Concentration bounds and applications

Huy L. Nguyễn

1 Chernoff bound

Consider independent random variables $X_1, X_2, \ldots, X_m \in \{0,1\}$ with $X = X_1 + X_2 + \cdots + X_m$ and $\mathbb{E}[X] = \mu$. For all $\delta > 0$, Chernoff bound (the upper tail) states that

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

Question: use Chernoff bound to show that with probability at least $1 - 1/n^2$, the max load is bounded by $O\left(\frac{\ln n}{\ln \ln n}\right)$.

There is also the lower tail version of Chernoff bound, which states that

$$\Pr[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\mu$$

Proof. Surprisingly we will prove this powerful inequality using Markov's inequality, abeit with a different variable. Consider the random variable e^{tX} where t is a constant that will be specified later. To apply Markov's inequality, we first need to compute the expectation.

$$\begin{aligned}\mathbb{E}[e^{tX}] &= \mathbb{E}\left[e^{t\sum_{i=1}^m X_i}\right] \\ &= \prod_{i=1}^m \mathbb{E}\left[e^{tX_i}\right]\end{aligned}$$

In the last inequality, we used the fact that X_i's are independent. Next we bound each term of the product. Suppose that X_i is 1 with probability p_i and 0 with probability $1 - p_i$. We have

$$\mathbb{E}\left[e^{tX_i}\right] = (1-p_i) + p_i e^t = 1 + p_i(e^t - 1) \leq \exp(p_i(e^t - 1))$$

In the inequality, we used the fact that $1 + x \leq e^x \ \forall x$. Multiplying all the bounds together we obtain

$$\mathbb{E}[e^{tX}] \leq \prod_{i=1}^m \exp(p_i(e^t - 1)) = \exp(\mu(e^t - 1))$$

For positive t, by Markov's inequality,

$$\Pr[X \geq (1+\delta)\mu] = \Pr[e^{tX} \geq \exp(t(1+\delta)\mu)] \leq \frac{\exp(\mu(e^t - 1))}{\exp(t(1+\delta)\mu)}$$

Now we choose t to obtain the strongest bound possible. That is, we choose t in order to minimize the right hand side. By taking derivative and setting it to 0, we obtain $t = \ln(1+\delta)$. The claim follows from substituting in this value of t.

To derive the lower tail version, we instead use a negative value for t. By Markov's inequality,

$$\Pr[X \leq (1-\delta)\mu] = \Pr[e^{tX} \geq \exp(t(1-\delta)\mu)] \leq \frac{\exp(\mu(e^t - 1))}{\exp(t(1+\delta)\mu)}$$

The lower tail bound now follows from setting $t = \ln(1-\delta)$. □

2 Better load balancing with more balls

We saw that with n bins and n balls, the expected load per bin is 1 but the maximum load is $\Theta(\log n / \log\log n)$. Now suppose we still have n bins but $n \log n$ balls. What do we get from the Chernoff bounds on the maximum load?

It turns out that in this case, the expected load of each bin is $\log n$ balls and with high probability $1 - 1/n^2$, the maximum load is bounded by $O(\log n)$.

This example shows that we can obtain significantly better load balancing with a lot of small tasks rather than fewer big tasks.

3 Polling

Let's use our tail bounds to analyze the performance of opinion polls. Suppose there are n people and we would like to estimate the fraction $p \in [0,1]$ of people with a certain opinion. We would like to sample t of them (with replacement) and poll their opinion. Let $\hat{p}$ be the fraction of the sampled people with the opinion. In order to guarantee that $|p - \hat{p}| \leq \varepsilon$ with probability $1-\delta$, how large do we need t to be?

The *expected* number of samples with the opinion is pt. The number of samples with the opinion is $\hat{p}t$. Using the Chernoff bounds, it turns out that we need $t = \Theta\left(\frac{1}{\varepsilon^2}\log(1/\delta)\right)$.