ETH Zurich Guarantees for Machine Learning 23 **(Due: 12/10/23)**

Homework #1: Concentration bounds

Name: XXX, *Student ID:* XXX *Students discussed with:* None

Problem 1: Optional Warm-up Optimality of polynomial Markov

Solution:
(a) Simply we let $X = a$ with probability 1, which is equivalent to the statement $\mathbb{E}(X \geq a) = 1$. $\mathbb{E}(X) = \sum_i i * \mathbb{P}(X = i) = a * \mathbb{P}(X = a) = a$ since $\mathbb{P}(x \neq a) = 0$ according to the definition of random variable X. Therefore, $\mathbb{E}(X \geq a) = 1 = \frac{a}{a} = \frac{\mathbb{E}(X)}{a}$, which meets the equality condition of Markov's inequality at a point $a > 0$.

(b) Recall that the Taylor expansion for $e^{\lambda X} = \sum_{i=0} \frac{(\lambda X)^i}{i!}$, then we relate $\mathbb{E}(e^{\lambda X})$ to δ with numerator and denominator multiplying δ^i:

$$\begin{aligned}
\mathbb{E}(e^{\lambda X}) &= \sum_{i=0} \frac{(\lambda X)^i}{i!} \\
&= \sum_{i=0} \frac{(\lambda X)^i}{i!} * \frac{\delta^i}{\delta^i} \\
&= \sum_{i=0} \frac{(\lambda \delta)^i}{i!} * \frac{X^i}{\delta^i} \\
&\geq \sum_{i=0} \frac{(\lambda \delta)^i}{i!} * \inf_{k=0,1,\ldots} \frac{\mathbb{E}|X|^k}{\delta^k}, \text{this holds since } X > 0 \\
&= e^{\lambda \delta} * \inf_{k=0,1,\ldots} \frac{\mathbb{E}|X|^k}{\delta^k}
\end{aligned} \tag{0.1}$$

Therefore, we achieve

$$\frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda \delta}} \geq \inf_{k=0,1,\ldots} \frac{\mathbb{E}|X|^k}{\delta^k}$$

Take infimum on both sides, we get:

$$\inf_{\lambda > 0} \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda \delta}} \geq \inf_{k=0,1,\ldots} \frac{\mathbb{E}|X|^k}{\delta^k}$$

Problem 2: Concentration and kernel density estimation

Solution: Let $g_n(X_1, X_2, ..., X_n) := \|f_n(x) - f\|_1$. We would like to bound $g_n(X_1, X_2, ..., X_n)$ so that we could use McDiarmid inequality to infer $\mathbb{P}(g_n(X_1, X_2, ..., X_n) - \mathbb{E}(g_n(X_1, X_2, ..., X_n)) \geq \delta)$ if such g_n satisfies bounded difference condition.

Given

$$\begin{cases} f_n(x) := \dfrac{1}{nh}\displaystyle\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \\ \|f_n(x) - f\|_1 = \displaystyle\int_{-\infty}^{+\infty} |f_n(t) - f(t)|dt \end{cases}$$

We bound $\epsilon = |g_n(X_1, X_2, ..., X_k, ..., X_n) - g_n(X_1, X_2, ..., X'_k, ..., X_n)|$:

$$\begin{aligned}
\epsilon &= \int_{-\infty}^{+\infty} |f_n(t) - f(t)|dt - \int_{-\infty}^{+\infty} |f'_n(t) - f(t)|dt \\
&\leq \int_{-\infty}^{+\infty} |f_n(t) - f'_n(t)|dt \qquad \text{(triangle inequality)} \\
&= \int_{-\infty}^{+\infty} |\frac{1}{nh}\left(\sum_{i=1}^{k-1} K\left(\frac{t - X_i}{h}\right) + K\left(\frac{t - X_k}{h}\right) + \sum_{i=k+1}^{n} K\left(\frac{t - X_i}{h}\right)\right) \\
&- \frac{1}{nh}\left(\sum_{i=1}^{k-1} K\left(\frac{t - X_i}{h}\right) + K\left(\frac{t - X'_k}{h}\right) + \sum_{i=k+1}^{n} K\left(\frac{t - X_i}{h}\right)\right)|dt \\
&= \int_{-\infty}^{+\infty} |\frac{1}{nh}\left(K\left(\frac{t - X_k}{h}\right) - K\left(\frac{t - X'_k}{h}\right)\right)|dt \\
&\leq \frac{2}{nh}\int_{-\infty}^{+\infty} K\left(\frac{t - X_k}{h}\right) dt \qquad \text{Remove absolute since K: } \mathbb{R} \to [0, \infty] \\
&= \frac{2h}{nh}\int_{-\infty}^{+\infty} K\left(w\right) dw \qquad \text{Change of variable } w = \frac{t - X_k}{h} \to dw = \frac{1}{h}dt \\
&= \frac{2}{n} \qquad \text{because } \int_{-\infty}^{+\infty} K\left(w\right) dw = 1
\end{aligned} \tag{0.2}$$

