Machine Learning Notes

This note is based on Machine Learning at DIKU

4 December 2021

Chapter 2. Concentration of Measure Inequalities

Concentration of measure inequalities are one of the main tools for analysing learning algorithms. This chapter is devoted to a number of concentration of measure inequalities that form the basis for the results discussed in later chapters.

2.1 Markov's Inequality

Markov's inequality is the simplest and relatively weak concentration 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:

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

NB. $Y = \mathbb{I}\left({}^{X}/_{\varepsilon} \geqslant 1\right) \in [0, 1]$, and that ${}^{X}/_{\varepsilon} \in [0, +\infty)$ 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 $\mathbf{Pr}\left(X \geqslant \frac{1}{\delta}\mathbf{E}[X]\right) \leqslant \frac{\mathbf{E}[X]}{\delta \mathbf{E}[X]} = \delta$. ■

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,

$$\mathbf{Pr}\left(X \geqslant \frac{1}{\delta}\mathbf{E}[X]\right) \leqslant \varepsilon. \tag{2.3}$$

We note that even though Markov's inequality is weak, there are situations in which it is tight.

2.2 Chebyshev's Inequality

Our next stop is Chebyshev's inequality, which exploits variance to obtain tighter concentration.

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

$$\mathbf{Pr}(|X - \mathbf{E}[X]| \geqslant \varepsilon) \leqslant \frac{\mathbf{Var}[X]}{\varepsilon^2}. \tag{2.4}$$

Proof. The proof uses a transformation of a random variable. We have that $\mathbf{Pr}(|X - \mathbf{E}[X]| \geqslant \varepsilon) = \mathbf{Pr}\big((X - \mathbf{E}[X])^2 \geqslant \varepsilon^2\big)$, 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

$$\mathbf{Pr}(|X - \mathbf{E}[X]| \geqslant \varepsilon) = \mathbf{Pr}\left((X - \mathbf{E}[X])^2 \geqslant \varepsilon^2\right) \leqslant \frac{\mathbf{E}\left[(X-\mathbf{E}[X])^2\right]}{\varepsilon^2} = \frac{\mathbf{Var}[X]}{\varepsilon^2}\,. \tag{2.5}$$

NB. By definition, for a random variable X, its expectation is given by $\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x\mathbf{Pr}(X = x)$, and its variance is defined by $\mathbf{Var}[X] = \mathbf{E}\left[\left(X - \mathbf{E}[X]\right)^2\right] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$. ■

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[1] we have $\mathbf{Var}[\hat{\mu}_n] = \frac{1}{n^2}\mathbf{Var}[n\hat{\mu}_n] = \frac{1}{n^2}\sum_{i=1}^n \mathbf{Var}[X_i] = \frac{1}{n}\mathbf{Var}[X_1]$. 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

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

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.

[1] We cite Theorem B.26 here.

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

$$\mathbf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{Var}[X_i]\,. \tag{2.6}$$

The proof is based on Theorem B.23 and the result does not necessarily hold when X_i-s are not independent.

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

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]\,. \tag{2.7}$$

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).

Theorem (B.22 Linearity). *For any pair of random variables X and Y, not necessarily independent,*

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

Machine Learning Notes

This note is based on Machine Learning at DIKU

5 December 2021

Chapter 2. Concentration of Measure Inequalities

2.3 Hoeffding's Inequality

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

Theorem 2.3 (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 $\varepsilon > 0$,*

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

and

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

Corollary 2.4. *Under the assumptions of Theorem 2.3,*

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

By taking a union bound of the events in Eqs. (2.11) and (2.12) we obtain Corollary 2.4. Note that Eqs. (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 Corollary 2.5.

NB. Let the event $\mathcal{A}$ represent $\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]$, then we have $\mathbf{Pr}(\mathcal{A}_+) \stackrel{\text{def}}{=} \mathbf{Pr}(\mathcal{A} \geqslant \varepsilon)$ and $\mathbf{Pr}(\mathcal{A}_-) \stackrel{\text{def}}{=} \mathbf{Pr}(\mathcal{A} \leqslant -\varepsilon)$. Therefore,

$$\mathbf{Pr}(|\mathcal{A}| \geqslant \varepsilon) = \mathbf{Pr}\big((\mathcal{A} \geqslant \varepsilon) \vee (\mathcal{A} \leqslant -\varepsilon)\big) = \mathbf{Pr}(\mathcal{A} \geqslant \varepsilon) + \mathbf{Pr}(\mathcal{A} \leqslant -\varepsilon) \leqslant 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.14}$$

Corollary 2.5. *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

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

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.

NB. To be specific, if $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ for all i. Let $Y_i = \frac{1}{n}X_i$ for all i, then the event

$$\mathcal{A} \stackrel{\text{def}}{=} \sum_{i=1}^n Y_i - \mathbf{E}\left[\sum_{i=1}^n Y_i\right] = \frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mathbf{E}[X_i] = \frac{1}{n}\sum_{i=1}^n X_i - \mu\,, \tag{2.17}$$

where $\mathbf{E}\left[\sum_{i=1}^n \left(\frac{1}{n}X_i\right)\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[X_i] = \mu$. Thus $Y_i \in [0, \frac{1}{n}]$ for all i, and we could obtain that

$$\mathbf{Pr}(\mathcal{A} \geqslant \varepsilon) = \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (\frac{1}{n}-0)^2}\right) = \exp\left(\frac{-2\varepsilon^2}{\frac{1}{n^2}\sum_{i=1}^n 1}\right) = e^{-2n\varepsilon^2}\,, \tag{2.18a}$$

$$\mathbf{Pr}(\mathcal{A} \leqslant -\varepsilon) = \mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (\frac{1}{n}-0)^2}\right) = \exp\left(\frac{-2\varepsilon^2}{\frac{1}{n}}\right) = \exp(-2n\varepsilon^2)\,. \tag{2.18b}$$

Lemma 2.6 (Hoeffding's Lemma)**.** *Let X be a random variable, such that $X \in [a,b]$. Then for any $\lambda \in \mathbb{R}$,*

$$\mathbf{E}\left[e^{\lambda X}\right] \leqslant \exp\left(\lambda \mathbf{E}[X] + \frac{\lambda^2(b-a)^2}{8}\right)\,. \tag{2.19}$$

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 2.3.

Proof of Theorem 2.3. We prove the first inequality in Theorem 2.3. The second inequality follows by applying the first inequality to $-X_1, ..., -X_n$. The proof is based on Chernoff's bounding technique. For any $\lambda > 0$, the following holds:

$$\begin{aligned}\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) &= \mathbf{Pr}\left(\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right) \geqslant e^{\lambda\varepsilon}\right) \\ &\leqslant \frac{\mathbf{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)\right]}{e^{\lambda\varepsilon}}, \end{aligned} \tag{2.20}$$

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:

$$\begin{aligned}\mathbf{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)\right] &= \mathbf{E}\left[\exp\left(\sum_{i=1}^{n} \lambda\big(X_i - \mathbf{E}[X_i]\big)\right)\right] = \mathbf{E}\left[\prod_{i=1}^{n} e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \\ &= \prod_{i=1}^{n} \mathbf{E}\left[e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \qquad (2.21) \\ &\leqslant \prod_{i=1}^{n} e^{\lambda^2 (b_i - a_i)^2/8} \qquad (2.22) \\ &= \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right), \end{aligned}$$

where (2.21) holds since $X_1, ..., X_n$ are independent and (2.22) 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!*

NB. Note that $e^x > 0$ for any $x \in \mathbb{R}$ and $e^0 = 1$. For a random variable X, let $Y = X - \mathbf{E}[X]$ such that $Y \in [a, b]$. Thus we get $\mathbf{E}[Y] = \mathbf{E}[X - \mathbf{E}[X]] = \mathbf{E}[X] - \mathbf{E}[X] = 0$. Therefore, for any $\lambda \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda\left(X - \mathbf{E}[X]\right)}\right] \leqslant \exp\left(\lambda\mathbf{E}[Y] + \tfrac{\lambda^2(b-a)^2}{8}\right) = e^0 \exp\left(\tfrac{\lambda^2(b-a)^2}{8}\right) = \exp\left(\tfrac{\lambda^2(b-a)^2}{8}\right). \tag{2.23}$$

Let $Y = \sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]$ and $Y \in [a, b]$, then

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mathbf{E}[X_i]\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] - \sum_{i=1}^{n}\mathbf{E}[X_i] = 0, \tag{2.24}$$

due to the independence of $X_1, ..., X_n$. Then for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbf{E}[e^{\lambda Y}] = \mathbf{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}[\sum_{i=1}^{n} X_i]\right)}\right] &= \mathbf{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mathbf{E}[X_i]\right)\right)\right] \\ &= \mathbf{E}\left[\exp\left(\sum_{i=1}^{n}\lambda\big(X_i - \mathbf{E}[X_i]\big)\right)\right] = \mathbf{E}\left[\prod_{i=1}^{n} e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \\ &= \prod_{i=1}^{n}\mathbf{E}\left[e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \leqslant \prod_{i=1}^{n}\exp\left(\lambda\mathbf{E}\left[X_i - \mathbf{E}[X_i]\right] + \tfrac{\lambda^2(b_i - a_i)^2}{8}\right) \qquad (2.25a) \\ &= \prod_{i=1}^{n} e^{\lambda\cdot 0}\cdot\exp\left(\frac{\lambda^2\left((b_i - \mu_i) - (a_i - \mu_i)\right)^2}{8}\right) = \prod_{i=1}^{n} e^{\lambda^2(b_i - a_i)^2/8} \qquad (2.25b) \\ &= \exp\left(\sum_{i=1}^{n}\tfrac{\lambda^2(b_i - a_i)^2}{8}\right) = \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right), \end{aligned}$$

where $\mu_i \stackrel{\text{def}}{=} \mathbf{E}[X_i]$ and $X_i - \mathbf{E}[X_i] \in [a_i - \mu_i, b_i - \mu_i]$ for all i.

To complete the proof we substitute the bound on the expectation into the previous calculation and obtain:

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right). \tag{2.26}$$

This expression is minimised by

$$\lambda^* = \operatorname*{argmin}_{\lambda} \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right) = \operatorname*{argmin}_{\lambda}\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right) = \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}. \tag{2.27}$$

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.

NB. According to Theorem 2.1,

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) = \mathbf{Pr}\left(e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)} \geqslant e^{\lambda\varepsilon}\right) \leqslant \frac{\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)}{e^{\lambda\varepsilon}} \tag{2.28a}$$

$$\leqslant \frac{1}{e^{\lambda\varepsilon}}\exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right) = \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right). \tag{2.28b}$$

Let $f(\lambda) \stackrel{\text{def}}{=} \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon$ and $AB \stackrel{\text{def}}{=} \sum_{i=1}^{n}(b_i - a_i)^2$ for brevity, then

$$8f(\lambda) = \sum_{i=1}^{n}(b_i - a_i)^2\lambda^2 - 8\varepsilon\lambda = \sum_{i=1}^{n}(b_i - a_i)^2\left(\lambda^2 - \frac{8\varepsilon}{AB}\lambda\right) = AB\left(\lambda^2 - \frac{8\varepsilon}{AB}\lambda + \frac{16\varepsilon^2}{AB^2} - \frac{16\varepsilon^2}{AB^2}\right)$$

$$= \sum_{i=1}^{n}(b_i - a_i)^2\left(\lambda - \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)^2 - \frac{16\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}, \tag{2.29a}$$

$$\lambda^* = \operatorname*{argmin}_{\lambda} f(\lambda) = \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2} \quad \text{where} \quad \min_{\lambda} f(\lambda) = \frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}. \tag{2.29b}$$

Finally,

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right) = \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.30}$$

By substituting λ^* into the calculation we obtain the result of the theorem. ■

Proof of Lemma 2.6. Note that

$$\mathbf{E}\left[e^{\lambda X}\right] = \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)+\lambda\mathbf{E}[X]}\right] = e^{\lambda\mathbf{E}[X]} \times \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right]. \tag{2.31}$$

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

$$\mathbf{E}\left[e^{\lambda Z}\right] \leqslant e^{\lambda^2(b-a)^2/8}. \tag{2.32}$$

NB. Because X is a random variable such that $X \in [a,b]$, let $\mu \stackrel{\text{def}}{=} \mathbf{E}[X]$ for brevity . Then $\mu \in [a,b]$, and for any $\lambda \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda X}\right] = \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]+\mathbf{E}[X]\right)}\right] = \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)} \cdot e^{\lambda\mathbf{E}[X]}\right] = e^{\lambda\mathbf{E}[X]}\mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right]. \tag{2.33}$$

Let $Z \stackrel{\text{def}}{=} X - \mathbf{E}[X]$, then $\mathbf{E}[Z] = \mathbf{E}[X - \mathbf{E}[X]] = \mathbf{E}[X] - \mathbf{E}[X] = 0$. As $X \in [a,b]$, we have $Z \in [a-\mu, b-\mu]$, and $\lambda Z \in [\lambda(a-\mu), \lambda(b-\mu)]$. Therefore, let $\lambda z \stackrel{\text{def}}{=} t\lambda(a-\mu) + (1-t)\lambda(b-\mu)$, that is, $z = t(a-\mu) + (1-t)(b-\mu) = t(a-b) + (b-\mu)$ we could obtain that $t = \frac{b-\mu-z}{b-a} = \frac{b-\mu-z}{b-a}$ and that $1-t = \frac{b-\mu-(a-\mu)}{b-a} - \frac{b-\mu-z}{b-a} = \frac{z-(a-\mu)}{b-a}$. Therefore, let $\hat{a} = a-\mu$ and $\hat{b} = b-\mu$, then

$$e^{\lambda z} = e^{\lambda\left(t(a-\mu)+(1-t)(b-\mu)\right)} \leqslant te^{\lambda(a-\mu)} + (1-t)e^{\lambda(b-\mu)} = \tfrac{b-\mu-z}{b-a}e^{\lambda(a-\mu)} + \tfrac{z-(a-\mu)}{b-a}e^{\lambda(b-\mu)}, \tag{2.34a}$$

$$\mathbf{E}\left[e^{\lambda z}\right] = \tfrac{\hat{b}-z}{\hat{b}-\hat{a}}e^{\lambda\hat{a}} + \tfrac{z-\hat{a}}{\hat{b}-\hat{a}}e^{\lambda\hat{b}} = \tfrac{\hat{b}-\mathbf{E}[z]}{b-a}e^{\lambda\hat{a}} + \tfrac{\mathbf{E}[z]-\hat{a}}{b-a}e^{\lambda\hat{b}} = \tfrac{b-\mu}{b-a}e^{\lambda(a-\mu)} + \tfrac{-(a-\mu)}{b-a}e^{\lambda(b-\mu)}. \tag{2.34b}$$

Let $p = \frac{-\hat{a}}{b-a}$ and $u = \lambda(b-a)$, then $1-p = \frac{\hat{b}}{b-a}$, and

$$\begin{aligned}\mathbf{E}\left[e^{\lambda z}\right] &= (1-p)e^{\lambda\hat{a}} + pe^{\lambda\hat{b}} = e^{\lambda\hat{a}}\left(1-p+pe^{\lambda\hat{b}-\lambda\hat{a}}\right) = e^{\lambda\frac{-\hat{a}}{b-a}(-1)(b-a)}\left(1-p+pe^{\lambda(\hat{b}-\hat{a})}\right)\\ &= e^{-p\lambda(b-a)}\left(1-p+pe^{\lambda(b-a)}\right) = e^{-pu}\left(1-p+pe^{u}\right) = e^{-pu+\ln(1-p+pe^u)} = e^{\phi(u)} \leqslant e^{\frac{\lambda^2(b-a)^2}{8}}.\end{aligned} \tag{2.35}$$

To summarise, $\mathbf{E}\left[e^{\lambda X}\right] = e^{\lambda\mathbf{E}[X]}\mathbf{E}\left[e^{\lambda Z}\right] \leqslant e^{\lambda\mathbf{E}[X]+\lambda^2(b-a)^2/8}$. Q.E.D.

