EE 325: Probability and Random Processes
Module 4: Bounds and Inequalities

D Manjunath

Networking Lab
Department of Electrical Engineering
Indian Institute of Technology

October 13, 2020

Topics in Module 4

- Recap of Union Bound and Bonferroni Inequalities
- Markov Inequality
- Chebyshev Inequality
- Chernoff Bound
- Cauchy-Schwarz Inequality
- Jensen's Inequality
- Hoeffding Bound

Recap Inequalities

- Often exact probabilities cannot be calculated and bounds may be more easily accessible. In fact, in many cases they suffice.
- Recall the **Union Bound:** If A_i, $i = 1, \ldots, K$ are a set of K events, then

$$\mathsf{Pr}\left(\cup_{i=1}^{K} A_i\right) \leq \sum_{i=1}^{K} \mathsf{Pr}(A_i)$$

- This is very useful when we know the individual probabilities but the union is hard to determine. Many a time, we may be just as happy with the upper bound from the Union Bound.
- From the Union Bound we can also write the following bound

$$\mathsf{Pr}\left(\cap_{i=1}^{K} A_i\right) \; \geq \; \sum_{i=1}^{K} (1 - \mathsf{Pr}(A_i))$$

- With a little more work, we can also obtain the lower following bound for the union of events.

$$\mathsf{Pr}\left(\cup_{i=1}^{K} A_i\right) \geq \sum_{i=1}^{K} \mathsf{Pr}(A_i) \; - \; \sum_{i<j} \mathsf{Pr}(A_i \cap A_j)$$

- This idea is useful: When you cannot obtain the exact probability.

An Identity

- Let X be discrete non negative random variable with pmf $p_{\mathsf{X}}(i)$.
- Obviously, $p_{\mathsf{X}}(i) = 0$ for $i < 0$.
- Let us evaluate $\sum_{i=0}^{\infty} \Pr(\mathsf{X} > i)$

$$
\begin{aligned}
\sum_{i=0}^{\infty} \Pr(\mathsf{X} > i) &= p_{\mathsf{X}}(1) + p_{\mathsf{X}}(2) + \cdots \\
&\quad\; p_{\mathsf{X}}(2) + p_{\mathsf{X}}(3) + \cdots \\
&\quad\quad p_{\mathsf{X}}(3) + p_{\mathsf{X}}(4) + \cdots \\
&= \sum_{i=0}^{\infty} i p_{\mathsf{X}}(i) \\
&= \mathsf{E}(\mathsf{X})
\end{aligned}
$$

- This is also true for non negative continuous random variables

$$
\mathsf{E}(\mathsf{X}) = \int_0^{\infty} (1 - F_{\mathsf{X}}(x))\, dx
$$

An Identity

- Let X be discrete non negative random variable with pmf $p_{\mathsf{X}}(i)$.
- Obviously, $p_{\mathsf{X}}(i) = 0$ for $i < 0$.
- Let us evaluate $\sum_{i=0}^{\infty} \Pr(\mathsf{X} > i)$

$$\begin{aligned}
\sum_{i=0}^{\infty} \Pr(\mathsf{X} > i) &= p_{\mathsf{X}}(1) + p_{\mathsf{X}}(2) + \cdots \\
&\quad\; p_{\mathsf{X}}(2) + p_{\mathsf{X}}(3) + \cdots \\
&\quad\quad\; p_{\mathsf{X}}(3) + p_{\mathsf{X}}(4) + \cdots \\
&= \sum_{i=0}^{\infty} i p_{\mathsf{X}}(i) \\
&= \mathsf{E}(\mathsf{X})
\end{aligned}$$

- This is also true for non negative continuous random variables

$$\mathsf{E}(\mathsf{X}) = \int_0^{\infty} (1 - F_{\mathsf{X}}(x))\, dx$$

An Identity

- Let X be discrete non negative random variable with pmf $p_{\mathsf{X}}(i)$.
- Obviously, $p_{\mathsf{X}}(i) = 0$ for $i < 0$.
- Let us evaluate $\sum_{i=0}^{\infty} \mathsf{Pr}(\mathsf{X} > i)$

$$
\begin{aligned}
\sum_{i=0}^{\infty} \mathsf{Pr}(\mathsf{X} > i) &= p_{\mathsf{X}}(1) + p_{\mathsf{X}}(2) + \cdots \\
& \qquad p_{\mathsf{X}}(2) + p_{\mathsf{X}}(3) + \cdots \\
& \qquad\quad p_{\mathsf{X}}(3) + p_{\mathsf{X}}(4) + \cdots \\
&= \sum_{i=0}^{\infty} i p_{\mathsf{X}}(i) \\
&= \mathsf{E}(\mathsf{X})
\end{aligned}
$$

- This is also true for non negative continuous random variables

$$
\mathsf{E}(\mathsf{X}) = \int_0^{\infty} (1 - F_{\mathsf{X}}(x))\, dx
$$

Markov Inequality

- Clearly,

$$\mathsf{E}(\mathsf{X}) \; = \; \int (1 - F_{\mathsf{X}}(x))\, dx \; \geq \; x\,\mathsf{Pr}(\mathsf{X} > x)$$

- This gives us the first key inequality.
- **Markov Inequality:** If X is a non negative random variable, then

$$\mathsf{Pr}(\mathsf{X} > x) \leq \frac{\mathsf{E}(X)}{x}$$

Chebyshev Inequality

- Consider $\mathsf{X} = (\mathsf{Y} - \mu_{\mathsf{Y}})^2$ where Y is a random variable.
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$
\begin{aligned}
\Pr(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\
\Pr\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2 > x\right) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > \sqrt{x}) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} \quad (\text{writing } y\sqrt{x}) \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} = \frac{\sigma_{\mathsf{Y}}^2}{y^2}
\end{aligned}
$$

Chebyshev Inequality

- Consider $\mathsf{X} = (\mathsf{Y} - \mu_{\mathsf{Y}})^2$ where Y is a random variable.
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$
\begin{aligned}
\Pr(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\
\Pr\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2 > x\right) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > \sqrt{x}) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} \quad (\text{writing } y\sqrt{x}) \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} = \frac{\sigma_{\mathsf{Y}}^2}{y^2}
\end{aligned}
$$

Chebyshev Inequality

- Consider $\mathsf{X} = (\mathsf{Y} - \mu_{\mathsf{Y}})^2$ where Y is a random variable.
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$
\begin{aligned}
\Pr(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\
\Pr\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2 > x\right) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > \sqrt{x}) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} \quad (\text{writing } y \sqrt{x}) \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} = \frac{\sigma_{\mathsf{Y}}^2}{y^2}
\end{aligned}
$$

Chebyshev Inequality

- Consider $\mathsf{X} = (\mathsf{Y} - \mu_{\mathsf{Y}})^2$ where Y is a random variable.
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$
\begin{aligned}
\Pr(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\
\Pr\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2 > x\right) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > \sqrt{x}) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} \quad \text{(writing } y\sqrt{x}\text{)} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} = \frac{\sigma_{\mathsf{Y}}^2}{y^2}
\end{aligned}
$$