We let $\sigma_k = \frac{2}{n}$. According to the McDiarmid theorem, if g_n satisfies the bounded difference condition, and X is a random vector with n independent entries, then

$$\begin{aligned}
\mathbb{P}(g_n(X) - \mathbb{E}(g_n(X)) \geq \delta) &\leq e^{-\frac{2\delta^2}{\sum_{k=1}^{n} \sigma_k^2}} \\
&= e^{-\frac{2\delta^2}{\frac{2}{n}^2 * n}} \\
&= e^{-\frac{n\sigma^2}{2}} \\
&\leq e^{-\frac{n\sigma^2}{18}}
\end{aligned} \tag{0.3}$$

Hence we get into the conclusion that:

$$\mathbb{P}[\|f_n - f\|_1 \geq \mathbb{E}[\|f_n - f\|_1] + \delta] \leq e^{-\frac{n\sigma^2}{18}} \tag{0.4}$$

Problem 3: Sub-Gaussian maxima

Solution:

(a) let $Y = \max_{i=1,\ldots,n} X_i$. According to Jensen's inequality applied to the convex function $e^{\lambda Y}$,

$$e^{\lambda \mathbb{E}(\max_{i=1,\ldots,n} X_i)} \leq \mathbb{E}(e^{\lambda \max_{i=1,\ldots,n} X_i})$$

Obviously, max function could be upper bounded by sum of X_1 to X_n:

$$e^{\lambda \max_{i=1,\ldots,n} X_i} \leq e^{\lambda \sum_{i=1}^{n} X_i}$$

therefore:

$$\mathbb{E}\left(e^{\lambda \max_{i=1,\ldots,n} X_i}\right) \leq \mathbb{E}\left(e^{\lambda \sum_{i=1}^{n} X_i}\right)$$

Use the linearity of expectation, we have:

$$\mathbb{E}\left(e^{\lambda \sum_{i=1}^{n} X_i}\right) = \sum_{i=1}^{n} \mathbb{E}\left(e^{\lambda X_i}\right)$$

X_i serves subgaussian with parameter σ:

$$\mathbb{E}(e^{\lambda X_i}) \leq e^{\frac{\lambda^2 \sigma^2}{2}}$$

Combining those facts, we could bound $e^{\lambda \mathbb{E}(\max_{i=1,\ldots,n} X_i)}$:

$$\begin{aligned} e^{\lambda \mathbb{E}(\max_{i=1,\ldots,n} X_i)} &\leq \mathbb{E}(e^{\lambda \sum_{i=1}^{n} X_i}) \\ &= \sum_{i=1}^{n} \mathbb{E}\left(e^{\lambda X_i}\right) \\ &\leq \sum_{i=1}^{n} e^{\frac{\lambda^2 \sigma^2}{2}} = n e^{\frac{\lambda^2 \sigma^2}{2}} \\ \rightarrow \mathbb{E}(\max_{i=1,\ldots,n} X_i) &\leq \frac{1}{\lambda} \log n e^{\frac{\lambda^2 \sigma^2}{2}} \\ &= \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2} \end{aligned} \tag{0.5}$$

where it holds for all λ. We could minimize $\frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2}$ w.r.t. λ. Then solve $\frac{\log n}{\lambda} = \frac{\lambda \sigma^2}{2}$, we get $\lambda = \frac{\sqrt{2 \log n}}{\sigma}$. Then

$$\begin{aligned} \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2} &= \frac{\log n}{\frac{\sqrt{2 \log n}}{\sigma}} + \frac{\frac{\sqrt{2 \log n}}{\sigma} \sigma^2}{2} \\ &= \frac{\sqrt{2 \log n} \sigma}{2} * 2 \\ &= \sqrt{2 \sigma^2 \log n} \end{aligned} \tag{0.6}$$