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

$$e^{\lambda z} \leqslant \frac{z-a}{b-a}e^{\lambda b} + \frac{b-z}{b-a}e^{\lambda a}. \tag{2.36}$$

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

$$\begin{aligned}\mathbf{E}\left[e^{\lambda Z}\right] &\leqslant \mathbf{E}\left[\frac{Z-a}{b-a}e^{\lambda b} + \frac{b-Z}{b-a}e^{\lambda a}\right] = \frac{\mathbf{E}[Z]-a}{b-a}e^{\lambda b} + \frac{b-\mathbf{E}[Z]}{b-a}e^{\lambda a}\\ &= \frac{-a}{b-a}e^{\lambda b} + \frac{b}{b-a}e^{\lambda a} = \left(1-p+pe^{\lambda(b-a)}\right)e^{-p\lambda(b-a)} = e^{\phi(u)},\end{aligned} \tag{2.37a}$$

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

NB. By convexity[2] of the exponential function e^z, for $z \in [a,b]$, we have: for $t \in (0,1)$, let $z \stackrel{\text{def}}{=} ta + (1-t)b = t(a-b) + b$, then $t = \frac{b-z}{b-a}$ and $1-t = \frac{b-a}{b-a} - \frac{b-z}{b-a} = \frac{z-a}{b-a}$, thus,

$$e^z = e^{ta+(1-t)b} \leqslant te^a + (1-t)e^b = \tfrac{b-z}{b-a}e^a + \tfrac{z-a}{b-a}e^b. \tag{2.38}$$

[2] Wikipedia (Convex function). Let X be a convex subset of a real vector space and let $f : X \to \mathbb{R}$ be a function. Then f is called *convex* if and only if any of the following equivalent conditions hold:

1) For all $0 \leqslant t \leqslant 1$ and all $x_1, x_2 \in \mathcal{X}$: $f(tx_1 + (1-t)x_2) \leqslant tf(x_1) + (1-t)f(x_2)$. The right hand side represents

Let $z' = \lambda z \in [\lambda a, \lambda b]$, then $z' \overset{\text{def}}{=} t(\lambda a) + (1-t)(\lambda b) = \lambda\big(t(a-b)+b\big) = \lambda z$. Thus, $t = \frac{b-\frac{1}{\lambda}z'}{b-a} = \frac{b-z}{b-a}$, and $1-t = \frac{\frac{1}{\lambda}z'-a}{b-a} = \frac{z-a}{b-a}$. Then we obtain that

$$e^{z'} = e^{\lambda z} = e^{t\lambda a+(1-t)\lambda b} \leqslant te^{\lambda a} + (1-t)e^{\lambda b} = \tfrac{b-z}{b-a}e^{\lambda a} + \tfrac{z-a}{b-a}e^{\lambda b}\,, \tag{2.39a}$$

$$\begin{aligned}\mathbf{E}\left[e^{\lambda z}\right] &\leqslant \mathbf{E}\left[\tfrac{b-z}{b-a}e^{\lambda a}\right] + \mathbf{E}\left[\tfrac{z-a}{b-a}e^{\lambda b}\right] = \mathbf{E}\left[\tfrac{b-z}{b-a}\right]e^{\lambda a} + \mathbf{E}\left[\tfrac{z-a}{b-a}\right]e^{\lambda b} \\ &= \tfrac{\mathbf{E}[b-z]}{b-a}e^{\lambda a} + \tfrac{\mathbf{E}[z-a]}{b-a}e^{\lambda b} = \tfrac{b-\mathbf{E}[z]}{b-a}e^{\lambda a} + \tfrac{\mathbf{E}[z]-a}{b-a}e^{\lambda b} = \tfrac{b}{b-a}e^{\lambda a} + \tfrac{-a}{b-a}e^{\lambda b}\,,\end{aligned} \tag{2.39b}$$

because of $\mathbf{E}[Z] = 0$. Let $p = -a/(b-a)$, then $1-p = 1-\frac{-a}{b-a} = b/(b-a)$, and

$$\mathbf{E}\left[e^{\lambda z}\right] \leqslant \frac{b}{b-a}e^{\lambda a} + \frac{-a}{b-a}e^{\lambda b} = (1-p)e^{\lambda a} + pe^{\lambda b} = e^{\lambda a}\left((1-p) + pe^{\lambda(b-a)}\right) \tag{2.40a}$$

$$= e^{-\frac{-a}{b-a}\lambda(b-a)}\left(1-p+pe^{\lambda(b-a)}\right) = e^{-p\lambda(b-a)}\left(1-p+pe^{\lambda(b-a)}\right)\,. \tag{2.40b}$$

Let $u = \lambda(b-a)$ and $\phi(u) = -pu + \ln(1-p+pe^u)$, then

$$\phi'(u) = -p + \tfrac{1}{1-p+pe^u}(pe^u) = -p + \tfrac{p}{p+(1-p)e^{-u}}\,, \tag{2.41a}$$

$$\phi''(u) = p(-1)\big(p+(1-p)e^{-u}\big)^{-2}(1-p)e^{-u}(-1) = \tfrac{p(1-p)e^{-u}}{\left(p+(1-p)e^{-u}\right)^2}\,, \tag{2.41b}$$

$$\phi(u) = -p\lambda(b-a) + \ln\left(1-p+pe^{\lambda(b-a)}\right)\,, \tag{2.41c}$$

$$\begin{aligned}e^{\phi(u)} &= e^{-p\lambda(b-a)}\cdot\left(1-p+pe^{\lambda(b-a)}\right) = e^{-\frac{-a}{b-a}\lambda(b-a)}\cdot\left(1-\tfrac{-a}{b-a}+\tfrac{-a}{b-a}e^{\lambda(b-a)}\right) \\ &= e^{\lambda a}\left(\tfrac{b}{b-a}+\tfrac{-a}{b-a}e^{\lambda(b-a)}\right) = \tfrac{b}{b-a}e^{\lambda a} + \tfrac{-a}{b-a}e^{\lambda b} = e^{\lambda a}\left(1-p+pe^{\lambda b-\lambda a}\right) = (1-p)e^{\lambda a} + pe^{\lambda b}\,.\end{aligned} \tag{2.41d}$$

---

the straight line between $(x_1, f(x_1))$ and $(x_2, f(x_2))$ in the graph of f as a function of t; increasing t from 0 to 1 or decreasing t from 1 to 0 sweeps this line. Similarly, the argument of the function f in the left hand side represents the straight line between x_1 and x_2 in X or the x-axis of the graph of f. So, this condition requires that the straight line between any pair of points on the curve of f to be above or just meets the graph.

2) For all $0 < t < 1$ and all $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$:

$$f\big(tx_1 + (1-t)x_2\big) \leqslant tf(x_1) + (1-t)f(x_2)\,.$$

The difference of this second condition with respect to the first condition above is that this condition does not include the intersection points (e.g., $(x_1, f(x_1))$ and $(x_2, f(x_2))$) between the straight line passing through a pair of points on the curve of f (the straight line is represented by the right hand side of this condition) and the curve of f; the first condition includes the intersection points as it becomes $f(x_1) \leqslant f(x_1)$ or $f(x_2) \leqslant f(x_2)$ at $t = 0$ or 1, or $x_1 = x_2$. In fact, the intersection points do not need to be considered in a condition of convex using $f(tx_1 + (1-t)x_2) \leqslant tf(x_1) + (1-t)f(x_2)$ because $f(x_1) \leqslant f(x_1)$ and $f(x_2) \leqslant f(x_2)$ are always true (so not useful to be a part of a condition).

The second statement characterising convex functions that are valued in the real line $\mathbb{R}$ is also the statement used to define *convex functions* that are valued in the extended real number line $[-\infty, \infty] = \mathbb{R} \cup \{\pm\infty\}$, where such a function f is allowed to (but is not required to) take $\pm\infty$ as a value. The first statement is not used because it permits t to take 0 or 1 as a value, in which case, if $f(x_1) = \pm\infty$ or $f(x_2) = \pm\infty$, respectively, then $tf(x_1) + (1-t)f(x_2)$ would be undefined (because the multiplications $0 \cdot \infty$ and $0 \cdot (-\infty)$ are undefined). Then sum $-\infty + \infty$ is also undefined so a convex extended real-valued function is typically only allowed to take exactly one of $-\infty$ and $+\infty$ as a value.

The second statement can also be modified to get the definition of *strict convexity*, where the latter is obtained by replacing $\leqslant$ with the strict inequality $<$. Explicitly, the map f is called *strictly convex* if and only if for all real $0 < t < 1$ and all $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$: $f\big(tx_1 + (1-t)x_2\big) < tf(x_1) + (1-t)f(x_2)$. A strictly convex function f is a function that the straight line between any pair of points on the curve f is above the curve f except for the intersection points between the straight line and the curve. The function f is said to be *concave* (resp. *strictly concave*) if $-f$ (f multiplied by -1) is convex (resp. strictly convex).

It is easy to verify that the derivative of ϕ is

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}}, \tag{2.42}$$

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

$$\phi''(u) = \frac{p(1-p)e^{-u}}{\left(p + (1-p)e^{-u}\right)^2} \leqslant \frac{1}{4}. \tag{2.43}$$

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:

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) = \frac{u^2}{2}\phi''(\theta) \leqslant \frac{u^2}{8} = \frac{\lambda^2(b-a)^2}{8}. \tag{2.44}$$

NB. As $u = \lambda(b-a)$, we have known that if let $u = 0$, then

$$\phi(0) = -pu + \ln(1 - p + pe^u)\big|_{u=0} = 0 + \ln(1 - p + p) = 0, \tag{2.45a}$$

$$\phi'(0) = -p + \frac{p}{p+(1-p)e^{-u}}\bigg|_{u=1} = -p + \frac{p}{p+(1-p)} = 0, \tag{2.45b}$$

$$\phi''(0) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2}\bigg|_{u=0} = \frac{p(1-p)}{(p+(1-p))^2} = p(1-p), \tag{2.45c}$$

$$\begin{aligned}\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} &= \frac{p(1-p)}{\left((1-p)\sqrt{e^{-u}} + \frac{p}{\sqrt{e^{-u}}}\right)^2} \leqslant \frac{p(1-p)}{\left(2\sqrt{(1-p)\sqrt{e^{-u}}\frac{p}{\sqrt{e^{-u}}}}\right)^2} \\ &= \frac{p(1-p)}{\left(2\sqrt{(1-p)p}\right)^2} = \frac{p(1-p)}{4(1-p)p} = \frac{1}{4},\end{aligned} \tag{2.45d}$$

due to $a + b \geqslant 2\sqrt{ab}$ for any $a, b \geqslant 0$.[3] At last, by Taylor's theorem[4,5] we have

$$\phi(u) = \phi(0) + u\phi'(0) + \tfrac{u^2}{2}\phi''(0) + \tfrac{u^3}{3!}\phi'''(\theta'), \quad \text{for some } \theta' \in [0, u] \tag{2.51a}$$

$$= \phi(0) + u\phi'(0) + \tfrac{u^2}{2}\phi''(\theta), \quad \text{for some } \theta \in [0, u] \tag{2.51b}$$

$$= 0 + u \cdot 0 + \tfrac{u^2}{2}\phi''(\theta) \leqslant 0 + \tfrac{u^2}{2} \cdot \tfrac{1}{4} = \tfrac{u^2}{8} = \tfrac{(\lambda(b-a))^2}{8} = \tfrac{\lambda^2(b-a)^2}{8}. \tag{2.51c}$$

[3] Note that $a + b = (\sqrt{a})^2 + (\sqrt{b})^2 = (\sqrt{a} + \sqrt{b})^2 - 2\sqrt{ab} = (\sqrt{a} - \sqrt{b})^2 + 2\sqrt{ab}$, therefore, $a + b \geqslant 2\sqrt{ab}$ for any $a, b \geqslant 0$.

[4] Wikipedia (Taylor's theorem). Let $k \geqslant 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be k-times differentiable at the point $a \in \mathbb{R}$. Then there exists a function $h_k : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k, \tag{2.46}$$

and $\lim_{x \to a} h_k(x) = 0$. This is called the Peano form of the remainder.

The polynomial appearing in Taylor's theorem is the *k-th order Taylor polynomial*

$$P_k(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k, \tag{2.47}$$

of the function f at the point a. The Taylor polynomial is the unique "asymptotic best fit" polynomial in the sense that if there exists a function $h_k : \mathbb{R} \to \mathbb{R}$ and a k-th order polynomial p such that $f(x) = p(x) + h_k(x)(x-a)^k$, $\lim_{x \to a} h_k(x) = 0$,then $p = P_k$. Taylor's theorem describes the asymptotic behavior of the *remainder term* $R_k(x) = f(x) - P_k(x)$,which is the approximation error when approximating f with its Taylor polynomial. Using the little-o notation, the statement in Taylor's theorem reads as $R_k(x) = \mathbf{o}(|x-a|^k)$, $x \to a$.

[5] **Explicit formulas for the remainder** Under stronger regularity assumptions on f there are several precise formulas for the remainder term R_k of the Taylor's polynomial, the most common ones being the following.