Chebyshev Inequality

- Consider $\mathsf{X} = (\mathsf{Y} - \mu_{\mathsf{Y}})^2$ where Y is a random variable.
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$
\begin{aligned}
\Pr(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\
\Pr\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2 > x\right) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > \sqrt{x}) &\leq \frac{\mathsf{E}\left((\mathsf{Y} - \mu_{\mathsf{Y}})^2\right)}{x} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} \quad \text{(writing } y\sqrt{x}\text{)} \\
\Pr(|\mathsf{Y} - \mu_{\mathsf{Y}}| > y) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} = \frac{\sigma_{\mathsf{Y}}^2}{y^2}
\end{aligned}
$$

Chebyshev Inequality

- The event $\{|\mathsf{Y} - \mu_\mathsf{Y}| > y\}$ is the same as $\{\mathsf{Y} - \mu_\mathsf{Y} < -y\} \cup \{\mathsf{Y} - \mu_\mathsf{Y} > y\}$.
- This in turn is the same as $\{\mathsf{Y}\mu_\mathsf{Y} < \mu_\mathsf{Y} - y\} \cup \{\mathsf{Y} > \mu_\mathsf{Y} + y\}$.
- Thus we have the Chebyshev Inequality

$$\begin{aligned}
\mathsf{Pr}(\mathsf{Y} \notin (\mu_\mathsf{Y} - y, \mu_\mathsf{Y} + y)) &\leq \frac{\mathsf{VAR}(\mathsf{Y})}{y^2} = \frac{\sigma_\mathsf{Y}^2}{y^2} \\
\mathsf{Pr}(\mu_\mathsf{Y} - y \leq \mathsf{Y} \leq \mu_\mathsf{Y} + y)) &\geq 1 - \frac{\sigma_\mathsf{Y}^2}{y^2}
\end{aligned}$$

Chernoff Bound

- Consider $\mathsf{X} = e^{s\mathsf{Z}}$ where Z is a random variable. Assume $s > 0$,
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$
\begin{aligned}
\mathsf{Pr}(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\
\mathsf{Pr}\left(e^{s\mathsf{Z}} > x\right) &\leq \frac{\mathsf{E}\left(e^{s\mathsf{Z}}\right)}{x} \\
\mathsf{Pr}\left(e^{s\mathsf{Z}} > e^{sz}\right) &\leq \frac{\mathsf{E}\left(e^{s\mathsf{Z}}\right)}{e^{sz}} \quad (\text{writing } x = e^{sz})
\end{aligned}
$$

- This gives us the Chernoff Bound: For $s > 0$,

$$\mathsf{Pr}(\mathsf{Z} > z) \leq \phi_{\mathsf{X}}(s)\, e^{-sz}$$

- Note that this is true for all values of s! Thus we can obtain very tight bounds by choosing the value of s for which $\phi_{\mathsf{X}}(s)\, e^{-sz}$ is minimum.

Chernoff Bound

- Consider $\mathsf{X} = e^{s\mathsf{Z}}$ where Z is a random variable. Assume $s > 0$,
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$\begin{aligned} \Pr(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\ \Pr\left(e^{s\mathsf{Z}} > x\right) &\leq \frac{\mathsf{E}\left(e^{s\mathsf{Z}}\right)}{x} \\ \Pr\left(e^{s\mathsf{Z}} > e^{sz}\right) &\leq \frac{\mathsf{E}\left(e^{s\mathsf{Z}}\right)}{e^{sz}} \quad \text{(writing } x = e^{sz}\text{)} \end{aligned}$$

- This gives us the Chernoff Bound: For $s > 0$,

$$\Pr(\mathsf{Z} > z) \leq \phi_{\mathsf{X}}(s)\, e^{-sz}$$

- Note that this is true for all values of s! Thus we can obtain very tight bounds by choosing the value of s for which $\phi_{\mathsf{X}}(s)\, e^{-sz}$ is minimum.

Chernoff Bound

- Consider $\mathsf{X} = e^{s\mathsf{Z}}$ where Z is a random variable. Assume $s > 0$,
- Clearly X is a non negative random variable. Hence we can apply Markov Inequality to X

$$\begin{aligned}
\Pr(\mathsf{X} > x) &\leq \frac{\mathsf{E}(X)}{x} \\
\Pr\left(e^{s\mathsf{Z}} > x\right) &\leq \frac{\mathsf{E}\left(e^{s\mathsf{Z}}\right)}{x} \\
\Pr\left(e^{s\mathsf{Z}} > e^{sz}\right) &\leq \frac{\mathsf{E}\left(e^{s\mathsf{Z}}\right)}{e^{sz}} \quad (\text{writing } x = e^{sz})
\end{aligned}$$

- This gives us the Chernoff Bound: For $s > 0$,

$$\Pr(\mathsf{Z} > z) \leq \phi_{\mathsf{X}}(s)\, e^{-sz}$$

- Note that this is true for all values of s! Thus we can obtain very tight bounds by choosing the value of s for which $\phi_{\mathsf{X}}(s)\, e^{-sz}$ is minimum.

Cauchy Schwarz Inequality

- The vector calculus version of the Cauchy Schwarz Inequality is: If **u** and **y** are vectors, then

$$|\mathbf{u} \cdot \mathbf{v}| \leq |\mathbf{u}|\ |\mathbf{v}|$$

- Equality holds when the vectors collinear, i.e., $\mathbf{u} = a\mathbf{v}$. Furthermore, the LHS is zero if the vectors are orthhogonal.
- A proof is obtained by considering the following quadratic function in a.

$$\begin{aligned} f(a) &= |au + v|^2 \\ &= \sum_i (au_i + v_i)^2 = \sum_i (a^2 u_i^2 + v_i^2 + 2au_i v_i) \\ &= a^2 \left(\sum_i u_i^2\right) + 2a \left(\sum_i u_i v_i\right) + \sum_i v_i^2 \geq 0 \end{aligned}$$

- Since this is non negative, the discriminant is non negative, i.e.,

$$4\left(\sum_i u_i v_i\right)^2 - 4\left(\sum_i u_i^2\right)\left(\sum_i u_i^2\right) \geq 0$$

and the identity follows.

Cauchy Schwarz Inequality

- The vector calculus version of the Cauchy Schwarz Inequality is: If $\mathbf{u}$ and $\mathbf{y}$ are vectors, then
$$|\mathbf{u}\cdot\mathbf{v}| \leq |\mathbf{u}|\ |\mathbf{v}|$$
- Equality holds when the vectors collinear, i.e., $\mathbf{u} = a\mathbf{v}$. Furthermore, the LHS is zero if the vectors are orthhogonal.
- A proof is obtained by considering the following quadratic function in a.

$$\begin{aligned} f(a) &= |au+v|^2 \\ &= \sum_i (au_i+v_i)^2 = \sum_i (a^2u_i^2+v_i^2+2au_iv_i) \\ &= a^2\left(\sum_i u_i^2\right)+2a\left(\sum_i u_iv_i\right)+\sum_i v_i^2 \geq 0 \end{aligned}$$

- Since this is non negative, the discriminant is non negative, i.e.,

$$4\left(\sum_i u_iv_i\right)^2 - 4\left(\sum_i u_i^2\right)\left(\sum_i u_i^2\right) \geq 0$$

and the identity follows.