Therefore, we have:

$$\mathbb{E}(\max_{i=1,\ldots,n} X_i) \leq \sqrt{2 \sigma^2 \log n}$$

(b) $\max_{i=1,\ldots,n}|X_i| = \max\{X_1, -X_1, X_2, -X_2, ..., X_n, -X_n\}$. We know that $-X_i$ is also σ-subgaussian with zero mean. X_i itself is σ-subgaussian with zero mean. Therefore, it is equivalent to say that we create a 2n-long X_i with each X_i zero mean and σ-subgaussian. Since we have 1) that

$$\mathbb{E}(\max_{i=1,\ldots,n} X_i) \leq \sqrt{2\sigma^2 \log n}$$

, substituting 2n instead of n leads to the answer:

$$\mathbb{E}(\max_{i=1,\ldots,n} |X_i|) \leq \sqrt{2\sigma^2 \log 2n} \leq 2\sqrt{\sigma^2 \log n}$$

, since $n \geq 2 \rightarrow 2n \leq n^2 \rightarrow \log 2n \leq \log n^2 \rightarrow \log 2n \leq 2 \log n$, which leads to the result.

Problem 4 Bonus: Sharper tail bounds for bounded variables: Bennett's inequality

Solution: (a) We could rewrite $\mathbb{E}e^{\lambda X_i}$ as:

$$\mathbb{E}e^{\lambda X_i} = \mathbb{E}(1 + \lambda X_i + e^{\lambda X_i} - 1 - \lambda X_i)$$

With the linearity of expectation and $\mathbb{E}X_i = 0$ as given,

$$\begin{aligned}\mathbb{E}e^{\lambda X_i} &= \mathbb{E}(1) + \lambda\mathbb{E}(X_i) + \mathbb{E}(e^{\lambda X_i} - 1 - \lambda X_i) \\ &= 1 + \mathbb{E}\left(\lambda^2 X_i^2 \frac{e^{\lambda X_i} - 1 - \lambda X_i}{\lambda^2 X_i^2}\right)\end{aligned} \tag{0.7}$$

We let $g(x) = \frac{e^x - 1 - x}{x^2}$. It can be rewritten in the form of Talyer expansion to observe the trend of g(x):

$$g(x) = \frac{e^x - 1 - x}{x^2} = \sum_{i=2} \frac{x^i}{i!} * x^{-2}$$

$$g'(x) = \sum_{i=2} (i-2)\frac{x^{i-3}}{i!} > 0$$

, when $x > 0$, the gradient is above zero, so the maximum of g(x) is bounded by $g(\lambda \max_i X_i) = g(\lambda b)$ Therefore,

$$\begin{aligned}1 + \mathbb{E}\left(\lambda^2 X_i^2 \frac{e^{\lambda X_i} - 1 - \lambda X_i}{\lambda^2 X_i^2}\right) &\leq 1 + \mathbb{E}\left(\lambda^2 X_i^2 \frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2}\right) \\ &= 1 + \mathbb{E}(\lambda^2 X_i^2)\frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2}\end{aligned} \tag{0.8}$$

$\mathbb{E}(X_i^2) = V(X_i) + \mathbb{E}(X_i)^2 = \sigma_i^2 \rightarrow$

$$1 + \mathbb{E}(\lambda^2 X_i^2)\frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2} = 1 + \lambda^2 \sigma_i^2 \frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2}$$

$$\begin{aligned}\rightarrow \log \mathbb{E}e^{\lambda X_i} &\leq \log\left(1 + \lambda^2 \sigma_i^2 \frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2}\right) \\ &\leq \lambda^2 \sigma_i^2 \frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2} \quad \text{since } \log(1+x) \leq x\end{aligned} \tag{0.9}$$

Finish the proof.

(b) Bound $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \delta\right) \leftrightarrow$ bound $\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}X \geq n\delta\right)$, where we let $X = \sum_{i=1}^n X_i$, where the expectation is zero as X_i has zero mean. Then we could use Chernoff bound:

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}X \geq n\delta\right) \leq \frac{\mathbb{E}e^{\lambda \sum_{i=1}^n X_i}}{e^{\lambda n\delta}}$$

$$\begin{aligned}
\log \mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}X \geq n\delta\right) &\leq \sum_{i=1}^n \log \mathbb{E}e^{\lambda X_i} - \lambda n\delta \\
&\leq \sum_{i=1}^n \lambda^2\sigma_i^2 \frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2} - \lambda n\delta \\
&= n\lambda^2\sigma^2 \frac{e^{\lambda b} - 1 - \lambda b}{\lambda^2 b^2} - \lambda n\delta \\
&= -\frac{n\sigma^2}{b^2}\left(\lambda n\delta * \frac{b^2}{n\sigma^2} - e^{\lambda b} + 1 + \lambda b\right) \\
&= -\frac{n\sigma^2}{b^2}\left(\frac{b\delta}{\sigma^2} * \lambda b - e^{\lambda b} + 1 + \lambda b\right)
\end{aligned} \tag{0.10}$$

