Group Assignment 2
Introduction to Data Science

Oskar Bäcklin, Elvis Schmidt, Edward Karlsson, Sam Farkhooi and Linus Jacobsson

October 2023

Proof of corollary 3.7 (Sam)

We want to prove the follwing corollary:

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

Proof. We assume that $n > 0$. Let $S_n := \sum_{i=1}^{n} X_i = n\overline{X}_n$

$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) = \mathbb{P}(n\overline{X}_n - n\mathbb{E}[\overline{X}_n] \leq -n\epsilon)$. Since $n\mathbb{E}[\overline{X}_n] = \mathbb{E}[n\overline{X}_n]$ we get
$\mathbb{P}(n\overline{X}_n - n\mathbb{E}[\overline{X}_n] \leq -n\epsilon) = \mathbb{P}(S_n - \mathbb{E}[S_n] \leq -n\epsilon) = \mathbb{P}(\mathbb{E}[S_n] - S_n \geq n\epsilon)$

Let $m > 0$ be an arbitrary constant, then: $\mathbb{P}(\mathbb{E}[S_n] - S_n \geq n\epsilon) = \mathbb{P}(e^{m(\mathbb{E}[S_n]-S_n)} \geq e^{nm\epsilon})$

Then using Markov's inequality we get that:

$$\mathbb{P}(e^{m(\mathbb{E}[S_n]-S_n)} \geq e^{nm\epsilon}) \leq \frac{\mathbb{E}(e^{m(\mathbb{E}(S_n)-S_n)})}{e^{nm\epsilon}} = e^{-nm\epsilon}\mathbb{E}(e^{m(\mathbb{E}(\sum_{i=1}^{n} X_i)-\sum_{i=1}^{n} X_i)})$$

Since $X_1, \ldots, X_n$ are IID it follows that

$$e^{-nm\epsilon}\mathbb{E}(e^{m(\mathbb{E}(\sum_{i=1}^{n} X_i)-\sum_{i=1}^{n} X_i)}) = e^{-nm\epsilon}\Pi_{i=1}^{n}\mathbb{E}(e^{m(\mathbb{E}(X_i)-X_i)}) = e^{-nm\epsilon}\Pi_{i=1}^{n}\mathbb{E}(e^{-m(X_i-\mathbb{E}(X_i))})$$

Why -m?

Now if we apply lemma 3.5 with $\lambda = -m$ we get that

$$\mathbb{E}(e^{-m(X_i-\mathbb{E}(X_i))}) \leq e^{\frac{(-m)^2(b-a)^2}{8}} = e^{\frac{m^2(b-a)^2}{8}} \quad \forall i \in \{1, \ldots, n\} \implies$$

$$\Pi_{i=1}^{n}\mathbb{E}(e^{-m(X_i-\mathbb{E}(X_i))}) \leq \Pi_{i=1}^{n} e^{\frac{m^2(b-a)^2}{8}} = \left(e^{\frac{m^2(b-a)^2}{8}}\right)^n = e^{\frac{nm^2(b-a)^2}{8}} \implies$$

$$e^{-nm\epsilon}\Pi_{i=1}^{n}\mathbb{E}(e^{-m(X_i-\mathbb{E}(X_i))}) \leq e^{-nm\epsilon} e^{\frac{nm^2(b-a)^2}{8}} = e^{\frac{nm^2(b-a)^2}{8} - nm\epsilon}$$

We have an upper bound with m arbitrary, so we can choose it to make the upper limit as tight as possible, to do that we should minimize the function:

$$h(m) = \frac{nm^2(b-a)^2}{8} - nm\epsilon$$

We'll do that using the derivative:

$$h'(m) = \frac{nm(b-a)^2}{4} - n\epsilon, \qquad h''(m) = \frac{n(b-a)^2}{4} > 0 \ \ \forall m \ \text{ (convex function)}$$

$$h'(m) = 0 \implies \frac{nm(b-a)^2}{4} - n\epsilon = 0 \implies \frac{nm(b-a)^2}{4} = n\epsilon \implies \frac{m(b-a)^2}{4} = \epsilon \implies m = \frac{4\epsilon}{(b-a)^2}$$

So the m value that minimizes the function is $m^* = \frac{4\epsilon}{(b-a)^2}$ and plugging this in we get:

$$h(m^*) = \frac{16n\epsilon^2(b-a)^2}{8(b-a)^4} - \frac{4n\epsilon^2}{(b-a)^2} = \frac{2n\epsilon^2}{(b-a)^2} - \frac{4n\epsilon^2}{(b-a)^2} = -\frac{2n\epsilon^2}{(b-a)^2}$$

So in conclusion:

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{-nm\epsilon}\Pi_{i=1}^n \mathbb{E}(e^{m(\mathbb{E}[X_i]-X_i)}) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}} \implies$$

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

□

Now for the second part of the proof where the goal is to show that

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

Proof.