Cauchy Schwarz Inequality

- The vector calculus version of the Cauchy Schwarz Inequality is: If $\mathbf{u}$ and $\mathbf{y}$ are vectors, then
$$|\mathbf{u} \cdot \mathbf{v}| \leq |\mathbf{u}| \, |\mathbf{v}|$$
- Equality holds when the vectors collinear, i.e., $\mathbf{u} = a\mathbf{v}$. Furthermore, the LHS is zero if the vectors are orthhogonal.
- A proof is obtained by considering the following quadratic function in a.

$$\begin{aligned} f(a) &= |au + v|^2 \\ &= \sum_i (au_i + v_i)^2 = \sum_i (a^2 u_i^2 + v_i^2 + 2au_i v_i) \\ &= a^2 \left(\sum_i u_i^2 \right) + 2a \left(\sum_i u_i v_i \right) + \sum_i v_i^2 \geq 0 \end{aligned}$$

- Since this is non negative, the discriminant is non negative, i.e.,

$$4 \left(\sum_i u_i v_i \right)^2 - 4 \left(\sum_i u_i^2 \right) \left(\sum_i u_i^2 \right) \geq 0$$

and the identity follows.

Cauchy Schwarz Inequality

- The vector calculus version of the Cauchy Schwarz Inequality is: If $\mathbf{u}$ and $\mathbf{y}$ are vectors, then
$$|\mathbf{u} \cdot \mathbf{v}| \leq |\mathbf{u}|\ |\mathbf{v}|$$
- Equality holds when the vectors collinear, i.e., $\mathbf{u} = a\mathbf{v}$. Furthermore, the LHS is zero if the vectors are orthhogonal.
- A proof is obtained by considering the following quadratic function in a.

$$\begin{aligned} f(a) &= |au + v|^2 \\ &= \sum_i (au_i + v_i)^2 \ = \ \sum_i (a^2 u_i^2 + v_i^2 + 2au_i v_i) \\ &= a^2 \left(\sum_i u_i^2 \right) + 2a \left(\sum_i u_i v_i \right) + \sum_i v_i^2 \ \geq \ 0 \end{aligned}$$

- Since this is non negative, the discriminant is non negative, i.e.,

$$4 \left(\sum_i u_i v_i \right)^2 - 4 \left(\sum_i u_i^2 \right) \left(\sum_i u_i^2 \right) \ \geq \ 0$$

and the identity follows.

Cauchy Schwarz Inequality

- We have an analog of Cauchy Schwarz Inequality for random variables

$$\mathsf{E}((\mathsf{X}-\mu_{\mathsf{X}})(\mathsf{Y}-\mu_{\mathsf{Y}})) \leq \sqrt{\mathsf{VAR}(X)\,\mathsf{VAR}(Y)}$$

- For zero-mean random variables X and Y, this means

$$\mathsf{E}(\mathsf{XY}) \leq \sqrt{\mathsf{E}(\mathsf{X}^2)\,\mathsf{E}(\mathsf{Y}^2)}$$

- In fact we seen this before! Same as $-1 \leq \rho_{\mathsf{XY}} \leq 1$
- A proof is exactly along the same lines as before

$$\begin{aligned} f(a) &= \mathsf{E}\Big((a(\mathsf{X}-\mu_{\mathsf{X}})+(\mathsf{Y}-\mu_{\mathsf{Y}}))^2\Big) \\ &= a^2\mathsf{E}((\mathsf{X}-\mu_{\mathsf{X}})^2)+\mathsf{E}((\mathsf{Y}-\mu_{\mathsf{Y}})^2)+2\mathsf{E}((\mathsf{X}-\mu_{\mathsf{X}})(\mathsf{Y}-\mu_{\mathsf{Y}})) \\ &= a^2\mathsf{VAR}(\mathsf{X})+a\,(2\mathsf{COV}(\mathsf{X},\mathsf{Y}))+\mathsf{VAR}(\mathsf{Y}) \end{aligned}$$

- $f(a)$ is non negative because it is the expectation of non negative random variable and the final arguments are like before.
- Equality holds when $\mathsf{Y} = a\mathsf{X}$
- Using the analogy from vectors, we say that X and Y are orthogonal if $\mathsf{COV}(\mathsf{X},\mathsf{Y}) = 0$.

Cauchy Schwarz Inequality

- We have an analog of Cauchy Schwarz Inequality for random variables

$$\mathsf{E}((\mathsf{X}-\mu_\mathsf{X})(\mathsf{Y}-\mu_\mathsf{Y})) \leq \sqrt{\mathsf{VAR}(X)\,\mathsf{VAR}(Y)}$$

- For zero-mean random variables X and Y, this means

$$\mathsf{E}(\mathsf{XY}) \leq \sqrt{\mathsf{E}(\mathsf{X}^2)\,\mathsf{E}(\mathsf{Y}^2)}$$

- In fact we seen this before! Same as $-1 \leq \rho_{\mathsf{XY}} \leq 1$
- A proof is exactly along the same lines as before

$$\begin{aligned} f(a) &= \mathsf{E}\Big((a(\mathsf{X}-\mu_\mathsf{X})+(\mathsf{Y}-\mu_\mathsf{Y}))^2\Big) \\ &= a^2\mathsf{E}\big((\mathsf{X}-\mu_\mathsf{X})^2\big)+\mathsf{E}\big((\mathsf{Y}-\mu_\mathsf{Y})^2\big)+2\mathsf{E}((\mathsf{X}-\mu_\mathsf{X})(\mathsf{Y}-\mu_\mathsf{Y})) \\ &= a^2\mathsf{VAR}(\mathsf{X})+a\,(2\mathsf{COV}(\mathsf{X},\mathsf{Y}))+\mathsf{VAR}(\mathsf{Y}) \end{aligned}$$

- $f(a)$ is non negative because it is the expectation of non negative random variable and the final arguments are like before.
- Equality holds when $\mathsf{Y} = a\mathsf{X}$
- Using the analogy from vectors, we say that X and Y are orthogonal if $\mathsf{COV}(\mathsf{X},\mathsf{Y}) = 0$.

