MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.265/15.070J Lecture 7 — Mar 6, SP17

Lecturer: Yury Polyanskiy — Scribe notes by Younhun Kim

Concentration of Measure

Content.

1. Introduction
2. Precursor to C.o.M. : Large Deviations Principle
3. Hoeffding's / Azuma-Hoeffding Lemma
4. Subgaussian Random Variables

1 Introduction

1.1 An analogy

To help guide our intuition, here are a few different areas of math and general meta-principles whose relationships are all analogous to each other.

- **Discrete Probability**: If Z is a RV which depends "nicely" on a bunch of independents, then

$$\mathbb{P}[|Z - \mathbb{E}Z| > t] \lesssim e^{-t^2/2\nu}$$

  where $\nu \approx \mathrm{var} Z$.

- **Geometry**: High-Dimensional spaces under product measures "look star-shaped", with tendril lengths $\sim n$ and core size $\sim \sqrt{n}$.

- **Functional Analysis**:

$$\int f^2 \log f d\nu^n \lesssim \int \|\nabla f\|^2$$

1.2 Longest Common Subsequence

Let $LCS(a, b)$ denote the length of the *longest common subsequence* of two strings a and b.
(*Note:* This is not to be confused with the similarly defined *longest common substring*. For example, the longest common *substring* of `01010` and `00001` is `01`, but the longest common *subsequence* is `001` or `000` since it allows characters to be skipped in a and b.)

Consider $X_i \sim Ber(1/2)$ iid and $Y_i \sim Ber(1/2)$ iid. Is there something we can say about the expected value of LCS, normalized by n? On one hand, using superadditivity (ie. $\mathbb{E}[X^{m+n}, Y^{m+n}] \geq \mathbb{E}[X^m, Y^m] + \mathbb{E}[X^n, Y^n]$), it is not too much work to realize that there is a constant ν such that

$$\mathbb{E}\left[\tfrac{1}{n} LCS(X^n, Y^n)\right] \to \nu$$

Historically, it has been difficult to compute ν precisely or to compute the distribution of LCS, but we can still compute bounds on the tail probabilities. (For example, we might want to try and estimate ν using a sampling method, for which we would like to determine how many samples we might need.) One way to do this is to apply Chebyshev's inequality by bounding the variance by $1/n$, which results in

$$\mathbb{P}\left[\left|\frac{1}{n} LCS - \mu\right| > t\right] < \frac{1}{nt^2}$$

More generally, using the ideas from this section, we will be able to find exponential bounds such as

$$\mathbb{P}\left[\left|\frac{1}{n} LCS - \mu\right| > t\right] \leq 2e^{-nt^2/2}$$

2 Precursor to Concentration of Measure: Large Deviations Principle

Theorem 1. *Let $X_i \sim P_X$ iid taking (discrete) values in $\mathbb{R}$. If $t > \mathbb{E}X$, then*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i > t\right] = \exp(-nE(t) + o(n))$$

and if $t < \mathbb{E}X$, then

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i < t\right] = \exp(-nE(t) + o(n))$$

where $E(t) = \sup_\lambda \lambda t - \Psi_X(\lambda)$ and $\Psi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}]$, the log MGF.

Proof. (We only prove the theorem for the case $t > \mathbb{E}X$. The other is analogous.)

Without loss of generality, let $P_X = \sum_{i=1}^{m} p_i \delta_{a_i}$; i.e. X takes value a_i with probability p_i.
($\leq$): Use the Chernoff bound

$$\mathbb{P}\left[\frac{1}{n}\sum X_i > t\right] \leq \mathbb{E}[\exp(n(\Psi_X(\lambda) - \lambda t))]$$

by minimizing over all λ, we get the upper bound $e^{-nE(t)}$.
($\geq$): The technique here is referred to as the "method of types". We introduce a few notations:

- As shorthand for the n-tuple $(x_1, \ldots, x_n)$, we write $x^{(n)}$.
- $\hat{p}$ denotes the distribution on $[n]$ with rational probabilities such that $n\hat{p}_i \in \mathbb{Z}$ for all $i \in [m]$. This distribution is called an **n-type**.
- For all n-types $\hat{p}$, let $T_{\hat{p}}$ denote the set of all outcomes of size n whose empirical distribution is exactly $\hat{p}$:

$$T_{\hat{p}} = \{x^{(n)} : \#\{i : x_i = a_j\} = n\hat{p}_j\}$$

From the above, we see that

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum X_i > t\right] &= \sum_{\text{n-types } \hat{p}:\sum \hat{p}_i q_i > t} \mathbb{P}[X^{(n)} \in T_{\hat{p}}] \\ &\geq \mathbb{P}[X^{(n)} \in T_{\hat{p}^*}]\end{aligned}$$

where $\hat{p}^*$ is the argmax of the probabilities in the summand. However, $T_{\hat{p}}$ is permutation-invariant, so in fact

$$\begin{aligned}\mathbb{P}[X^{(n)} \in T_{\hat{p}}] &= \binom{n}{n_1, \ldots, n_m} p_1^{n_1} \cdots p_m^{n_m} \\ &= \exp\left(nH(\hat{p}) + o(n) + n\sum \hat{p}_i \ln p_i\right)\end{aligned}$$

In the second line in the above calculations, we made use of Stirling's approximation, and the definition of entropy $H(\pi) = \sum \pi_i \ln \frac{1}{\pi_i}$.

