Let $a_1, \ldots, a_n \in \mathbb{R}$ and let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables: $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 0.5$.

Theorem 7.1 (Hoeffding)**.** *For $t \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right).$$

Proof. Similarly to the proof of Bennett's inequality (Lecture 5),

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right) \leq e^{-\lambda t}\mathbb{E}\exp\left(\lambda\sum_{i=1}^n \varepsilon_i a_i\right) = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\exp\left(\lambda\varepsilon_i a_i\right).$$

Using inequality $\frac{e^x+e^{-x}}{2} \leq e^{x^2/2}$ (from Taylor expansion), we get

$$\mathbb{E}\exp\left(\lambda\varepsilon_i a_i\right) = \frac{1}{2}e^{\lambda a_i} + \frac{1}{2}e^{-\lambda a_i} \leq e^{\frac{\lambda^2 a_i^2}{2}}.$$

Hence, we need to minimize the bound with respect to $\lambda > 0$:

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right) \leq e^{-\lambda t}e^{\frac{\lambda^2}{2}\sum_{i=1}^n a_i^2}.$$

Setting derivative to zero, we obtain the result. □

Now we change variable: $u = \frac{t^2}{2\sum_{i=1}^n a_i^2}$. Then $t = \sqrt{2u\sum_{i=1}^n a_i^2}$.

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq \sqrt{2u\sum_{i=1}^n a_i^2}\right) \leq e^{-u}$$

and

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \leq \sqrt{2u\sum_{i=1}^n a_i^2}\right) \geq 1 - e^{-u}.$$

Here $\sum_{i=1}^n a_i^2 = \mathrm{Var}(\sum_{i=1}^n \varepsilon_i a_i)$.

Rademacher sums will play important role in future. Consider again the problem of estimating $\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f$. We will see that by the Symmetrization technique,

$$\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f \sim \frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n f(X_i').$$

In fact,

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right| \le \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n f(X_i')\right| \le 2\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|.$$

The second inequality above follows by adding and subtracting $\mathbb{E}f$:

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n f(X_i')\right| &\le \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right| + \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i') - \mathbb{E}f\right| \\ &= 2\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|\end{aligned}$$

while for the first inequality we use Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right| &= \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i')\right| \\ &\le \mathbb{E}_X\mathbb{E}_{X'}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i')\right|.\end{aligned}$$

Note that $\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i')$ is equal in distribution to $\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f(X_i'))$.

We now prove Hoeffding-Chernoff Inequality:

Theorem 7.2. *Assume* $0 \le X_i \le 1$ *and* $\mu = \mathbb{E}X$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \mu \ge t\right) \le e^{-n\mathcal{D}(\mu+t,\mu)}$$

where the KL-divergence $\mathcal{D}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$.

Proof. Note that $\phi(x) = e^{\lambda x}$ is convex and so $e^{\lambda x} = e^{\lambda(x\cdot 1 + (1-x)\cdot 0)} \le xe^{\lambda} + (1-x)e^{\lambda\cdot 0} = 1 - x + xe^{\lambda}$. Hence,

$$\mathbb{E}e^{\lambda X} = 1 - \mathbb{E}X + \mathbb{E}Xe^{\lambda} = 1 - \mu + \mu e^{\lambda}.$$

Again, we minimize the following bound with respect to $\lambda > 0$:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \ge n(\mu+t)\right) &\le e^{-\lambda n(\mu+t)}\mathbb{E}e^{\lambda\sum X_i} \\ &= e^{-\lambda n(\mu+t)}\left(\mathbb{E}e^{\lambda X}\right)^n \\ &\le e^{-\lambda n(\mu+t)}\left(1-\mu+\mu e^{\lambda}\right)^n\end{aligned}$$

Take derivative w.r.t. λ:

$$-n(\mu+t)e^{-\lambda n(\mu+t)}(1-\mu+\mu e^{\lambda})^n + n(1-\mu+\mu e^{\lambda})^{n-1}\mu e^{\lambda}e^{-\lambda n(\mu+t)} = 0$$

$$-(\mu+t)(1-\mu+\mu e^{\lambda}) + \mu e^{\lambda} = 0$$

$$e^{\lambda} = \frac{(1-\mu)(\mu+t)}{\mu(1-\mu-t)}.$$

Substituting,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n X_i \ge n(\mu+t)\right) &\le \left(\left(\frac{\mu(1-\mu-t)}{(1-\mu)(\mu+t)}\right)^{\mu+t}\left(1-\mu+\frac{(1-\mu)(\mu+t)}{1-\mu-t}\right)\right)^n \\
&= \left(\left(\frac{\mu}{\mu+t}\right)^{\mu+t}\left(\frac{1-\mu}{1-\mu-t}\right)^{1-\mu-t}\right)^n \\
&= \exp\left(-n\left((\mu+t)\log\frac{\mu+t}{\mu} + (1-\mu-t)\log\frac{1-\mu-t}{1-\mu}\right)\right),
\end{aligned}$$

completing the proof. Moreover,

$$\mathbb{P}\left(\mu - \sum_{i=1}^n X_i \ge t\right) = \mathbb{P}\left(\sum_{i=1}^n Z_i - \mu_Z \ge t\right) \le e^{-n\mathcal{D}(\mu_z+t,\mu_Z)} = e^{-n\mathcal{D}(1-\mu_X+t,1-\mu_X)}$$

where $Z_i = 1 - X_i$ (and thus $\mu_Z = 1-\mu_X$). $\square$

If $0 < \mu \le 1/2$,

$$\mathcal{D}(1-\mu+t, 1-\mu) \ge \frac{t^2}{2\mu(1-\mu)}.$$

Hence, we get

$$\mathbb{P}\left(\mu - \sum_{i=1}^n X_i \ge t\right) \le e^{-\frac{nt^2}{2\mu(1-\mu)}} = e^{-u}.$$

Solving for t,

$$\mathbb{P}\left(\mu - \sum_{i=1}^n X_i \ge \sqrt{\frac{2\mu(1-\mu)u}{n}}\right) \le e^{-u}.$$

If $X_i = 0, 1$, then $\mu = \mathbb{E}X = \mathbb{P}(X=1)$ and $\mathrm{Var}(X) = \mu(1-\mu)$.