Lecture Notes 9: Information Measure Entropy

Professor: Zhihua Zhang *Scribe:*

9 Probability Inequality

9.1 Jensen Inequality

If g is convex, then $\mathbb{E}[g(X)] \geq g(\mathbb{E}X)$.

Proof. Since g is convex, we can find a linear function $L(x) = a + bx$ such that the only intersection point is $\mathbb{E}X$ and $L'(\mathbb{E}X) = g'(\mathbb{E}X)$. So

$$\begin{aligned} g(x) &\geq L(x) \\ \mathbb{E}[g(X)] &\geq \mathbb{E}[L(X)] \\ &= a + b\mathbb{E}X \\ &= L(\mathbb{E}X) \\ &= g(\mathbb{E}X) \end{aligned}$$

□

9.2 Cauchy-Schwartz Inequality

If X and Y have finite variances, then

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

Proof. Consider vector variable $\begin{bmatrix} X \\ Y \end{bmatrix}$, its variance is

$$var(\begin{bmatrix} X \\ Y \end{bmatrix}) = \begin{bmatrix} var(X) & cov(X,Y) \\ cov(Y,X) & var(Y) \end{bmatrix}$$

Since variance is semi-definite, so $var(X)var(Y) \geq cov(X,Y)cov(Y,X)$. Now let $\mathbb{E}X = \mathbb{E}Y = 0$, we can get the inequality. □

9.3 Markov Inequality

For all $t > 0$,

$$\begin{aligned} Y1_{\{y \geq t\}} &\geq t1_{\{y \geq t\}} \\ \mathbb{E}(Y1_{\{y \geq t\}}) &\geq \mathbb{E}(t1_{\{y \geq t\}}) \\ Pr(\{y \geq t\}) &\leq \frac{\mathbb{E}(Y1_{\{y \geq t\}})}{t} \end{aligned}$$

If $Y > 0$, then $Pr(\{y \geq t\}) \leq \frac{\mathbb{E}Y}{t}$

Corollary 9.1. *Let* $Y = |Z - \mathbb{E}Z|$, *then* $Pr(\{|Z - \mathbb{E}Z| \geq t\}) \leq \frac{\mathbb{E}|Z-\mathbb{E}Z|}{t}$

Corollary 9.2. *If* ϕ *denotes a nondecreasing and nonnegative function of* Z *on a (possibly infinite) interval* $I \subset \mathbf{R}$. *Let* Y *and* t *take values in* I, $t \in \mathbb{R}$, *then*

$$\begin{aligned} Pr(\{Y \geq t\}) &\leq Pr(\{\phi(Y) \geq \phi(t)\}) \\ &\leq \frac{\mathbb{E}[\phi(Y)]}{\phi(t)} \end{aligned}$$

Example 9.1. *Let* $\phi(t) = t^2$, $I = (0, +\infty)$, $Y = |Z - \mathbb{E}Z|$. *Then*

$$Pr(\{|Z - \mathbb{E}Z| \geq t\}) \leq \frac{var(Z)}{t^2}$$

which is called Chebyshev's inequality.

More generally, $\phi(t) = t^q$, then for some $q > 0$, we have

$$Pr(\{|Z - \mathbb{E}Z| \geq t\})\ leq \frac{\mathbb{E}(\{||Z - \mathbb{E}Z||^q\})}{t^q}$$

Example 9.2. Z *is a sum of independent of random variables* $Z = X_1 + X_2 + ... + X_n$, *so* $var(Z) = \sum_{i=1}^n var(X_i)$. *Then we have*

$$Pr(\{\frac{1}{n}|\sum_{i=1}^n (X_i - \mathbb{E}X_i)| \geq t\}) \leq \frac{\sigma^2}{nt^2}$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n var(X_i)$.

Let $\phi(t) = e^{\lambda t}$ where λ is a positive number, then we will get

$$Pr(\{Z \geq t\}) \leq \frac{\mathbb{E}[e^{\lambda Z}]}{e^{\lambda t}}$$

Note: $M(\lambda) = \mathbb{E}e^{\lambda Z}$, $\lambda \in \mathbb{R}$ is called the moment generating function.

