

Homework Assignment 3

Due in class, Wednesday March 7th

SDS 384-11 Theoretical Statistics

- Suppose that X_1 and X_2 are zero-mean and sub-Gaussian with parameters σ_1 and σ_2 respectively. **Assume that the variance parameters are equal to the sub-gaussian parameters, i.e. $E[X_1^2] = \sigma_1^2$ and $E[X_2^2] = \sigma_2^2$. This is needed for part (a) and (c) uses part (a).**
 - Show that the MGF of $V := X_1^2 - E[X_1^2]$ can be bounded as $E[e^{tV}] \leq e^{2\sigma_1^4 t^2}$ for $0 \leq t \leq 1/4\sigma_1^2$. *Hint: write the mgf in terms of X_1 and an independent standard normal.*
 - If X_1 and X_2 are not independent, show that $X_1 + X_2$ is sub-Gaussian with parameter at most $\sqrt{2(\sigma_1^2 + \sigma_2^2)}$.
 - If X_1 and X_2 are independent, show that $X_1 X_2$ is sub-exponential with parameters $(\sqrt{2}\sigma_1\sigma_2, \frac{1}{\sqrt{2}\sigma_1\sigma_2})$.
- Let X_1, X_2, \dots, X_n be i.i.d. samples of random variable with density f on the real line. A standard estimate of f is the kernel density estimate

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

where $K : \mathbb{R} \rightarrow [0, \infty)$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(t)dt = 1$, and h is a bandwidth parameter. We will measure the quality of \hat{f} using

$$\|\hat{f} - f\|_1 := \int_{-\infty}^{\infty} |\hat{f}(t) - f(t)|dt.$$

Prove that:

$$P(\|\hat{f} - f\|_1 \geq E\|\hat{f} - f\|_1 + \delta) \leq e^{-cn\delta^2},$$

where c is some constant.

- Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of Bernoulli variables with parameter $\alpha \in (0, 1/2]$, and consider the binomial random variable $Z_n = \sum_i X_i$. We want to prove for any $\delta \in (0, \alpha)$,

$$P(Z_n \leq \delta n) \leq \exp(-nKL(\delta|\alpha)) \quad KL(\delta|\alpha) := \delta \log \frac{\delta}{\alpha} + (1 - \delta) \log \frac{1 - \delta}{1 - \alpha}$$

where $KL(p, q)$ is the Kullback-Leibler divergence between two bernoullis with parameters p, q respectively. Show that the above is strictly better than Hoeffding's inequality.

4. Now we will prove a lower bound on the binomial tail to show that indeed what you derived in the last question is sharp upto polynomial factors. Define $m = \lfloor n\delta \rfloor$ and $\delta' = \frac{m}{n}$.

- (a) Prove $\frac{1}{n} \log P(Z_n \leq \delta n) \geq \frac{1}{n} \log \binom{n}{m} + \delta' \log \alpha + (1 - \delta') \log(1 - \alpha)$.
 (b) Show that

$$\frac{1}{n} \log \binom{n}{m} \geq -\delta' \log \delta' - (1 - \delta') \log(1 - \delta') - \frac{\log(n+1)}{n}$$

Hint: Use the fact that for $Y \sim \text{Bin}(n, m/n)$ $P(Y = k)$ is maximized at $k = m$.

- (c) Now show that

$$P(Z_n \leq \delta n) \geq \frac{1}{n+1} \exp(-KL(\delta||\alpha))$$