqslant 8) = \tag{2.40i}$$
$$\mathbf{Pr}(A \geqslant 9) = \tag{2.40j}$$
$$\mathbf{Pr}(A \geqslant 10) = \tfrac{C_{10}^{10}}{2^{10}} = \tfrac{1}{1024} \tag{2.40k}$$
$$\tag{2.40l}$$

2.4 Basics of Information Theory: Entropy, Relative Entropy, and the Method of Types

In this section we briefly introduce a number of basic concepts from information theory that are very useful for deriving concentration inequalities. Specifically, we introduce the notions of entropy and relative entropy and some basic tools from the method of types. We start with some definitions.

Definition 2.9 (Entropy)**.** *Let $p(x)$ be a distribution of a discrete random variable X taking values in a finite set $\mathcal{X}$. We define the entropy of p as:*

$$\mathbf{H}(p) = -\sum_{x \in \mathcal{X}} p(x) \ln p(x)\,. \tag{2.41}$$

We use the convention that $0 \ln 0 = 0$ (which is justified by continuity of $z \ln z$, since $z \ln z \to 0$ as $z \to 0$).

We have special interest in Bernoulli random variables.

Definition 2.10 (Bernoulli random variable)**.** *X is a Bernoulli random variable with bias p if X accepts values in $\{0, 1\}$ with $\mathbf{Pr}(X = 0) = 1 - p$ and $\mathbf{Pr}(X = 1) = p$.*

Note that expectation of a Bernoulli random variable is equal to its bias:

$$\mathbf{E}[X] = 0 \times \mathbf{Pr}(X = 0) + 1 \times \mathbf{Pr}(X = 1) = \mathbf{Pr}(X = 1) = p\,. \tag{2.42}$$

With a slight abuse of notation we specialise the definition of entropy to Bernoulli random variables.

Definition 2.11 (Binary entropy)**.** *Let p be a bias of Bernoulli random variable X. We define the entropy of p as*

$$\mathbf{H}(p) = -p \ln p - (1 - p) \ln(1 - p)\,. \tag{2.43}$$

Note that when we talk about Bernoulli random variables p denotes the bias of the random variable and when we talk about more general random variables p denotes the complete distribution.

Entropy is one of the central quantities in information theory and it has numerous applications. We start by using binary entropy to bound binomial coefficients.

Lemma 2.12.

$$\frac{1}{n+1}e^{n\mathbf{H}\left(\frac{k}{n}\right)} \leqslant \binom{n}{k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)}. \tag{2.44}$$

(Note that $\frac{k}{n} \in [0,1]$ and $\mathbf{H}(\frac{k}{n})$ in the lemma is the binary entropy.)

Proof. By the binomial formula we know that for any $p \in [0,1]$:

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1. \tag{2.45}$$

We start with the upper bound. Take $p = \frac{k}{n}$. Since the sum is larger than any individual term, for the k-th term of the sum we get:

$$\begin{aligned} 1 \geqslant \tbinom{n}{k} p^k (1-p)^{n-k} &= \binom{n}{k} \left(\tfrac{k}{n}\right)^k \left(1-\tfrac{k}{n}\right)^{n-k} = \binom{n}{k} \left(\tfrac{k}{n}\right)^k \left(\tfrac{n-k}{n}\right)^{n-k} \\ &= \tbinom{n}{k} e^{k \ln \frac{k}{n} + (n-k)\ln \frac{n-k}{n}} = \tbinom{n}{k} e^{n\left(\frac{k}{n}\ln\frac{k}{n} + \frac{n-k}{n}\ln\frac{n-k}{n}\right)} = \tbinom{n}{k} e^{-n\mathbf{H}\left(\frac{k}{n}\right)}. \end{aligned} \tag{2.46a}$$

By changing sides of the inequality we obtain the upper bound.