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) = \mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon \ \vee \ \overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) =$$
$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) + \mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon)$$

The last equality holds since the two events are mutually exclusive.
Theorem 3.6 tells us that:

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

And above we proved that:

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

So

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) + \mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}} + e^{-\frac{2n\epsilon^2}{(b-a)^2}} = 2e^{-\frac{2n\epsilon^2}{(b-a)^2}} \implies$$

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

□

A proof of lemma 3.15(Oskar)

We prove the following properties for sub-Gaussian and sub-exponential variables.

Lemma. *Let X be a random variable, $\alpha \in \mathbb{R}$ and $(\Omega, \mathcal{F}, \mathbb{P})$ a probability triple.*

1. *If X is sub-Gaussian with parameter λ, then αX is sub-Gaussian with parameter $|\alpha|\lambda$.*
2. *If X is sub-exponential with parameter λ, then αX is sub-exponential with parameter $|\alpha|\lambda$.*
3. *If X is sub-Gaussian with parameter λ, then X is also sub-exponential with parameter λ.*
4. *If X is bounder s.t $\mathbb{P}(X \in [a,b]) = 1$, $a < b$, then X is sub-Gaussian with parameter $(b-a)/2$. Specifically, if X is Bernoulli distributed then X has parameter $1/2$.*

Proof. 1. Recall that X is sub-Gaussian if for all $t \in \mathbb{R}$ we have $\mathbb{E}[e^{t(X-\mathbb{E}X)}] \leq e^{\lambda^2 t^2/2}$. Let $Z := \alpha X$, then $\mathbb{E}Z = \mathbb{E}\alpha X = \alpha \mathbb{E}X$. Thus we have

$$\mathbb{E}[e^{s(Z-\mathbb{E}Z)}] = \mathbb{E}[e^{s(\alpha X - \alpha \mathbb{E}X)}] = \mathbb{E}[e^{s\alpha(X-\mathbb{E}X)}].$$

Now since X is sub-Gaussian and the inequality is valid for every t, and in particular for $t = s\alpha$, we have that

$$\mathbb{E}[e^{t(X-\mathbb{E}X)}] \leq e^{\lambda^2 t^2/2} = e^{\lambda^2 (s\alpha)^2/2} = e^{(|\alpha|\lambda)^2 s^2/2}.$$

Hence we have shown that

$$\mathbb{E}[e^{s(\alpha X - \alpha \mathbb{E}X)}] \leq e^{(|\alpha|\lambda)^2 s^2/2}, \text{ for all } s,$$

and conclude that αX is sub-Gaussian with parameter $|\alpha|\lambda$

2. We proceed as in 1. Recall now instead that X is called sub-exponential if for all $|t| < 1/\lambda$ we have that $\mathbb{E}[e^{t(X-\mathbb{E}X)}] \leq e^{\lambda^2 t^2/2}$. In particular for $t = \alpha s$ we have that

$$\mathbb{E}[e^{s(\alpha X - \alpha \mathbb{E}X)}] \leq e^{(|\alpha|\lambda)^2 s^2/2} \quad \text{for all } s \text{ such that} \quad |s| \leq \frac{1}{|\alpha|\lambda}$$

The bound on s we get using the fact that $|t| = |\alpha s| \leq 1/\lambda \implies |s| \leq \frac{1}{\alpha|\lambda}$

3. Again, note that if X is sub-Gaussian then the inequality $\mathbb{E}[e^{t(X-\mathbb{E}X)}] \leq e^{\lambda^2 t^2/2}$ holds *for all* t, then in particular it must hold for all $|t| \leq 1/\lambda$. Hence X is also sub-Exponential.

