1 Hoeffding’s Bound

We say $X$ is a sub-Gaussian random variable if it has quadratically bounded logarithmic moment generating function, e.g.

$$\ln E e^{\lambda (X - \mu)} \leq \frac{\lambda^2}{2} b.$$

For a sub-Gaussian random variable, we have

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-n\epsilon^2/2b}.$$

Similarly,

$$P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-n\epsilon^2/2b}.$$

2 Chernoff Bound

For a binary random variable, recall the Kullback-Leibler divergence is

$$KL(p||q) = p \ln(p/q) + (1 - p) \ln((1 - p)/(1 - q)).$$

**Theorem 2.1.** (Relative Entropy Chernoff Bound) Assume that $X \in [0, 1]$ and $EX = \mu$. We have the following inequality

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-nKL(\mu + \epsilon||\mu)}$$

and

$$P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-nKL(\mu - \epsilon||\mu)},$$

First, let us understand the worst case MGF for $X$.

**Lemma 2.2.** Assume that $X \in [0, 1]$ and $EX = \mu$. We have the following inequality

$$E e^{\lambda X} \leq (1 - \mu) e^0 + \mu e^\lambda.$$  

This shows that the maximum logarithmic moment generating function is achieved with a $\{0, 1\}$ valued random variable, i.e.

$$E e^{\lambda X} \leq EX' \sim \mu [e^{\lambda X'}]$$

where $X'$ is a $\{0, 1\}$ valued random variable which takes the value 1 with probability $\mu$.

**Proof.** Let $M_X(\lambda) = E e^{\lambda X}$ and $M'_{X'}(\lambda) = (1 - \mu) e^0 + \mu e^\lambda$. Then $M_X(0) = M_{X'}(0)$. Moreover,  

$$M'_X(\lambda) = EX e^{\lambda X} \leq EX e^{\lambda X} = \mu e^\lambda = M'_{X'}(\lambda)$$

which completes the proof.
Now we are ready to provide the proof.

**Proof.** By the previous lemma, we only need to prove the result for binary $X \in \{0, 1\}$, with mean 1. Recall from Lemma 1.4 in the previous lecture that,

$$I(\mu + \epsilon) = KL(P_{\mu + \epsilon} || P)$$

where $P_{\mu + \epsilon}$ was the “variational” distribution $P_{\lambda}$ where $\lambda$ was is set such that $E_{X \sim P_{\lambda}}[X] = \mu + \epsilon$.

Since $X$ is binary, it must be that $P_{\mu + \epsilon}$ is just distribution which is 1 with probability $\mu + \epsilon$. Hence $KL(P_{\mu + \epsilon} || P)$ is just the KL between two binary distributions with means $\mu + \epsilon$ and $\mu$, which completes the proof. \qed

### 2.1 Useful Forms of the Chernoff Bound

Note that by Hoeffding’s lemma (as $X$ is sub-Gaussian), we have (from Lecture 5) that

$$-KL(\mu + \epsilon || \mu) = \inf_{\lambda > 0} \left[ -\lambda (\mu + \epsilon) + \ln((1 - \mu)e^\lambda + \mu e^\lambda) \right] \leq 2\epsilon^2$$

Define $Var_p$ be the variance of a $X$ which is 1 with probability $p$ and 0 with probability $1 - p$. It is straightforward to show that the second derivative with respect to $\delta$ is:

$$KL''(\mu + \delta || \mu) = 1/Var_\delta$$

Define

$$\text{MaxVar}[\mu, \mu + \epsilon] = \max_{p \in [\mu, \mu + \epsilon]} Var_p$$

which provides a lower bound on the second derivative for $\delta$ between 0 and $\epsilon$.

Hence, we have that:

$$KL(\mu + \epsilon || \mu) \geq \frac{1}{2} \epsilon^2 / \text{MaxVar}[\mu, \mu + \epsilon]$$

which leads to a nicer version of the Chernoff bound.

**Theorem 2.3. (Nicer Form of the Chernoff Bound)** Assume that $X \in [0, 1]$ and $EX = \mu$. Fix $\epsilon$. Define:

$$\text{MaxVar}[\mu, \mu + \epsilon] = \max_{p \in [\mu, \mu + \epsilon]} Var_p$$

as before (i.e. it is the maximal variance (of $\{0, 1\}$ variable) between $\mu$ and $\mu + \epsilon$).