For the lower bound it is possible to show that if we fix $p = \frac{k}{n}$ then $\binom{n}{k} p^k (1-p)^{n-k} \geqslant \binom{n}{i} p^i (1-p)^{n-i}$ for any $i \in \{0, ..., n\}$. We also note that there are $(n+1)$ elements in the sum in Eq. (2.45). Again, take $p = \frac{k}{n}$, then

$$1 \leqslant (n+1) \max_i \binom{n}{i} \left(\tfrac{k}{n}\right)^2 \left(\tfrac{n-k}{n}\right)^{n-i} = (n+1)\binom{n}{k} \left(\tfrac{k}{n}\right)^k \left(\tfrac{n-k}{n}\right)^{n-k} = (n+1)\binom{n}{k} e^{-n\mathbf{H}\left(\frac{k}{n}\right)}, \tag{2.47}$$

where the last step follows the same steps as in the derivation of the upper bound. □

A Lemma 2.12:

$$e^{k\ln\frac{k}{n} + (n-k)\ln\frac{n-k}{n}} = e^{\ln\left(\frac{k}{n}\right)^k + \ln\left(\frac{n-k}{n}\right)^{n-k}} = e^{\ln\left(\frac{k}{n}\right)^k} e^{\ln\left(\frac{n-k}{n}\right)^{n-k}} = \left(\tfrac{k}{n}\right)^k \left(\tfrac{n-k}{n}\right)^{n-k} \tag{2.48a}$$

$$\left(\tfrac{k}{n}\right)^k \left(1-\tfrac{k}{n}\right)^{n-k} = \left(\tfrac{k}{n}\right)^k \left(\tfrac{n-k}{n}\right)^{n-k} = e^{n\left(\frac{k}{n}\ln\frac{k}{n} + \frac{n-k}{n}\ln\frac{n-k}{n}\right)} = e^{-\mathbf{H}\left(\frac{k}{n}\right)\cdot n} \tag{2.48b}$$

Lemma 2.12 shows that the number of configurations of chosing k out of n objects is directly related to the entropy of the imbalance $\frac{k}{n}$ between the number of objects that are selected (k) and the number of objects that are left out $(n-k)$.

We now introduce one additional quantity, the *Kullback-Leibler (KL) divergence*, also known as *Kullback-Leibler distance* and as *relative entropy*.

Definition 2.13 (Relative entropy or Kullback-Leibler divergence)**.** *Let $p(x)$ and $q(x)$ be two probability distributions of a random variable X (or two probability density functions, if X is a continuous random variable), the* Kullback-Leibler divergence *or* relative entropy *is defined as:*

$$\mathbf{KL}(p\|q) = \mathbf{E}_p\left[\ln\frac{p(X)}{q(X)}\right] = \begin{cases} \sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}\,, & \text{if } \mathcal{X} \text{ is discrete}\,; \\ \int_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}dx\,, & \text{if } \mathcal{X} \text{ is continuous}\,. \end{cases} \tag{2.49}$$

We use the convention that $0\ln\frac{0}{0} = 0$ and $0\ln\frac{0}{q} = 0$ and $p\ln\frac{p}{0} = \infty$.

We specialize the definition to Bernoulli distributions.

Definition 2.14 (Binary kl-divergence)**.** *Let p and q be biases of two Bernoulli random variables. The* binary kl divergence *is defined as:*

$$\mathbf{kl}(p\|q) = \mathbf{KL}([1-p,p]\|[1-q,q]) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}\,. \tag{2.50}$$

KL divergence is the central quantity in information theory. Although it is not a distance measure, because it does not satisfy the triangle inequality, it is the right way of measuring distances between probability distributions. This is illustrated by the following example.

Example 2.15. *Let $X_1, ..., X_n$ be an i.i.d. sample of n Bernoulli random variables with bias p and let $\frac{1}{n}\sum_{i=1}^n X_i$ be the empirical bias of the sample. (Note that $\frac{1}{n}\sum_{i=1}^n X_i \in \{0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n}\}$.) Then by Lemma 2.12:*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) = \binom{n}{k}p^k(1-p)^{n-k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)}e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)} = e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,, \tag{2.51}$$

and

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) \geqslant \frac{1}{n+1}e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,. \tag{2.52}$$

Thus, $\mathbf{kl}(\frac{k}{n}\|p)$ governs the probability of observing empirical bias $\frac{k}{n}$ when the true bias is p. It is easy to verify that $\mathbf{kl}(p\|p) = 0$ and it is also possible to show that $\mathbf{kl}(\hat{p}\|p)$ is convex in $\hat{p}$ and that $\mathbf{kl}(\hat{p}\|p) \geqslant 0$. Thus, the probability of empirical bias is maximised when it coincides with the true bias.

