Lecture 2 :- Concentration and (saving) memory.

Variance reduction by taking average of copies :

Let $X_1, X_2, \ldots X_s$ be (pairwise) independent random variables. Define

$$X = \frac{1}{s}\sum_{i=1}^{s} X_i .$$

Then $E[X] = \frac{1}{s}\sum_{i=1}^{s} E[X_i]$ (Linearity of expectation)

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

Pf :-

$$E[X^2] = \frac{1}{s^2} E\left[\left(\sum_{i=1}^{s} X_i\right)^2\right]$$

$$= \frac{1}{s^2}\sum_{i=1}^{s} E[X_i^2] + \frac{2}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{i-1} E[X_i X_j]$$

$$= \cdots\cdots + \frac{2}{s^2} \cdots\cdots E[X_i]E[X_j]$$

(Pairwise independence)

$$E[X]^2 = \frac{1}{s^2}\left(\sum_{i=1}^{s} E[X_i]\right)^2$$

$$= \frac{1}{s^2}\sum_{i=1}^{s} E[X_i]^2 + \frac{2}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{i-1} E[X_i]E[X_j]$$

Subtract

$$Var[X] = \frac{1}{s^2}\sum_{i=1}^{s}\left(E[X_i^2] - E[X_i]^2\right)$$

$$= \frac{1}{s^2}\sum_{i=1}^{s} Var[X_i] .$$

In particular if $E[X_i] = \mu$ for every i
$\mathrm{Var}[X_i] = \sigma^2$ for every i, then.

$$E[X] = \mu, \qquad \mathrm{Var}[X] = \frac{\sigma^2}{s}.$$

So if we have
s (pairwise) independent saples $Y_1, \ldots, Y_s$ of Y
then $G := \frac{1}{s}\sum_{i=1}^{s} Y_i$ satisfies

$$E[G] = F_2$$

$$\mathrm{Var}[G] \leq \frac{2F_2^2}{s}.$$

Therefore if $s \geq \left\lceil \frac{16}{\varepsilon^2} \right\rceil$

$$\Pr[\,|G - F_2| > \varepsilon F_2\,] \leq \frac{\mathrm{Var}[G]}{\varepsilon^2 F_2^2}$$

$$\leq \frac{2F_2^2}{s\cdot\varepsilon^2 F_2^2} = \frac{2}{s\varepsilon^2}$$

$$\leq \frac{1}{8},$$

as $s \geq \frac{16}{\varepsilon^2}$.

Now suppose we want our probability of error to be at most $\delta > 0$. Then the above argument would lead to the requirement:

$$s \geq \left\lceil \frac{2}{\delta\varepsilon^2} \right\rceil.$$

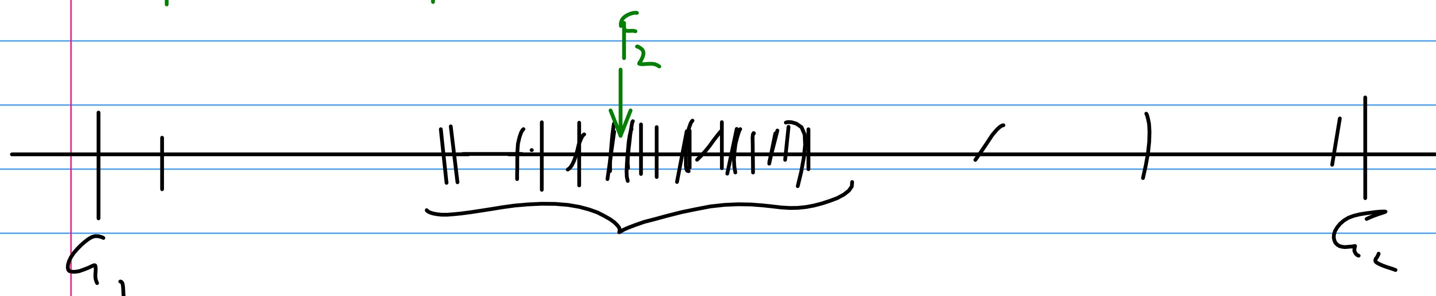
[Q]: Can we do something better than this $1/\delta$ dependence?

Suppose we take k independent samples of G

$G_1 \quad \ldots\ldots \quad G_k$

- Call G_i good if $|G_i - F_2| < \varepsilon F_2$.
  - By looking at G_i we cannot tell whether it's good [unless we know F_2].
  - But we know is that, independently, each G_i is good with probability $\geq \frac{7}{8}$.

We expect the picture to look at this.



We coul try to use

$$H = \frac{1}{k} \sum_{i=1}^{s} G_i .$$

Currently, we haven't quantified how 'bad' a 'bad' G_i can be.

- A 'bad' G_i that is really far from F_2 can 'destroy' the mean.
- Typically to show that the mean concentrates (to be defined later) we will need some control over how 'bad' the 'bad' samples can be.

Therefore we consider the median of $G_1 \ldots G_k$;

which at least superficially does not seem to get affected much by outlier G_i's.

