Lecture 3: CS395T Numerical Optimization for Graphics and AI — Probability

Qixing Huang
The University of Texas at Austin
huangqx@cs.utexas.edu

1 Basic

You need to be familiar with basics Probability theory. A good Wikipedia page for review is https://en.wikipedia.org/wiki/Probability. The focus of this lecture is on concentration inequalities.

2 Markov Inequality

Markov's inequality gives an upper bound for the probability that a non-negative function of a random variable is greater than or equal to some positive constant. It is named after the Russian mathematician Andrey Markov. Markov's inequality (and other similar inequalities) relate probabilities to expectations, and provide (frequently loose but still useful) bounds for the cumulative distribution function of a random variable.

Fact 2.1. *Formally speaking, if X is a non-negative random variable and choose $a > 0$, then the probability that X is no less than a is no greater than the expectation of X divided by a:*

$$P(X \geq a) \leq \frac{E[X]}{a}. \tag{1}$$

Proof: For any event E, let I_E be the indicator random variable of E, that is, $I_E = 1$ if E occurs and $I_E = 0$ otherwise.

Using this notation, we have $I(X \geq a) = 1$ if the event $X \geq a$ occurs, and $I(X \geq a) = 0$ if $X < a$. Then, given $a > 0$,

$$aI(X \geq a) \leq X,$$

which is clear if we consider the two possible values of $X \geq a$. If $X < a$, then $I(X \geq a) = 0$, and so $aI(X \geq a) = 0 \leq X$. Otherwise, we have $X \geq a$, for which $I(X \geq a) = 1$, and so $aI(X \geq a) = (a \leq X)$.

Since E is a monotonically increasing function, taking expectation of both sides of an inequality cannot reverse it. Therefore,

$$E(aI(X \geq a)) \leq E(X).$$

Now, using linearity of expectations, the left side of this inequality is the same as

$$aE(I(X \geq a)) = a(1 \cdot P(X \geq a) + 0 \cdot P(X < a)) = aP(X \geq a).$$

Thus we have

$$aP(X \geq a) \leq E(X).$$

and since $a > 0$, we can divide both sides by a. □

Markov's inequality has a version in the language of measure theory:

Fact 2.2. *In the language of measure theory, Markov's inequality states that if (X, Σ, μ) is a measure space, f is a measurable extended real-valued function, and $\epsilon > 0$, then*

$$\mu(\{x \in X : |f(x)| \geq \epsilon\}) \leq \frac{1}{\epsilon} \int_X |f| d\mu. \tag{2}$$

This measure theoretic definition is sometimes referred to as Chebyshev's inequality (Andrew Markov's teacher).

Markov's inequality is the foundation for deriving provable bounds. However, it is usually not applied on random variables directly, but rather transformations of random variables. For example, if we apply Markov's inequality to $(X - E[X])^2$, we obtain Chebyshev's inequality:

$$P(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}.$$

2.1 Power Moment

In this section, we are interested in derivating a concentration bound on Rademacher variables. Specifically, suppose we have n independent random variables $X_1, \cdots, X_n$. Each variable X_i takes 1 or -1 with equal probability. It is clear that

$$E(X_i) = 0, \quad E(X_i^2) = 1, \qquad 1 \leq i \leq n.$$

We want to estimate the value of

$$s = \sum_{i=1}^{n} X_i.$$

We look at what the so-called power moment offers.

The basic idea is to look at $E(s^{2k})$. The Markov's inequality then gives

$$P(|s| \geq a) \leq \frac{E(s^{2k})}{a^{2k}}. \tag{3}$$

As you will see later, we will get tighter and tighter bounds by varying k. Our goal is to look for a small upper bound on $P(|s| \geq a)$ when $a = O(\sqrt{n \log(n)})$.

When $k = 1$. We have

$$\begin{aligned} E(s^2) &= E\big((\sum_{i=1}^{n} X_i)^2\big) \\ &= n, \end{aligned} \tag{4}$$

where we have used the fact that for different i and j, X_i and X_j are independent, so $E(X_i X_j) = 0$. Applying (3), we have

$$P(|s| \geq c\sqrt{n \log(n)}) \leq \frac{1}{c^2 \log(n)}. \tag{5}$$

This is actually not bad. But let us look at what $k = 2$ offers. In fact,

$$\begin{aligned} E(s^4) &= E((\sum_{i=1}^{n} X_i)^4) \\ &= \sum_{1 \leq i_1, i_2, i_3, i_4 \leq n} E[X_{i_1} X_{i_2} X_{i_3} X_{i_4}]. \end{aligned} \tag{6}$$

In order for $E[X_{i_1} X_{i_2} X_{i_3} X_{i_4}]$ to be non-zero. There are two possibilities

- **Type I.** $i_1 = i_2 = i_3 = i_4$. There are n such possibilities, and the total contribution to $E(s^4)$ is n.

- **Type II.** $i_{s_1} = i_{t_1} \neq i_{s_2} = i_{t_2}, \{s_1, t_1, s_2, t_2\} = \{1, 2, 3, 4\}$. First of all, there are 3 such configurations. For each configuration, there are $n(n-1)$ different $E[X_{i_1}X_{i_2}X_{i_3}X_{i_4}]$, and the contribution of each term is 1. So the total contribution is $3n(n-1)$.

As a summary, we have

$$E(s^4) = 3n(n-1) + n.$$

This means

$$P(|s| \geq c\sqrt{n\log(n)}) \leq \frac{n(3n-2)}{c^4 n^2 \log^2(n)} \leq \frac{3}{c^4 \log^2(n)}. \tag{7}$$

(7) improves from (5) by a $\log(n)$ factor, this motivates us to look at bigger values of k. Before proceeding, we do a big relaxation by just counting Type 2 while allow $i_{s_1} = i_{s_2}$ when enumerating i_{s_1} and i_{s_2}. This will, however, count Type 1 multiple times (6 times when $k = 2$). Nevertheless, as we will see later, this relaxation will not incur any change in the order of the approximation.

Generally speaking, there are $\frac{(2k)!}{k!2^k}$ configurations of 2-pairs from $2k$ elements. The contribution of each configuration is upper bounded by n^k. Please be aware that different configurations multiple times. This can be easily understood via recursion. So we have

$$E(s^{2k}) \leq \frac{(2k)!}{k!2^k} n^k.$$

This means

$$P(|s| \geq c\sqrt{n\log(n)}) \leq \frac{(2k)!}{k! \cdot 2^k \cdot c^{2k} \cdot \log^k(n)}. \tag{8}$$

Using Stirling's approximation $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$, we have

$$P(|s| \geq c\sqrt{n\log(n)}) \leq \sqrt{2}\big(\frac{2k}{ec^2\log(n)}\big)^k. \tag{9}$$

Let $k = c_2\log(n)$, we have

$$P(|s| \geq c\sqrt{n\log(n)}) \leq \big(\frac{2c_2}{ec^2}\big)^{c_2\log(n)} = n^{c_2\log(\frac{2c_2}{ec^2})}. \tag{10}$$

We can optimize c_2 to minimize the right-hand side of (10), the optimal c_2 is given by

$$c_2 = \frac{ec^2}{2} e^{-\frac{ec^2}{2}}.$$

In other words, we have

$$P(|s| \geq c\sqrt{n\log(n)}) \leq n^{-\frac{c^2}{2}}.$$

2.2 Exponential Moment

We will cover Pages 12-16 of `https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`.