A Example 2.15: According to Lemma 2.12, we get $\frac{1}{n+1}e^{n\mathbf{H}\left(\frac{k}{n}\right)} \leqslant \binom{n}{k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)}$, then we already know $\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} = e^{k\ln\frac{k}{n}+(n-k)\ln\frac{n-k}{n}} = e^{-n\mathbf{H}\left(\frac{k}{n}\right)}$. Let $p = \frac{k}{n}$, then

$$\binom{n}{k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)} = e^{-n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)} \tag{2.53a}$$

$$e^{n\left(\frac{k}{n}\ln p+\frac{n-k}{n}\ln(1-p)\right)} = e^{\ln p^k+\ln(1-p)^{n-k}} = p^k(1-p)^{n-k} \tag{2.53b}$$

$$\mathbf{kl}\left(\tfrac{k}{n}\|p\right) = \tfrac{k}{n}\left(\ln\tfrac{k}{n} - \ln p\right) + \tfrac{n-k}{n}\left(\ln\tfrac{n-k}{n} - \ln(1-p)\right) \tag{2.53c}$$

$$\binom{n}{k}p^k(1-p)^{n-k} \leqslant e^{-n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)+n\left(\frac{k}{n}\ln p+\frac{n-k}{n}\ln(1-p)\right)} = e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \tag{2.53d}$$

$$\binom{n}{k}p^k(1-p)^{n-k} \geqslant \frac{1}{n+1}e^{-n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)+n\left(\frac{k}{n}\ln p+\frac{n-k}{n}\ln(1-p)\right)} = \frac{1}{n+1}e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \tag{2.53e}$$

2.5 kl Inequality

Example 2.15 shows that **kl** can be used to bound the empirical bias when the true bias is known. But in machine learning we are usually interested in the inverse problem — how to infer the true bias p when the empirical bias $\hat{p}$ is known. Next we demonstrate that this is also possible and that it leads to an inequality, which in most cases is tighter than Hoeffding's inequality. We start with the following lemma.

Lemma 2.16. *Let $X_1, ..., X_n$ be i.i.d. Bernoulli with bias p and $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias. Then*

$$\mathbf{E}\left[e^{n\,\mathbf{kl}(\hat{p}\|p)}\right] \leqslant n+1\,. \tag{2.54}$$

Proof.

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbf{Pr}\left(\hat{p} = \tfrac{k}{n}\right) e^{n\mathbf{kl}(\frac{k}{n}\|p)} \leqslant \sum_{k=0}^{n} e^{-n\mathbf{kl}(\frac{k}{n}\|p)} e^{n\mathbf{kl}(\frac{k}{n}\|p)} = \sum_{k=0}^{n} 1 = n+1\,, \tag{2.55}$$

where the inequality was derived in Eq. (2.51). □

A Lemma 2.16: We have already known that $\mathbf{E}[X] = \sum_{k=1}^{\infty} x_k p_k = \int_{-\infty}^{\infty} x f(x) dx$.

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbf{Pr}\left(\hat{p} = \tfrac{k}{n}\right) e^{n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right)} \overset{\text{def}}{=} A \tag{2.56a}$$

$$\tfrac{1}{n+1} e^{-n\mathbf{kl}(\frac{k}{n}\|p)} \leqslant \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) \leqslant e^{-n\mathbf{kl}(\frac{k}{n}\|p)} \tag{2.56b}$$

$$A \leqslant \sum_{k=0}^{n} e^{-n\mathbf{kl}(\frac{k}{n}\|p)} e^{n\mathbf{kl}(\frac{k}{n}\|p)} = \sum_{k=0}^{n} e^{0} = \sum_{k=0}^{n} 1 = n+1 \tag{2.56c}$$

$$A \geqslant \sum_{k=0}^{n} \tfrac{1}{n+1} e^{-n\mathbf{kl}(\frac{k}{n}\|p)} e^{n\mathbf{kl}(\frac{k}{n}\|p)} = \sum_{k=0}^{n} \tfrac{1}{n+1} e^{0} = \tfrac{1}{n+1}\sum_{k=0}^{n} 1 = \tfrac{1}{n+1}(n+1) = 1 \tag{2.56d}$$

We combine this lemma with Markov's inequality to obtain the following result.

