Last time we proved Bennett's inequality: $\mathbb{E}X = 0$, $\mathbb{E}X^2 = \sigma^2$, $|X| < M = const$, $X_1, \cdots, X_n$ independent copies of X, and $t \geq 0$. Then

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{n\sigma^2}{M^2}\phi\left(\frac{tM}{n\sigma^2}\right)\right),$$

where $\phi(x) = (1+x)\log(1+x) - x$.

If X is small, $\phi(x) = (1+x)(x - \frac{x^2}{2} + \cdots) - x = x + x^2 - \frac{x^2}{2} - x + \cdots = \frac{x^2}{2} + \cdots$.

If X is large, $\phi(x) \sim x \log x$.

We can weaken the bound by decreasing $\phi(x)$. Take[1] $\phi(x) = \frac{x^2}{2+\frac{2}{3}x}$ to obtain **Bernstein's inequality**:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) &\leq \exp\left(-\frac{n\sigma^2}{M^2}\left(\frac{\left(\frac{tM}{n\sigma^2}\right)^2}{2+\frac{2}{3}\frac{tM}{n\sigma^2}}\right)\right)\\ &= \exp\left(-\frac{t^2}{2n\sigma^2+\frac{2}{3}tM}\right)\\ &= e^{-u}\end{aligned}$$

where $u = \frac{t^2}{2n\sigma^2+\frac{2}{3}tM}$. Solve for t:

$$t^2 - \frac{2}{3}uMt - 2n\sigma^2 u = 0$$

$$t = \frac{1}{3}uM + \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2 u}.$$

Substituting,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2 u} + \frac{uM}{3}\right) \leq e^{-u}$$

or

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leq \sqrt{\frac{u^2M^2}{9} + 2n\sigma^2 u} + \frac{uM}{3}\right) \geq 1 - e^{-u}$$

Using inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leq \sqrt{2n\sigma^2 u} + \frac{2uM}{3}\right) \geq 1 - e^{-u}$$

[1] exercise: show that this is the best approximation

For non-centered X_i, replace X_i with $X_i - \mathbb{E}X$ or $\mathbb{E}X - X_i$. Then $|X_i - \mathbb{E}X| \le 2M$ and so with high probability

$$\sum (X_i - \mathbb{E}X) \le \sqrt{2n\sigma^2 u} + \frac{4uM}{3}.$$

Normalizing by n,

$$\frac{1}{n}\sum X_i - \mathbb{E}X \le \sqrt{\frac{2\sigma^2 u}{n}} + \frac{4uM}{3n}$$

and

$$\mathbb{E}X - \frac{1}{n}\sum X_i \le \sqrt{\frac{2\sigma^2 u}{n}} + \frac{4uM}{3n}.$$

Whenever $\sqrt{\frac{2\sigma^2 u}{n}} \ge \frac{4uM}{3n}$, we have $u \le \frac{n\sigma^2}{8M^2}$. So, $|\frac{1}{n}\sum X_i - \mathbb{E}X| \lesssim \sqrt{\frac{2\sigma^2 u}{n}}$ for $u \lesssim n\sigma^2$ (range of normal deviations). This is predicted by the Central Limit Theorem (condition for CLT is $n\sigma^2 \to \infty$). If $n\sigma^2$ does not go to infinity, we get Poisson behavior.

Recall from the last lecture that the we're interested in concentration inequalities because we want to know $\mathbb{P}(f(X) \ne Y)$ while we only observe $\frac{1}{n}\sum_{i=1}^n I(f(X_i) \ne Y_i)$. In Bernstein's inequality take $''X_i''$ to be $I(f(X_i) \ne Y_i)$. Then, since $2M = 1$, we get

$$\mathbb{E}I(f(X_i) \ne Y_i) - \frac{1}{n}\sum_{i=1}^n I(f(X_i) \ne Y_i) \le \sqrt{\frac{2\mathbb{P}(f(X_i) \ne Y_i)(1 - \mathbb{P}(f(X_i) \ne Y_i))u}{n}} + \frac{2u}{3n}$$

because $\mathbb{E}I(f(X_i) \ne Y_i) = \mathbb{P}(f(X_i) \ne Y_i) = \mathbb{E}I^2$ and therefore $\mathrm{Var}(I) = \sigma^2 = \mathbb{E}I^2 - (\mathbb{E}I)^2$. Thus,

$$\mathbb{P}(f(X_i) \ne Y_i) \le \frac{1}{n}\sum_{i=1}^n I(f(X_i) \ne Y_i) + \sqrt{\frac{2\mathbb{P}(f(X_i) \ne Y_i)\,u}{n}} + \frac{2u}{3n}$$

with probability at least $1 - e^{-u}$. When the training error is zero,

$$\mathbb{P}(f(X_i) \ne Y_i) \le \sqrt{\frac{2\mathbb{P}(f(X_i) \ne Y_i)\,u}{n}} + \frac{2u}{3n}.$$

If we forget about $2u/3n$ for a second, we obtain $\mathbb{P}(f(X_i) \ne Y_i)^2 \le 2\mathbb{P}(f(X_i) \ne Y_i)\,u/n$ and hence

$$\mathbb{P}(f(X_i) \ne Y_i) \le \frac{2u}{n}.$$

The above *zero-error rate*is better than $n^{-1/2}$ predicted by CLT.