let $h(t) = \inf_x \left(t * x - e^x + 1 + x\right)$. Set the gradient w.r.t. x equal to $0 \rightarrow t - e^x + 1 = 0 \rightarrow x = \log(t+1)$. So $h(t) = t\log(t+1) - 1 - t + 1 + \log(t+1) = (t+1)\log(t+1) - t$, which leads: $\left(\frac{b\delta}{\sigma^2} * \lambda b - e^{\lambda b} + 1 + \lambda b\right) = h(\frac{b\delta}{\sigma^2})$. Therefore:

$$\log \mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}X \geq n\delta\right) \leq -\frac{n\sigma^2}{b^2} h(\frac{b\delta}{\sigma^2})$$

, which is equivalently the Bennett's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \delta\right) \leq e^{-\frac{n\sigma^2}{b^2}h(\frac{b\delta}{\sigma^2})}$$

, where $h(t) = (t+1)\log(t+1) - t$ for $t \geq 0$

(c) remind the Bernstein's inequality for this question:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \delta\right) \leq e^{-\frac{t^2}{2(\sigma^2+bt)}}$$

To show that Bennett's inequality is at least as good as Bernstein's inequality:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \delta\right) \leq e^{-\frac{n\sigma^2}{b^2}h(\frac{b\delta}{\sigma^2})} \leq e^{-\frac{\delta^2}{2(\sigma^2+b\delta)}}$$

We must have

$$\begin{aligned} -\frac{\sigma^2}{b^2}h\left(\frac{b\delta}{\sigma^2}\right) &\leq -\frac{\delta^2}{2(\sigma^2+b\delta)} \\ &= -\frac{\sigma^2}{b^2}\frac{b^2\delta^2}{2(\sigma^2+b\delta)\sigma^2} \\ &= -\frac{\sigma^2}{b^2}\frac{(b\delta/\sigma^2)^2}{2(1+b\delta/\sigma^2)} \end{aligned} \tag{0.11}$$

Let $g(x) = \frac{x^2}{2(1+x)}$, then it is equivalently to prove that:

$$h\left(\frac{b\delta}{\sigma^2}\right) \geq g\left(\frac{b\delta}{\sigma^2}\right)$$

Let $f(t) = h(t) - g(t) = (t+1)\log(t+1) - t - \frac{t^2}{2(1+t)}$. We first check that $f(0) = 0$. Then we check the first order derivative of f(t) at t = 0: $f'(t) = 1 + \log(t+1) - 1 - \frac{2t}{2(1+t)} + \frac{t^2}{2(t+1)^2} = 0$. So we additionally check the second order derivative of f(t): $f''(t) = \frac{1}{t+1} - \frac{1}{(1+t)^2} - \frac{t}{(t+1)^3} = \frac{t^2}{(1+t)^3} \geq 0$ when $t \geq 0$. Therefore,

$$f\left(\frac{b\delta}{\sigma^2}\right) \geq 0 \rightarrow h\left(\frac{b\delta}{\sigma^2}\right) \geq g\left(\frac{b\delta}{\sigma^2}\right)$$

Then we could equivalently state that Bennett's inequality is at least as good as Bernstein's inequality.

Problem 5 Sharp upper bound on binomial tails

Solution: (a) We can use the Chernoff bound:

$$\begin{aligned}
\mathbb{P}[Z_n \le \delta n] &= \mathbb{P}[e^{-\lambda Z_n} \ge e^{-\lambda \delta n}] \\
&\le e^{\lambda\delta n}\mathbb{E}[e^{-\lambda Z_n}] \quad \text{Chernoff bound} \\
&= e^{\lambda\delta n}\mathbb{E}[e^{-\lambda \sum_{i=1}^n X_i}] \quad \text{Definition of } Z_i \\
&= e^{\lambda\delta n}\prod_{i=1}^n \mathbb{E}[e^{-\lambda X_i}] \quad \text{i.i.d of } X_i \\
&= e^{\lambda\delta n}\prod_{i=1}^n [\alpha e^{-\lambda} + (1-\alpha)] \quad \text{Bernoulli variable } X_i \text{ with parameter } \alpha \\
&= e^{\lambda\delta n}\left(\alpha e^{-\lambda} + (1-\alpha)\right)^n \\
\rightarrow \log \mathbb{P}[Z_n \le \delta n] &\le n\log\left(\alpha e^{-\lambda} + (1-\alpha)\right) + \lambda\delta n
\end{aligned}$$