We have the following inequality

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-n \frac{\epsilon^2}{2 \text{MaxVar}[\mu, \mu + \epsilon]}}$$

and

$$P(\bar{X}_n \geq \mu - \epsilon) \leq e^{-n \frac{\epsilon^2}{2 \text{MaxVar}[\mu - \epsilon, \mu]}}$$

The following corollary (while always true) is much sharper bound than Hoeffding’s bound when $\mu \approx 0$.

**Corollary 2.4.** We have the following bound:

$$P(\bar{X}_n \geq \mu + \epsilon) \leq \exp[-n\epsilon^2/2(\mu + \epsilon)]$$

and thus

$$P(\bar{X}_n \leq \mu - \epsilon) \leq \exp[-n\epsilon^2/2\mu].$$
This implies a multiplicative form of the Chernoff bound since:

\[ P(\bar{X}_n \geq (1 + \delta)\mu) \leq \exp[-n\mu \frac{\delta^2}{2(1 + \delta)}] \]

and

\[ P(\bar{X}_n \leq (1 - \delta)\mu) \leq \exp[-n\mu \delta^2/2] \]

Similar results for Bernstein and Benett inequalities are available.

3 Benett Inequality

In Benett inequality, we assume that the variable is upper bounded, and want to estimate its moment generating function using variance information.

Lemma 3.1. If \( X - EX \leq 1 \), then \( \forall \lambda \geq 0 \):

\[ \ln E e^{\lambda(X-\mu)} \leq (e^\lambda - \lambda - 1) Var(X). \]

where \( \mu = EX \)

Proof. It suffices to prove the lemma when \( \mu = 0 \). Using \( \ln z \leq z - 1 \), we have

\[
\ln E e^{\lambda X} = \ln E e^{\lambda X} \\
\leq E e^{\lambda X} - 1 \\
= \lambda^2 E \frac{e^{\lambda X} - \lambda X - 1}{(\lambda X)^2} (X)^2 \\
\leq \lambda^2 E \frac{e^{\lambda X} - \lambda - 1}{\lambda^2} (X)^2,
\]

where the second inequality follows from the fact that the function \( (e^z - z - 1)/z^2 \) is non-decreasing and \( \lambda X \leq \lambda \). \( \square \)

Lemma 3.2. We have

\[
\inf_{\lambda > 0} [-\lambda \epsilon + (e^\lambda - \lambda - 1)Var(X)] = -Var(X)\phi(\epsilon/Var(X)) \leq \frac{\epsilon^2}{2(Var(X) + \epsilon/3)}.
\]

where \( \phi(z) = (1 + z)\ln(1 + z) - z \).

Proof. Take derivative with respect to \( \lambda \), we obtain

\[ -\epsilon + (e^\lambda - 1)Var(X) = 0. \]

Therefore \( \lambda = \ln(1 + \epsilon/Var(X)) \). Plug in, we obtain the equality.

It is easy to verify using Taylor expansion of the exponential function that for \( \lambda \in (0, 3) \):

\[ e^\lambda - \lambda - 1 \leq \frac{\lambda^2}{2} \sum_{m=0}^{\infty} (\lambda/3)^m = \frac{\lambda^2}{2(1 - \lambda/3)}. \]

Now by picking \( \lambda = \epsilon/(Var(X) + \epsilon/3) \), we have

\[ -\lambda \epsilon + \frac{\lambda^2}{2(1 - \lambda/3)} = -\epsilon^2/[2Var(X) + 2\epsilon/3]. \]
This proves the desired bound. □

The above bound implies the following bound: If $X - EX \leq b$, for some $b > 0$, then

$$P[X \geq EX + \epsilon] \leq \exp[-n\epsilon^2/(2\text{Var}(X) + 2eb/3)].$$

This is similar to the Gaussian result, except for the term $2eb/3$. Behaves similar to Gaussian tail bound when $eb \ll \text{Var}(X)$.