Theorem 2.17 (kl inequality)**.** *Let $X_1, ..., X_n$ be i.i.d. Bernoulli with bias p and let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias. Then*

$$\mathbf{Pr}(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon) \leqslant (n+1)e^{-n\varepsilon}\,. \tag{2.57}$$

Proof. By Markov's inequality and Lemma 2.16:

$$\mathbf{Pr}\big(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\big) = \mathbf{Pr}\left(e^{n\,\mathbf{kl}(\hat{p}\|p)} \geqslant e^{n\varepsilon}\right) \leqslant \frac{\mathbf{E}\left[e^{n\,\mathbf{kl}(\hat{p}\|p)}\right]}{e^{n\varepsilon}} \leqslant \frac{n+1}{e^{n\varepsilon}}\,. \tag{2.58}$$

□

A Theorem 2.17: Markov's inequality: $\mathbf{Pr}[X \geqslant \varepsilon] \leqslant \frac{\mathbf{E}[X]}{\varepsilon}$. And then it's obviously demonstrated.

2.5.1 Relaxations of the kl-inequality: Pinsker's and refined Pinsker's equations

By denoting the right hand side of kl inequality (2.57) by δ, we obtain that with probability greater than $(1-\delta)$:

$$\mathbf{kl}(\hat{p}\|p) \leqslant \frac{\ln \frac{n+1}{\delta}}{n}. \tag{2.59}$$

This leads to an implicit bound on p, which is not very intuitive and not always convenient to work with. In order to understand the behavior of the kl inequality better, we use a couple of its relaxations. The first relaxation is known as Pinsker's inequality.

A: Because $\mathbf{Pr}\big(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\big) \leqslant (n+1)e^{-n\varepsilon} \stackrel{\text{def}}{=} \delta$, we get that with the probability greater than $(1-\delta)$, we have $e^{-n\varepsilon} = \frac{\delta}{n+1}$ and $-n\varepsilon = \ln \frac{\delta}{n+1}$, and then

$$\mathbf{kl}(\hat{p}\|p) \leqslant \varepsilon = -\tfrac{1}{n}\ln\tfrac{\delta}{n+1} = \tfrac{1}{n}\ln\tfrac{n+1}{\delta}. \tag{2.60}$$

Lemma 2.18 (Pinsker's inequality)**.**

$$\mathbf{KL}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2, \tag{2.61}$$

where $\|p-q\|_1 = \sum_{x\in\mathcal{X}} |p(x)-q(x)|$ *is the* L_1*-norm.*

Corollary 2.19 (Pinsker's inequality for the binary kl divergence)**.**

$$\mathbf{kl}(p\|q) \geqslant \tfrac{1}{2}\big(|p-q| + |(1-p)-(1-q)|\big)^2 = 2(p-q)^2. \tag{2.62}$$

A: We have already known that

$$\mathbf{KL}(p\|q) = \mathbf{E}_p\left[\ln\frac{p(X)}{q(X)}\right] = \begin{cases} \sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}, & \text{discrete} \\ \int_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}dx, & \text{continuous} \end{cases} \tag{2.63}$$

$$\mathbf{kl}(p\|q) = \mathbf{KL}([1-p,p]\|[1-q,q]) = p\ln\tfrac{p}{q} + (1-p)\ln\tfrac{1-p}{1-q} \tag{2.64}$$

$$\mathbf{kl}(p\|q) \geqslant \tfrac{1}{2}\big(|p-q| + |(1-p)-(1-q)|\big)^2 = \tfrac{1}{2}\big(|p-q| + |q-p|\big)^2 = \tfrac{1}{2}\big(2|p-q|\big)^2 = 2(p-q)^2 \tag{2.65a}$$