Then we solve

$$\inf_{\lambda>0} n\log\left(\alpha e^{-\lambda} + (1-\alpha)\right) + \lambda\delta n$$

$$\begin{aligned}
\frac{d\left(n\log\left(\alpha e^{-\lambda} + (1-\alpha)\right) + \lambda\delta n\right)}{d\lambda} &= \frac{-n\alpha e^{-\lambda}}{(\alpha e^{-\lambda} + (1-\alpha))} + \delta n = 0 \\
\rightarrow \delta n\left(\alpha e^{-\lambda} + (1-\alpha)\right) &= n\alpha e^{-\lambda} \\
n\alpha(1-\delta)e^{-\lambda} &= \delta n(1-\alpha) \\
e^{\lambda} &= \frac{1-\delta}{\delta} * \frac{\alpha}{1-\alpha} \\
\lambda &= \log\left(\frac{1-\delta}{\delta} * \frac{\alpha}{1-\alpha}\right)
\end{aligned} \tag{0.12}$$

Then we take $e^{-\lambda}$ and λ back to the RHS of the log-inequality above, we'll achieve:

$$\begin{aligned}
\text{RHS} &= n\log\left(\alpha\frac{\delta}{1-\delta} * \frac{1-\alpha}{\alpha} + (1-\alpha)\right) + \delta n\log\left(\frac{1-\delta}{\delta} * \frac{\alpha}{1-\alpha}\right) \\
&= n\log\frac{1-\alpha}{1-\delta} + \delta n\log\left(\frac{1-\delta}{\delta} * \frac{\alpha}{1-\alpha}\right) \\
&= -n\left(\delta\log\frac{\delta}{\alpha} + (1-\delta)\log\frac{1-\delta}{1-\alpha}\right) \\
&= -nD(\delta||\alpha)
\end{aligned} \tag{0.13}$$

where $D(\delta||\alpha) = \left(\delta\log\frac{\delta}{\alpha} + (1-\delta)\log\frac{1-\delta}{1-\alpha}\right)$ represents the KL divergence between the Bernoulli distributions with parameters δ and α. Therefore,

$$\mathbb{P}[Z_n \le \delta n] \le e^{-nD(\delta||\alpha)}$$

(b) We know that a bounded variable within range [a, b] is $\frac{b-a}{2}$-subgaussian, therefore we quickly know that Bernoulli random variable X_i is $\frac{1}{2}$-subgaussian. Therefore the sum of Bernoulli random variable Z_i is $\frac{n}{2}$-subgaussian. From probability theory we know that the expectation of Bernoulli random variable X_i is equal to its parameter which means that $\mathbb{E}[X_i] = \alpha$, then we could write the Hoeffding bound on random variable Z_i:

$$\mathbb{P}[\sum_{i=1}^{n} Z_i \leq \delta n] = \mathbb{P}[\sum_{i=1}^{n} X_i - \mathbb{E}X_i \leq (\delta - \alpha)n]$$

. According to the symmetric property of Hoeddfing bound, we could let $t = (\alpha - \delta)n$, since from symmetry and Hoeffding, we know that

$$\mathbb{P}[\sum_{i=1}^{n} X_i - \mathbb{E}X_i \leq -t] = \mathbb{P}[\sum_{i=1}^{n} X_i - \mathbb{E}X_i \geq t] \leq e^{-\frac{nt^2}{2\sigma^2}}$$

, where here $\sigma = \frac{n}{2}$ being analyzed before. Then,

$$\mathbb{P}[\sum_{i=1}^{n} X_i - \mathbb{E}X_i \geq t] \leq e^{-\frac{n(\alpha-\delta)^2/n^2}{2n^2/2^2}} = e^{-2n(\alpha-\delta)^2}$$

If the bound from part (a) is strictly better than Hoeffding bound as shown above, we must have:

$$\mathbb{P}[Z_n \leq \delta n] \leq e^{-nD(\delta||\alpha)} \leq e^{-2n(\alpha-\delta)^2} \tag{0.14}$$

which means:

$$D(\delta||\alpha) \geq 2n(\alpha - \delta)^2 \tag{0.15}$$

Proof of $D(\delta||\alpha) \geq 2n(\alpha - \delta)^2$:
Let $f(x) = \delta \log x + (1 - \delta) \log(1 - x)$, then