4. For this we use Hoeffdings lemma; If X satisfies $\mathbb{P}(X \in [a,b]) = 1$ for $a < b$ and X maps to $\mathbb{R}$. Then for all $t \in \mathbb{R}$, $\mathbb{E}[e^{t(X-\mathbb{E}X)}] \leq e^{t^2(b-a)^2/8}$. It's straightforward to rewrite this in the following way.

$$\mathbb{E}[e^{t(X-\mathbb{E}X)}] \leq e^{t^2(b-a)^2/8} = e^{t^2(\frac{b-a}{2})^2/2}$$

Evidently X is sub-Gaussian with parameter $(b-a)/2$.

Now if X is Bernoulli distributed then for every $\omega \in \Omega$ we have that $X(\omega) = 1$ or $X(\omega) = 0$. So clearly $\mathbb{P}(X \in [0,1]) = 1$. Hence by the above($a = 0, b = 1$), X is sub-Gaussian with parameter $1/2$.

This concludes the proof of properties 1-4 of lemma 3.15. □

Exercise 3.16 (Linus)

For the Poisson distribution, we have $E[e^{sX}] = e^{\lambda(e^s - 1)}$. Is this sub-Gaussian, sub-exponential, or neither?

Sub-Gaussian

A random variable X is sub-Gaussian with parameter λ if its centered moment-generating function (MGF) satisfies the following condition for all $s \in \mathbb{R}$:

$$E[e^{s(X-E[X])}] \leq e^{\frac{s^2\lambda^2}{2}}$$

For a Poisson distributed random variable X with mean λ, the centered MGF is:

$$M_X(s) = E[e^{s(X-\lambda)}] = e^{-\lambda s} \cdot E[e^{sX}] = e^{-\lambda s} \cdot e^{\lambda(e^s - 1)}$$

After simplification, we get:

$$M_X(s) = e^{\lambda(e^s-1-s)}$$

We focus on comparing the expressions $e^s - 1 - s$ and s^2 directly, rather than comparing the entire exponentials, because the exponential function is monotonic. This means that if $e^s - 1 - s$ grows faster than s^2, then the exponential of $e^s - 1 - s$ will grow faster than the exponential of s^2. Thus, it's sufficient to compare these expressions to draw conclusions about the behavior of their exponentials.

To compare the growth rates, we consider the limits of their ratio as s approaches infinity:

$$\lim_{s\to\infty} \frac{e^s - 1 - s}{s^2}$$

Applying L'Hôpital's rule:

$$\lim_{s\to\infty} \frac{e^s - 1}{2s}$$

Again applying L'Hôpital's rule:

$$\lim_{s\to\infty} \frac{e^s}{2} = \infty$$

As e^s grows exponentially, the limit goes to infinity, indicating that $e^s - 1 - s$ grows faster than s^2. Therefore, the inequality does not hold for all $s \in \mathbb{R}$

Sub-exponential

We will explore this by constructing the function

$$h(s) = \lambda(e^s - 1 - s) - \frac{\lambda^2 s^2}{2}.$$

Our goal is to show that there exists some $s \neq 0$ in the interval $|s| \leq \frac{1}{\lambda}$ such that $h(s) > 0$. This would prove that the Poisson distribution is not a sub-exponential.

Dividing $h(s)$ by λ (assuming $\lambda \neq 0$), we can simplify our analysis to the function

$$g(s) = e^s - 1 - s - \frac{\lambda s^2}{2},$$

since $h(s) > 0$ if and only if $g(s) > 0$.

Taking derivatives with respect to s, we find

$$g'(s) = e^s - 1 - \lambda s,$$

and

$$g''(s) = e^s - \lambda.$$

The critical point $s = 0$ satisfies $g'(0) = 0$. To determine the nature of this critical point, we examine $g''(s)$:

- For $0 < \lambda < 1$, $g''(0) = 1 - \lambda > 0$, indicating that $g(s)$ is convex near $s = 0$ and $s = 0$ is a local minimum. Hence, for small positive s, $g(s) > 0$.
- For $\lambda \geq 1$, while $g''(0) \leq 0$, $g''(s)$ becomes positive as s increases past $s = \ln(\lambda)$, where $e^s = \lambda$. Thus, $g(s)$ will eventually be greater than 0 for some $s > \ln(\lambda)$.