$$\mathbf{kl}(\hat{p}\|p) = \hat{p}\ln\tfrac{\hat{p}}{p} + (1-\hat{p})\ln\tfrac{1-\hat{p}}{1-p} \leqslant \tfrac{1}{n}\ln\tfrac{n+1}{\delta} \tag{2.65b}$$

$$2(\hat{p}-p)^2 \leqslant \mathbf{kl}(\hat{p}\|p) \leqslant \tfrac{1}{n}\ln\tfrac{n+1}{\delta} \tag{2.65c}$$

$$|\hat{p}-p|^2 \leqslant \tfrac{\mathbf{kl}(\hat{p}\|p)}{2} \leqslant \tfrac{1}{2n}\ln\tfrac{n+1}{\delta} \tag{2.65d}$$

$$|\hat{p}-p| \leqslant \sqrt{\tfrac{\mathbf{kl}(\hat{p}\|p)}{2}} \leqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{n+1}{\delta}} \tag{2.65e}$$

By applying Corollary 2.19 to inequality (2.59), we obtain that with probability greater than $(1-\delta)$,

$$|p-\hat{p}| \leqslant \sqrt{\tfrac{\mathbf{kl}(\hat{p}\|p)}{2}} \leqslant \sqrt{\tfrac{\ln\frac{n+1}{\delta}}{2n}}. \tag{2.66}$$

Recall that Hoeffding's inequality assures that with probability greater than $(1-\delta)$,

$$p \leqslant \hat{p} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\,. \tag{2.67}$$

Thus, in the worst case the kl inequality is only weaker by the $\ln(n+1)$ factor and in fact the $\ln(n+1)$ factor can be reduced by a more careful analysis. Next we show that the kl inequality can actually be significantly tighter than Hoeffding's inequality. For this we use refined Pinsker's inequality.

A: We know that $\delta = (n+1)e^{-n\varepsilon}$, and $\mathbf{Pr}(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon) \leqslant \delta$, then with the probability greater than $(1-\delta)$, we have $\mathbf{kl}(\hat{p}\|p) \leqslant \frac{\ln \frac{n+1}{\delta}}{n}$. We also know that Hoeffding's inequality

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) \qquad \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2} \tag{2.68a}$$

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \leqslant -\varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) \qquad \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leqslant -\varepsilon\right) \leqslant e^{-2n\varepsilon^2} \tag{2.68b}$$

$$\mathbf{Pr}\left(\left|\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) \qquad \mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2} \tag{2.68c}$$

We know that

$$\hat{p} - \sqrt{\frac{\ln \frac{n+1}{\delta}}{2n}} \leqslant p \leqslant \hat{p} + \sqrt{\frac{\ln \frac{n+1}{\delta}}{2n}} \tag{2.69}$$

Is it right in the following? If let $\delta = e^{-2n\varepsilon^2}$, then with the probability greater than $(1-\delta)$, we have $\ln\delta = -2n\varepsilon^2$ and $\varepsilon^2 = -\frac{\ln\delta}{2n} = \frac{\ln\frac{1}{\delta}}{2n}$, then we get

$$\hat{p} - p \leqslant \varepsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \tag{2.70a}$$

$$\hat{p} - p \geqslant -\varepsilon = -\sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \tag{2.70b}$$

$$\hat{p} - \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \leqslant p \leqslant \hat{p} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \tag{2.70c}$$

Note that $\ln\frac{1}{\delta} = \ln\delta^{-1} = \ln e^{2n\varepsilon^2} = 2n\varepsilon^2$, and $\sqrt{\frac{\ln\frac{1}{\delta}}{2n}} = \varepsilon$.

