

SML Recitation Notes Week 1

Concentration of Measure:

Min Xu

January 13, 2011

1 Moments and C. of M.

Theorem 1 (Markov Inequality). *Let X be a non-negative random variable. Assume $\mathbb{E}X < \infty$. Then*

$$P(X > \lambda) \leq \frac{\mathbb{E}X}{\lambda}$$

Example 1. Let X_1, \dots, X_n be random variables. Assume:

1. they are I.I.D.
2. mean zero $\mathbb{E}[X_i] = 0$
3. bounded by one $|X_i| \leq 1$
4. same variance $Var(X_i) = \sigma^2$

Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then by Markov Inequality:

$$P(|\bar{X}_n| \geq \epsilon) \leq \frac{\mathbb{E}|\bar{X}_n|}{\epsilon} \tag{1.1}$$

$$P(|\bar{X}_n| \geq \epsilon) = P(|\bar{X}_n|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[|\bar{X}_n|^2]}{\epsilon^2} \tag{1.2}$$

$$P(|\bar{X}_n| \geq \epsilon) = P(|\bar{X}_n|^k \geq \epsilon^k) \leq \frac{\mathbb{E}[|\bar{X}_n|^k]}{\epsilon^k} \tag{1.3}$$

We want the upper bound $\frac{\mathbb{E}[|\bar{X}_n|^k]}{\epsilon^k}$ to be as small as possible. However, $\mathbb{E}|\bar{X}_n|^k \approx \frac{k^{k/2}}{n^{k/2}}$ (not easy to show) Thus, we want to minimize

$$\frac{\mathbb{E}[|\bar{X}_n|^k]}{\epsilon^k} \approx \frac{k^{k/2}}{\epsilon^k} \frac{1}{n^{k/2}}$$

First let us assume $\epsilon < 1$.

1. If we use higher moments (higher k), then $\frac{k^{k/2}}{\epsilon^k}$ is **bigger** ☹, but $\frac{1}{n^{k/2}}$ gets **smaller** ☺
2. As n gets larger, we want to use higher moment bounds
3. As ϵ gets larger, we want to use higher moment bounds

These observations tell us the interesting fact that no single moment-bound is good; we should choose the moment depending on n and ϵ . Wouldn't it be great to have a theorem that uses ALL the moments and automatically choose the right one for us? This is exactly Hoeffding Inequality, Bernstein Inequality, and anything derived using the Chernoff Trick.

Theorem 2. (Hoeffding Inequality) Using the same X_i 's as defined above and the fact that $b_i - a_i \leq 2$ and so $\frac{1}{n} \sum_{i=1}^n (b_i - a_i)^2 = 4$ for our example:

$$P(|\bar{X}_n| \geq \epsilon) \leq 2e^{-\frac{1}{2}n\epsilon^2}$$

Proof. (Sketch)

- Step 1: bound moment-generating function $\mathbb{E}e^{tX_i}$, which is equivalent to bounding $\mathbb{E}e^{t\bar{X}_n}$
- Step 2: (Chernoff Trick) Use Markov Inequality

$$P(|\bar{X}_n| \geq \epsilon) = P(e^{t\bar{X}_n} \geq e^{t\epsilon}) \leq \frac{\mathbb{E}e^{t\bar{X}_n}}{e^{t\epsilon}}$$

- Step 3: Apply bound from Step 1. Then use optimize t

□

To reveal the intuition behind the Chernoff Trick and the mysterious t parameter, we will use Taylor Expansion:

$$\frac{\mathbb{E}e^{t\bar{X}_n}}{e^{t\epsilon}} = \frac{1 + t\mathbb{E}\bar{X}_n + \frac{t^2\mathbb{E}[\bar{X}_n^2]}{2!} + \dots + \frac{t^k\mathbb{E}[\bar{X}_n^k]}{k!} + \dots}{1 + t\epsilon + \frac{t^2\epsilon^2}{2!} + \dots + \frac{t^k\epsilon^k}{k!} + \dots}$$

If t is large, then the Taylor Series is dominated by the higher-degree terms, this is like choosing a high moment. In fact, in the proof, the optimal t is chosen to be $t = \epsilon n$, confirming our observation that as ϵ or n increases, we should use higher moments.

2 Confidence-Interval form

How much better is the bound derived with Hoeffding than say the bound derived from Markov? We can put the concentration of measure inequalities in Confidence-Interval form to see. The following statements says with some probability at least $1 - \delta$, the sample mean is concentrated around the actual mean where the interval of concentration depends on n, δ .