Cauchy Schwarz Inequality

- We have an analog of Cauchy Schwarz Inequality for random variables

$$\mathsf{E}((\mathsf{X}-\mu_\mathsf{X})(\mathsf{Y}-\mu_\mathsf{Y})) \leq \sqrt{\mathsf{VAR}(X)\,\mathsf{VAR}(Y)}$$

- For zero-mean random variables X and Y, this means

$$\mathsf{E}(\mathsf{XY}) \leq \sqrt{\mathsf{E}(\mathsf{X}^2)\,\mathsf{E}(\mathsf{Y}^2)}$$

- In fact we seen this before! Same as $-1 \leq \rho_{\mathsf{XY}} \leq 1$
- A proof is exactly along the same lines as before

$$\begin{aligned} f(a) &= \mathsf{E}\left((a(\mathsf{X}-\mu_\mathsf{X}) + (\mathsf{Y}-\mu_\mathsf{Y}))^2\right) \\ &= a^2\mathsf{E}\left((\mathsf{X}-\mu_\mathsf{X})^2\right) + \mathsf{E}\left((\mathsf{Y}-\mu_\mathsf{Y})^2\right) + 2\mathsf{E}((\mathsf{X}-\mu_\mathsf{X})(\mathsf{Y}-\mu_\mathsf{Y})) \\ &= a^2\mathsf{VAR}(\mathsf{X}) + a\,(2\mathsf{COV}(\mathsf{X},\mathsf{Y})) + \mathsf{VAR}(\mathsf{Y}) \end{aligned}$$

- $f(a)$ is non negative because it is the expectation of non negative random variable and the final arguments are like before.
- Equality holds when $\mathsf{Y} = a\mathsf{X}$
- Using the analogy from vectors, we say that X and Y are orthogonal if $\mathsf{COV}(\mathsf{X},\mathsf{Y}) = 0$.

Convex functions; Jensen's Inequality

- Recall that the expection of $g(\mathsf{X})$ is

$$\mathsf{E}(g(\mathsf{X})) = \int g(x) f_{\mathsf{X}}(x) dx$$

- $g(x)$ is a convex function if, for $0 \leq a \leq 1$,

$$g(ax_1 + (1-a)x_2) \leq ag(x_1) + (1-a)g(x_2)$$

- Using induction, this can be generalized as follows. Let $a_1, a_2, \ldots a_n$ be such that $a_i \geq 0$ for all i and $\sum_{i=1}^n a_i = 1$. If $g(x)$ is convex, then

$$g(a_1x_1 + a_2x_2 + \cdots + a_nx_n) \leq a_1g(x_1) + a_2g(x_2) + \cdots a_ng(x_n)$$

- We can now use this as follows.
  - Assume X is a discrete random variable and it takes values $\{x_1, \ldots, x_n\}$. Note the generalisation from X being integer valued.
  - Let $a_i = p_{\mathsf{X}}(x_i)$. This gives us

$$\begin{aligned} & g(p_{\mathsf{X}}(x_1)x_1 + p_{\mathsf{X}}(x_2)x_2 + \cdots + p_{\mathsf{X}}(x_n)x_n) \\ \leq\ & p_{\mathsf{X}}(x_1)g(x_1) + p_{\mathsf{X}}(x_2)g(x_2) + \cdots p_{\mathsf{X}}(x_n)g(x_n) \end{aligned}$$

  - LHS is $g(\mathsf{E}(\mathsf{X}))$ and RHS is $\mathsf{E}(g(\mathsf{X}))$. This gives **Jensen's Inequality**

$$g(\mathsf{E}(\mathsf{X})) \leq \mathsf{E}(g(\mathsf{X}))$$

Convex functions; Jensen's Inequality

- Recall that the expection of $g(\mathsf{X})$ is

$$\mathsf{E}(g(\mathsf{X})) = \int g(x) f_{\mathsf{X}}(x) dx$$

- $g(x)$ is a convex function if, for $0 \leq a \leq 1$,

$$g(ax_1 + (1-a)x_2) \leq a g(x_1) + (1-a) g(x_2)$$

- Using induction, this can be generalized as follows. Let $a_1, a_2, \ldots a_n$ be such that $a_i \geq 0$ for all i and $\sum_{i=1}^n a_i = 1$. If $g(x)$ is convex, then

$$g(a_1 x_1 + a_2 x_2 + \cdots + a_n x_n) \leq a_1 g(x_1) + a_2 g(x_2) + \cdots a_n g(x_n)$$

- We can now use this as follows.
  - Assume X is a discrete random variable and it takes values $\{x_1, \ldots, x_n\}$. Note the generalisation from X being integer valued.
  - Let $a_i = p_{\mathsf{X}}(x_i)$. This gives us

$$\begin{aligned} & g(p_{\mathsf{X}}(x_1)x_1 + p_{\mathsf{X}}(x_2)x_2 + \cdots + p_{\mathsf{X}}(x_n)x_n) \\ \leq \; & p_{\mathsf{X}}(x_1)g(x_1) + p_{\mathsf{X}}(x_2)g(x_2) + \cdots p_{\mathsf{X}}(x_n)g(x_n) \end{aligned}$$

  - LHS is $g(\mathsf{E}(\mathsf{X}))$ and RHS is $\mathsf{E}(g(\mathsf{X}))$. This gives **Jensen's Inequality**

$$g(\mathsf{E}(\mathsf{X})) \;\leq\; \mathsf{E}(g(\mathsf{X}))$$

Jensen's Inequality

- There are several other proofs. Let us consider one more, which is also applicable for continuous random variables.
- Assume that $g(x)$ is convex.
- Now consider the tangent to $g(x)$ at $\mathsf{E}(\mathsf{X})$. Let $mx + c$ be the equation describing this line.
  - Note that since this line passes through the point $(\mathsf{E}(\mathsf{X}), g(\mathsf{E}(\mathsf{X})))$, the following is true
  $$g(\mathsf{E}(\mathsf{X})) = m\mathsf{E}(\mathsf{X}) + c.$$
  - Convexity of $g(x)$ means that the line obtained above is 'below' $g(x)$, i.e., $g(x) > mx + c$ for $a \leq x \leq b$.
  - We thus have

$$\mathsf{E}(g(\mathsf{X})) \geq \mathsf{E}(m\mathsf{X} + c) = m\mathsf{E}(\mathsf{X} + c) = g(\mathsf{E}(\mathsf{X}))$$

Hoeffding's Lemma

- Let X be a bounded zero-mean random variable, i.e., $\infty < a \leq \mathsf{X} \leq b < \infty$ and $\mathsf{E}(\mathsf{X}) = 0$.
- Consider the function e^{sx} for $a \leq x \leq b$. This is a convex function. Hence
$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$
It helps to visualise this with a figure. Observe that $\frac{x-a}{b-a} + \frac{b-x}{b-a} = 1$ for all $a \leq x \leq b$.
- Applying this identity to the random variable X, and taking expectations, we get

$$\begin{aligned}
\mathsf{E}(e^{s\mathsf{X}}) &\leq \mathsf{E}\left(\frac{\mathsf{X}-a}{b-a} e^{sb}\right) + \mathsf{E}\left(\frac{b-\mathsf{X}}{b-a} e^{sa}\right) \\
&= \mathsf{E}\left(\frac{\mathsf{X}}{b-a} e^{sb}\right) - \mathsf{E}\left(\frac{a}{b-a} e^{sb}\right) + \mathsf{E}\left(\frac{b}{b-a} e^{sa}\right) - \mathsf{E}\left(\frac{\mathsf{X}}{b-a} e^{sa}\right) \\
&= \mathsf{E}(\mathsf{X}) \frac{1}{b-a} e^{sb} - \frac{a}{b-a} e^{sb} + \frac{b}{b-a} e^{sa} - \mathsf{E}(\mathsf{X}) \frac{1}{b-a} e^{sa} \\
\mathsf{E}(e^{s\mathsf{X}}) &\leq \frac{be^{sa} - ae^{sb}}{b-a}
\end{aligned}$$

Hoeffding's Lemma

- Let X be a bounded zero-mean random variable, i.e., $\infty < a \leq \mathsf{X} \leq b < \infty$ and $\mathsf{E}(\mathsf{X}) = 0$.
- Consider the function e^{sx} for $a \leq x \leq b$. This is a convex function. Hence

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

It helps to visualise this with a figure. Observe that $\frac{x-a}{b-a} + \frac{b-x}{b-a} = 1$ for all $a \leq x \leq b$.