Lemma 2.20 (Refined Pinsker's inequality)**.**

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}}\,. \tag{2.71}$$

Corollary 2.21 (Refined Pinsker's inequality)**.** *If $q > p$ then*

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2q}\,. \tag{2.72}$$

Corollary 2.22 (Refined Pinsker's inequality)**.** *If $\mathbf{kl}(p\|q) \leqslant \varepsilon$, then*

$$q \leqslant p + \sqrt{2p\varepsilon} + 2\varepsilon\,. \tag{2.73}$$

By applying Corollary 2.22 to inequality (2.59) we obtain that with probability greater than $(1-\delta)$:

$$p \leqslant \hat{p} + \sqrt{\frac{2\hat{p}\ln\frac{n+1}{\delta}}{n}} + \frac{2\ln\frac{n+1}{\delta}}{n}\,. \tag{2.74}$$

Note that when $\hat{p}$ is close to zero, the latter inequality is much tighter than Hoeffding's inequality. Finally, we note that although there is no analytic inversion of $\mathbf{kl}(\hat{p}\|p)$ it is possible to invert it numerically to obtain even tighter bounds than the relaxations above. Additionally, the bound in Theorem 2.17 can be improved slightly.

A: According to Eq. (2.59) we know that $\delta = (n+1)e^{-n\varepsilon}$ and with probability greater than $(1-\delta)$, we have $\mathbf{kl}(\hat{p}\|p) \leqslant \frac{\ln \frac{n+1}{\delta}}{n} = \frac{\ln e^{n\varepsilon}}{n} = \varepsilon$. Therefore, we have $e^{-n\varepsilon} = \frac{\delta}{n+1}$ and $-n\varepsilon = \ln \frac{\delta}{n+1}$ and $\varepsilon = -\frac{1}{n}\ln\frac{\delta}{n+1} = \frac{1}{n}\ln\frac{n+1}{\delta}$. Thus, if $\mathbf{kl}(\hat{p}\|q) \leqslant \varepsilon$, then

$$p \leqslant \hat{p} + \sqrt{2\hat{p}\tfrac{1}{n}\ln\tfrac{n+1}{\delta}} + 2\tfrac{1}{n}\ln\tfrac{n+1}{\delta} = \hat{p} + \sqrt{\tfrac{2\hat{p}\ln\frac{n+1}{\delta}}{n}} + \tfrac{2\ln\frac{n+1}{\delta}}{n}. \tag{2.75}$$

We have known that

$$\mathbf{KL}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2 \tag{2.76a}$$

$$\mathbf{kl}(p\|q) \geqslant 2(p-q)^2 = \tfrac{(p-q)^2}{\frac{1}{2}} \tag{2.76b}$$

$$\because x \in [0,1] \qquad \therefore \tfrac{1}{x} \in [1,+\infty) \tag{2.76c}$$

$$\tfrac{1}{\max\{p,q\}},\ \tfrac{1}{\max\{(1-p),(1-q)\}} \in [1,+\infty) \tag{2.76d}$$

$$\tfrac{(p-q)^2}{2\max\{p,q\}} + \tfrac{(p-q)^2}{2\max\{(1-p),(1-q)\}} \geqslant 2\tfrac{(p-q)^2}{2} = (p-q)^2 \tag{2.76e}$$

due to $p, q, (1-p), (1-q) \in [0,1]$.

2.6 Sampling Without Replacement

Let $X_1, ..., X_n$ be a sequence of random variables *sampled without replacement* from a finite set of values $\mathcal{X} = \{x_1, ..., x_N\}$ of size N. The random variables $X_1, ..., X_n$ are dependent. For example, if $\mathcal{X} = \{-1, +1\}$ and we sample two values then $X_1 = -X_2$. Since $X_1, ..., X_n$ are dependent, the concentration results from previous sections do not apply directly. However, the following result by Hoeffding, which we cite without a proof, allows to extend results for sampling with replacement to sampling without replacement.

Lemma 2.23. *Let $X_1, ..., X_n$ denote a random sample without replacement from a finite set $\mathcal{X} = \{x_1, ..., x_N\}$ of N real values. Let $Y_1, ..., Y_n$ denote a random sample with replacement from $\mathcal{X}$. Then for any continuous and convex function $f : \mathbb{R} \to \mathbb{R}$,*

$$\mathbf{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \leqslant \mathbf{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right]. \tag{2.77}$$

In particular, the lemma can be used to prove Hoeffding's inequality for sampling without replacement.