Remark (Mean-value forms of the remainder). *Let $f : \mathbb{R} \to \mathbb{R}$ be $(k+1)$ times differentiable on the open interval with*

To sum up, we have

$$\mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right] = \mathbf{E}[e^{\lambda z}] \leqslant e^{\phi(u)} \leqslant e^{\frac{\lambda^2(b-a)^2}{8}}, \tag{2.52a}$$

$$\mathbf{E}\left[e^{\lambda X}\right] = e^{\lambda \mathbf{E}[X]} \cdot \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right] \leqslant e^{\lambda \mathbf{E}[X]} e^{\frac{\lambda^2(b-a)^2}{8}} = e^{\lambda \mathbf{E}[X] + \frac{\lambda^2(b-a)^2}{8}}, \tag{2.52b}$$

that is, the inequality in Lemma 2.6 is demonstrated. ■

$f^{(k)}$ continuous on the closed interval between a and x. Then

$$R_k(x) = \frac{f^{(k+1)}(\xi_L)}{(k+1)!}(x-a)^{k+1}, \tag{2.48}$$

for some real number ξ_L between a and x. This is the Lagrange form *of the remainder. Similarly,*

$$R_k(x) = \frac{f^{(k+1)}(\xi_C)}{k!}(x-\xi_C)^k(x-a), \tag{2.49}$$

for some real number ξ_C between a and x. This is the Cauchy form *of the remainder.*

These refinements of Taylor's theorem are usually proved using the mean value theorem, whence the name. Also other similar expressions can be found. For example, if $G(t)$ is continuous on the closed interval and differentiable with a non-vanishing derivative on the open interval between a and x, then

$$R_k(x) = \frac{f^{(k+1)}(\xi)}{k!}(x-\xi)^k \frac{G(x)-G(a)}{G'(\xi)},$$

for some number ξ between a and x. This version covers the Lagrange and Cauchy forms of the remainder as special cases, and is proved below using Cauchy's mean value theorem. The statement for the integral form of the remainder is more advanced than the previous ones, and requires understanding of Lebesgue integration theory for the full generality. However, it holds also in the sense of Riemann integral provided the $(k+1)$-th derivative of f is continuous on the closed interval $[a, x]$.

Remark (Integral form of the remainder). *Let $f^{(k)}$ be absolutely continuous on the closed interval between a and x. Then*

$$R_k(x) = \int_a^x \frac{f^{(k+1)}(t)}{k!}(x-t)^k dt. \tag{2.50}$$

Due to absolute continuity of $f^{(k)}$ on the closed interval between a and x, its derivative $f^{(k+1)}$ exists as an L^1-function, and the result can be proven by a formal calculation using fundamental theorem of calculus and integration by parts.

Estimmates for the remainder It is often useful in practice to be able to estimate the remainder term appearing in the Taylor approximation, rather than having an exact formula for it. Suppose that f is $(k+1)$-times continuously differentiable in an interval ℓ containing a. Suppose that there are real constants q and Q such that $q \leqslant f^{(k+1)}(x) \leqslant Q$ throughout ℓ. Then the remainder term satisfies the inequality

$$q\frac{(x-a)^{k+1}}{(k+1)!} \leqslant R_k(x) \leqslant Q\frac{(x-a)^{k+1}}{(k+1)!},$$

if $x > a$, and a similar estimate if $x < a$. This is a simple consequence of the Lagrange form of the remainder. In particular, if $|f^{(k+1)}(x)| \leqslant M$ on an interval $\ell = (a-r, a+r)$ with some $r > 0$, then

$$|R_k(x)| \leqslant M\frac{|x-a|^{k+1}}{(k+1)!} \leqslant M\frac{r^{k+1}}{(k+1)!},$$

for all $x \in (a-r, a+r)$. The second inequality is called a uniform estimate, because it holds uniformly for all x on the interval $(a-r, a+r)$.

Machine Learning Notes

This note is based on Machine Learning at DIKU

5 December 5 2021

Chapter 2. Concentration of Measure Inequalities

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

$$\delta = 2\exp\left(-2n\varepsilon^2\right), \tag{2.53a}$$

$$\varepsilon = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}, \tag{2.53b}$$

$$n = \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}. \tag{2.53c}$$

Overall, Hoeffding's inequality tells by how much the empirical 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 analyse 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 $\frac{1}{n}\sum_{i=1}^{n} X_i$.

NB. According to Theorem 2.3–Lemma 2.6 [6], we have: let $\delta \overset{\text{def}}{=} 2e^{-2n\varepsilon^2}$, then $-2n\varepsilon^2 = \ln(\delta/2)$ and $2n\varepsilon^2 = -\ln(\delta/2) = \ln(\delta/2)^{-1} = \ln(2/\delta)$. Thus $\varepsilon = \left(\frac{1}{2n}\ln\frac{2}{\delta}\right)^{1/2}$ and $n = \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}$. As ε increases,

[6] We re-list them here.

Theorem (2.3 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 $\varepsilon > 0$: "one-sided Hoeffding's inequalities" hold, i.e.,*

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right), \tag{2.54}$$

and

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \leqslant -\varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.55}$$

Remark (Corollary 2.4). *Under the assumptions of Theorem 2.3: "two-sided Hoeffding's inequality" holds, that is,*

$$\mathbf{Pr}\left(\left|\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.56}$$

confidence δ will decrease; as δ increases, precision ε will decrease.

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 (2.57) and (2.58) with a fixed confidence in mind, thus we have

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leqslant \delta\,, \tag{2.60a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leqslant \delta\,, \tag{2.60b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \leqslant \delta\,. \tag{2.60c}$$

(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.)

NB. Let $\delta \overset{\text{def}}{=} 2e^{-2n\varepsilon^2}$, then $\varepsilon = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$ and $n = \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}$. If we use them as substitutes in Eqs. (2.57–2.58), we could obtain that

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}\right) \leqslant \frac{1}{2}\delta\,, \tag{2.61a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}\right) \leqslant \frac{1}{2}\delta\,, \tag{2.61b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}\right) \leqslant \delta\,. \tag{2.61c}$$

If let $\delta \overset{\text{def}}{=} e^{-2n\varepsilon^2}$, then $n\varepsilon^2 = -\frac{1}{2}\ln\delta = \ln(\frac{1}{\delta})^{1/2} = \ln\sqrt{\frac{1}{\delta}}$. Thus $\varepsilon = \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}$ and $n = \frac{1}{2\varepsilon^2}\ln\frac{1}{\delta}$. Then we could obtain that

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.62a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.62b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\right) \leqslant 2\delta\,. \tag{2.62c}$$

Remark (Corollary 2.5). *Let $X_1, \ldots, X_n$ be independent random variables, such that $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$:*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2} = \exp\left(-2n\varepsilon^2\right)\,, \tag{2.57}$$

and

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leqslant -\varepsilon\right) = \mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2} = \exp\left(-2n\varepsilon^2\right)\,. \tag{2.58}$$

Remark (Lemma 2.6, Hoeffding's lemma). *Let X be a random variable, such that $X \in [a,b]$. Then for any $\lambda \in \mathbb{R}$:*

$$\mathbf{E}\left[e^{\lambda X}\right] \leqslant e^{\lambda\mathbf{E}[X] + \frac{\lambda^2(b-a)^2}{8}} = e^{\lambda\mathbf{E}[X] + \lambda^2(b-a)^2/8} = \exp\left(\lambda\mathbf{E}[X] + \lambda^2(b-a)^2/8\right)\,. \tag{2.59}$$

The function $f(\lambda) = \mathbf{E}\left[e^{\lambda X}\right]$ is known as the *moment generating function* of X, since $f'(0) = \mathbf{E}[X]$, $f''(0) = \mathbf{E}[X^2]$, and, more generally, $f^{(k)}(0) = \mathbf{E}[X^k]$. *NB.* Since $f(\lambda) = \mathbf{E}\left[e^{\lambda X}\right]$ and $e^{\lambda X}|_{\lambda=0} = e^0 = 1$, then $f'(\lambda)|_{\lambda=0} = \mathbf{E}\left[e^{\lambda X}X\right]\big|_{\lambda=0} = \mathbf{E}[X]$, and $f''(\lambda)|_{\lambda=0} = \mathbf{E}\left[e^{\lambda X}X^2\right]\big|_{\lambda=0} = \mathbf{E}[X^2]$, and, more generally, $f^{(k)}(\lambda)|_{\lambda=0} = \mathbf{E}\left[e^{\lambda X}X^k\right]\big|_{\lambda=0} = \mathbf{E}\left[X^k\right]$.

Then the complimentary events of the inequality (2.62b), that is, (2.60b), will become the inequality (2.63).

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

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geqslant (1-\delta)\,. \tag{2.63}$$

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 advantage 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

$$\mu \leqslant \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,, \tag{2.64}$$

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 $(1-\delta)$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \mu - \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,, \tag{2.65}$$

i.e., 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.

NB. According to the equality (2.63) we will get the inequality (2.64) directly, with the equivalent (2.65). The complimentary events of the inequality (2.61c/2.60c) will become the inequality

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leqslant \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \geqslant (1-\delta)\,, \tag{2.66}$$

then with probability at least $(1-\delta)$, we have

$$-\sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \leqslant \mu - \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\,, \tag{2.67a}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \leqslant \mu \leqslant \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\,, \tag{2.67b}$$

$$\mu - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \leqslant \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \mu + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\,. \tag{2.67c}$$

Machine Learning Notes

This note is based on Machine Learning at DIKU

5 December 2021

Chapter 2. Concentration of Measure Inequalities

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 (Cover and Thomas 2006, Chapter 2)[7,8,9] and some basic tools from

[7] We first introduce the concept of *entropy*, which is a measure of the uncertainty of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \mathbf{Pr}\{X = x\}, x \in \mathcal{X}$. We denote the probability mass function by $p(x)$ rather than $p_X(x)$, for convenience. Thus, $p(x)$ and $p(y)$ refer to two different random variables and are in fact different probability mass functions, $p_X(x)$ and $p_Y(y)$, respectively.

Remark (Definition). *The entropy* $\mathbf{H}(X)$ *of a discrete random variable* X *is defined by* $\mathbf{H}(X) = -\sum_{x\in\mathcal{X}} p(x) \log p(x)$.

We also write $H(p)$ for the above quantity. The $\log(\cdot)$ is to the base 2 and entropy is expressed in bits. For example, the entropy of a fair coin toss is 1 bit. We will use the convention that $0 \log 0 = 0$, which is easily justfied by continuity since $x \log x \to 0$ as $x \to 0$. Adding terms of zero probability does not change the entropy.

If the base of the logarithm is b, we denote the entropy as $\mathbf{H}_b(X)$. If the base of the logarithm is e, the entropy is measured in *nats*. Unless otherwise specified, we will take all logarithms to base 2, and hence all the entropies will be measured in bits. Note that entropy is a functional of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities.

We denote expectation by $\mathbf{E}[\cdot]$. Thus, if $X \sim p(x)$, the expected value of the random variable $g(X)$ is written as $\mathbf{E}_p[g(X)] = \sum_{x\in\mathcal{X}} g(x)p(x)$or more simply as $\mathbf{E}[g(X)]$ when the probability mass function is understood from the context. We shall take a peculiar interest in the eerily self-referential expectation of $g(X)$ under $p(x)$ when $g(X) = \log \frac{1}{p(X)}$. *NB.* When $g(X) = \log \frac{1}{p(X)}$, then $\mathbf{E}[g(X)] = \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} = -\mathbf{H}(p)$. When $f(x) = \log \frac{1}{x}$, then its derivates are $f'(x) = \frac{1}{x^{-1}}(-x^{-2}) = -x^{-1} = -\frac{1}{x}$. Note that $(e^x)' = e^x$ and $(\ln x)' = 1/x$.

Remark. *The entropy of* X *can also be interpreted as the expected value of the random variable* $\log(1/p(X))$*, where* X *is drawn according to probability mass function* $p(x)$*. Thus,* $\mathbf{H}(X) = \mathbf{E}_p\left[\log\left(1/p(X)\right)\right]$.

This definition of entropy is related to the definition of entropy in thermodynamics; some of the connections are explored later. It is possible to derive the definition of entropy axiomatically by defining certain properties that the entropy of a random variable must satisfy. This approach is illustrated in Problem 2.46. We do not use the axiomatic approach to justify the definition of entropy; instead, we show that it arises as the answer to a number of natural questions, such as "What is the average length of the shortest description of the random variable?"

[8] Wikipedia (nat (unit)). The *natural unit of information* (symbol: nat), sometimes also nit or nepit, is a unit of information, based on natural logarithms and powers of e, rather than the powers of 2 and base 2 logarithms, which define the shannon. This unit is also known by its unit symbol, the nat. One nat is the information content of an event when the probability of that event occurring is $1/e$, where e is the Euler's number. One nat is equal to $\frac{1}{\ln 2}$ shannons ≈ 1.44 Sh or, equivalently, $\frac{1}{\ln 10}$ hartleys ≈ 0.434 Hart.

[9] First, we derive some immediate consequences of the definition.

Remark (Lemma 2.1.1). $\mathbf{H}(X) \geqslant 0$. *Proof.* $0 \leqslant p(x) \leqslant 1$ *implies that* $\log(1/p(x)) \geqslant 0$ *due to* $1/p(x) \in [1, +\infty)$.

Remark (Lemma 2.1.2). $\mathbf{H}_b(X) = (\log_b a)\mathbf{H}_a(X)$. *Proof.* $\log_b p = \log_b a \log_a p$ *dule to* $\log_b a = \ln a / \ln b$.

the method of types (Cover and Thomas 2006, Chapter 11). We start with some definitions, and we have special interest in Bernoulli random variables.

Definition 2.7 (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*

$$\mathbf{H}(p) = -\sum_{x\in\mathcal{X}} p(x)\ln p(x)\,. \tag{2.68}$$

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 $z \to 0$).

Definition 2.8 (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, that is,

$$\mathbf{E}[X] = 0\times\mathbf{Pr}(X=0) + 1\times\mathbf{Pr}(X=1) = \mathbf{Pr}(X=1) = p\,. \tag{2.69}$$

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

Definition 2.9 (Binary entropy). *Let p be a bias of Bernoulli random variable X. We define the entropy of p as*

$$\mathbf{H}(p) = -p\ln p - (1-p)\ln(1-p)\,. \tag{2.70}$$

NB. As p is the bias of Bernoulli random variable X, we have that $\mathbf{E}[X] = p$, where $\mathbf{Pr}(X=0) = 1-p$, and $\mathbf{Pr}(X=1) = p$. Then $\mathbf{H}(p) = -\sum_{x=0}^{1}\mathbf{Pr}(X=x)\ln\mathbf{Pr}(X=x)$, that is, Eq. (2.70).