$M := \text{Median}(G_1 \ldots G_k) = G_{\frac{k+1}{2}}$ when G_i are sorted in ascending order (k assumed odd).

What is the event $M > (1+\varepsilon)F_2$?

$$\{M > (1+\varepsilon)F_2\} \Leftrightarrow \{G_{\frac{k+1}{2}} > (1+\varepsilon)F_2\}$$

$$\Leftrightarrow \{G_{\frac{k+1}{2}}, \ldots, G_k > (1+\varepsilon)F_2\}.$$

$\Leftrightarrow$ { $\frac{k+1}{2}$ of the k G_i's are greater than $(1+\varepsilon)F_2$ }.

What is the event $M < (1-\varepsilon)F_2$?

$$\{M < (1-\varepsilon)F_2\} \Leftrightarrow \{G_{\frac{k+1}{2}} < (1-\varepsilon)F_2\}$$

$$\Leftrightarrow \{G_1, \ldots, G_{\frac{k+1}{2}} < (1-\varepsilon)F_2\}.$$

$\Leftrightarrow$ { $\frac{k+1}{2}$ of the k G_i's are less than $(1-\varepsilon)F_2$ }.

<u>Upshot of this</u>

$\{|M - F_2| > \varepsilon F_2\} \Rightarrow$ { at least $\frac{k+1}{2}$ of the G_i are bad }.

Let's define $X_i = I[G_i \text{ is bad}]$, then,

X_i are independent !! $E[X_i] \leq \frac{1}{8}$. $X_i \in \{0, 1\}$

then, we can write the above as.

$$\{|M - F_2| > \varepsilon F_2\} \Rightarrow \left\{ \sum_{i=1}^{k} X_i \geq \frac{k+1}{2} \right\}.$$

$$\Pr\left[|M - F_2| > \varepsilon F_2 \right] \leq \underbrace{\Pr\left[\sum_{i=1}^{k} X_i \geq \frac{k+1}{2} \right]}_{\text{This is what we want to bound.}}$$

$(X_i)_{i=1}^{n}$ are Bernoulli random variable.
$[X_i \in \{0,1\}]$

$$E[X_i] = p \ , \qquad Var(X_i) = p(1-p).$$

X_i are independent. We want to bound

$$\Pr\left[\sum_{i=1}^{n} X_i > (p+\varepsilon)n \right] \begin{cases} \leq \dfrac{pn}{(p+\varepsilon)n} = \dfrac{p}{p+\varepsilon} & \text{(Markov)} \\[2ex] \leq \dfrac{p(1-p)n}{\varepsilon^2 n^2} \leq \dfrac{p}{\varepsilon^2 n} & \text{(Chebyshev)} \end{cases}$$

At least Chebyshev inequality gives some improvement using the number of samples and is also able to exploit independence (at least pairwise indepence).

↳ This was because we were using the 'test statistic' $\left(\sum_{i=1}^{n} X_i\right)^2$, whose mean 'behaves differently' when X_i are pairwise independent.

→ This suggest using 'higher moments', i.e. using $\left(\sum_{i=1}^{n} X_i\right)^m$ for large m can give even better results.

(In fact, it does!).

- One can try to find a best m (this is specially useful when one only has limited independence).

But for full-independence, it is easier to just use all the moments at once. For example look at

$$\exp(tS) \qquad \left[S := \sum_{i=1}^{n} X_i\right]$$

$$\exp(tS) = \sum_{i=0}^{\infty} \frac{(tS)^i}{i!}$$

A reason for preferring this is that when X_i are independent,

$$E[\exp(tS)] = E\left[\prod_{i=1}^{n} \exp(tX_i)\right]$$

$$\overset{\text{indep.}}{=\!=} \prod_{i=1}^{n} E[\exp(tX_i)]$$

$f_X(t) := E[\exp(tX)]$ is called the moment generating function. [the accurate 'generating function' theory name for this would be 'exponential generating function of the moments'] of X, and also the Laplace transform of X.

But what about concentration?