Thus we can find a lower bound for the lambdas, λ_* that make the distribution sub-exponential by solving:

$$\ln(\lambda) = \frac{1}{\lambda}.$$

This is a transcendental equation and does not have a closed-form solution. However, we can approximate the value of λ numerically.

$$\lambda_* \approx 1.763.$$

Note however that this is not guaranteed to be the absolute lowest λ, but we can be sure that every λ larger than this will fulfill the requirement of sub-exponentiality. Thus, we find that the Poisson distribution is sub-exponential, for $\lambda \geq 1.763$

Exercise 4.7 (Edward)

For the pattern recognition problem we want find a λ that minimizes

$$R(\lambda) = \mathbb{E}[L(Y, g_\lambda(X)].$$

where $L(Y, g_\lambda(X))$ is the loss function for $z = (x, y)$

$$L(z, u) = \begin{cases} 0 & \text{if } u = y \\ 1 & \text{if } u \neq y \end{cases}$$

and g_λ is a decision function.
This means that given y we want to minimize $R(\lambda)$. For a statistical model we then want to find a family of distributions that generates the values $R(\lambda)$ takes for y and λ. We start by noting that given λ the probability of $L(Y, g_\lambda(X)) = 1$ is

$$\mathbb{P}(\{g_\lambda(X) \neq Y\})$$

which means that it's reasonable to assume that the data is generated by probabilities of the form

$$\mathbb{P}(g_\lambda(x) \neq y) = Ber(p),\ p \in [0, 1]$$

then $\mathbb{E}[L(y, g_\lambda)]$ is generated by Bernoulli distributions giving us for distributions of the generators of $R(\lambda)$, f that

$$\{f(\lambda) = Ber(p);\ p \in [0, 1]\}$$

which is a statisitcal model that generates the values of $R(\lambda)$.

The details of the proof in 4.9 (Elvis)

Define

$$R(g) = \mathbb{E}[L(Y, g(x))]$$

where L is the 0-1-Loss function, namely

$$L(y, g(x)) = \begin{cases} 0 & y = g(x) \\ 1 & y \neq g(x) \end{cases}$$

We moreover define

$$r(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x)$$

and

$$h^*(x) = \begin{cases} 1 & r(x) > \frac{1}{2} \\ 0 & r(x) \leq \frac{1}{2} \end{cases}$$

The theorem in the lecture notes states that for any decision function $g(x)$ taking values in $\{0, 1\}$, it holds that $R(h^*) \leq R(g)$.

Proof. In the first step, we use the tower property to conclude that

$$\mathbb{E}[L(Y, g(X))] = \mathbb{E}\left[\mathbb{E}[L(Y, g(X))|X]\right]$$

In the next step we use the following properties:

1. $L(Y, g(X)) = 1 - \mathbb{1}_{Y=g(X)}$ and expectation is linear.
2. $\mathbb{1}_{Y=g(X)} = 1$ if and only if $(Y = 1 \wedge g(X) = 1) \vee (Y = 0 \wedge g(X) = 0)$. Thus
$$\mathbb{1}_{Y=g(X)} = \mathbb{1}_{Y=1}\mathbb{1}_{g(X)=1} + \mathbb{1}_{Y=0}\mathbb{1}_{g(X)=0}$$
3. $(f(X)|X = x) = f(x)$ deterministicly. Thus once again by linearity:
$$\mathbb{E}[\mathbb{1}_{g(X)=1}\mathbb{1}_{Y=1}|X = x] = \mathbb{E}[(\mathbb{1}_{g(X)=1}|X = x)(\mathbb{1}_{Y=1}|X = x)] = \mathbb{1}_{g(x)=1}\mathbb{E}[\mathbb{1}_{y=1}|X = x]$$
The same of course holds for the case "$= 0$"
4. $$\mathbb{E}[\mathbb{1}_{Y=1}|X = x] = \sum_{k=0}^{1} k \cdot \mathbb{P}(\mathbb{1}_{Y=1} = k|X = x) = \mathbb{P}(\mathbb{1}_{Y=1} = 1|X = x) = r(x)$$
Also, since $\mathbb{1}_{Y=0} = 1 - \mathbb{1}_{Y=1}$ we get that
$$\mathbb{E}[\mathbb{1}_{Y=0}|X = x] = \mathbb{P}(\mathbb{1}_{Y=0} = 1|X = x) = \mathbb{P}(1 - \mathbb{1}_{Y=1} = 1|X = x) = \mathbb{P}(-\mathbb{1}_{Y=1} = |X = x)$$
$$= \mathbb{P}(\mathbb{1}_{Y=1} = 0|X = x) = 1 - \mathbb{P}(\mathbb{1}_{Y=1} = 1|X = x) = 1 - r(x)$$