9.4 The Cramer-Chernoff Method

Let Z be a real-valued random variable. For all $\lambda \geq 0$, we have

$$Pr(\{Z \geq t\}) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda Z}]$$

We want to minimize the upper bound, so

$$\begin{aligned} &\inf_{\lambda \geq 0} e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] \\ \Leftrightarrow &\inf_{\lambda \geq 0} -\lambda t + \log \mathbb{E}[e^{\lambda Z}] \end{aligned}$$

Define $\psi_Z(\lambda) = \log \mathbb{E}e^{\lambda Z}$. Let $\psi_Z^*(t) \triangleq \sup_{\lambda \geq 0} \lambda t - \psi_Z(\lambda)$, which is called the cramer transform of Z.

If $\lambda = 0$, then $\psi_Z(0) = \log \mathbb{E}e^0 = 0$. So we can get $\psi_Z^* \geq 0$.

1. $\mathbb{E}Z \leq t \leq +\infty$.

$$\begin{aligned}\psi_Z(\lambda) &= \log \mathbb{E}(e^{\lambda Z}) \\ &\geq \log e^{\lambda \mathbb{E} Z} \\ &= \lambda \mathbb{E} Z.\end{aligned}$$

   If $\lambda < 0$, then $\lambda t - \psi_Z(\lambda) \leq 0$. So $\sup_{\lambda \geq 0} \lambda t - \psi_Z(\lambda) = \sup_{\lambda \in \mathbb{R}} \lambda t - \psi_Z(\lambda)$.

   Note: $\psi_Z^*(t) = \sup_{\lambda \in \mathbb{R}} \lambda t - \psi_Z(\lambda)$ is called Fenchel-Legendre dual function and convex conjugate.

   SO if $t \geq \mathbb{E}Z$, we only need to compute the dual function.

2. $t \leq \mathbb{E}Z$, To get the maximum value of $\lambda t - \psi_Z(\lambda)$, we compute its deriatives.

$$\psi_Z'(\lambda) = \frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}e^{\lambda Z}}$$

$$\psi_Z''(\lambda) = \frac{\mathbb{E}[Z^2 e^{\lambda Z}]\mathbb{E}[e^{\lambda Z}] - \mathbb{E}[Ze^{\lambda Z}]\mathbb{E}[Ze^{\lambda Z}]}{(\mathbb{E}[e^{\lambda Z}])^2}$$

   According to Cauchy-Schwartz inequality, $\psi_Z''(\lambda) \geq 0$. So,

$$\psi_Z'(\lambda) \geq \psi_Z'(0) = \mathbb{E}Z$$

$$t - \psi_Z'(\lambda) \leq t - \mathbb{E}Z \leq 0$$

   Then $\lambda t - \psi_Z(\lambda)$ gets its maximum value at $\lambda = 0$. In this case, we will get $\psi_Z^*(t) = 0$, which means $Pr(Z \geq t) \leq 1$.

In the following, we only care about $t \geq \mathbb{E}Z$. We will get

$$\psi_Z^*(t) = \lambda_t t - \psi_Z(\lambda_t)$$

where λ_t is the solution of $t - \psi_Z'(\lambda) = 0$, i.e. $\lambda_t = ({\psi_Z'}^{-1})(t)$.

Example 9.3. *Let $Z \sim N(0, \sigma^2)$, then we have*

$$\begin{aligned}\psi_Z(\lambda) &= \log \int e^{\lambda z} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{z^2}{2\sigma^2}) dz \\ &= \log \int \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{z^2 - 2\lambda\sigma^2 z}{2\sigma^2}) dz \\ &= \log \int \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{(z - \lambda z^2)^2 - \lambda^2\sigma^4}{2\sigma^2}) \\ &= \frac{\lambda^2\sigma^2}{2}\end{aligned}$$

$$\psi_Z^*(t) = \sup_\lambda \lambda t - \psi_Z(\lambda)$$

$t - \lambda\sigma^2 = 0 \Rightarrow \lambda_t = \frac{t}{\sigma^2}$, so

$$Pr(Z \geq t) \leq \exp(-\frac{t^2}{2\sigma^2})$$

, where $t \geq \mathbb{E}Z = 0$.

Note: If $\psi_Y(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$, then we call Y is sub-Gaussian.

Homework: Given that $\psi_Y(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$, prove $var(Y) \leq \sigma^2$.

