HW03 — Solution.

1) Let $X \sim U([a,b])$.

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a}\,dx = \left.\frac{x^2}{2(b-a)}\right|_{x=a}^{b} = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}.$$

$$\mathbb{E}[X^2] = \int_a^b \frac{x^2}{b-a}\,dx = \left.\frac{x^3}{3(b-a)}\right|_{x=a}^{b} = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}.$$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{b^2+ab+a^2}{3} - \frac{(b+a)^2}{4} = \frac{\cancel{ab}\,(b-a)^2}{\cancel{12}\ 12}$$

2) Let $X_1, X_2, \dots, X_k$ be independent uniform r.v.'s over $[0,1]$.

$Y = \min(X_1, \dots, X_k)$.

In class, we know that the probability density function of y is $f(y) = k(1-y)^{k-1}$.

$$\mathbb{E}[Y^2] = \int_0^1 y^2 k(1-y)^{k-1}\,dy.$$

$u = y^2$ $dv = k(1-y)^{k-1}dy$

$du = 2y\,dy$ $v = -(1-y)^k$

integration by parts

$$= \left. -y^2(1-y)^k \right|_{y=0}^{1} + \int_0^1 2y(1-y)^k\,dy.$$

$$= 2\int_0^1 y(1-y)^k\,dy.$$

$u = y$ $dv = (1-y)^k dy$

$du = dy$ $v = -\frac{(1-y)^{k+1}}{k+1}$

$$= 2\left[-\left.\frac{y(1-y)^{k+1}}{k+1}\right|_{y=0}^{1} + \int_0^1 \frac{(1-y)^{k+1}}{k+1}\,dy \right]$$

$$= \frac{2}{k+1}\int_0^1 (1-y)^{k+1}\,dy$$

$$= \frac{2}{k+1}\cdot\left[-\frac{(1-y)^{k+2}}{k+2}\right]_{y=0}^{1} = \frac{2}{(k+1)(k+2)} \leq \frac{2}{(k+1)^2}.$$

3) $\mathbb{E}[X] = 0$ and $Var[X] = 1$.

a) $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{Var[X]}{\varepsilon^2} = \frac{1}{\varepsilon^2}$.

This bound is not helpful when $\varepsilon \in (0,1)$ since $\frac{1}{\varepsilon^2} > 1$.

b) Let $X_1, \dots, X_n$ be independent copies of X.

Let $Z = \frac{1}{n} \sum_{i=1}^{n} X_i$. $\Rightarrow$ $\mathbb{E}[Z] = \mathbb{E}[X] = 0$

and $Var[Z] = \frac{1}{n^2} \sum_{i=1}^{n} Var[X_i] = \frac{1}{n} Var[X] = \frac{1}{n}$.

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \varepsilon) \leq \frac{Var[Z]}{\varepsilon^2} = \frac{1}{n\varepsilon^2},$$

4) Let f_i be the frequency of the i symbol in the data stream.

$$X_i = \begin{cases} 1 & \text{with probability } 1/2. \\ -1 & \text{with probability } 1/2. \end{cases}$$

Let $V = \left(\sum_{i=1}^{m} X_i f_i \right)^2 = \cancel{\sum X_i f_i} \left(\sum_{i=1}^{m} Y_i \right)^2$, where $Y_i = X_i f_i$.

Then $V^2 = \left(\sum_{i=1}^{m} Y_i \right)^4 = \sum_{i,j,l,k} Y_i Y_j Y_l Y_k$.

$$\Rightarrow \mathbb{E}[V^2] = \sum_{i,j,l,k} \mathbb{E}[Y_i Y_j Y_l Y_k] \qquad (*)$$

Let's consider possibilities of $\mathbb{E}[Y_i Y_j Y_l Y_k]$:

i) When all indices i, j, l, k are the sames:

$$\mathbb{E}[Y_i Y_j Y_l Y_k] = \mathbb{E}[Y_i^4] = \mathbb{E}[X_i^4 f_i^4] = f_i^4.$$

in the sum of $(*)$, there are m such terms.

5) a) $X_j^i = \begin{cases} 1 & \text{if job } j \text{ is assigned to server } i, \text{ wp } 1/k \\ 0 & \text{otherwise, wp } 1 - \frac{1}{k}. \end{cases}$

$X^i = \sum_{j=1}^{n} X_j^i$ the load on machine i.

Then,

a) $\mathbb{E}[X^i] = \sum_{j=1}^{n} \mathbb{E}[X_j^i] = \frac{n}{k}$.

b) $\mathbb{P}\left(X^i > \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right)$?