Remark 1. (Probability-bound form vs. Confidence-Interval form)

- (First Moment) Let $c \equiv \mathbb{E}[|\bar{X}_n|]$

$$P(|\bar{X}_n| \geq \epsilon) \leq \frac{c}{\epsilon}$$

implies that with probability at least $1 - \delta$

$$|\bar{X}_n| \leq \frac{c}{\delta}$$

- (Second Moment) Let $\sigma^2 \equiv \mathbb{E}[|X_i|^2]$, then $\mathbb{E}[|\bar{X}_n|^2] = \frac{\sigma^2}{n}$

$$P(|\bar{X}_n| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

implies that with probability at least $1 - \delta$

$$|\bar{X}_n| \leq \sqrt{\frac{\sigma^2}{n\delta}}$$

- (Moment-Generating Function)

$$P(|\bar{X}_n| \geq \epsilon) \leq 2e^{-\frac{1}{2}n\epsilon^2}$$

implies that with probability at least $1 - \delta$

$$|\bar{X}_n| \leq \sqrt{\frac{2}{n} \log \frac{2}{\delta}}$$

Proof. We will prove only the second statement; the rest are proven in an identical manner.

Let $\delta > 0$. Suppose that $P(|\bar{X}_n| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} = \delta$, then, from the inequality $\frac{\sigma^2}{n\epsilon^2} = \delta$, we get that $\epsilon = \sqrt{\frac{\sigma^2}{n\delta}}$.

Thus, $P(|\bar{X}_n| \geq \sqrt{\frac{\sigma^2}{n\delta}}) \leq \delta$ and by elementary probability, $P(|\bar{X}_n| \leq \sqrt{\frac{\sigma^2}{n\delta}}) \geq 1 - \delta$

□

Let us compare intervals of size $\sqrt{\frac{\sigma^2}{n\delta}}$ and $\sqrt{\frac{2}{n} \log \frac{2}{\delta}}$. Suppose we want to be more confident; we want a confidence of $\frac{\delta}{2}$ instead of δ . The two new intervals become $\sqrt{2\frac{\sigma^2}{n\delta}}$ and $\sqrt{\frac{2}{n} \log \frac{2}{\delta} + \frac{2}{n} \log 2}$.

The interval we derived from second moment grew by a multiplicative factor of $\sqrt{2}$ while the interval we derived from Hoeffding grew by only an additive factor. Thus the exponentially-concentrated Hoeffding bound is much more powerful; it allows us to increase our confidence dramatically without blowing up the interval.

3 O_p and C. of M.

Concentration of measure sharpness has a curious relationship with O_p rates of convergence.

Definition 3. Let $\{X_n\}_{n=1,\dots}$ be a sequence of random variables. Then we say the sequence $X_n = O_p(1)$ if the X_n are uniformly concentrated, i.e.: For all $\epsilon > 0$, there exist M such that for all n , $P(|X_n| \geq M) \leq \epsilon$.

We say that sequence $X_n = O_p(r_n)$ if there exist a sequence of random variables $Y_n = O_p(1)$ such that $X_n = r_n Y_n$. That is, X_n is the scaled version of some $O_p(1)$ sequence. Equivalently, we could define $X_n = O_p(r_n)$ if $\frac{X_n}{r_n} = O_p(1)$.

We will analyze the convergence rate of the sequence of sample means: $\{\bar{X}_n\}_{n=1,\dots}$.

- The first moment confidence interval proves directly that $\bar{X}_n = O_p(1)$.

- The second moment confidence interval show that with probability at least $1 - \delta$, $|\sqrt{n}\bar{X}_n| \leq \sqrt{\frac{\sigma^2}{\delta^2}}$ and hence $\bar{X}_n = O_p(\frac{1}{\sqrt{n}})$.
- The Hoeffding confidence interval show that with probability at least $1 - \delta$, $|\sqrt{n}\bar{X}_n| \leq \sqrt{2 \log \frac{2}{\delta}}$ and hence $\bar{X}_n = O_p(\frac{1}{\sqrt{n}})$.

From these results, we can make several important observations:

1. Although Hoeffding concentration is much sharper than Second moment concentration, it did NOT give a better O_p convergence rate.
2. Higher moment concentrations will not give us better rate than $O_p(\frac{1}{\sqrt{n}})$ either.
3. The O_p rate depends on the relationship between ϵ and n in the probability-bound form of the concentration results. Hoeffding and Second moment both yielded the same O_p rate because the probability decreased with $n\epsilon^2$.

Thus, O_p rates doesn't tell the whole story. You should look at both the O_p rate and the sharpness of concentration when you evaluate theoretical properties of a statistical/machine learning method.