At this point we are basically done. We compute $\mathbb{P}[X^{(n)} \in T_{\hat{p}^*}]$ by maximizing the above over all $\hat{p}$, subject to $\sum \hat{p}_i q_i > t$. This can be accomplished using Lagrange multipliers, which yields $-E(t)$. (The computation is omitted.) $\square$

There are a few problems with this result, which we will address shortly.

1. $E(t)$ is hard to evaluate.
2. What if (X_i) are independent, but not identically distributed?
3. What if they are not even independent?

(1) and (2) will be resolved by Hoeffding's lemma, and (3) will be addressed by another, related lemma called Azuma-Hoeffding.

3 Hoeffding and Azuma-Hoeffding Lemmas

Lemma 2 (Hoeffding's Lemma)**.** *If $a \leq X \leq b$ almost surely, then*

$$\Psi_X''(\lambda) \leq \left(\frac{b-a}{2}\right)^2$$

An immediate, useful corollary is

Corollary 3. *If (X_i) are independent and $a_i \leq X_i \leq b_i$ for all i, then*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}X_i\right| > t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right)$$

Here, we present the proof of both.

Proof of Lemma 2.

$$\Psi'(x) = \frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}}$$

$$\begin{aligned}
\Rightarrow \Psi''(x) &= \frac{\mathbb{E}X^2e^{\lambda X}}{\mathbb{E}e^{\lambda X}} - \left(\frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}}\right)^2 \\
&= \operatorname{var}(X_\lambda) \\
&\leq \left(\frac{b-a}{2}\right)^2
\end{aligned}$$

where $X_\lambda \sim P_X(dx)\frac{e^{\lambda x}}{\mathbb{E}e^{\lambda x}}$ is the specified exponential tilt of X. □

Proof of Corollary 3. Without loss of generality, assume $\mathbb{E}X_i = 0$ for all i. Observe that

$$\Psi_{X_1+\cdots X_n}(\lambda) = \Psi_{X_1}(\lambda) + \cdots + \Psi_{X_n}(\lambda)$$

By taylor expanding each individial Ψ, we see that

$$\Psi_{X_i}(\lambda) \leq \frac{1}{2}\left(\frac{b_i - a_i}{2}\right)^2 \lambda^2$$

$$\Rightarrow \Psi_{\sum X_i} \leq \frac{\lambda^2}{2}\sum_i \left(\frac{b_i - a_i}{2}\right)^2$$

Applying Chernoff's bound completes the proof. □

Theorem 4 (Azuma-Hoeffding)**.** *Let $(S_k)_{k=1,\ldots,n}$ be a martingale with respect to the filtration $(\mathcal{F}_k)_{k=1,\ldots,n}$, with $|S_k - S_{k-1}| \leq c_k$ for all k. Then*

$$\mathbb{P}\left[|S_n - \mathbb{E}S_n| > t\right] \leq \exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right)$$

Proof. Without loss of generality, assume $S_0 = 0$. Let M be the martingale difference of S: $M_i = S_i - S_{i-1}$. (Note that $S_n = \sum M_i$.)

By hypothesis, we know that $|M_i| \leq c_i$ for all i, and $\mathbb{E}[M_i \mid \mathcal{F}_{i-1}] = 0$ by definition of a MG. Observe that

$$\mathbb{E}[e^{\lambda S_n} \mid \mathcal{F}_{n-1}] = e^{\lambda S_{n-1}}\mathbb{E}[e^{\lambda M_n} \mid \mathcal{F}_{n-1}]$$

so to get a bound on the conditional expectation of $e^{\lambda S_n}$, it suffices to get a bound on that of $e^{\lambda M_n}$. This is straightforward using a linear interpolation: since $\exp$ is convex, the function $e^\lambda x$ is upper bounded by $(x+c)\cdot\frac{e^{\lambda c}-e^{-\lambda c}}{2c} + e^{-\lambda c}$ in the region $x \in [-c, c]$.

$$\begin{aligned}
\mathbb{E}[e^{\lambda M_n} \mid \mathcal{F}_{n-1}] &\leq (\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] + c_n)\cdot\left(\frac{e^{\lambda c_n} - e^{-\lambda c_n}}{2c_n}\right) + e^{-\lambda c_n} \\
&= \frac{e^{\lambda c_n} + e^{-\lambda c_n}}{2} \\
&= \sum_{m\geq 0}\frac{(\lambda c_n)^{2m}}{(2m)!} \\
&\leq \sum_{m\geq 0}\frac{(\lambda^2 c_n^2/2)^m}{m!} = e^{\lambda^2 c_n^2/2}
\end{aligned}$$

and therefore $\mathbb{E}[e^{\lambda S_n} \mid \mathcal{F}_{n-1}] \leq e^{\lambda S_{n-1}} e^{\lambda^2 c_n^2/2}$. Applying this argument inductively yields the bound

$$\mathbb{E}e^{\lambda S_n} \leq e^{\frac{1}{2}\lambda^2 \sum c_i^2}$$

From here, using a Chernoff bound completes the proof. □

4 Sub-gaussian Random Variables

Frequently, we have been proving bounds in the form

$$\mathbb{P}[|Z - \mathbb{E}Z| > t] \lesssim e^{-t^2/2\nu}$$

The following definition generalizes this.