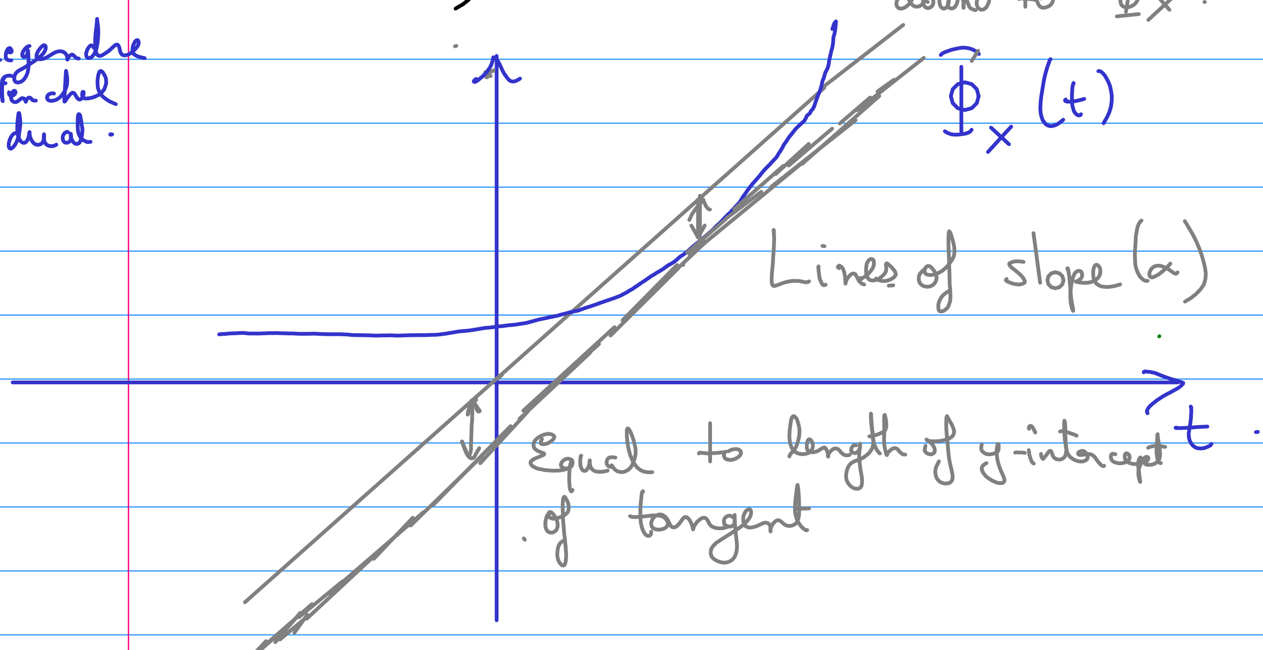
$$\phi_X(t) := \log E[\exp(tX)].$$ [log moment gen. fn. - cumulant generation fn.].

$$\Pr[X > \alpha]$$
$$= \Pr[tX > t\alpha] \quad \forall t > 0.$$
$$= \Pr[\exp(tX) > \exp(t\alpha)] \quad \forall t > 0$$
$$\leq \frac{E[\exp(tX)]}{\exp(t\alpha)} \quad \forall t > 0$$
$$= \exp(\phi_X(t) - t\alpha) \quad \forall t > 0.$$

$$\Pr[X > \alpha] \leq \exp\left(-\sup_{t>0}\left(t\alpha - \Phi_X(t)\right)\right)$$

$$\Phi^*(\alpha) := \sup_t \left(t\alpha - \Phi_X(t)\right)$$ } Maximum vertical distance from the line of slope α, down to Φ_X. (Cramer, Chernoff, Hoeffding, Azuma)

Legendre - Fenchel dual.



$$\Pr[X > \alpha] \leqslant \exp\left(-\sup_{t>0}\left(\alpha t - \Phi_X(t)\right)\right)$$

Subgaussian random variables :- X is said to be subgaussian with parameter $\nu > 0$ if

$$\Phi_X(t) \leqslant \frac{\nu t^2}{2}.$$

For such variables, (which include Bernoulli/Gaussian bdd rvs etc.)

$$\sup_{t>0} \alpha t - \Phi_X(t) \qquad (\alpha > 0).$$

$$\geqslant \sup_{t>0} \alpha t - \frac{\nu t^2}{2}$$

$$= \frac{\alpha^2}{2\nu}.$$

If X is Subgaussian with parameter ν then

$$\Pr[X > \alpha] \leqslant \exp\left(-\frac{\alpha^2}{2\nu}\right). \qquad (\alpha > 0)$$

A random variable which takes values only in the interval, $[a, b]$ is Sub-Gaussian with parameter $\frac{(b-a)^2}{4}$. [Hoeffding's bound]

So if X is Bernoulli(p)

$X - p$ is sub-Gaussian with parameter $\frac{1}{2}$, and has mean 0.

By this method only.

$$\sum_{i=1}^{n} (X_i - p) \qquad (X_i \text{ iid Bernoulli } p)$$

would be subGaussian with parameter $\frac{n^2}{4}$.

- But this fails to make use of independence!!

Crucial point: If $X_1, \ldots, X_n$ are sub-Gaussian with parameters $\nu_1, \ldots, \nu_n$ and are independent then $\sum_{i=1}^{n} X_i$ is sub-Gaussian with parameter $\sum_{i=1}^{n} \nu_i$. (Homework). $\left[\sum_{i=1}^{n}(X_i - p)\right.$ is sub-gaussian w/ param. $\frac{n}{4}$.

Thing to try out :- What do we get if we just apply Hoeffding's bound to Y or G directly ??

$$\Pr\left[\sum_{i=1}^{n} X_i > (p+\varepsilon)n\right]$$

$$= \Pr\left[\sum_{i=1}^{n}(X_i - p) > \varepsilon n\right]$$

$$\leq \exp\left(-\frac{\varepsilon^2 n^2}{2\cdot(n/4)}\right) = \exp\left(-2n\varepsilon^2\right).$$