Lemma 2.10.

$$\frac{1}{n+1}e^{n\,\mathbf{H}\left(\frac{k}{n}\right)} \leqslant \binom{n}{k} \leqslant e^{n\,\mathbf{H}\left(\frac{k}{n}\right)}\,. \tag{2.71}$$

(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 $p\in[0,1]$,

$$\sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i} = 1\,. \tag{2.72}$$

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

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

The second property of entropy enables us to change the base of the logarithm in the definition. Entropy can be changed from one base to another by multiplying by the appropriate factor.

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

NB. Eq. (2.72) holds because the sum of all possibilities is one. Besides, for any $k \in \{0, 1, ..., n\}$,

$$\sum_{k'=0}^{n} \binom{n}{k'} p^{k'}(1-p)^{n-k'} \geqslant \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} = \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \tag{2.74a}$$

$$= \binom{n}{k} e^{\ln\left(\left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}\right)} = \binom{n}{k} e^{k \ln \frac{k}{n} + (n-k) \ln \frac{n-k}{n}} \tag{2.74b}$$

$$= \binom{n}{k} e^{n\left(\frac{k}{n} \ln \frac{k}{n} + \frac{n-k}{n} \ln \frac{n-k}{n}\right)} = \binom{n}{k} e^{n\left(\frac{k}{n} \ln \frac{k}{n} + \left(1 - \frac{k}{n}\right) \ln \left(1 - \frac{k}{n}\right)\right)} \tag{2.74c}$$

$$= \binom{n}{k} e^{n \cdot \left(-\mathbf{H}\left(\frac{k}{n}\right)\right)} = \binom{n}{k} e^{-n\mathbf{H}\left(\frac{k}{n}\right)}, \tag{2.74d}$$

therefore,

$$\binom{n}{k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)} = \exp\left(n\mathbf{H}\left(\frac{k}{n}\right)\right). \tag{2.75}$$

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 $i \in \{0, ..., n\}$, see (Cover and Thomas 2006, Example 11.1.3)[10,11,12] for details. We

[10] See the mentioned example (pp. 353) as follows. We give a slightly better approximation for the binary case. These bounds can be proved using Stirling's approximation for the factorial function (Lemma 17.5.1).

Remark (Example 11.1.3, Binary alphabet). *In this case, the type is defined by the number of* 1*'s in the sequence, and the size of the type class is therefore* $\binom{n}{k}$. *We show that* $\frac{1}{n+1} 2^{n\mathbf{H}\left(\frac{k}{n}\right)} \leqslant \binom{n}{k} \leqslant 2^{n\mathbf{H}\left(\frac{k}{n}\right)}$.

Remark (Lemma 17.5.1). *For* $0 < p < 1$, $q = 1 - p$, *such that* np *is an integer,* $\frac{1}{\sqrt{8npq}} \leqslant \binom{n}{np} 2^{-n\mathbf{H}(p)} \leqslant \frac{1}{\sqrt{\pi npq}}$.

Remark (Theorem 11.1.3, Size of a type class $T(P)$). *For any type* $P \in \mathcal{P}_n$, $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n\mathbf{H}(P)} \leqslant |T(P)| \leqslant 2^{n\mathbf{H}(P)}$.

[11] The AEP for discrete random variables (Chapter 3) focuses our attention on a small subset of typical sequences. The method of types is an even more powerful procedure in which we consider sequences that have the same empirical distribution. With this restriction, we can derive strong bounds on the number of sequences with a particular empirical distribution and the probability of each sequence in this set. It is then possible to derive strong error bounds for the channel coding theorem and prove a variety of rate distortion results. The method of types was fully developed by (Csiszár and Körner 2011), who obtained most of their results from this point of view. (see Chapter 11, pp. 347)

Let $X_1, X_2, ..., X_n$ be a sequence of n symbols from an alphabet $\mathcal{X} = \{a_1, a_2, ..., a_{|\mathcal{X}|}\}$. We use the notation x^n and $\mathbf{x}$ interchangeably to denote a sequence $x_1, x_2, ..., x_n$. The type of a sequence $\mathbf{x}$ is denoted as $P_{\mathbf{x}}$. It is a probability mass function on $\mathcal{X}$. (Note that in this chapter, we will use capital letters to denote types and distributions. We also loosely use the word *distribution* to mean a probability mass function.)

Remark (Definition). *The type* P_x *(or empirical probability distribution) of a sequence* $x_1, x_2, ..., x_n$ *is the relative proportion of occurrences of each symbol of* $\mathcal{X}$ *(i.e.,* $P_{\mathbf{x}}(a) = N(a|\mathbf{x})/n$ *for all* $a \in \mathcal{X}$, *where* $N(a|\mathbf{x})$ *is the number of times the symbol* a *occurs in the sequence* $\mathbf{x} \in \mathcal{X}^n$*).*

[12] In information theory, the analog of the law of large numbers is the asymptotic equipartition property (AEP). It is a direct consequence of the weak law of large numbers. The *law of large numbers* states that for independent, identically distributed (i.i.d.) random variables, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is close to its expected value $\mathbf{E}[X]$ for large values of n. The AEP states that $\frac{1}{n} \log \frac{1}{p(X_1, X_2, ..., X_n)}$ is close to the entropy $\mathbf{H}(\cdot)$, where $X_1, X_2, ..., X_n$ are i.i.d. random variables and $p(X_1, X_2, ..., X_n)$ is the probability of observing the sequence $X_1, X_2, ..., X_n$. Thus, the probability $p(X_1, X_2, ..., X_n)$ assigned to an observed sequence will be close to $2^{-n\mathbf{H}}$. (see Chapter 3, pp. 57)

This enables us to divide the set of all sequences into two sets, the *typical set*, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences. Most of our attention will be on the typical sequences. Any property that is proved for the typical sequences will then be true with high probability and will determine the average behaviour of a large sample.

also note that there are $(n+1)$ elements in the sum in Eq. (2.72). Again, take $p = \frac{k}{n}$, then

$$1 \leqslant (n+1)\max_i \binom{n}{i}\left(\frac{k}{n}\right)^i\left(\frac{n-k}{n}\right)^{n-i} = (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} = (n+1)\binom{n}{k}e^{-n\mathbf{H}\left(\frac{k}{n}\right)}, \tag{2.76}$$

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

NB. According to (Cover and Thomas 2006, Example 11.1.3, pp. 353),[13] we have that:

$$\frac{\mathbf{Pr}(S=i+1)}{\mathbf{Pr}(S=i)} = \frac{C_n^{i+1}p^{i+1}(1-p)^{n-(i+1)}}{C_n^i p^i(1-p)^{n-i}} = \frac{\frac{n!}{(n-(i+1))!(i+1)!}p}{\frac{n!}{(n-i)!i!}(1-p)} = \frac{(n-i)!i!}{(n-i-1)!(i+1)!}\frac{p}{1-p} = \frac{n-i}{i+1}\frac{p}{1-p}, \tag{2.80a}$$

$$1 = \sum_{k'=0}^{n}\binom{n}{k'}p^{k'}(1-p)^{n-k'} \leqslant (n+1)\max_{k'}\binom{n}{k'}p^{k'}(1-p)^{n-k'} = (n+1)\binom{n}{k}p^k(1-p)^{n-k}, \tag{2.80b}$$

$$\text{s.t. } k \overset{\text{def}}{=} \operatorname*{argmax}_{k'\in\{0,1,\ldots,n\}}\binom{n}{k'}p^{k'}(1-p)^{n-k'}. \tag{2.80c}$$

Therefore,

$$1 \leqslant (n+1)\binom{n}{k}p^k(1-p)^{n-k} = (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} = (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} \tag{2.81a}$$

$$= (n+1)\binom{n}{k}e^{\ln\left(\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}\right)} = (n+1)\binom{n}{k}e^{k\ln\frac{k}{n}+(n-k)\ln\frac{n-k}{n}} \tag{2.81b}$$

$$= (n+1)\binom{n}{k}e^{n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)} = (n+1)\binom{n}{k}e^{n\left(\frac{k}{n}\ln\frac{k}{n}+\left(1-\frac{k}{n}\right)\ln\left(1-\frac{k}{n}\right)\right)} \tag{2.81c}$$

$$= (n+1)\binom{n}{k}e^{n\left(-\mathbf{H}\left(\frac{k}{n}\right)\right)} = (n+1)\binom{n}{k}e^{-n\mathbf{H}\left(\frac{k}{n}\right)}, \tag{2.81d}$$

and consequently,

$$\binom{n}{k} \geqslant \frac{1}{n+1}e^{n\mathbf{H}\left(\frac{k}{n}\right)}. \tag{2.82}$$

■

[13] For the lower bound, let S be a random variable with a binomial distribution with parameters n and p. The most likely value of S is $S = \langle np\rangle$. This can easily be verified from the fact that

$$\frac{\mathbf{Pr}(S=i+1)}{\mathbf{Pr}(S=i)} = \frac{n-i}{i+1}\cdot\frac{p}{1-p}, \tag{2.77}$$

and considering the cases when $i < np$ and when $i > np$. Then, since there are $(n+1)$ terms in the binomial sum,

$$1 = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} \leqslant (n+1)\max_k\binom{n}{k}p^k(1-p)^{n-k} \tag{2.78a}$$

$$= (n+1)\binom{n}{\langle np\rangle}p^{\langle np\rangle}(1-p)^{n-\langle np\rangle}. \tag{2.78b}$$

Now let $p = k/n$. Then we have $1 \leqslant (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}$, which by the arguments in the upper bound is equivalent to $\frac{1}{n+1} \leqslant \binom{n}{k}2^{-n\mathbf{H}(k/n)}$, or $\binom{n}{k} \geqslant \frac{1}{n+1}2^{n\mathbf{H}(k/n)}$. Combining the two results, we see that $\binom{n}{k} \approx 2^{n\mathbf{H}\left(\frac{k}{n}\right)}$. A more precise bound can be found in theorem 17.5.1 when $k \neq 0$ or n.

Remark (Proof of Lemma 17.5.1). *We begin with a strong form of Stirling's approximation (Feller 1957), which states that*

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}. \tag{2.79}$$

Applying this to find an upper bound, and the lower bound is obtained similarly.

Lemma 2.10 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 2.11 (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*

$$\mathbf{KL}(p\|q) = \mathbf{E}_p\left[\ln\frac{p(X)}{q(X)}\right] = \begin{cases} \sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}\,, & \textit{if } \mathcal{X} \textit{ is discrete}\,; \\ \int_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}dx\,, & \textit{if } \mathcal{X} \textit{ is continuous}\,. \end{cases} \tag{2.83}$$

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

Definition 2.12 (Binary kl-divergence). *Let p and q be biases of two Bernoulli random variables. The* binary kl divergence *is defined as*

$$\mathbf{kl}(p\|q) = \mathbf{KL}\big([1-p,p]\big\|[1-q,q]\big) = p\ln\tfrac{p}{q} + (1-p)\ln\tfrac{1-p}{1-q}\,. \tag{2.84}$$

We specialise the definition to Bernoulli distributions. 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 2.13. *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 2.10,*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) = \binom{n}{k}p^k(1-p)^{n-k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)}e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)} = e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,, \tag{2.85}$$

and

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) \geqslant \frac{1}{n+1}e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,. \tag{2.86}$$

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.

NB. We have known that by Lemma 2.10, $\binom{n}{k}e^{-n\mathbf{H}\left(\frac{k}{n}\right)} \in [\frac{1}{n+1}, 1]$, and $-\mathbf{H}\left(\frac{k}{n}\right) = \frac{k}{n}\ln\frac{k}{n} + \frac{n-k}{n}\ln\frac{n-k}{n}$. Then,

$$\binom{n}{k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)} = e^{-n\left(\frac{k}{n}\ln\frac{k}{n} + \frac{n-k}{n}\ln\frac{n-k}{n}\right)} = e^{-k\ln\frac{k}{n} - (n-k)\ln\frac{n-k}{n}}\,, \tag{2.87a}$$

$$p^k(1-p)^{n-k} = e^{\ln\left(p^k(1-p)^{n-k}\right)} = e^{k\ln p + (n-k)\ln(1-p)}\,, \tag{2.87b}$$

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) = \binom{n}{k}p^k(1-p)^{n-k} \leqslant e^{-k\left(\ln\frac{k}{n} - \ln p\right) - (n-k)\left(\ln\frac{n-k}{n} - \ln(1-p)\right)} \tag{2.87c}$$

$$= e^{-n\left(\frac{k}{n}\ln\frac{k/n}{p} + \frac{n-k}{n}\ln\frac{(n-k)/n}{1-p}\right)} = e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,. \tag{2.87d}$$

Similarly,

$$\binom{n}{k} \geqslant \frac{1}{n+1} e^{n\mathbf{H}\left(\frac{k}{n}\right)} = \frac{1}{n+1} e^{-n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)} = \frac{1}{n+1} e^{-k\ln\frac{k}{n}-(n-k)\ln\frac{n-k}{n}}, \tag{2.88a}$$

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k} \geqslant \frac{1}{n+1} e^{-k\left(\ln\frac{k}{n}-\ln p\right)-(n-k)\left(\ln\frac{n-k}{n}-\ln(1-p)\right)} \tag{2.88b}$$

$$= \frac{1}{n+1} e^{-n\left(\frac{k}{n}\ln\frac{k/n}{p}+\frac{n-k}{n}\ln\frac{(n-k)/n}{1-p}\right)} = \frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right)}. \tag{2.88c}$$

Overall,

$$\frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right)} \leqslant \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) \leqslant e^{-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right)}. \tag{2.89}$$

Additionally, we could obtain that

$$\mathbf{kl}(p\|p) = p\ln\frac{p}{p} + (1-p)\ln\frac{1-p}{1-p} = p\ln 1 + (1-p)\ln 1 = 1\ln 1 = 0, \tag{2.90a}$$

$$\mathbf{kl}(\hat{p}\|p) = \hat{p}\ln\frac{\hat{p}}{p} + (1-\hat{p})\ln\frac{1-\hat{p}}{1-p}. \tag{2.90b}$$

Question: How to explain $\mathbf{kl}(\hat{p}\|p)$?