### 4 Bernstein Inequality

In Bernstein inequality, we obtain a result similar to the simplified Bennett bound but with a moment condition. There are different forms. We consider one form.

**Lemma 4.1.** If $X$ satisfies the moment condition with $b > 0$ for integers $m \geq 2$:

$$EX^m \leq m!b^{m-2}\text{Var}/2,$$

then when $\lambda \in (0, 1/b)$:

$$\ln Ee^{\lambda X} \leq \lambda EX + 0.5\lambda^2\text{Var}(1 - \lambda b)^{-1},$$

and thus

$$P[\bar{X}_n \geq EX + \epsilon] \leq \exp[-n\epsilon^2/(2\text{Var} + 2eb)].$$

**Proof.** We have the following estimation of logarithmic moment generating function:

$$\ln Ee^{\lambda X} \leq Ee^{\lambda X} - 1 \leq \lambda EX + 0.5V\lambda^2 \sum_{m=2}^{\infty} \frac{b^{m-2}\lambda^{m-2}}{m-2} = \lambda EX + 0.5\lambda^2\text{Var}(1 - \lambda b)^{-1}.$$ 

The last inequality is similar to the proof of Bennett inequality. Exercise: finish the proof. □

### 5 Independent but non-iid random variables

If $X_1, \ldots, X_n$ are independent but not iid. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\mu = E\bar{X}_n$, then we have

$$P(\bar{X}_n \geq \mu + \epsilon) \leq \inf_{\lambda > 0} \left[-\lambda n(\mu + \epsilon) + \sum_{i=1}^{n} \ln Ee^{\lambda X_i}\right].$$

In particular, we have the following results:

**Lemma 5.1.** If $X_i$ are sub-Gaussians with $Ee^{\lambda X_i} \leq \lambda EX_i + 0.5\lambda^2V_i$, then

$$P(\bar{X}_n \geq \mu + \epsilon) \leq \exp\left[-\frac{n^2\epsilon^2}{2\sum_{i=1}^{n} V_i}\right].$$

An example is Radamacher average: let $\sigma_i = \{\pm 1\}$ be independent random Bernoulli variables, and $a_i$ be fixed numbers, then

$$P(\frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \geq \epsilon) \leq \exp\left[-\frac{n\epsilon^2}{2(n-1)\sum_{i=1}^{n} a_i^2}\right].$$

Similarly one can derive bounds for Bennett and Bernstein inequalities.
Lemma 5.2. If $X_i - EX_i \leq b$ for all $i$, then
\[ P(\bar{X}_n \geq \mu + \varepsilon) \leq \exp \left[ -\frac{n^2 \varepsilon^2}{2 \sum_{i=1}^{n} Var(X_i) + 2nb/3} \right]. \]

6 Alternative Expression

Tail inequality: $P(\text{deviation} \geq \varepsilon) \leq \delta(\varepsilon)$. Equivalent expression: with probability $1 - \delta$: deviation \leq \varepsilon(\delta)$, where $\varepsilon(\delta)$ is the inverse function of $\delta(\varepsilon)$.

For example the Chernoff bound
\[ P(\bar{X}_n - \mu \geq \varepsilon) \leq \exp(-2n\varepsilon^2) = \delta, \]
means with probability $1 - \delta$: $\bar{X}_n - EX \leq \sqrt{\ln(1/\delta)/(2n)}$.

For Bennet inequality,
\[ P[\bar{X}_n \geq EX + \varepsilon] \leq \exp[-n\varepsilon^2/(2Var(X) + 2eb/3)], \]
we set
\[ \delta = \exp[-n\varepsilon^2/(2Var(X) + 2eb/3)], \]
and thus using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$:
\[ \varepsilon = \sqrt{2Var(X) \ln(1/\delta)/n + b^2 \ln(1/\delta)^2/(9n^3)} + \frac{b \ln(1/\delta)}{3n} \leq \sqrt{2Var(X) \ln(1/\delta)/n + \frac{2b \ln(1/\delta)}{3n}} \]
That is, with probability at least $1 - \delta$, we have
\[ \bar{X}_n - EX \leq \sqrt{2Var(X) \ln(1/\delta)/n + \frac{2b \ln(1/\delta)}{3n}}. \]