- Applying this identity to the random variable X, and taking expectations, we get

$$\begin{aligned}
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq \mathsf{E}\left(\frac{\mathsf{X}-a}{b-a}e^{sb}\right) + \mathsf{E}\left(\frac{b-\mathsf{X}}{b-a}e^{sa}\right) \\
&= \mathsf{E}\left(\frac{\mathsf{X}}{b-a}e^{sb}\right) - \mathsf{E}\left(\frac{a}{b-a}e^{sb}\right) + \mathsf{E}\left(\frac{b}{b-a}e^{sa}\right) - \mathsf{E}\left(\frac{\mathsf{X}}{b-a}e^{sa}\right) \\
&= \mathsf{E}(\mathsf{X})\frac{1}{b-a}e^{sb} - \frac{a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} - \mathsf{E}(\mathsf{X})\frac{1}{b-a}e^{sa} \\
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq \frac{be^{sa} - ae^{sb}}{b-a}
\end{aligned}$$

Hoeffding's Lemma

- Let X be a bounded zero-mean random variable, i.e., $\infty < a \leq \mathsf{X} \leq b < \infty$ and $\mathsf{E}(\mathsf{X}) = 0$.
- Consider the function e^{sx} for $a \leq x \leq b$. This is a convex function. Hence
$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$
It helps to visualise this with a figure. Observe that $\frac{x-a}{b-a} + \frac{b-x}{b-a} = 1$ for all $a \leq x \leq b$.
- Applying this identity to the random variable X, and taking expectations, we get

$$\begin{aligned}
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq \mathsf{E}\left(\frac{\mathsf{X}-a}{b-a} e^{sb}\right) + \mathsf{E}\left(\frac{b-\mathsf{X}}{b-a} e^{sa}\right) \\
&= \mathsf{E}\left(\frac{\mathsf{X}}{b-a} e^{sb}\right) - \mathsf{E}\left(\frac{a}{b-a} e^{sb}\right) + \mathsf{E}\left(\frac{b}{b-a} e^{sa}\right) - \mathsf{E}\left(\frac{\mathsf{X}}{b-a} e^{sa}\right) \\
&= \mathsf{E}(\mathsf{X}) \frac{1}{b-a} e^{sb} - \frac{a}{b-a} e^{sb} + \frac{b}{b-a} e^{sa} - \mathsf{E}(\mathsf{X}) \frac{1}{b-a} e^{sa}
\end{aligned}$$

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \leq \frac{be^{sa} - ae^{sb}}{b-a}$$

Hoeffding's Lemma

- Let X be a bounded zero-mean random variable, i.e., $\infty < a \leq \mathsf{X} \leq b < \infty$ and $\mathsf{E}(\mathsf{X}) = 0$.
- Consider the function e^{sx} for $a \leq x \leq b$. This is a convex function. Hence
$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$
It helps to visualise this with a figure. Observe that $\frac{x-a}{b-a} + \frac{b-x}{b-a} = 1$ for all $a \leq x \leq b$.
- Applying this identity to the random variable X, and taking expectations, we get

$$\begin{aligned}
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq \mathsf{E}\left(\frac{\mathsf{X}-a}{b-a}e^{sb}\right) + \mathsf{E}\left(\frac{b-\mathsf{X}}{b-a}e^{sa}\right) \\
&= \mathsf{E}\left(\frac{\mathsf{X}}{b-a}e^{sb}\right) - \mathsf{E}\left(\frac{a}{b-a}e^{sb}\right) + \mathsf{E}\left(\frac{b}{b-a}e^{sa}\right) - \mathsf{E}\left(\frac{\mathsf{X}}{b-a}e^{sa}\right) \\
&= \mathsf{E}(\mathsf{X})\,\frac{1}{b-a}e^{sb} - \frac{a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} - \mathsf{E}(\mathsf{X})\,\frac{1}{b-a}e^{sa} \\
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq \frac{be^{sa} - ae^{sb}}{b-a}
\end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \leq \frac{be^{sa} - ae^{sb}}{b-a} = xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^{y}}{x + (1-x)e^{y}} \implies f'(0) = 0 \\ f''(y) &= \frac{x(1-x)e^{y}}{(x+(1-x)e^{y})^2} = \left(\frac{x}{(x+(1-x)e^{y})}\right)\left(1 - \frac{x}{(x+(1-x)e^{y})}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}(e^{s\mathsf{X}}) \leq \frac{be^{sa} - ae^{sb}}{b-a} = xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^y}{x+(1-x)e^y} \implies f'(0) = 0 \\ f''(y) &= \frac{x(1-x)e^y}{(x+(1-x)e^y)^2} = \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1 - \frac{x}{(x+(1-x)e^y)}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}(e^{s\mathsf{X}}) \leq \frac{be^{sa} - ae^{sb}}{b-a} = xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^{y}}{x + (1-x)e^{y}} \implies f'(0) = 0 \\ f''(y) &= \frac{x(1-x)e^{y}}{(x+(1-x)e^{y})^2} = \left(\frac{x}{(x+(1-x)e^{y})}\right)\left(1 - \frac{x}{(x+(1-x)e^{y})}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}(e^{s\mathsf{X}}) \le \frac{be^{sa} - ae^{sb}}{b-a} = xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^y}{x + (1-x)e^y} \implies f'(0) = 0 \\ f''(y) &= \frac{x(1-x)e^y}{(x+(1-x)e^y)^2} = \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1 - \frac{x}{(x+(1-x)e^y)}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}(e^{s\mathsf{X}}) \leq \frac{be^{sa} - ae^{sb}}{b-a} = xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^y}{x + (1-x)e^y} \implies f'(0) = 0 \\ f''(y) &= \frac{x(1-x)e^y}{(x+(1-x)e^y)^2} = \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1 - \frac{x}{(x+(1-x)e^y)}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \leq \frac{be^{sa} - ae^{sb}}{b-a} \; = \; xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^{y}}{x + (1-x)e^{y}} \implies f'(0) \; = \; 0 \\ f''(y) &= \frac{x(1-x)e^{y}}{(x + (1-x)e^{y})^2} = \left(\frac{x}{(x + (1-x)e^{y})}\right)\left(1 - \frac{x}{(x + (1-x)e^{y})}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}(e^{s\mathsf{X}}) \le \frac{be^{sa} - ae^{sb}}{b-a} = xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^{y}}{x + (1-x)e^{y}} \implies f'(0) = 0 \\ f''(y) &= \frac{x(1-x)e^{y}}{(x+(1-x)e^{y})^2} = \left(\frac{x}{(x+(1-x)e^{y})}\right)\left(1 - \frac{x}{(x+(1-x)e^{y})}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \leq \frac{be^{sa} - ae^{sb}}{b-a} = xe^{sa} + (1-x)e^{sb}$$

- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider

$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$

- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^{y}}{x + (1-x)e^{y}} \implies f'(0) = 0 \\ f''(y) &= \frac{x(1-x)e^{y}}{(x + (1-x)e^{y})^2} = \left(\frac{x}{(x + (1-x)e^{y})}\right)\left(1 - \frac{x}{(x + (1-x)e^{y})}\right) \end{aligned}$$

Hoeffding's Lemma

- From the previous slide
$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \le \frac{be^{sa} - ae^{sb}}{b-a} \;=\; xe^{sa} + (1-x)e^{sb}$$
- Here we use $x = b/(b-a)$. Let $y = (b-a)s$ and consider
$$\begin{aligned} f(y) &:= \log\left(xe^{sa} + (1-x)e^{sb}\right) \\ &= sa + \log\left(x + (1-x)e^{s(b-a)}\right) \\ &= (x-1)y + \log(x + (1-x)e^{y}) \end{aligned}$$
- Note $f(0) = 0$;. Performing Taylor series expansion around $y = 0$ we get, for some $z = \psi(y)$,

$$\begin{aligned} f(y) &= f(0) + f'(0)y + \frac{1}{2}f''(z(y))y^2 \\ f'(y) &= (x-1) + \frac{(1-x)e^y}{x + (1-x)e^y} \implies f'(0) \;=\; 0 \\ f''(y) &= \frac{x(1-x)e^y}{(x + (1-x)e^y)^2} = \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1 - \frac{x}{(x+(1-x)e^y)}\right) \end{aligned}$$

Hoeffding's Lemma

- Thus

$$
\begin{aligned}
f''(y) &= \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1-\frac{x}{(x+(1-x)e^y)}\right) \le \frac{1}{4} \text{ for all } y. \\
f(y) &= f(0)+f'(0)y+\frac{1}{2}f''(z)y^2 = \frac{1}{2}f''(z)y^2 \\
&\le 0+0+\frac{1}{2}\,\frac{1}{4}\,y^2 = \frac{y^2}{8} \\
\mathsf{E}(e^{s\mathsf{X}}) &\le e^{y^2/8} = e^{\frac{s^2(b-a)^2}{8}} \text{ recall that } y=(b-a)s
\end{aligned}
$$

- This is Hoeffding's Lemma: If X is a bounded in (a, b) then

$$\mathsf{E}(e^{s\mathsf{X}}) \le e^{\frac{s^2(b-a)^2}{8}}$$

Hoeffding's Lemma

- Thus

$$
\begin{aligned}
f''(y) &= \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1-\frac{x}{(x+(1-x)e^y)}\right) \leq \frac{1}{4} \text{ for all } y. \\
f(y) &= f(0)+f'(0)y+\frac{1}{2}f''(z)y^2 = \frac{1}{2}f''(z)y^2 \\
&\leq 0+0+\frac{1}{2}\,\frac{1}{4}\,y^2 = \frac{y^2}{8} \\
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq e^{y^2/8} = e^{\frac{s^2(b-a)^2}{8}} \text{ recall that } y=(b-a)s
\end{aligned}
$$

- This is Hoeffding's Lemma: If X is a bounded in (a,b) then

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \leq e^{\frac{s^2(b-a)^2}{8}}$$

Hoeffding's Lemma

- Thus

$$\begin{aligned} f''(y) &= \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1-\frac{x}{(x+(1-x)e^y)}\right) \le \frac{1}{4} \text{ for all } y. \\ f(y) &= f(0)+f'(0)y+\frac{1}{2}f''(z)y^2 = \frac{1}{2}f''(z)y^2 \\ &\le 0+0+\frac{1}{2}\,\frac{1}{4}\,y^2 = \frac{y^2}{8} \\ \mathsf{E}\left(e^{s\mathsf{X}}\right) &\le e^{y^2/8} = e^{\frac{s^2(b-a)^2}{8}} \text{ recall that } y=(b-a)s \end{aligned}$$

- This is Hoeffding's Lemma: If X is a bounded in (a, b) then

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \le e^{\frac{s^2(b-a)^2}{8}}$$

Hoeffding's Lemma

- Thus

$$
\begin{aligned}
f''(y) &= \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1-\frac{x}{(x+(1-x)e^y)}\right) \leq \frac{1}{4} \text{ for all } y. \\
f(y) &= f(0)+f'(0)y+\frac{1}{2}f''(z)y^2 = \frac{1}{2}f''(z)y^2 \\
&\leq 0+0+\frac{1}{2}\,\frac{1}{4}\,y^2 = \frac{y^2}{8} \\
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq e^{y^2/8} = e^{\frac{s^2(b-a)^2}{8}} \text{ recall that } y=(b-a)s
\end{aligned}
$$

- This is Hoeffding's Lemma: If X is a bounded in (a,b) then

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \leq e^{\frac{s^2(b-a)^2}{8}}$$

Hoeffding's Lemma

- Thus

$$
\begin{aligned}
f''(y) &= \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1-\frac{x}{(x+(1-x)e^y)}\right) \leq \frac{1}{4} \text{ for all } y. \\
f(y) &= f(0)+f'(0)y+\frac{1}{2}f''(z)y^2 = \frac{1}{2}f''(z)y^2 \\
&\leq 0+0+\frac{1}{2}\,\frac{1}{4}\,y^2 = \frac{y^2}{8} \\
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\leq e^{y^2/8} = e^{\frac{s^2(b-a)^2}{8}} \text{ recall that } y=(b-a)s
\end{aligned}
$$

- This is Hoeffding's Lemma: If X is a bounded in (a,b) then

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \leq e^{\frac{s^2(b-a)^2}{8}}$$

Hoeffding's Lemma

- Thus

$$\begin{aligned}
f''(y) &= \left(\frac{x}{(x+(1-x)e^y)}\right)\left(1-\frac{x}{(x+(1-x)e^y)}\right) \le \frac{1}{4} \text{ for all } y. \\
f(y) &= f(0)+f'(0)y+\frac{1}{2}f''(z)y^2 = \frac{1}{2}f''(z)y^2 \\
&\le 0+0+\frac{1}{2}\,\frac{1}{4}\,y^2 = \frac{y^2}{8} \\
\mathsf{E}\left(e^{s\mathsf{X}}\right) &\le e^{y^2/8} = e^{\frac{s^2(b-a)^2}{8}} \text{ recall that } y=(b-a)s
\end{aligned}$$