As we all know, a convex function means that for all $0 < t < 1$ and all $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$, it holds that $f(tx_1 + (1-t)x_2) \leqslant tf(x_1) + (1-t)f(x_2)$. Besides, $\mathbf{kl}(\hat{p}\|p) = \mathbf{kl}\left(\frac{k}{n}\|p\right)$ where $k/n, p \in [0,1]$, thus, we have $1 - k/n, 1-p \in [0,1]$. Because

$$\frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \leqslant \mathbf{Pr}\left(\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) \leqslant 1, \tag{2.91}$$

then we get

$$-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right) \leqslant \ln(n+1), \tag{2.92a}$$

$$\mathbf{kl}\left(\frac{k}{n}\middle\|p\right) \leqslant \frac{1}{n}\ln(n+1) = \ln\sqrt[n]{n+1}. \tag{2.92b}$$

Because $e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \geqslant 0$, then we have $\mathbf{Pr}(\hat{p} = k/n) \geqslant \frac{1}{n+1} e^{-n\mathbf{kl}(\hat{p}\|p)} \geqslant 0$.

Question: But I still have no idea how to demonstrate $\mathbf{kl}(\hat{p}\|p)$ is convex in $\hat{p}$.

If we would like to demonstrate that $\mathbf{kl}(\hat{p}\|p)$ is convex in $\hat{p}$, we need to demonstrate that: for all $0 < t < 1$ and all x_1, x_2 such that $x_1 \neq x_2$, it holds

$$\mathbf{kl}(tx_1 + (1-t)x_2\|p) \leqslant t\mathbf{kl}(x_1\|p) + (1-t)\mathbf{kl}(x_2\|p), \tag{2.93}$$

that is to say, for all $0 < t < 1$ and all $k_i \in [0,n], k_i \in \mathbb{Z}$, $\forall i \in \{1,2\}$ such that $k_1 \neq k_2$, it holds

$$\mathbf{kl}\left(t\frac{k_1}{n} + (1-t)\frac{k_2}{n}\|p\right) \leqslant t\mathbf{kl}\left(\frac{k_1}{n}\|p\right) + (1-t)\mathbf{kl}\left(\frac{k_2}{n}\|p\right). \tag{2.94}$$

Machine Learning Notes

This note is based on Machine Learning at DIKU

6 December 2021

Chapter 2. Concentration of Measure Inequalities

2.5 kl Inequality

Example 2.13 shows that $\mathbf{kl}(\cdot)$ 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 and then combine it with Markov's inequality to obtain the following result in Theorem 2.15.

Lemma 2.14. *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*

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] \leqslant n+1\,. \tag{2.95}$$

Proof.

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbf{Pr}\left(\hat{p}=\frac{k}{n}\right) e^{n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \leqslant \sum_{k=0}^{n} e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} e^{n\mathbf{kl}\left(\frac{k}{n}\|p\right)} = n+1\,, \tag{2.96}$$

where the inequality was derived in Eq. (2.85). ■

NB. We have known that by definition, $\mathbf{E}[X] = \sum_k = 1^{\infty} x_k p_k = \int_{-\infty}^{\infty} x f(x) dx$. Then according to Example 2.13, we have $\frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \leqslant \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p} = \frac{k}{n}\right) \leqslant e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}$. Therefore,

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbf{Pr}\left(\hat{p}=\frac{k}{n}\right) e^{n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \stackrel{\text{def}}{=} A\,, \tag{2.97a}$$

$$A \leqslant \sum_{k=0}^{n} e^{-n\mathbf{kl}(\hat{p}\|p)} e^{n\mathbf{kl}(\hat{p}\|p)} = \sum_{k=0}^{n} e^{0} = \sum_{k=0}^{n} 1 = n+1\,, \tag{2.97b}$$

$$A \geqslant \sum_{k=0}^{n} \frac{1}{n+1} e^{-n\mathbf{kl}(\hat{p}\|p)} e^{n\mathbf{kl}(\hat{p}\|p)} = \frac{1}{n+1}\sum_{k=0}^{n} e^{0} = \frac{1}{n+1}(n+1) = 1\,. \tag{2.97c}$$

Theorem 2.15 (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*

$$\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) \leqslant (n+1)e^{-n\varepsilon}\,. \tag{2.98}$$

Proof. By Markov's inequality and Lemma 2.14,

$$\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) = \mathbf{Pr}\left(e^{n\mathbf{kl}(\hat{p}\|p)} \geqslant e^{n\varepsilon}\right) \leqslant \frac{e^{n\mathbf{kl}(\hat{p}\|p)}}{e^{n\varepsilon}} \leqslant \frac{n+1}{e^{n\varepsilon}}\,. \tag{2.99}$$

■

Machine Learning Notes

This note is based on Machine Learning at DIKU

10 December 2021

Chapter 2. Concentration of Measure Inequalities

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

By denoting the right hand side of kl inequality (2.98) by δ, we obtain that with probability greater than $(1-\delta)$,

$$\mathbf{kl}(\hat{p}\|p) \leqslant \frac{\ln \frac{n+1}{\delta}}{n}\,. \tag{2.100}$$

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, see (Cover and Thomas 2006, Lemma 11.6.1).

NB. According to the right side of kl inequality (2.98)[14], and let $\delta \overset{\text{def}}{=} (n+1)e^{n\varepsilon}$ [15], therefore, (2.98) will become

$$\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \tfrac{1}{n}\ln\tfrac{n+1}{\delta}\right) \leqslant \delta\,, \tag{2.101}$$

that is to say, with probability greater than $(1-\delta)$, we have $\mathbf{kl}(\hat{p}\|p) \leqslant \frac{1}{n}\ln\frac{n+1}{\delta}$.

Lemma 2.16 (Pinsker's inequality)**.**

$$\mathbf{KL}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2\,, \tag{2.102}$$

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

Corollary 2.17 (Pinsker's inequality for the binary kl divergence)**.**

$$\mathbf{kl}(p\|q) \geqslant \tfrac{1}{2}\big(|p-q|+|(1-p)-(1-q)|\big)^2 = 2(p-q)^2\,. \tag{2.103}$$

NB. We have known that by definition,

$$\mathbf{KL}(p\|q) = \mathbf{E}_p\left[\ln\frac{p(X)}{q(X)}\right] = \begin{cases}\sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}\,, & \text{if } \mathcal{X} \text{ is discrete;}\\ \int_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}dx\,, & \text{if } \mathcal{X} \text{ is continuous,}\end{cases} \tag{2.104a}$$

$$\mathbf{kl}(p\|q) = \mathbf{KL}([1-p,p]\|[1-q,q]) = p\ln\tfrac{p}{q} + (1-p)\ln\tfrac{1-p}{1-q}\,. \tag{2.104b}$$

Back to the inequality in Lemma 2.16, therefore, we have: for the binary kl divergence,

$$\mathbf{KL}(p\|q) = \textstyle\sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)} = \sum_{x\in\mathcal{X}} p(x)\left(\ln p(x) - \ln q(x)\right) \overset{\text{def}}{=} \text{LHS}\,, \tag{2.105a}$$

$$\tfrac{1}{2}\|p-q\|_1^2 = \tfrac{1}{2}\big(\textstyle\sum_{x\in\mathcal{X}}|p(x)-q(x)|\big)^2 \overset{\text{def}}{=} \text{RHS}\,, \tag{2.105b}$$

$$\mathbf{kl}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2 = \tfrac{1}{2}\left(|p-q|+|(1-p)-(1-q)|\right)^2 = \tfrac{1}{2}\left(2|p-q|\right)^2 = 2|p-q|^2 = 2(p-q)^2\,. \tag{2.105c}$$

[14] that is, $\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) \leqslant (n+1)e^{-n\varepsilon}$.

[15] then we obtain that $-n\varepsilon = \ln\frac{\delta}{n+1}$ and $\varepsilon = \frac{1}{n}\ln\frac{n+1}{\delta}$.

Qustion: Unlike Corollary 2.17, I'm not very sure how to demonstrate Lemma 2.16.

By applying Corollary 2.17 to (2.100) we obtain that with probability greater than $(1-\delta)$,

$$|p-\hat{p}| \leqslant \sqrt{\frac{\mathbf{kl}(\hat{p}\|p)}{2}} \leqslant \sqrt{\frac{\ln\frac{n+1}{\delta}}{2n}}\,. \tag{2.106}$$

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

$$p \leqslant \hat{p} + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,. \tag{2.107}$$

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, see (Maurer 2004; Langford and Schapire 2005). Next we show that the kl inequality can actually be significantly tighter than Hoeffding's inequality. For this we use refined Pinsker's inequality, see (Marton 1996; Samson 2000), (Boucheron, Lugosi, and Massart 2013, Lemma 8.4).

NB. Let $X_1, ..., X_n$ be i.i.d. Bernoulli with bias p and q and let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias, then $\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) \leqslant (n+1)e^{-n\varepsilon}$. If we let $\delta \overset{\text{def}}{=} (n+1)e^{-n\varepsilon}$, we will get $\varepsilon = -\frac{1}{n}\ln\frac{\delta}{n+1}$, which means, with probability greater than $(1-\delta)$, it holds $\mathbf{kl}(\hat{p}\|p) \leqslant \frac{1}{n}\ln\frac{n+1}{\delta}$. According to the combination of Corollary 2.17 and (2.100), we have: with probability greater than $(1-\delta)$,

$$2(\hat{p}-p)^2 \leqslant \mathbf{kl}(\hat{p}\|p) \leqslant \tfrac{1}{n}\ln\tfrac{n+1}{\delta}\,, \tag{2.108a}$$

$$|\hat{p}-p| \leqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{n+1}{\delta}}\,, \tag{2.108b}$$

$$p \leqslant \hat{p} + \sqrt{\tfrac{1}{2n}\ln\tfrac{n+1}{\delta}} = \hat{p} + \sqrt{\tfrac{1}{2}\varepsilon}\,. \tag{2.108c}$$

Recall that Hoeffding's inequality[16] states that: Let $X_1, ..., X_n$ be independent real-valued random variables, then for every $\varepsilon > 0$, that is, $\delta = e^{-2n\varepsilon^2} \in (0,1)$,

$$\mathbf{Pr}\left(\hat{p} - p \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.112a}$$

$$\mathbf{Pr}\left(\hat{p} - p \leqslant -\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.112b}$$

[16] According to Theorem 2.3, Corollary 2.4–2.5, we have: Let $X_i, ..., X_n$ be independent random variables, such that $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$, it holds that

$$\mathbf{Pr}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)\,, \tag{2.109a}$$

$$\mathbf{Pr}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu \leqslant -\varepsilon\right) = \mathbf{Pr}\left(\mu - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)\,, \tag{2.109b}$$

$$\mathbf{Pr}\left(\left|\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(-2n\varepsilon^2\right)\,. \tag{2.109c}$$

If we let $\delta \overset{\text{def}}{=} e^{-2n\varepsilon^2}$, we will have $n\varepsilon^2 = -\frac{1}{2}\ln\delta = \frac{1}{2}\ln\frac{1}{\delta}$. In this case, the Hoeffding's inequalities will become

$$\mathbf{Pr}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.110a}$$

$$\mathbf{Pr}\left(\mu - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,. \tag{2.110b}$$

If we let $\delta \overset{\text{def}}{=} 2e^{-2n\varepsilon^2}$, we will have $n\varepsilon^2 = -\frac{1}{2}\ln\frac{\delta}{2} = \frac{1}{2}\ln\frac{2}{\delta}$, and then the Hoeffding's inequality will hold

$$\mathbf{Pr}\left(\left|\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\right) \leqslant \delta\,. \tag{2.111}$$

which means, it assures that with probability greater than $(1-\delta)$, it holds

$$-\sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}} < -\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}} \leqslant \hat{p}-p \leqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}} < \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\,, \tag{2.113a}$$

$$\hat{p}-\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}} \leqslant \quad p \quad \leqslant \hat{p}+\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\,. \tag{2.113b}$$

Therefore, the only difference between Hoeffding's inequality and the kl inequality is the $\ln(n+1)$ factor, in the worst case. To demonstrate that the kl inequality can actually be significantly tighter than Hoeffding's inequality, we also present some details about refined Pinsker's inequality (Marton 1996; Samson 2000) (Boucheron, Lugosi, and Massart 2013, Lemma 8.4).

As presented in (Marton 1996),[17] it gives a proof of Talagrand's inequality, which admits an extension to contracting Markov chains. The proof is based on a new asymmetric notion of distance between probability measures, and bounding this distance by informational divergence. The author also analyses the bin packing problem for Markov chains as an application. Note that the distance that the author proposed is $d_2(p,r)$, and the d_2-distance can be extended to a distance between probability measures p^n and r^n defined on the product space $\mathcal{X}^n$. (We get an analogue of the $\bar{d}$-distance.)

[17](Marton 1996) Let $\{X_i\}_{i=-\infty}^{\infty}$ be a Markov chain (not necessarily homogeneous) taking values in a complete separable metric space $\mathcal{X}$, endowed with the Borel σ-algebra $\mathcal{B}$. We assume that the diagonal of the product space $\mathcal{X}\times\mathcal{X}$ is measurable with respect to the product σ-algebra $\mathcal{B}\times\mathcal{B}$. We introduce an asymmetric notion of how much a probability distribution p differs from another one, r. The Radon-Nikodym derivative $\frac{dr}{dp}$ is assumed to be ∞ on the set of singularity of r with respect to p.

Remark (Definition). *If p and r are probability measures on $\mathcal{X}$ then*

$$d_2(p,r) = \left[\int \left|1-\tfrac{dr}{dp}(y)\right|_+^2 dp(y)\right]^{1/2}. \tag{2.114}$$

Notice that if μ is a probability measure on $\mathcal{X}$, and both p and r are absolutely continuous with respect to μ with Radon-Nikodym derivatives $f=\frac{dp}{d\mu}, g=\frac{dr}{d\mu}$, respectively, then $d_2(p,r)$ can be written as

$$d_2(p,r) = \left[\int \left|1-\tfrac{g}{f}\right|_+^2 f d\mu\right]^{1/2}. \tag{2.115}$$

The functional $d_2(p,r)$ is analogous to variational distance (divided by 2), which is defined as

$$|p-r| = \frac{1}{2}\int \left|1-\tfrac{dr}{dp}(y)\right| dp(y) = \int \left|1-\tfrac{dr}{dp}(y)\right|_+ dp(y)\,. \tag{2.116}$$

Obviously, $|p-r| \leqslant d_2(p,r) \leqslant |p-r|^{1/2}$. It is easy to see that $d_2(p,r) = \inf\left[\int \mathbf{Pr}(Z\neq y|Y=y)^2 dp(y)\right]^{1/2}$, where the infinum is taken over all joint distributions $\mathrm{dist}(Y,Z)$ with marginals $p=\mathrm{dist}(Y), r=\mathrm{dist}(Z)$. This is again analogous to the fact that $|p-r| = \inf \mathbf{Pr}(Z\neq Y)$, with the infimum taken over the same set of joint distributions.