Theorem 2.24 (Hoeffding's inequality for sampling without replacement)**.** *Let $X_1, ..., X_n$ denote a random sample without replacement from a finite set $\mathcal{X} = \{x_1, ..., x_N\}$ of N values, where each element x_i is in the $[0,1]$ interval. Let $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the average of the values in $\mathcal{X}$. Then for all $\varepsilon > 0$,*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2}, \tag{2.78a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) = \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leqslant -\varepsilon\right) \leqslant e^{-2n\varepsilon^2}. \tag{2.78b}$$

The proof is a minor adaptation of the proof of Hoeffding's inequality for sampling with replacement using Lemma 2.23 and is left as an exercise. (Note that it requires a small modification inside the proof, because Lemma 2.23 cannot be applied directly to the statement of Hoeffding's inequality.)

While formal proof requires a bit of work, intuitively the result is quite expected. Imagine the process of sampling without replacement. If the average of points sampled so far starts deviating from the mean of the values in $\mathcal{X}$, the average of points that are left in $\mathcal{X}$ deviates in the opposite direction and "applies extra force" to new samples to bring the average back to μ. In the limit when $n = N$ we are guaranteed to have the average of X_i-s being equal to μ.

A: I don't quite understand that. What does it mean? opposite direction

3 Appendix B: Probability Theory Basics

Theorem 3.1 (Theorem B.26). *If $X_1, ..., X_n$ are independent random variables then*

$$\mathbf{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{V}[X_i]\,. \tag{3.1}$$

Proof. We have already known that

$$\mathbf{V}[X] = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2\,, \tag{3.2a}$$
$$\mathbf{E}[cX] = c\mathbf{E}[X]\,, \tag{3.2b}$$
$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]\,, \tag{3.2c}$$
$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]\,. \tag{3.2d}$$

Then we could obtain that

$$\begin{aligned}
\mathbf{V}\left[\sum_{i=1}^{n} X_i\right] &= \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)^2\right] \\
&= \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbf{E}[X_i]\right)^2\right] = \mathbf{E}\left[\left(\sum_{i=1}^{n} \Big(X_i - \mathbf{E}[X_i]\Big)\right)^2\right] && \text{(3.3a)} \\
&= \mathbf{E}\Big[(g)^2\Big] && \text{(3.3b)} \\
\mathbf{V}\left[\sum_{i=1}^{n} X_i\right] &= \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)^2 = \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\sum_{i=1}^{n} \mathbf{E}[X_i]\right)^2 \\
\sum_{i=1}^{n} \mathbf{V}[X_i] &= \sum_{i=1}^{n} \mathbf{E}\Big[\big(X_i - \mathbf{E}[X_i]\big)^2\Big] = \sum_{i=1}^{n} \mathbf{E}\Big[X_i^2 - 2X_i\mathbf{E}[X_i] + \big(\mathbf{E}[X_i]\big)^2\Big] \\
&= \sum_{i=1}^{n} \Big(\mathbf{E}[X_i^2] - \big(\mathbf{E}[X_i]\big)^2\Big) && \text{(3.3c)}
\end{aligned}$$

Eq. (3.3a) not necessarily independent.