Chernoff's bound.

$$= \mathbb{P}\Big(\sum_{j=1}^{n} X_j^i > \underbrace{\frac{n}{k}}_{\mu}\Big(1 + \underbrace{3\frac{\sqrt{\ln k}}{\sqrt{\frac{n}{k}}}}_{\varepsilon}\Big)\Big) \leq e^{-\frac{n}{k}\cdot\left(\frac{3\sqrt{\ln k}}{\sqrt{n}}\cdot\sqrt{k}\right)^2/4}$$

$$= e^{-9\ln k/4}$$

$$= \frac{1}{k^{9/4}}.$$

$\therefore \quad \mathbb{P}\left(X^i > \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right) \leq \frac{1}{k^{9/4}}$.

b) Let the (bad) event $\left\{X^i > \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right\}$ be Bad_i

Then $\mathbb{P}\left(M \geq \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right)$

$$= \mathbb{P}\left(\left\{X_1 \geq \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right\} \cup \dots \cup \left\{X_k \geq \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}}\right\}\right)$$

$$= \mathbb{P}(Bad_1 \cup \dots \cup Bad_k)$$

$$\underset{\text{union bound}}{\leq} \sum_{i=1}^{k} \mathbb{P}(Bad_i) \leq \sum_{i=1}^{k} \frac{1}{k^{9/4}} = \frac{k}{k^{9/4}} = \frac{1}{k^{5/4}}.$$

ii) ~~2)~~ when three indices are the same, and one is different.

$$\cancel{\mathbb{E}[X} \quad \mathbb{E}[Y_i^3 Y_k] = \mathbb{E}[Y_i^3]\,\mathbb{E}[Y_k] \quad \text{since } Y_i \text{ and } Y_k \text{ are independent.}$$

$$= 0 \quad \text{since } \mathbb{E}[Y_k] = \mathbb{E}[X_k f_k] = \mathbb{E}[X_k] f_k = 0 \cdot f_k$$

iii) ~~3)~~ When two indices are the same, and the other two are different.

$$\mathbb{E}[Y_i^2 Y_l Y_k] = 0.$$

iv) ~~4)~~ when $i=j$ and $l=k$, but $i \neq l$:

$$\mathbb{E}[Y_i^2 Y_l^2] = \mathbb{E}[X_i^2 f_i^2 X_l^2 f_l^2] = f_i^2 f_l^2.$$

in (i, j, l, k), there are $\binom{4}{2} = 6$ ~~ways~~ choices.

v) ~~5)~~ when all indices are different.

$$\rightarrow \quad \mathbb{E}[Y_i Y_j Y_l Y_k] = 0.$$

$$\therefore \quad \mathbb{E}[V^2] = \sum_{i=1}^{m} \mathbb{E}[Y_i^4] + 6 \sum_{i \neq j} \mathbb{E}[Y_i^2 Y_j^2]$$

$$= \sum_{i=1}^{m} f_i^4 + 6 \sum_{i \neq j} f_i^2 f_j^2$$

$$\leq 3 \sum_{i=1}^{m} f_i^4 + 6 \sum_{i \neq j} f_i^2 f_j^2.$$

$$= 3 \left(\sum_{i=1}^{m} f_i^4 + 2 \sum_{i \neq j} f_i^2 f_j^2 \right).$$

$$= 3 \left(\sum_{i=1}^{m} f_i^2 \right)^2.$$

$$= 3\, \mathbb{E}[V]^2.$$

$\therefore \quad \cancel{\mathbb{P}(M \geq n}$

$$\therefore \quad \mathbb{P}\left(M \leq \frac{n}{k} + 3\sqrt{\ln k}\sqrt{\frac{n}{k}} \right) \geq 1 - \frac{1}{k^{5/4}} .$$

6) Let $X_1, \dots, X_n$ be independent copies of X.

Let $Y = \dfrac{X_1 + \dots + X_n}{n}$. Then

$$\mathbb{E}[Y] = \mathbb{E}[X].$$

We can use the following variants of Chernoff's bound (called Chernoff-Hoeffding bounds).

Thm: Let $X_1, X_2, \dots, X_n$ be independent random variables such that $X_i \in [0,1]$ and $\mathbb{E}[X_i] = p$. Let $X = \sum_{i=1}^{n} X_i$, then $\mathbb{E}[X] = pn$ and for $\varepsilon > 0$,

$$\left.\begin{aligned} &\mathbb{P}(X - pn > \varepsilon n) < e^{-2\varepsilon^2 n} \\ \text{and } &\mathbb{P}(X - pn < -\varepsilon n) < e^{-2\varepsilon^2 n}. \end{aligned}\right\} \Rightarrow \mathbb{P}(|X - pn| > \varepsilon n) < 2e^{-2\varepsilon^2 n}.$$

Applying the theorem,

$$\begin{aligned} \mathbb{P}(|Y - \mathbb{E}[Y]| > \varepsilon) &= \mathbb{P}\left(\left| \frac{\sum_i X_i}{n} - \frac{\sum \mathbb{E}[X_i]}{n} \right| > \varepsilon \right). \\ &= \mathbb{P}\left(\left| \sum_i X_i - \sum \mathbb{E}[X_i] \right| > \varepsilon \cancel{p}\, n \right). \\ &= \mathbb{P}\left(\left| \sum X_i - n\mathbb{E}[X] \right| > \varepsilon n \right) \\ &\leq 2e^{-2\varepsilon^2 n} = \delta. \end{aligned}$$

$\Rightarrow$ take $n = \dfrac{1}{2\varepsilon^2} \ln\left(\dfrac{2}{\delta}\right)$ samples $X_1, \dots, X_n$ and use Y to estimate $\mathbb{E}[X]$