- This is Hoeffding's Lemma: If X is a bounded in (a,b) then

$$\mathsf{E}\left(e^{s\mathsf{X}}\right) \le e^{\frac{s^2(b-a)^2}{8}}$$

Hoeffding's Inequality

- This is an extremely important inequality widely used in proving properties of machine learning algorithms.
- Consider X_i, $i = 1, \ldots, n$ be independent bounded random variables with $a_i \leq \mathsf{X}_i \leq b_i$. This means that $f_{\mathsf{X}_i}(x) = 0$ for $x \in [a_i, b_i]$.
- We are interested in the tail distribution of $\mathsf{Y}_n := \sum_{i=1}^n \mathsf{X}_i$, i.e., $\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y)$. Use the Chernoff Bound with $s > 0$,

$$
\begin{aligned}
\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y) &\leq e^{-sy} \mathsf{E}\left(e^{s(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n))}\right) \\
&= e^{-sy} \prod_{i=1}^n \mathsf{E}\left(e^{s(\mathsf{Y}_i - \mathsf{E}(\mathsf{Y}_i))}\right) \\
&\leq e^{-sy} \prod_{i=1}^n e^{s^2 (b_i - a_i)^2 / 8} \\
&= e^{-sy + \frac{s^2}{8} \sum_{i=1}^n (b_i - a_i)^2}
\end{aligned}
$$

- Since this is true for all $s > 0$, we can find the s for which the RHS is the minimum value, i.e., $s = \frac{4y}{\sum_{i=1}^n (b_i - a_i)^2}$. Substituting, we get Hoeffding's Inequality.

Hoeffding's Inequality

- This is an extremely important inequality widely used in proving properties of machine learning algorithms.
- Consider X_i, $i = 1, \ldots, n$ be independent bounded random variables with $a_i \leq \mathsf{X}_i \leq b_i$. This means that $f_{\mathsf{X}_i}(x) = 0$ for $x \in [a_i, b_i]$.
- We are interested in the tail distribution of $\mathsf{Y}_n := \sum_{i=1}^{n} \mathsf{X}_i$, i.e., $\mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y)$. Use the Chernoff Bound with $s > 0$,

$$
\begin{aligned}
\mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y) &\leq e^{-sy} \mathsf{E}\left(e^{s(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n))}\right) \\
&= e^{-sy} \prod_{i=1}^{n} \mathsf{E}\left(e^{s(\mathsf{Y}_i - \mathsf{E}(\mathsf{Y}_i))}\right) \\
&\leq e^{-sy} \prod_{i=1}^{n} e^{s^2 (b_i - a_i)^2/8} \\
&= e^{-sy + \frac{s^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2}
\end{aligned}
$$

- Since this is true for all $s > 0$, we can find the s for which the RHS is the minimum value, i.e., $s = \frac{4y}{\sum_{i=1}^{n} (b_i - a_i)^2}$. Substituting, we get Hoeffding's Inequality.

Hoeffding's Inequality

- This is an extremely important inequality widely used in proving properties of machine learning algorithms.
- Consider X_i, $i = 1, \ldots, n$ be independent bounded random variables with $a_i \leq \mathsf{X}_i \leq b_i$. This means that $f_{\mathsf{X}_i}(x) = 0$ for $x \in [a_i, b_i]$.
- We are interested in the tail distribution of $\mathsf{Y}_n := \sum_{i=1}^n \mathsf{X}_i$, i.e., $\mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y)$. Use the Chernoff Bound with $s > 0$,

$$\begin{aligned}
\mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y) &\leq e^{-sy}\mathsf{E}\left(e^{s(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n))}\right) \\
&= e^{-sy}\prod_{i=1}^{n}\mathsf{E}\left(e^{s(\mathsf{Y}_i - \mathsf{E}(\mathsf{Y}_i))}\right) \\
&\leq e^{-sy}\prod_{i=1}^{n} e^{s^2(b_i - a_i)^2/8} \\
&= e^{-sy + \frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2}
\end{aligned}$$

- Since this is true for all $s > 0$, we can find the s for which the RHS is the minimum value, i.e., $s = \frac{4y}{\sum_{i=1}^n (b_i - a_i)^2}$. Substituting, we get Hoeffding's Inequality.

Hoeffding's Inequality

- This is an extremely important inequality widely used in proving properties of machine learning algorithms.
- Consider X_i, $i = 1, \ldots, n$ be independent bounded random variables with $a_i \leq \mathsf{X}_i \leq b_i$. This means that $f_{\mathsf{X}_i}(x) = 0$ for $x \in [a_i, b_i]$.
- We are interested in the tail distribution of $\mathsf{Y}_n := \sum_{i=1}^n \mathsf{X}_i$, i.e., $\mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y)$. Use the Chernoff Bound with $s > 0$,

$$
\begin{aligned}
\mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) > y) &\leq e^{-sy} \mathsf{E}\left(e^{s(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n))}\right) \\
&= e^{-sy} \prod_{i=1}^n \mathsf{E}\left(e^{s(\mathsf{Y}_i - \mathsf{E}(\mathsf{Y}_i))}\right) \\
&\leq e^{-sy} \prod_{i=1}^n e^{s^2 (b_i - a_i)^2/8} \\
&= e^{-sy + \frac{s^2}{8} \sum_{i=1}^n (b_i - a_i)^2}
\end{aligned}
$$

- Since this is true for all $s > 0$, we can find the s for which the RHS is the minimum value, i.e., $s = \frac{4y}{\sum_{i=1}^n (b_i - a_i)^2}$. Substituting, we get Hoeffding's Inequality.

Hoeffding's Inequality

- Hoeffding Inequality: For bounded random variables X_i, for $i = 1, \ldots, n$, and $a_i \leq \mathsf{X}_i \leq b_i$,

$$\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) \leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

$$\Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) \leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

- If the X_i are iid, then

$$\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) \leq e^{\frac{-2y^2}{n(b-a)^2}}$$

$$\Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) \leq e^{\frac{-2y^2}{n(b-a)^2}}$$

- If, in addition $a = 0$, and $b = 1$, then

$$\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) \leq e^{\frac{-2y^2}{n}}$$

$$\Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) \leq e^{\frac{-2y^2}{n}}$$

Hoeffding's Inequality

- Hoeffding Inequality: For bounded random variables X_i, for $i = 1, \ldots, n$, and $a_i \leq \mathsf{X}_i \leq b_i$,

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \end{aligned}$$

- If the X_i are iid, then

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \end{aligned}$$

- If, in addition $a = 0$, and $b = 1$, then

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n}} \end{aligned}$$

Hoeffding's Inequality

- Hoeffding Inequality: For bounded random variables X_i, for $i = 1, \ldots, n$, and $a_i \leq \mathsf{X}_i \leq b_i$,

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \end{aligned}$$

- If the X_i are iid, then

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \end{aligned}$$

- If, in addition $a = 0$, and $b = 1$, then

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n}} \end{aligned}$$

Hoeffding's Inequality

- Hoeffding Inequality: For bounded random variables X_i, for $i = 1, \ldots, n$, and $a_i \leq \mathsf{X}_i \leq b_i$,

$$\begin{aligned}\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}}\end{aligned}$$

- If the X_i are iid, then

$$\begin{aligned}\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}}\end{aligned}$$

- If, in addition $a = 0$, and $b = 1$, then

$$\begin{aligned}\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n}}\end{aligned}$$