Remark (Lemma 3.2 (Marton 1996)).

$$\begin{aligned} &(i) \quad d_2(p,r) \leqslant [2\mathbf{D}(p\|r)]^{1/2}\,,\\ &(ii) \quad d_2(r,p) \leqslant [2\mathbf{D}(p\|r)]^{1/2}\,. \end{aligned} \tag{2.117}$$

There is a simple inequality by Pinsker between variational distance and information divergence. Our next move is to prove a similar inequality for d_2. We shall have two different inequalities, since both d_2 and informational divergence are asymmetric. Pinsker's inequality says that $|p-r| \leqslant [\mathbf{D}(p\|r)/2]^{1/2}$. For d_2 we have the above bounds. On the right-hand side of these formulae we have a factor 2 instead of the $1/2$ of Pinsker's inequality. (It can be shown that the factor 2 cannot be improved.) Note, however, that the d_2-distance cannot be overbounded by a constant multiple of variational distance.

As presented in (Samson 2000),[18,19,20] it proves concentration inequalities for some classes of Markov chains and Φ-mixing processes, with constants independent of the size of the sample, that extend the inequalities for product measures of Talagrand. This work is based on information inequalities put forward by Marton in case of contracting Markov chains. Using a simple duality argument on entropy, the results in this paper also include the family of logarithmic Sobolev inequalities for convex functions. Additionally, applications to bounds on supremum

[18](Samson 2000) In a recent series of striking papers, Talagrand deeply analysed the concentration of measure phenomenon in product space, with applications to various areas of probability theory. A first result at the origin of his investigation concerns deviation inequalities for product measures $P = \mu_1 \otimes \cdots \otimes \mu_n$ on $[0,1]^n$. Namely, for every convex function f on $[0,1]^n$, with Lipschitz constant $\|f\|_{\mathrm{Lip}} \leqslant 1$, and for every $t \geqslant 0$,

$$P(|f - M| \geqslant t) \leqslant 4\exp(-t^2/4)\,, \tag{2.118}$$

where M is a median of f for P. This Gaussian-type bound may be considered as an important generalisation of the classical inequalities for sums of independent random variables. The deviation inequality (2.118) is a consequence of a concentration inequality on sets which takes the following form. To measure the "distance" of a point $\boldsymbol{x} \in \mathbb{R}^n$ to a set A, consider the functional (see "convex hull", (Talagrand 1995), Chapter 4),

$$f_{\mathrm{conv}}(A, \boldsymbol{x}) = \sup_{\alpha} \inf_{y \in A} \left(\sum_{i=1}^{n} \alpha_i \mathbb{I}(x_i \neq y_i) \right), \tag{2.119}$$

where the supremum is over all vectors $\alpha = (\alpha_i)_{1 \leqslant i \leqslant n}, \alpha_i \geqslant 0, \sum_{i=1}^n \alpha_i^2 = 1$. If we let $A_t^{\mathrm{conv}} = \{\boldsymbol{x} \in \mathbb{R}^n, f_{\mathrm{conv}}(A, \boldsymbol{x}) \leqslant t\}$, Talagrand shows that for every $t \geqslant \sqrt{2\log(1/P(A))}$,

$$P(A_t^{\mathrm{conv}}) \geqslant 1 - \exp\left[-\tfrac{1}{2} \left(t - \sqrt{2\log \tfrac{1}{\mathbf{Pr}(A)}} \right)^2 \right]. \tag{2.120}$$

Besides the convex hull approximation, Talagrand considers two other approximations on product spaces for which he proves similar concentration properties. One of the main features of these inequalities is that they are independent of the dimension of the product space, that is, of the size of the sample. We will be mainly concerned with extensions of the convex hull approximation in this work.

[19](Samson 2000) Recently, an alternative, simpler, approach to some of Talagrand's inequalities was suggested by Ledoux (Ledoux 1997) on the basis of log-Sobolev inequalities. Introduce, for every function g on $\mathbb{R}^n$, the entropy functional,

$$\mathrm{Ent}_P(g^2) = \int g^2 \log g^2 dP - \int g^2 dP \log \int g^2 dP\,. \tag{2.121}$$

Then, it can easily be shown that, for every product measure P on $[0,1]^n$ and for every separately convex function f, $\mathrm{Ent}_P(e^f) \leqslant \frac{1}{2}\int |\nabla f|^2 e^f dP$, where ∇f denotes the usual gradient of f on $\mathbb{R}^n$ and $|\nabla f|$ its Euclidean length. This inequality easily implies deviation inequalities of the type of (2.118). Indeed, the preceding log-Sobolev inequality may be turned into a differential inequality on the Laplace transform of convex Lipschitz functions, which then yields tail estimates by Chebyshev's inequality. This type of argument may be pushed further to recover most of Talagrand's deviation inequalities for functions (Ledoux 1997). It however does not seem to succeed for deviations under the median (or for concave functions).

[20](Samson 2000) Let now P denote the law of the sample X on $\mathbb{R}^n$. For every probability measures Q and R on $\mathbb{R}^n$, let $\mathcal{U}(Q,R)$ denote the set of all probability measures on $\mathbb{R}^n \otimes \mathbb{R}^n$ with marginals Q and R. Define

$$d_2(Q,R) = \inf_{\Pi \in \mathcal{U}(Q,R)} \sup_{\alpha} \iint \sum_{i=1}^{n} \alpha_i(y) \mathbb{I}(x_i \neq y_i)\, d\Pi(x,y)\,, \tag{2.122}$$

where the $\sup_\alpha$ is over all vectors of positive functions $\alpha = (\alpha_1, ..., \alpha_n)$, with $\int \sum_{i=1}^n \alpha_i^2(y) dR(y) \leqslant 1$. As a main result, we show in Theorem 1 below that, for every probability measure Q on $\mathbb{R}^n$ with Radon-Nikodym derivative dQ/dP with respect to the measure P, $d_2(Q,P) \leqslant \|\Gamma\| \sqrt{2\mathrm{Ent}_P\left(dQ/dP\right)}$. Furthermore, $d_2(P,Q) \leqslant \|\Gamma\| \sqrt{2\mathrm{Ent}_P\left(dQ/dP\right)}$. Such Pinsker type inequalities have already been investigated by Marton for contracting Markov chains, and then by Dembo in the independent case (Marton 1996). Recently, Marton also obtained related bounds with a parameter readily comparable to $\|\Gamma\|$ (Marton 1999). Following these works, we could easily derive concentration in the form of (2.120) [and thus (2.118)] from these information inequalities. We however take

of dependent empirical processes complete this work.

As presented in (Boucheron, Lugosi, and Massart 2013),[21] the authors take a step further and relax the bounded differences condition. They assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies $f(y) - f(x) \leqslant \sum_{i=1}^n c_i(x)\mathbb{I}(x_i \neq y_i)$, for some functions $c_i : \mathcal{X}^n \to [0,\infty), i = 1,\dots,n$. Instead of forcing the c_i to be bounded we only assume that they are bounded in "quadratic mean" in the sense that $v \stackrel{\text{def}}{=} \mathbf{E}\left[\sum_{i=1}^n c_i^2(X)\right]$, is finite. Under this assumption, the transportation method may be used as follows. Let Q be a probability distribution, absolutely continuous with respect to P, the distribution of X. Let $\mathbf{P}$ be a coupling of P and Q. Then

$$\mathbf{E}_Q[f] - \mathbf{E}_P[f] \leqslant \sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[c_i(X)\mathbf{P}(X_i \neq Y_i|X)\right], \tag{2.127}$$

which implies, by applying the Cauchy-Schwarz inequality twice,

$$\mathbf{E}_Q[f] - \mathbf{E}_P[f] \leqslant \left(\mathbf{E}_{\mathbf{P}}[c_i^2(X)]\right)^{1/2}\left(\mathbf{E}_{\mathbf{P}}[\mathbf{P}^2(X_i \neq Y_i|X)]\right)^{1/2} \tag{2.128a}$$

$$\leqslant \left(\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}[c_i^2(X)]\right)^{1/2}\left(\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[\mathbf{P}^2(X_i \neq Y_i|X)\right]\right)^{1/2}. \tag{2.128b}$$

Using their assumption on f, this implies

$$\mathbf{E}_Q[f] - \mathbf{E}_P[f] \leqslant \sqrt{v}\left(\inf_{\mathbf{P}\in\mathcal{P}(P,Q)}\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[\mathbf{P}^2(X_i \neq Y_i|X)\right]\right)^{1/2}. \tag{2.129}$$

Thus, by the road map laid down in the introduction of this chapter, if the authors can prove the inequality

$$\inf_{\mathbf{P}\in\mathcal{P}(P,Q)}\sum_{i=1}^n \mathbf{E}_{\mathbf{P}}\left[\mathbf{P}^2(X_i \neq Y_i|X)\right] \leqslant 2\mathbf{D}(Q\|P), \tag{2.130}$$

then Lemma 4.18 implies $\psi_{Z-\mathbf{E}[Z]}(\lambda) \leqslant v\lambda^2/2$ and the resulting sub-Gaussian tail inequality with variance factor v. Since Lemma 8.13 is applicable, it suffices to prove the transportation inequality above for $n = 1$. They solve the corresponding transportation cost problem first.

---

a somewhat different route related to exponential integrability and log-Sobolev inequalities. TBC.

[21] (Boucheron, Lugosi, and Massart 2013) The next step is an analog of Pinsker's inequality in which d_2 plays the role of the total variation distance.

Remark (Lemma 8.4 (Boucheron, Lugosi, and Massart 2013)). *Let P and Q be probability distributions on a common measure space $(\Omega, \mathcal{A})$. If Q is absolutely continuous with respect to P, then*

$$d_2^2(Q,P) + d_2^2(P,Q) \leqslant 2\mathbf{D}(Q\|P). \tag{2.123}$$

Proof. Since $Q \ll P$, setting $q = dQ/dP$ we may write

$$d_2^2(Q,P) + d_2^2(P,Q) = \mathbf{E}_P\left[(1-q(X))_+^2\right] + \mathbf{E}_P\left[\tfrac{(q(X)-1)_+^2}{q(X)}\right]. \tag{2.124}$$

Moreover, defining $h(t) = (1-t)\log(1-t) + t$ for $t < 1$ and $h(1) = 1$, we may write

$$\mathbf{D}(Q\|P) = \mathbf{E}_P\left[h\left(1-q(X)\right)\right] = \mathbf{E}_P\left[h\left((1-q(X))_+\right)\right] + \mathbf{E}_P\left[h\left(-(q(X)-1)_+\right)\right], \tag{2.125}$$

and the result follows by the inequalities

$$h(t) \geqslant t^2/2 \quad \text{for } t \in [0,1], \quad \text{and} \quad h(-t) \geqslant \tfrac{t^2}{2(1+t)} \quad \text{for } t \geqslant 0. \tag{2.126}$$

(recall Exercise 2.8). ∎

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

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}}\,. \tag{2.131}$$

Corollary 2.19 (Refined Pinsker's inequality)**.** *If $q > p$ then*

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2q}\,. \tag{2.132}$$

Corollary 2.20 (Refined Pinsker's inequality)**.** *If $\mathbf{kl}(p\|q) \leqslant \varepsilon$ then*

$$q \leqslant p + \sqrt{2p\varepsilon} + 2\varepsilon\,. \tag{2.133}$$

NB. **Qustion:** I'm not sure how to demonstrate Lemma 2.18.

We first present Pinsker's inequality[22] here, that is to say,

$$\mathbf{KL}(p\|q) \geqslant \tfrac{1}{2}\|p-q\|_1^2 = \tfrac{1}{2}\left(\textstyle\sum_{x\in\mathcal{X}}|p(x)-q(x)|\right)^2\,, \tag{2.138}$$

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

According to Lemma 2.18, if $q > p$, then $1-q < 1-p$, and because of $p, q \in [0,1]$, it holds

$$\mathbf{kl}(p\|q) \geqslant \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}} = \frac{(p-q)^2}{2q} + \frac{(p-q)^2}{2(1-p)} \geqslant \frac{(p-q)^2}{2q}\,. \tag{2.139}$$

---

[22] Wikipedia (Pinsker's inequality) In information theory, Pinsker's inequality, named after its inventor Mark Semenovich Pinsker, is an inequality that bounds the total variation distance (or statistical distance) in terms of the Kullback-Leibler divergence. The inequality is tight up to constant factors.

Formal statement. Pinsker's inequality states that, if P and Q are two probability distributions on a measurable space (X, Σ), then

$$\delta(P,Q) \leqslant \sqrt{\tfrac{1}{2}\mathbf{D}_{KL}(P\|Q)}\,, \tag{2.134}$$

where

$$\delta(P\|Q) = \sup\left\{|P(A)-Q(A)|\,\middle|\,A\in\Sigma \text{ is a measurable event}\right\}, \tag{2.135}$$

is the total variation distance (or statistical distance) between P and Q and

$$\mathbf{D}_{KL}(P\|Q) = \mathbf{E}_P\left(\log\tfrac{\mathrm{d}P}{\mathrm{d}Q}\right) = \int_X\left(\log\tfrac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P\,, \tag{2.136}$$

is the Kullback-Leibler divergence in nats. When the sample space X is a finite set, the Kullback-Leibler divergence is given by

$$\mathbf{D}_{KL}(P\|Q) = \sum_{i\in X}\left(\log\tfrac{P(i)}{Q(i)}\right)P(i)\,. \tag{2.137}$$

Note that in terms of the total variation norm $\|P-Q\|$ of the signed measure $(P-Q)$, Pinsker's inequality differs from the one given above by a factor of two, $\|P-Q\| \leqslant \sqrt{2\mathbf{D}_{KL}(P\|Q)}$. A proof of Pinsker's inequality uses the partition inequality for f-divergences.

Alternative version. There is an alternative statement of Pinsker's inequality in some literature that relates information divergence to variation distance, $\mathbf{D}(P\|Q) \geqslant \frac{1}{2\ln 2}\mathbf{V}^2(p,q)$, in which $\mathbf{V}(p,q) = \sum_{x\in\mathcal{X}}|p(x)-q(x)|$ is the variation distance between two probability density functions p and q on the same alphabet $\mathcal{X}$. This form of Pinsker's inequality shows that "convergence in divergence" is strong notion than "convergence in variation distance".