Definition 1. *A random variable Z is (b, ν)-subgaussian if for all $t > 0$, it satisfies*

$$\mathbb{P}[|Z - \mathbb{E}Z| > t] \leq be^{-t^2/2\nu}$$

The parameter ν here should be thought of as a placeholder for the variance of some gaussian distribution. Here is an easy property we can prove right away:

Proposition 5. *If Z is (b, ν)-subgaussian, then* $\mathrm{var}Z \leq 2b\nu$.

Proof. Without loss of generality, assume $\mathbb{E}Z = 0$. The derivation is straightforward:

$$\begin{aligned} \mathrm{var}Z = \mathbb{E}Z^2 &= \int_0^\infty \mathbb{P}[Z^2 > s]ds \\ &\leq \int_0^\infty be^{-s/2\nu}ds = 2b\nu \end{aligned}$$

□

In the context of Homework 1, this puts a bound on how different the mean and the median can be: $|\mathbb{E}Z - \mathrm{M}Z| \leq \sqrt{\mathrm{var}Z} = \sqrt{2\nu b}$.

In fact, there are many ways to characterize a subgaussian random variable. In fact, depending on the literature being referenced, one may see any of the following properties used in the definition of a subgaussian RV.

Proposition 6. *The following are equivalent, up to absolute constants that relate b, b' and ν, ν' in each of the statements separately.*

(1) Z is a (b, ν)-subgaussian RV.

(2) $\Psi_Z(\lambda) \leq \lambda \mathbb{E}Z + \frac{\nu'}{2}\lambda^2 + \ln b'$

(3) $\mathbb{E}[\exp(\frac{1}{2\nu'}(Z - \mathbb{E}Z)^2)] \leq b'$

(1') $\mathbb{P}[|Z - MZ|] \leq b' \exp\left(-\frac{t^2}{2\nu}\right)$

(2') $\Psi_Z(\lambda) \leq \lambda MZ + \frac{\nu'}{2}\lambda^2 + \ln b'$

(3') $\mathbb{E}[\exp(\frac{1}{2\nu'}(Z - MZ)^2)] \leq b'$

We sketch a proof for the equivalence of some of the above. The rest aren't too much more work.

Proof. $(2 \Rightarrow 1)$ with $b = 2b', \nu = \nu'$.

$$\begin{aligned}\mathbb{P}[|Z - \mathbb{E}Z| > t] &\leq 2\exp(\Psi_x(\lambda) - \lambda t) \\ &\leq 2b' \exp(\lambda(\mathbb{E}Z - t) + \nu'\lambda^2/2)\end{aligned}$$

WLOG assume $\mathbb{E}Z = 0$. Also, take $\lambda = t/\nu'$ to minimize the right hand side.

$$= 2b' \exp(-\lambda t + \nu' t^2/2) = 2b' e^{-t^2/2\nu'}$$

$(1 \Rightarrow 2)$ with $b' = b\sqrt{2\pi}$, and $\nu' = 2\nu$.

$$\begin{aligned}\mathbb{E}e^{\lambda Z} &= \int_0^\infty \mathbb{P}[e^{\lambda Z} > s]ds \\ &= \int_0^\infty \mathbb{P}[Z > \tfrac{1}{\lambda}\ln S]ds \\ &= \int_{-\infty}^\infty e^u \mathbb{P}[z > u/\lambda]du \\ &= b\sqrt{2\pi\lambda^2\nu}e^{\lambda^2\nu/2} \leq b\sqrt{2\pi}e^{\lambda^2\nu}\end{aligned}$$

$(3 \Rightarrow 1)$ with $b = b'$ and $\nu = \nu'$: The idea is to bound $\mathbb{P}[|Z - \mathbb{E}Z| > t] = \mathbb{P}[e^{(Z-\mathbb{E}Z)^2} > e^{t^2}] \leq b' \exp(-t^2/2\nu')$ via Chebyshev.
$(1 \Rightarrow 3)$ with $b = b'$ and $\nu = \nu'$. This is shown similarly to $(1 \Rightarrow 2)$: just bound $\int \mathbb{P}[e^{\frac{1}{2\nu'}(Z-\mathbb{E}Z)^2} > s]ds$.

Finally, to convert to $1'), 2'), 3')$ we use $|\mathbb{E}Z - \mathrm{M}Z| \leq \sqrt{2\nu b}$ and after simple algebra we get, for example, that $(1 \Rightarrow 1')$ with $b' = be^b$ and $\nu' = 2\nu$ and, similarly, $(1' \Rightarrow 1)$ with $b = b'e^{b'}$ and $\nu = 2\nu'$. The equivalence of all $1', 2', 3'$ is shown as above. □