Hoeffding's Inequality

- Hoeffding Inequality: For bounded random variables X_i, for $i = 1, \ldots, n$, and $a_i \leq \mathsf{X}_i \leq b_i$,

$$\begin{aligned} \mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \\ \mathsf{Pr}(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \end{aligned}$$

- If the X_i are iid, then

$$\begin{aligned} \mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \\ \mathsf{Pr}(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \end{aligned}$$

- If, in addition $a = 0$, and $b = 1$, then

$$\begin{aligned} \mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\ \mathsf{Pr}(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n}} \end{aligned}$$

Hoeffding's Inequality

- Hoeffding Inequality: For bounded random variables X_i, for $i = 1, \ldots, n$, and $a_i \leq \mathsf{X}_i \leq b_i$,

$$\begin{aligned} \mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^{n}(b_i - a_i)^2}} \\ \mathsf{Pr}(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^{n}(b_i - a_i)^2}} \end{aligned}$$

- If the X_i are iid, then

$$\begin{aligned} \mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \\ \mathsf{Pr}(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \end{aligned}$$

- If, in addition $a = 0$, and $b = 1$, then

$$\begin{aligned} \mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\ \mathsf{Pr}(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n}} \end{aligned}$$

Hoeffding's Inequality

- Hoeffding Inequality: For bounded random variables X_i, for $i = 1, \ldots, n$, and $a_i \leq \mathsf{X}_i \leq b_i$,

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{\sum_{i=1}^n (b_i - a_i)^2}} \end{aligned}$$

- If the X_i are iid, then

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n(b-a)^2}} \end{aligned}$$

- If, in addition $a = 0$, and $b = 1$, then

$$\begin{aligned} \Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\ \Pr(\mathsf{E}(\mathsf{Y}_n) - \mathsf{Y}_n \geq y) &\leq e^{\frac{-2y^2}{n}} \end{aligned}$$

Using Hoeffding's Inequality

- A coin has an unknown bias p.
- The coin has been tossed n, let Y_n be the number of times a head has been observed.
- Note that $\mathsf{E}(\mathsf{Y_n}) = np$ and $\mathsf{E}(\mathsf{Y}_n/n) = p$.
- Hoeffding's Inequality tells us that

$$
\begin{aligned}
\mathsf{Pr}(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\
\mathsf{Pr}(\mathsf{Y}_n - np \geq ny) &\leq e^{-2ny^2} \\
\mathsf{Pr}\left(p - \frac{\mathsf{Y}_n}{n} \geq ny\right) &\leq e^{-2ny^2} \\
\mathsf{Pr}\left(p - \frac{\mathsf{Y}_n}{n} \geq y\right) &\leq e^{-2ny^2}
\end{aligned}
$$

- This bounds the probability.
- Let us now see how to use this result in a practical setting.
- For example, let $e^{\frac{-2y^2}{n}} = 0.01$. This corresponds to $y \approx 2.14n$
- What can we say about the true value of p based on this observation?

Using Hoeffding's Inequality

- A coin has an unknown bias p.
- The coin has been tossed n, let Y_n be the number of times a head has been observed.
- Note that $\mathsf{E}(\mathsf{Y_n}) = np$ and $\mathsf{E}(\mathsf{Y}_n/n) = p$.
- Hoeffding's Inequality tells us that

$$
\begin{aligned}
\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\
\Pr(\mathsf{Y}_n - np \geq ny) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq ny\right) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq y\right) &\leq e^{-2ny^2}
\end{aligned}
$$

- This bounds the probability.
- Let us now see how to use this result in a practical setting.
- For example, let $e^{\frac{-2y^2}{n}} = 0.01$. This corresponds to $y \approx 2.14n$
- What can we say about the true value of p based on this observation?

Using Hoeffding's Inequality

- A coin has an unknown bias p.
- The coin has been tossed n, let Y_n be the number of times a head has been observed.
- Note that $\mathsf{E}(\mathsf{Y_n}) = np$ and $\mathsf{E}(\mathsf{Y}_n/n) = p$.
- Hoeffding's Inequality tells us that

$$
\begin{aligned}
\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\
\Pr(\mathsf{Y}_n - np \geq ny) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq ny\right) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq y\right) &\leq e^{-2ny^2}
\end{aligned}
$$

- This bounds the probability.
- Let us now see how to use this result in a practical setting.
- For example, let $e^{\frac{-2y^2}{n}} = 0.01$. This corresponds to $y \approx 2.14n$
- What can we say about the true value of p based on this observation?

Using Hoeffding's Inequality

- A coin has an unknown bias p.
- The coin has been tossed n, let Y_n be the number of times a head has been observed.
- Note that $\mathsf{E}(\mathsf{Y_n}) = np$ and $\mathsf{E}(\mathsf{Y}_n/n) = p$.
- Hoeffding's Inequality tells us that

$$\begin{aligned}
\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\
\Pr(\mathsf{Y}_n - np \geq ny) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq ny\right) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq y\right) &\leq e^{-2ny^2}
\end{aligned}$$

- This bounds the probability.
- Let us now see how to use this result in a practical setting.
- For example, let $e^{\frac{-2y^2}{n}} = 0.01$. This corresponds to $y \approx 2.14n$
- What can we say about the true value of p based on this observation?

Using Hoeffding's Inequality

- A coin has an unknown bias p.
- The coin has been tossed n, let Y_n be the number of times a head has been observed.
- Note that $\mathsf{E}(\mathsf{Y}_\mathsf{n}) = np$ and $\mathsf{E}(\mathsf{Y}_n/n) = p$.
- Hoeffding's Inequality tells us that

$$\begin{aligned}
\Pr(\mathsf{Y}_n - \mathsf{E}(\mathsf{Y}_n) \geq y) &\leq e^{\frac{-2y^2}{n}} \\
\Pr(\mathsf{Y}_n - np \geq ny) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq ny\right) &\leq e^{-2ny^2} \\
\Pr\left(p - \frac{\mathsf{Y}_n}{n} \geq y\right) &\leq e^{-2ny^2}
\end{aligned}$$

- This bounds the probability.
- Let us now see how to use this result in a practical setting.
- For example, let $e^{\frac{-2y^2}{n}} = 0.01$. This corresponds to $y \approx 2.14n$
- What can we say about the true value of p based on this observation?

A learning problem

- There are three coins, named A, B, and C, each with unknownb biases possibly different. You are allowed a total of N tosses. For each toss you can choose any one of the three coins using any algorithm. And when you toss the chosen coin, you get a reward of one unit if it comes up heads.
- You have to 'learn' which of them is the more profitable coin and use it maximally.
- Before the n-th toss, let n_A, n_B, and n_C, be the number of times coins A, B, and C, respectively, have been tossed and let k_A, k_B, and k_C, be the number of heads for these coins.
- The preceding slide allows us to claim that value of p_A is less than $UCB_A = k_A/n_A + X_A$ with probability 0.01. X from n_A and the discussion in the previous slide. Similarly, for coins B and C.
- For the n-toss choosing the coin with the highest UCB has some very nice properties. We will investigate these in a computational experiment.