Inverse problem. A precise inverse of the inequality cannot hold: for every $\varepsilon > 0$, there are distributions P_ε, Q with $\delta(P_\varepsilon, Q) \leqslant \varepsilon$ but $\mathbf{D}_{KL}(P_\varepsilon\|Q) = \infty$. An easy example given by the two-point space $\{0,1\}$ with $Q(0) = 0, Q(1) = 1$ and $P_\varepsilon(0) = \varepsilon, P_\varepsilon(1) = 1-\varepsilon$. However, an inverse inequality holds on finite spaces X with a constant depending on Q. More specifically, it can be shown that with the definition $\alpha_Q \overset{\text{def}}{=} \min_{x\in X:Q(x)>0} Q(x)$ we have for any measure P which is absolutely continuous to Q, $\frac{1}{2}\mathbf{D}_{KL}(P\|Q) \leqslant \frac{1}{\alpha_Q}\delta(P,Q)^2$. As a consequence, if Q has full support (i.e., $Q(x) > 0$ for all $x \in X$), then $\delta(P,Q)^2 \leqslant \frac{1}{2}\mathbf{D}(P\|Q) \leqslant \delta(P,Q)^2/\alpha_Q$.

Therefore, Corollary 2.19 is demonstrated.

Question: It seems that we still require one more condition, which is $q = 1 - p$, before we reach the result in Corollary 2.19. Never mind, I was wrong, it doesn't have to.

Then if $\mathbf{kl}(p\|q) \leqslant \varepsilon$, we will get that

$$\text{RHS} = p + \sqrt{2p\varepsilon} + 2\varepsilon = \left(\sqrt{p} + \sqrt{\tfrac{\varepsilon}{2}}\right)^2 + \tfrac{3}{2}\varepsilon^2 = \left(\sqrt{2\varepsilon} + \tfrac{1}{2}p\right)^2 + \tfrac{3}{4}p^2; \tag{2.140a}$$

$$\tfrac{(p-q)^2}{2q} \leqslant \mathbf{kl}(p\|q) \leqslant \varepsilon, \tag{2.140b}$$

$$(p-q)^2 = p^2 - 2pq + q^2 \leqslant 2q\varepsilon, \quad \because p, q \in [0,1], \tag{2.140c}$$

$$-\sqrt{2q\varepsilon} \leqslant p - q \leqslant \sqrt{2q\varepsilon}, \quad q - \sqrt{2q\varepsilon} \leqslant p \leqslant q + \sqrt{2q\varepsilon}, \tag{2.140d}$$

$$q^2 - 2(p+\varepsilon)q + p^2 \leqslant 0 \tag{2.140e}$$

$$[q - (p+\varepsilon)]^2 \leqslant (p+\varepsilon)^2 - p^2 = 2p\varepsilon + \varepsilon^2 \tag{2.140f}$$

$$p + \varepsilon - \sqrt{2p\varepsilon + \varepsilon^2} \leqslant q \leqslant p + \varepsilon + \sqrt{2p\varepsilon + \varepsilon^2} \tag{2.140g}$$

Question: I'm not sure how to demonstrate Corollary 2.20 as well.

By applying Corollary 2.20 to (2.100) we obtain that with probability greater than $(1-\delta)$,

$$p \leqslant \hat{p} + \sqrt{\frac{2\hat{p}\ln\frac{n+1}{\delta}}{n}} + \frac{2\ln\frac{n+1}{\delta}}{n}. \tag{2.141}$$

Note that when $\hat{p}$ is close to zero, then 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 2.15 can be improved slightly, see (Maurer 2004; Langford and Schapire 2005).

NB. Note that (2.100) means that: with probability greater than $(1-\delta)$, it holds that $\mathbf{kl}(\hat{p}\|p) \leqslant \varepsilon = \frac{1}{n}\ln\frac{n+1}{\delta}$. Besides, Corollary 2.20 gives that $q \leqslant p + \sqrt{2p\varepsilon} + 2\varepsilon$, if $\mathbf{kl}(p\|q) \leqslant \varepsilon$. Therefore,

$$p \leqslant \hat{p} + \sqrt{2\hat{p}\varepsilon} + 2\varepsilon = \hat{p} + \sqrt{2\hat{p}\tfrac{1}{n}\ln\tfrac{n+1}{\delta}} + 2\tfrac{1}{n}\ln\tfrac{n+1}{\delta} = \hat{p} + \sqrt{\tfrac{2\hat{p}}{n}\ln\tfrac{n+1}{\delta}} + \tfrac{2}{n}\ln\tfrac{n+1}{\delta}. \tag{2.142}$$

We already know that Hoeffding's inequality means that $p \leqslant \hat{p} + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}$ with probability greater than $(1-\delta)$. So when $\hat{p}$ is close to zero, the latter inequality is much tighter than Hoeffding's inequality.

Question: Doesn't it mean that smaller is tighter? Where am I wrong?

The bound in Theorem 2.15 [23] is $\mathbf{Pr}(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon) \leqslant (n+1)e^{-n\varepsilon}$, which can be improved slightly, see (Maurer 2004; Langford and Schapire 2005).

According to (Maurer 2004),[24] the author proves general exponential moment inequalities for averages of $[0,1]$-valued iid random variables and use them to tighten the PAC Bayesian Theo-

[23] where $X_1, ..., X_n$ are i.i.d. Bernoulli with bias p and $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ is the empirical bias.

[24] Throughout this note $X_1, ..., X_n$ are assumed to be IID random variables with values in $[0,1]$ and expectation $\mathbf{E}[X_i] = \mu$. We use $\mathbf{X}$ to denote the corresponding random vector $\mathbf{X} = (X_1, ..., X_n)$ with values in $[0,1]^n$ and $M(\mathbf{X})$ to denote its arithmetic mean $M(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$. For any $[0,1]$-valued random variables X use X' to denote the unique Bernoulli ($\{0,1\}$-valued) random variable with $\mathbf{Pr}(X' = 1) = \mathbf{E}[X'] = \mathbf{E}[X]$. Evidently $X'' = X', \forall X$. For $\mathbf{X} = (X_1, ..., X_n)$ we denote $\mathbf{X}' = (X_1', ..., X_n')$. We restate our principal bounds in a slightly more general way.

Theorem (1). *For all $n \geqslant 2$, $\mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}), \mu)\right)\right] \leqslant \mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}'), \mu)\right)\right] \leqslant e^{1/(12n)}\left(\frac{\pi n}{2}\right)^{1/2} + 2$. If the X_i are nontrivial Bernoulli variables (i.e., if $\mu \in (0,1)$) then there is a sequence c_n such that $1 \leqslant c_n \to \pi$ as $n \to \infty$ and $e^{-1/6}\left(\frac{n}{2\pi}\right)^{1/2} c_n + 2 \leqslant \mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}), \mu)\right)\right]$. In this case the expectation on the right is independent of μ.*

rem. The logarithmic dependence on the sample count in the enumerator of the PAC Bayesian bound is halved.

According to (Langford and Schapire 2005), the author discusses basic prediction theory and its impact on classification success evaluation, implications for learning algorithm design, and uses in learning algorithm execution. This tutorial is meant to be a comprehensive compilation of results which are both theoretically rigorous and quantitatively useful.

The right side (i.e., $e^{1/(12n)}(\frac{\pi n}{2})^{1/2}+2$) is bounded above by $2\sqrt{n}$ for $n \geqslant 8$ and the left side (i.e., $e^{-1/6}(\frac{n}{2\pi})^{1/2}c_n+2$) is bounded below by $\sqrt{n}$ for $n \geqslant 2$, thus giving the simpler bounds ($\mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}),\mu)\right)\right] \leqslant 2\sqrt{n}$) and ($\sqrt{n} \leqslant \mathbf{E}\left[\exp\left(n\mathbf{KL}(M(\mathbf{X}),\mu)\right)\right]$) of the introduction. To prove Theorem 1 we need some auxilliary results. The first is Stirling's Formula:

Theorem (2). *For $n \in \mathbb{N}$, $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(g(n)/(12n)\right)$, with $0 < g(n) < 1$.*

We will use Theorem 2 in form of the following inequalities $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n}\right)$. The simple Lemma 3 shows that the expectation of a convex function of iid variables can always be bounded by the expectation of the corresponding Bernoulli variables. The next Lemma 4 is concerned with a series which can be viewed as a Rieman sum approximation an instance of the Beta-function.

Remark (Lemma 3). *Suppose that $f : [0,1]^n \to \mathbb{R}$ is convex. Then $\mathbf{E}[f(\mathbf{X})] \leqslant \mathbf{E}[f(\mathbf{X}')]$. If f is permutation symmetric in its arguments and $\boldsymbol{\theta}(k)$ denotes the vector $\boldsymbol{\theta}(k) = (1,...,1,0,...,0)$ in $\{0,1\}^n$, whose first k coordinates are 1 and whose remaining $(n-k)$ coordinates are zero, we also have $\mathbf{E}[f(\mathbf{X}')] = \sum_{k=0}^{n}\binom{n}{k}(1-\mu)^{n-k}\mu^k f\left(\boldsymbol{\theta}(k)\right)$.*

Remark (Lemma 4). *For $n \geqslant 2$ the sequence $c_n = \sum_{k=1}^{n-1}\frac{1}{\sqrt{k(n-k)}}$, satisfies $1 \leqslant c_n \leqslant \pi$, and $c_n \to \pi$ as $n \to \infty$.*

Machine Learning Notes

This note is based on Machine Learning at DIKU

12 December 2021

Chapter 2. Concentration of Measure Inequalities

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 1963, Theorem 4),[25] which we cite without a proof, allows to extend results for sampling with replacement to sampling without replacement.

Lemma 2.21. *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: $f : \mathbb{R} \to \mathbb{R}$,*

$$\mathbf{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \leqslant \mathbf{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right] . \tag{2.145}$$

[25] In this section it will be shown that the inequalities of section 2 yield probability bounds for the sum of a random sample without replacement from a finite population. Let the population C consist of N values $c_1, c_2, ..., c_N$. Let $X_1, X_2, ..., X_n$ denote a random sample without replacement from C and let $Y_1, Y_2, ..., Y_n$ denote a random sample with replacement from C. The random variables $Y_1, ..., Y_n$ are independent and identically distributed with mean μ and variance σ^2, where

$$\mu = \frac{1}{N}\sum_{i=1}^{N} c_i , \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i - \mu)^2 . \tag{2.143}$$

If $a \leqslant c_i \leqslant b$, Theorems 1,2, and 3 give upper bounds for $\mathbf{Pr}(\overline{Y} - \mu \geqslant t)$, where $\overline{Y} = (Y_1 + \cdots + Y_n)/n$. It will now be shown that the same bounds, with μ and σ^2 defined by (2.143), are upper bounds for $\mathbf{Pr}(\overline{X} - \mu \geqslant t)$, where $\overline{X} = (X_1 + \cdots + X_n)/n$. (Note that $\mathbf{E}[\overline{X}] = \mathbf{E}[\overline{Y}] = \mu$ but $\mathbf{Var}[\overline{X}] = \frac{N-n}{N-1}\frac{\sigma^2}{n} < \frac{\sigma^2}{n} = \mathbf{Var}[\overline{Y}]$). This will be an immediate consequence of Theorem 4.

Theorem (4). *If the function $f(x)$ is continuous and convex then*

$$\mathbf{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \leqslant \mathbf{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right] . \tag{2.144}$$

Applied to $f(x) = \exp(hx)$ the theorem yields the claimed result if we recall that the bounds of Theorems 1 to 3 have been obtained from inequality (2.153). *(Note that the inequality $\mathbf{Var}[\overline{X}] \leqslant \mathbf{Var}[\overline{Y}]$ is a special case of* (2.144).*)*

To prove Theorem 4 we first observe that for an arbitrary function g of n variables we have, in the notation of (2.147), $\mathbf{E}[g(X_1, ..., X_n)] = \frac{1}{N^{(n)}}\sum_{N,n} g(c_{i_1}, \cdots, c_{i_n})$, $\mathbf{E}[g(Y_1, ..., Y_n)] = \frac{1}{N^n}\sum_{i_1=1}^{N}\cdots\sum_{i_n=1}^{N} g(c_{i_1}, ..., c_{i_n})$, The right-hand sides are of the same form as U in (2.147) and W in (2.148), respectively. It has been observed in section 5c that W can be written as U with g replaced by an arithmetic mean g^* of values of g. It follows that $\mathbf{E}[g(Y_1, ..., Y_n)] = \mathbf{E}[g^*(X_1, ..., X_n)]$. As mentioned after (2.151), the function g^* is not uniquely determined. The version of $g^*(x_1, ..., x_n)$ which is symmetric in $x_1, ..., x_n$ will be denoted by $\bar{g}(x_1, ..., x_n)$. Here we are concerned with the special case $g(x_1, ..., x_n) = f(x_1 + ... + x_n)$.

In particular, Lemma 2.21 can be used to prove Hoeffding's inequality for sampling without replacement.

NB. We list some knowledge that we need to know (Hoeffding 1963),[26,27] here.

Theorem 2.22 (Hoeffding's inequality for sampling without replacement). *Let $X_1, ..., X_n$ denote a*

[26]The inequalities of sections 2 and 4 can be used to obtain probability bounds for certain sums of dependent random variables. Suppose that T is a random variable which can be written in the form

$$T = p_1 T_1 + p_2 T_2 + \cdots + p_N T_N \,, \tag{2.146}$$

where each of $T_1, T_2, ..., T_N$ is a sum of independent random variables and $p_1, p_2, ..., p_N$ are nonnegative numbers, $p_1 + p_2 + \cdots + p_N = 1$. The random variables $T_1, T_2, ..., T_N$ need not be mutually independent. For $h > 0$, $\mathbf{Pr}(T \geqslant t) \leqslant e^{-ht}\mathbf{E}[e^{hT}]$. Since the exponential function is convex, we have by Jensen's inequality (2.154), $\exp(hT) = \exp\left(h\sum_{i=1}^N p_i T_i\right) \leqslant \sum_{i=1}^N p_i \exp(hT_i)$. Therefore, $\mathbf{Pr}(T \geqslant t) \leqslant \sum_{i=1}^N p_i \mathbf{E}[e^{h(T_i - t)}]$. Since each T_i is a sum of independent random variables, the expectations on the right can be bounded as in section 4. If the random variables T_i are identically distributed or if the upper bound for $\mathbf{E}[\exp(h(T_i - t))]$ is independent of i, then the upper bound we obtain for $\mathbf{Pr}(T \geqslant t)$ is also an upper bound for $\mathbf{Pr}(T_i \geqslant t)$. The bounds obtained in this way will be rather crude but may be useful. We now consider several types of random variables T which can be represented in the form (2.146).

5a. One-sample U statistics. Let $X_1, X_2, ..., X_N$ be independent random variables (real or vector valued). For $n \geqslant r$ consider a random variable of the form

$$U = \frac{1}{n^{(r)}} \sum_{n,r} g(X_{i_1}, \cdots, X_{i_r}) \,, \tag{2.147}$$

where $n^{(r)} = n(n-1)\cdots(n-r+1)$ and the sum $\sum_{n,r}$ is taken over all r-tuples $i_1, \cdots, i_r$ of distinct positive integers not exceeding n. Random variables of the form (2.147) have been called (one-sample) U statistics.