Example 9.4. *A random variable Y has Poisson distribution with parameter ν.*

$$Pr(Y = k) = \frac{e^{-\nu}\nu^k}{k!}$$

where $k = 0, 1, 2....$

Let $Z = Y - \nu$, then $\mathbb{E}Z = 0$.

$$\begin{aligned}
\mathbb{E}e^{\lambda Z} &= e^{-\lambda\nu} \sum_{k=0}^{\infty} e^{\lambda k} e^{-\nu} \frac{\nu^k}{k!} \\
&= e^{-\lambda\nu - \nu} \sum_{k=0}^{\infty} \frac{(\nu e^\lambda)^k}{k!} \\
&= e^{-\lambda\nu-\nu} e^{\nu e^\lambda}
\end{aligned}$$

Then $\psi(\lambda) = \nu(e^\lambda - \lambda - 1)$. So

$$\begin{aligned}
t - \psi'(\lambda) &= 0 \\
\Rightarrow \lambda_t &= \log(1 + \frac{t}{\nu})
\end{aligned}$$

So $\psi^*(t) = \nu[(1 + \frac{t}{\nu})\log(1 + \frac{t}{\nu}) - \frac{t}{\nu}]$

Example 9.5. *A random variable Y has Bernoulli distribution with parameter p.*

$$Pr(Y = 1) = 1 - Pr(Y = 0) = p.$$

Let $Z = Y - p$.

$$\begin{aligned}
\psi_Z(\lambda) &= \log \mathbb{E}e^{\lambda Z} \\
&= \log(pe^{\lambda(1-p)} + (1-p)e^{-\lambda p}) \\
&= -\lambda p + \log(pe^\lambda + 1 - p)
\end{aligned}$$

Since $(\lambda t - \psi_Z(\lambda))' = 0$, so $pe^\lambda(1 - t - p) = (t + p)(1 - p)$. Then $0 \leq 1 - t - p$. So $0 \leq t \leq 1 - p$.

$$\psi_Z^*(t) = (1 - p - t)\log\frac{1 - p - t}{1 - p} + (p + t)\log\frac{p + t}{p}$$

Let $a = p + t$, $p \leq a \leq 1$, then we get

$$\begin{aligned}\psi_Z^*(t) &= (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p} \\ &= D(P_a||P_p)\end{aligned}$$

$D(P_a||P_p)$ is the KL-Divergence between P_a and P_p, P_a means the Bernoulli distribution with paramter a, P_p means the Bernoulli distribution with parameter p.

Example 9.6. *Let $Y \sim Binomial(n,p)$, so $Y = Z_1 + Z_2 + .. + Z_n$, and $Z_i \sim Bernoulli(p)$ and Z_i's are independent.*

$$\begin{aligned}\psi_Y(\lambda) &= \log \mathbb{E}e^{\lambda\sum_{i=1}^n Z_i} \\ &= \log \Pi_{i=1}^n \mathbb{E}e^{\lambda Z_i} \\ &= \sum \log \mathbb{E}e^{\lambda Z_i} \\ &= n\psi_Z(\lambda)\end{aligned}$$

$$\begin{aligned}\lambda t - \psi_Y(\lambda) &= \lambda t - n\psi_Z(\lambda) \\ &= n(\frac{\lambda t}{den} - \psi_Z(\lambda))\end{aligned}$$

So, $\psi_Y^*(t) = n\psi_Z^*(\frac{t}{n})$

9.5 Hoeffding's Inequality

If $X_1, X_2, ...X_n$ are independent random variables with a finite mean value such that for some non-empty interval I, $\mathbb{E}e^{\lambda X_i}$ is finite, then define

$$S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$$

. And assume that X_i takes its values in a bounded interval $[a_i, b_i]$. Then

$$Pr(S \geq t) \leq \exp(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2})$$

for all $t > 0$.

Definition 9.1. *If $Pr(\epsilon_i = 1) = Pr(\epsilon_i = -1) = \frac{1}{2}$, then we call ϵ_i Rademacher random variable.*

Let $X_i = \epsilon_i a_i$, a_i is a real number. Then we will get $X_i \in [\min\{-a_i, a_i\}, \max\{-a_i, a_i\}]$. So the inequality above will be

$$Pr(S \geq t) \leq \exp(-\frac{t^2}{2\sum_{i=1}^n a_i^2})$$