$$\begin{aligned}
\mathbf{V}[X_i + X_i] &= \mathbf{E}\Big[\big(X_i + X_i - \mathbf{E}[X_i] - \mathbf{E}[X_i]\big)^2\Big] = \mathbf{E}\Big[4\,\big(X_i - \mathbf{E}[X_i]\big)^2\Big] = 4\mathbf{V}[X_i] && \text{(3.4a)} \\
\mathbf{V}[X_i + X_j] &= \mathbf{E}\Big[\big(X_i + X_j - \mathbf{E}[X_i] - \mathbf{E}[X_j]\big)^2\Big] = \mathbf{E}\Big[\big(\big(X_i - \mathbf{E}[X_i]\big) + \big(X_j - \mathbf{E}[X_j]\big)\big)^2\Big] \\
&= \mathbf{E}\left[\Big(X_i - \mathbf{E}[X_i]\Big)^2 + \Big(X_j - \mathbf{E}[X_j]\Big)^2 + 2\Big(X_i - \mathbf{E}[X_i]\Big)\Big(X_j - \mathbf{E}[X_j]\Big)\right] && \text{(3.4b)} \\
&= \mathbf{E}\left[\Big(X_i - \mathbf{E}[X_i]\Big)^2\right] + \mathbf{E}\left[\Big(X_j - \mathbf{E}[X_j]\Big)^2\right] + 2\mathbf{E}\left[X_iX_j - X_i\mathbf{E}[X_j] - X_j\mathbf{E}[X_i] + \mathbf{E}[X_i]\mathbf{E}[X_j]\right] \\
&= \mathbf{V}[X_i] + \mathbf{V}[X_j] + 2\left(\mathbf{E}[X_iX_j] - 2\mathbf{E}[X_i]\mathbf{E}[X_j] + \mathbf{E}[X_i]\mathbf{E}[X_j]\right) && \text{(3.4c)} \\
&= \mathbf{V}[X_i] + \mathbf{V}[X_j] + 2\left(\mathbf{E}[X_iX_j] - \mathbf{E}[X_i]\mathbf{E}[X_j]\right) = \mathbf{V}[X_i] + \mathbf{V}[X_j] && \text{(3.4d)}
\end{aligned}$$

Induction (1) When $n = 1$ and $n = 2$, the equality is satisfied. (2) If the equality satisfied when $n = t$, that is,

$$\mathbf{V}\left[\sum_{i=1}^{t} X_i\right] = \sum_{i=1}^{t} \mathbf{V}[X_i]\,, \tag{3.5}$$

then when $n = t + 1$,

$$\begin{aligned}
Eq.^{\text{left}} &= \mathbf{V}\left[\sum_{i=1}^{t} X_i + X_{t+1}\right] = \mathbf{E}\left[\left(\sum_{i=1}^{t} X_i + X_{t+1} - \mathbf{E}\left[\sum_{i=1}^{t} X_i + X_{t+1}\right]\right)^2\right] && \text{(3.6a)}\\
&= \mathbf{E}\left[\left(\sum_{i=1}^{t} X_i - \mathbf{E}\left[\sum_{i=1}^{t} X_i\right]\right)^2 + \left(X_{t+1} - \mathbf{E}[X_{t+1}]\right)^2 + 2\left(\sum_{i=1}^{t} X_i - \mathbf{E}\left[\sum_{i=1}^{t} X_i\right]\right)\left(X_{t+1} - \mathbf{E}[X_{t+1}]\right)\right]\\
&= \sum_{i=1}^{t} \mathbf{V}[X_i] + \mathbf{V}[X_{t+1}] + 2\mathbf{E}\left[X_{t+1}\sum_{i=1}^{t} X_i - \mathbf{E}[X_{t+1}]\sum_{i=1}^{t} X_i - \mathbf{E}\left[\sum_{i=1}^{t} X_i\right] X_{t+1} + \mathbf{E}\left[\sum_{i=1}^{t} X_i\right]\mathbf{E}[X_{t+1}]\right]\\
&= \sum_{i=1}^{t+1} \mathbf{V}[X_i] + 2\left(\mathbf{E}\left[X_{t+1}\sum_{i=1}^{t} X_i\right] - \mathbf{E}\left[\sum_{i=1}^{t} X_i\right]\mathbf{E}[X_{t+1}]\right) && \text{(3.6b)}\\
&= \sum_{i=1}^{t+1} \mathbf{V}[X_i] + 2\left(\mathbf{E}[X_{t+1}]\mathbf{E}\left[\sum_{i=1}^{t} X_i\right] - \mathbf{E}\left[\sum_{i=1}^{t} X_i\right]\mathbf{E}[X_{t+1}]\right) = \sum_{i=1}^{t+1} \mathbf{V}[X_i] && \text{(3.6c)}
\end{aligned}$$

Eq. (3.6b) because X_{t+1} and $\sum_{i=1}^{t} X_i$ are independent random variables (after combination). (3) In conclusion, the equality is satisfied for all $n \in \mathbb{Z}^+$. □