$$\begin{aligned}
D(\delta||\alpha) &= f(\delta) - f(\alpha) \\
&= \int_{\alpha}^{\delta} f'(x)dx \\
&= \int_{\alpha}^{\delta} \frac{\delta}{x} - \frac{1-\delta}{1-x}dx \\
&= \int_{\alpha}^{\delta} \frac{\delta - x}{x(1-x)}dx \\
&\geq 4 \int_{\alpha}^{\delta} \delta - xdx \quad \text{since } \frac{1}{x(1-x)} \geq 4, \forall x \in (0, \frac{1}{2}] \\
&= 4 \left(\left(-\frac{1}{2}(\delta - x)^2 \right) \Big|_{x=\delta} - \left(-\frac{1}{2}(\delta - x)^2 \right) \Big|_{x=\alpha} \right) \\
&= 4 \left(0 + \frac{1}{2}(\delta - \alpha)^2 \right) \\
&= 2(\delta - \alpha)^2
\end{aligned}$$

Therefore we prove the property for bounding $D(\delta||\alpha)$, where $D(\delta||\alpha) \geq 2(\delta - \alpha)^2$. Therefore we could get into the conclusion that: the bound from part (a) is strictly better than Hoeffding bound.

Problem 6 Robust estimation of the mean

Solution: We evaluate the **empirical** mean of these $X_1, X_2, ..., X_n$ for a simple sample case:

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then we can compute the variance of Z:

$$Var(Z) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

We investigate the probability that Z goes outside the region $[\mu - \epsilon, \mu + \epsilon]$:

$$\mathbb{P}[|Z-\mu| > \epsilon] = \mathbb{P}[|Z-\mu|^2 > \epsilon^2] \leq \frac{\mathbb{P}[|Z-\mu|^2]}{\epsilon^2} = \frac{Var(Z)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Let the above probability equal to $\frac{1}{4} = \frac{\sigma^2}{n\epsilon^2}$, therefore, $n = \frac{4\sigma^2}{\epsilon^2}$. Which means that if we want this probability less than a value e.g. $\frac{1}{4}$, we must at least $n = \lceil\frac{4\sigma^2}{\epsilon^2}\rceil = O(\frac{\sigma^2}{\epsilon^2})$ for a single sample.

Suppose that we require such K samples in order to achieve overall $O(\log(\frac{1}{\delta})\frac{\sigma^2}{\epsilon^2})$ samples to compute ϵ-accurate estimate of the mean with prob at least $1-\delta$, where a simple sample is $X_1, ..., X_n$ as given. Suppose that $\hat{\mu} = \text{median}\{\mu_1, \mu_2, ..., \mu_K\}$, where μ_i is the mean for the sample s_i. Here we bound this median with deciding whether μ_i is in the range:

$$\begin{aligned}
\mathbb{P}[|\hat{\mu}-\mu| > \epsilon] &= \mathbb{P}[\sum_{i=1}^{K} \mathbb{I}(|\hat{\mu_i}-\mu| \leq \epsilon) \geq \frac{K}{2}] \quad \text{according to definition of median prob.} \\
&= \mathbb{P}[\sum_{i=1}^{K} \theta_i \geq \frac{K}{2}] \quad\quad \theta_i = \mathbb{I}(|\hat{\mu_i}-\mu| \leq \epsilon) \\
&= \mathbb{P}[\sum_{i=1}^{K} \theta_i - \mathbb{E}[\theta_i] \geq \frac{K}{2} - \sum_{i=1}^{K}\mathbb{E}[\theta_i]]
\end{aligned} \tag{0.16}$$

In first part of the solution, we simply let $\mathbb{E}[\theta_i] = \mathbb{E}[\mathbb{I}(|\hat{\mu_i}-\mu| \leq \epsilon)] = \mathbb{P}[|\hat{\mu_i}-\mu| \leq \epsilon) = \frac{1}{4}$, then

$$\mathbb{P}[|\hat{\mu}-\mu| = \mathbb{P}[\sum_{i=1}^{K} \theta_i - \frac{1}{4} \geq \frac{K}{2} - K * \frac{1}{4}] \tag{0.17}$$

Using Hoeffding inequality, we get:

$$\begin{aligned}
\mathbb{P}[\sum_{i=1}^{K} \theta_i - \frac{1}{4} \geq \frac{K}{4}] &\leq e^{-\frac{2\frac{K}{4}^2}{\sum_{i=1}^{K}(1-0)^2}} \\
&= e^{-\frac{K}{8}} = \delta \\
&\rightarrow K = 8\log\frac{1}{\delta}
\end{aligned} \tag{0.18}$$

This means that it requires at least $\lceil 8\log\frac{1}{\delta}\rceil = O(\log\frac{1}{\delta})$ samples.