5b. Two-sample U statistics. Let $X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$ be $(m+n)$ independent random variables. For $m \geqslant r$ and $n \geqslant s$ consider a random variable of the form $U = \frac{1}{m^{(r)} n^{(s)}} \sum_{m,r;n,s} g(X_{i_1}, \cdots, X_{i_r}, Y_{j_1}, \cdots, Y_{j_s})$, where the sum $\sum_{m,r;n,s}$ is taken over all r-tuples $(i_1, ..., i_r)$ of distinct positive integers $\leqslant m$ and all s-tuples $(j_1, ..., j_s)$ of distinct positive integers $\leqslant n$. A random variable of the form (26) has been called a two-sample U statistic.

5c. Sums related to U statistics. Let again $X_1, X_2, ..., X_n$ be independent and consider the random variable

$$W = \frac{1}{n^r} \sum_{i_1=1}^n \cdots \sum_{i_r=1}^n g(X_{i_1}, \cdots, X_{i_r}) \,, \tag{2.148}$$

A random variable W of the form (2.148) can be written as a U statistic,

$$W = \frac{1}{n^{(r)}} \sum_{n,r} g^*(X_{i_1}, \cdots, X_{i_r}) \,, \tag{2.149}$$

where $g^*(x_1, ..., x_r)$ is a weighted arithmetic mean of certain values of g. For example, for $r = 2$ and $r = 3$ we have, respectively,

$$g^*(x_1, x_2) = \tfrac{n-1}{n} g(x_1, x_2) + \tfrac{1}{n} g(x_1, x_1) \,, \tag{2.150}$$

$$g^*(x_1, x_2, x_3) = \tfrac{(n-1)(n-2)}{n^2} g(x_1, x_2, x_3) + \tfrac{n-1}{n^2}\{g(x_1, x_1, x_2) + g(x_1, x_2, x_1) + g(x_2, x_1, x_1)\} + \tfrac{1}{n^2} g(x_1, x_1, x_1) \,. \tag{2.151}$$

(The function g^* for which (2.149) is satisfied is not uniquely determined. For example, in (2.150) the value $g(x_1, x_1)$ may be replaced by $\frac{1}{2}g(x_1, x_1) + \frac{1}{2}g(x_2, x_2)$.)

5d. Sums of m-dependent random variables. Let $S = Y_1 + Y_2 + \cdots + Y_n$, where the sequence of random variables $Y_1, Y_2, ..., Y_n$ is $(r-1)$-dependent; that is, the random vectors $(Y_1, \cdots, Y_i)$ and $(Y_j, \cdots, Y_n)$ are independent if $j - i \geqslant r$, where r is a positive integer. (Example: $S = X_1 X_r + X_2 X_{r+1} + \cdots + X_n X_{r+n-1}$, where $X_1, X_2, ...$ are independent.) Then the random variables $Y_i, Y_{r+i}, Y_{2r+i}, ...$ are independent. For $i = 1, ..., r$ let $S_i = Y_i + Y_{r+i} + Y_{2r+i} + \cdots + Y_{n_i r - r + i}$, $n_i = [(n - i + r)/r]$. Then $S = S_1 + S_2 + \cdots + S_r$ and S_i is a sum of n_i independent random variables.

[27]Let $X_1, X_2, ..., X_n$ be independent random variables with finite first and second moments,

$$S = X_1 + \cdots + X_n \,, \qquad \overline{X} = S/n \,, \tag{2.152a}$$

$$\mu = \mathbf{E}[\overline{X}] = \mathbf{E}[S/n] \,, \qquad \sigma^2 = n\mathbf{Var}(\overline{X}) = \mathbf{Var}(S)/n \,. \tag{2.152b}$$

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* $\varepsilon > 0$,

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right), \tag{2.162a}$$

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leqslant -\varepsilon\right) = \mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right). \tag{2.162b}$$

(Thus if the X_i have a common mean then its value is μ and if they have a common variance then its value is σ^2.) In section 2 upper bounds are given for the probability $\mathbf{Pr}(\overline{X} - \mu \geqslant t) = \mathbf{Pr}(S - \mathbf{E}[S] \geqslant nt)$, where $t > 0$, under the additional assumption that the range of each random variable X_i is bounded (or at least bounded from above). These upper bounds depends only on t, n, the endpoints of the ranges of the X_i, and on μ, or on μ and σ. Note that an upper bound for $\mathbf{Pr}(\overline{X} - \mu \geqslant t)$ implies in an obvious way an upper bound for $\mathbf{Pr}(-\overline{X} + \mu \geqslant t)$ and hence also for $\mathbf{Pr}(|\overline{X} - \mu| \geqslant t) = \mathbf{Pr}(\overline{X} - \mu \geqslant t) + \mathbf{Pr}(-\overline{X} + \mu \geqslant t)$.

The probability $\mathbf{Pr}(S - \mathbf{E}[S] \geqslant nt)$ is the expected value of the function which takes the values 0 and 1 according as $(S - \mathbf{E}[S] - nt)$ is either < 0 or $\geqslant 0$. This function does not exceed $\exp\{h(S - \mathbf{E}[S] - nt)\}$, where h is an arbitrary positive constant. Hence

$$\mathbf{Pr}(\overline{X} - \mu \geqslant t) = \mathbf{Pr}(S - \mathbf{E}[S] \geqslant nt) \leqslant \mathbf{E}\left[\exp\left(h(S - \mathbf{E}[S] - nt)\right)\right]. \tag{2.153}$$

If, as we here assume, the summands of S are independent, then $\mathbf{E}\left[e^{h(S-\mathbf{E}[S]-nt)}\right] = e^{-hnt}\prod_{i=1}^{n}\mathbf{E}\left[e^{h(X_i-\mathbf{E}[X_i])}\right]$.

The following facts about convex functions will be used; for proofs see reference (Hardy et al. 1952). A continuous function $f(x)$ is convex in the interval I if and only if $f(px + (1-p)y) \leqslant p(f(x)) + (1-p)f(y)$ for $0 < p < 1$ and all x and y in I. If this is true for all real x and y, the function is simply called convex. A continuous function is convex in I if it has a nonnegative second derivative in I. If $f(x)$ is continuous and convex in I then for any positive numbers $p_1, ..., p_N$ such that $p_1 + \cdots + p_N = 1$ and any numbers $x_1, \cdots, x_N$ in I,

$$f\left(\sum_{i=1}^{N} p_i x_i\right) \leqslant \sum_{i=1}^{N} p_i f(x_i). \tag{2.154}$$

This is known as Jensen's inequality.

Let $X_1, X_2, ..., X_n$ be independent random variables and let $S, \overline{X}, \mu$, and σ^2 be defined by (2.152a) and (2.152b). First we consider bounds which do not depend on σ^2. Then Theorem 2 gives an extension of bound (2.155) to the case where the ranges of the summands need not be the same.

Theorem (1). *If* $X_1, X_2, ..., X_n$ *are independent and* $0 \leqslant X_i \leqslant 1$ *for* $i = 1, ..., n$, *then for* $0 < t < 1 - \mu$,

$$\mathbf{Pr}(\overline{X} - \mu \geqslant t) \leqslant \left\{\left(\frac{\mu}{\mu+t}\right)^{\mu+t}\left(\frac{1-\mu}{1-\mu-t}\right)^{1-\mu-t}\right\}^n \leqslant \exp\left(-nt^2 g(\mu)\right) \leqslant \exp\left(-2nt^2\right) \tag{2.155}$$

where

$$g(\mu) = \begin{cases} \frac{1}{1-2\mu}\ln\frac{1-\mu}{\mu}, & \text{for } 0 < \mu < \frac{1}{2}; \\ \frac{1}{2\mu(1-\mu)}, & \text{for } \frac{1}{2} \leqslant \mu < 1. \end{cases} \tag{2.156}$$

Theorem (2). *If* $X_1, X_2, ..., X_n$ *are independent and* $a_i \leqslant X_i \leqslant b_i$ $(i = 1, 2, ..., n)$, *then for* $t > 0$,

$$\mathbf{Pr}(\overline{X} - \mu \geqslant t) \leqslant \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.157}$$

Remark (Corollary). *If* $Y_1, ..., Y_m, Z_1, ..., Z_n$ *are independent random variables with values in the interval* $[a, b]$, *and if* $\overline{Y} = (Y_1 + \cdots + Y_m)/m, \overline{Z} = (Z_1 + \cdots + Z_n)/n$, *then for* $t > 0$,

$$\mathbf{Pr}(\overline{Y} - \overline{Z} - (\mathbf{E}[\overline{Y}] - \mathbf{E}[\overline{Z}]) \geqslant t) \leqslant \exp\left(\frac{-2t^2}{(m^{-1}+n^{-1})(b-a)^2}\right). \tag{2.158}$$

Theorem (3). *If* $X_1, X_2, ..., X_n$ *are independent,* $\mathbf{E}[X_i] = 0$, $X_i \leqslant b$ $(i = 1, 2, ..., n)$, *then for* $0 < t < b$,

$$\mathbf{Pr}(\overline{X} \geqslant t) \leqslant \left\{\left(1 + \frac{bt}{\sigma^2}\right)^{-\left(1+\frac{bt}{\sigma^2}\right)\frac{\sigma^2}{b^2+\sigma^2}}\left(1 - \frac{t}{b}\right)^{-\left(1-\frac{t}{b}\right)\frac{b^2}{b^2+\sigma^2}}\right\}^n \leqslant \exp\left(-\frac{nt}{b}\left[\left(1 + \frac{\sigma^2}{bt}\right)\ln\left(1 + \frac{bt}{\sigma^2}\right) - 1\right]\right). \tag{2.159}$$

Let X be a random variable such that $a \leqslant X \leqslant b$. Since the exponential function $\exp(hX)$ is convex, its graph is bounded above on the interval $a \leqslant X \leqslant b$ by the straight line which connects its ordinates at $X = a$ and $X = b$. Thus $e^{hX} \leqslant \frac{b-X}{b-a}e^{ha} + \frac{X-a}{b-a}e^{hb}, a \leqslant X \leqslant b$.

The proof is a minor adaptation of the proof of Hoeffding's inequality for sampling with replacement using Lemma 2.21 and is left as an exercise. (Note that is requires a small modification inside the proof, because Lemma 2.21 cannot be applied directly to the statement of Hoeffding's inequality.)

NB. **Question:** I'm not very sure how to demonstrate it.

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, 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 $\mathcal{X}$. Let $Y_1, ..., Y_n$ denote a random sample with replacement from $\mathcal{X}$. Then we have $Y_i \in [0, 1]$. To use the statement of Hoeffding's inequality, we still need to demonstrate that $\mathbf{E}[Y_i] = \mu$ for all i. Is that true?

Well,

According to Hoeffding's inequality (Theorem 2.3, Corollaries 2.4–2.5), we have: for avery $\varepsilon > 0$,

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right), \tag{2.163a}$$

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu \leqslant -\varepsilon\right) = \mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} Y_i \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right), \tag{2.163b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(-2n\varepsilon^2\right). \tag{2.163c}$$

Then according to Lemma 2.21, we have: for any continuous and convex function: $f : \mathbb{R} \to \mathbb{R}$, $\mathbf{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \leqslant \mathbf{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right]$. Note that we also have for any random variable z,

$$\mathbf{E}[z] = \begin{cases} \sum_{z\in\mathcal{Z}} z\mathbf{Pr}(Z = z), & \text{if } z \text{ is discrete;} \\ \int_{z\in\mathcal{Z}} z\mathbf{Pr}(Z = z)dz, & \text{if } z \text{ is continuous.} \end{cases} \tag{2.164}$$

Therefore, we will have: for $f(z) = z$,

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \leqslant \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right], \tag{2.165a}$$

Question: I'm not very sure.

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.

Remark (Lemma 1). *If X is a random variable such that $a \leqslant X \leqslant b$, then for any real number h,*

$$\mathbf{E}\left[e^{hX}\right] \leqslant \frac{b-\mathbf{E}[X]}{b-a}e^{ha} + \frac{\mathbf{E}[X]-a}{b-a}e^{hb}. \tag{2.160}$$

Remark (Lemma 2). *If X is a random variable such that $\mathbf{E}[X] = 0, \mathbf{E}[X^2] = \sigma^2$ and $X \leqslant b$, then for any positive number h,*

$$\mathbf{E}\left[e^{hX}\right] \leqslant \frac{b^2}{b^2+\sigma^2}\exp\left(-\frac{\sigma^2}{b}h\right) + \frac{\sigma^2}{b^2+\sigma^2}\exp(bh). \tag{2.161}$$

Remark (Lemma 3). *If $c > 0$, the function $f(u) = \ln\left(\frac{1}{1+u}e^{-cu} + \frac{u}{1+u}e^{c}\right)$ has a negative second derivative for $u \geqslant 0$.*

References

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Cover, Thomas M and Joy A Thomas (2006). *Elements of Information Theory*. Second Edition. Wiley Series in Telecommunications and Signal Processing.

Csiszár, Imre and János Körner (2011). *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press.

Feller, William (1957). *An Introduction to Probability Theory and Its Applications*. Second Edition. New York: John Wiley & Sons.

Hardy, Godfrey Harold et al. (1952). *Inequalities*. Cambridge: University Press.

Hoeffding, Wassily (1963). "Probability Inequalities for Sums of Bounded Random Variables". In: *J Am Stat Assoc* 58.301, pp. 13–30.

Langford, John and Robert Schapire (2005). "Tutorial on Practical Prediction Theory for Classification." In: *J Mach Learn Res* 6.3.

Ledoux, Michel (1997). "On Talagrand's deviation inequalities for product measures". In: *ESAIM Probab Stat* 1, pp. 63–87.

Marton, K (1999). "On a measure concentration inequality of Talagrand for dependent random variables". In: *Preprint*.

Marton, Katalin (1996). "A measure concentration inequality for contracting Markov chains". In: *Geom Funct Anal* 6.3, pp. 556–571.

Maurer, Andreas (2004). "A note on the PAC Bayesian theorem". In: *arXiv preprint cs/0411099*.

Samson, Paul-Marie (2000). "Concentration of measure inequalities for markov chains and ϕ-mixing processes". In: *Ann Probab* 28.1, pp. 416–461.

Talagrand, Michel (1995). "Concentration of measure and isoperimetric inequalities in product spaces". In: *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques* 81.1, pp. 73–205.