We can now conclude that (this part is taken from the lecture notes)

$$\begin{aligned}
\mathbb{E}[L(Y, g(X))|X = x] &\overset{1.}{=} 1 - \mathbb{E}[\mathbb{1}_{Y=g(X)}|X = x] && (1)\\
&\overset{2.}{=} 1 - \mathbb{E}[\mathbb{1}_{g(X)=1}\mathbb{1}_{Y=1} + \mathbb{1}_{g(X)=0}\mathbb{1}_{Y=0}|X = x] && (2)\\
&\overset{3.}{=} 1 - \mathbb{1}_{g(x)=1}\mathbb{E}[\mathbb{1}_{Y=1}|X = x] - \mathbb{1}_{g(x)=0}\mathbb{E}[\mathbb{1}_{Y=0}|X = x] && (3)\\
&\overset{4.}{=} 1 - \mathbb{1}_{g(x)=0}r(x) - \mathbb{1}_{g(x)=0}(1 - r(x)) && (4)
\end{aligned}$$

Using that $(*) : \mathbb{1}_{Y=0} = 1 - \mathbb{1}_{Y=1}$ we get that (again from the lecture notes)

$$\mathbb{E}[L(Y, g(X)|X = x] - \mathbb{E}[L(Y, h^*(X)|X = x] =$$

$$\begin{aligned}
&= -\mathbb{1}_{g(x)=1}r(x) - \mathbb{1}_{g(x)=0}(1 - r(x)) + \mathbb{1}_{h^*(x)=1}r(x) + \mathbb{1}_{h^*(x)=0}(1 - r(x)) && (5)\\
&= r(x)(\mathbb{1}_{h^*(x)=1} - \mathbb{1}_{g(x)=1}) + (1 - r(x))(\mathbb{1}_{h^*(x)=0} - \mathbb{1}_{g(x)=0}) && (6)\\
&\overset{(*)}{=} r(x)(\mathbb{1}_{h^*(x)=1} - \mathbb{1}_{g(x)=1}) + (1 - r(x))(1 - \mathbb{1}_{h^*(x)=1} - (1 - \mathbb{1}_{g(x)=1})) && (7)\\
&= r(x)(\mathbb{1}_{h^*(x)=1} - \mathbb{1}_{g(x)=1}) - (1 - r(x))(\mathbb{1}_{h^*(x)=1} - \mathbb{1}_{g(x)=1})) && (8)\\
&= (2r(x) - 1)(\mathbb{1}_{h^*(x)=1} - \mathbb{1}_{g(x)=1}) && (9)\\
&= \begin{cases} (2r(x) - 1)(1 - \mathbb{1}_{g(x)=1}) & \text{if } r(x) > \frac{1}{2} \\ (2r(x) - 1)(0 - \mathbb{1}_{g(x)=1}) & \text{if } r(x) \leq \frac{1}{2} \end{cases} && (10)
\end{aligned}$$

In both cases, the expression is non-negative. Therefore we can conclude that $\mathbb{E}[L(Y, g(X)|X = x] \geq \mathbb{E}[L(Y, h^*(X))|X = x]$. Recall that $R(g) := \mathbb{E}[\mathbb{E}[L(Y, g(X)|X = x]]$. To prove the theorem it thus suffices to show that expectation is monotone, i.e.

$$X \geq Y \implies \mathbb{E}[X] \geq \mathbb{E}[Y]$$

Recall that by linearity of expectation $\mathbb{E}[X] - \mathbb{E}[Y] = \mathbb{E}[X - Y] := \mathbb{E}[Z]$. Let Ω be the sample space of Z. Since $Z \geq 0$ it follows that

$\mathbb{E}[Z] = \sum_{z\in\Omega} zf(z) \geq 0$ and consequently that $\mathbb{E}[X] \geq \mathbb{E}[Y]$. Hence $R(h^*) \leq R(g)$. □