Altogether we need $nK = \frac{4\sigma^2}{\epsilon^2} * 8\log\frac{1}{\delta} = 32\log\frac{1}{\delta}\frac{\sigma^2}{\epsilon^2} = O(\log\frac{1}{\delta}\frac{\sigma^2}{\epsilon^2})$ samples which suffices to ensure an ϵ-accurate estimate of the mean with prob. at least $1-\delta$.

Problem 7 Best-arm identification

Solution:

a) We have defined

$$\begin{cases} \epsilon = \bigcup_{k=1}^{K}\bigcup_{t=1}^{\infty}\{|\hat{\mu_{k,t}} - \mu_k| > U(t, \delta/K)\} \\ \mathbb{P}\left(\bigcup_{t=1}^{\infty}\{|\hat{\mu_{k,t}} - \mu_k| > U(t, \delta)\}\right) \leq \delta \end{cases}$$

Using the **union bound** and the above definition, we have:

$$\begin{aligned} \mathbb{P}(\epsilon) &= \mathbb{P}\left(\bigcup_{k=1}^{K}\bigcup_{t=1}^{\infty}\{|\hat{\mu_{k,t}} - \mu_k| > U(t, \delta/K)\}\right) && \text{definition} \\ &\leq \sum_{k=1}^{K}\mathbb{P}\left(\bigcup_{t=1}^{\infty}\{|\hat{\mu_{k,t}} - \mu_k| > U(t, \delta/K)\}\right) && \text{union bound} \\ &\leq \sum_{k=1}^{K}\frac{\delta}{K} = K * \frac{\delta}{K} \\ &= \delta \end{aligned} \tag{0.19}$$

Therefore, we prove that $\mathbb{P}(\epsilon) \leq \delta$

b) We know that if we want to drop i, there exists some k in S_{t-1} s.t.

$$\hat{\mu_{k,t}} - U(t, \delta/K) > \hat{\mu_{i,t}} + U(t, \delta/K)$$

We assume that ϵ holds such that arm i and k are contained in confidence interval $U(t, \delta/K)$, so that we have:

$$\begin{cases} \hat{\mu_{k,t}} - U(t, \delta/K) \leq \mu_k & \text{since } |\hat{\mu}_{k,t} - \mu_k| \leq U(t, \delta/K) \\ \hat{\mu_{i,t}} + U(t, \delta/K) \geq \mu_i & \text{since } |\hat{\mu}_{i,t} - \mu_i| \leq U(t, \delta/K) \end{cases}$$

Therefore, it is obvious that:

$$\mu_k \geq \hat{\mu_{k,t}} - U(t, \delta/K) > \hat{\mu_{i,t}} + U(t, \delta/K) \geq \mu_i$$

which means that given an i, we will always find a k in S_{i-1} s.t. $\mu_k > \mu_i$. The best arm is always in the set. Suppose that we want to drop the best arm k^*. There's no other possible arms k in S_t to make $\mu_k > \mu_{k^*}$ since k^* is defined as: $k^* = \max_k \mu_k$. Therefore the best arm k^* is contained in the set $S_t, \forall t \geq 1$.

c) Using the union bound, as well as the valid definition of Z_t for any-time confidence interval, we have:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{t=1}^{\infty} |\hat{\mu_{k,t}} - \mu_k| > U(t,\delta)\right) &\leq \sum_{i=1}^{\infty} \mathbb{P}\left(|\hat{\mu_{k,t}} - \mu_k| > U(t,\delta)\right) \\ &= \sum_{i=1}^{\infty} \mathbb{P}\left(|\frac{1}{t}\sum_{i=1}^{t} Z_t - \mathbb{E}[Z_t]| > U(t,\delta)\right)\end{aligned}$$

Since Z_t is bounded within [a, b], therefore we know that it is $\frac{b-a}{2}$-subgaussian. We also have defined $U(t,\delta)$ as:

$$U(t,\delta) = \sqrt{\frac{(b-a)^2 \log(4t^2/\delta)}{2t}}$$

. Therefore, we could then bound the prob. $\mathbb{P}\left(|\frac{1}{t}\sum_{i=1}^{t} Z_t - \mathbb{E}[Z_t]| > U(t,\delta)\right)$ using the Hoeffding bound:

$$\begin{aligned}\mathbb{P}\left(|\frac{1}{t}\sum_{i=1}^{t} Z_t - \mathbb{E}[Z_t]| > U(t,\delta)\right) &\leq 2e^{-\frac{tU(t,\delta)^2}{2\frac{b-a}{2}^2}} \\ &= 2e^{-\frac{t\frac{(b-a)^2\log(4t^2/\delta)}{2t}}{2\frac{b-a}{2}^2}} \\ &= 2e^{-\log(4t^2/\delta)} \\ &= 2 * \frac{\delta}{4t^2} = \frac{\delta}{2t^2}\end{aligned} \tag{0.20}$$

Therefore:

$$\begin{aligned}\mathbb{P}\left(\bigcup_{t=1}^{\infty} |\hat{\mu_{k,t}} - \mu_k| > U(t,\delta)\right) &\leq \sum_{i=1}^{\infty} \mathbb{P}\left(|\frac{1}{t}\sum_{i=1}^{t} Z_t - \mathbb{E}[Z_t]| > U(t,\delta)\right) \\ &\leq \sum_{t=1}^{\infty} \frac{\delta}{2t^2}\end{aligned} \tag{0.21}$$

We know that:

$$\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$$

so

$$\sum_{t=1}^{\infty} \frac{\delta}{2t^2} \leq \frac{\pi^2}{12}\delta \approx 0.82\delta \leq \delta$$

Therefore,

$$\mathbb{P}\left(\bigcup_{t=1}^{\infty} |\hat{\mu_{k,t}} - \mu_k| > U(t,\delta)\right) \leq \sum_{i=1}^{\infty} \mathbb{P}\left(|\frac{1}{t}\sum_{i=1}^{t} Z_t - \mathbb{E}[Z_t]| > U(t,\delta)\right) \leq \delta$$

Proof ends.

d) Removing i from S_{t-1} requires

$$\hat{\mu}_t^* - U(t, \delta/K) \geq \hat{\mu}_{k,t} + U(t, \delta/K)$$

Assume that the event ϵ holds. Therefore:

$$\begin{cases} \hat{\mu^*} - U(t, \delta/K) \leq \mu_t^* & \text{since } |\hat{\mu_t^*} - \mu^*| \leq U(t, \delta/K) \\ \hat{\mu_{k,t}} \leq U(t, \delta/K) + \mu_k & \text{since } |\hat{\mu}_{k,t} - \mu_k| \leq U(t, \delta/K) \end{cases}$$

Therfore: if we guarantees

$$\begin{aligned} \hat{\mu^*} - 2U(t, \delta/K) &\geq \mu_k + 2U(t, \delta/K) \\ \rightarrow \hat{\mu^*} - \mu_k &\geq 4U(t, \delta/K) \\ \rightarrow \Delta_k &\geq 4U(t, \delta/K) \quad \text{Define } \Delta_k := \mu^* - \mu_k \end{aligned}$$

Then the event ϵ always holds. In this question, we want to prove that after $\sum_{k \neq k^*} \lceil c\Delta_k^{-2} \log\left(K\Delta_k^{-1}\right) \rceil$ samples, the Successive Elimination algorithm terminates. We could let $T_k = c\Delta_k^{-2} \log\left(K\Delta_k^{-1}\right)$ s.t. $\Delta_k \geq 4U(T_k, \delta/K)$. We need to find that such c exists for T_k:

$$\begin{aligned} \Delta_k &\geq 4U(T_k, \delta/K) \\ &= 4\sqrt{\frac{(b-a)^2 \log(4KT_k^2/\delta)}{2T_k}} \\ &= 4\sqrt{\frac{(b-a)^2 \log(4K\left(c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\right)\right)^2/\delta)}{2c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\right)}} \\ \rightarrow \Delta_k^2 &\geq \frac{8\log(4K\left(c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\right)\right)^2/\delta)}{c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\right)}, \text{a} = 0, \text{ b} = 1 \\ \rightarrow \log(K\Delta_k^{-1}) &\geq \frac{8}{c}\log(4K\left(c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\right)\right)^2/\delta) \end{aligned} \tag{0.22}$$

We focus on RHS of this inequality. We know that $\frac{a}{c}\log bc$ ranges from $(-\infty, m]$ where m is some constant. Therefore we could always find such c that the above inequality holds.

If an arm k requires $T_k = c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\right)$ to be removed from the set S_{t-1}, the overall samples which is non-optimal need to be removed from S_{t-1} within:

$$\sum_{k \neq k^*}^{K} \lceil T_k \rceil = O(\sum_{k \neq k^*}^{K} \lceil c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\rceil\right)) = O(\sum_{k \neq k^*}^{K} c\Delta_k^{-2}\log\left(K\Delta_k^{-1}\right))$$

with $\Delta_k := \mu^* - \mu_k$. Proof ends.