MTH 995-003: Intro to CS and Big Data	Spring 2014

Lecture 13 — Feb 18th, 2014

Inst. Mark Iwen *Scribe: Ruochuan Zhang*

1 A Concentration Inequality for Subgaussians

Recall that the following fact was critical to our analysis of the LSH function we considered for Euclidean distance:

- If $\vec{g} \sim N(0, I_{D\times D})$, then $< \vec{g}, \vec{x} > \sim N(0, \|\vec{x}\|_2^2)$.

We can finally now get a similar result for any subgaussian random vector!

Theorem 1 (Stability of Subgaussians)**.** *Let $\mathbb{X}_1, ..., \mathbb{X}_m$ be i.i.d. mean zero subgaussian random variables. Let $\vec{a} \in \mathbb{R}^m$, and define $Z := \sum_{l=1}^m a_l \mathbb{X}_l$. Then Z is also subgaussian. More specifically, if $\mathbb{E}\left[\exp(\theta \mathbb{X}_l)\right] \leq \exp(c\theta^2)$, $\forall \theta, l$, then*
(i) $\mathbb{E}\left[\exp(\theta Z)\right] \leq \exp(c\|\vec{a}\|_2^2\theta^2)$, and
(ii) $\mathbb{P}\left[|Z| \geq t\right] \leq 2\exp\left(\frac{-t^2}{4c\|\vec{a}\|_2^2}\right)$, $\forall t > 0$.

Proof. For part (i)

$$\begin{aligned}
\mathbb{E}\left[\exp\left[\theta \sum_{l=1}^m a_l \mathbb{X}_l\right]\right] &= \prod_{l=1}^m \mathbb{E}\left[\exp(\theta a_l \mathbb{X}_l)\right] \quad (\textit{by independence of } \mathbb{X}_l\ 's) \\
&\leq \prod_{l=1}^m \exp(c\theta^2 a_l^2) \quad (\textit{by Thm } 1,\ \textit{part } 1\ \textit{of Lecture } 12) \\
&= \exp(c\|\vec{a}\|_2^2\theta^2)
\end{aligned}$$

Part(ii) follows from Thm1, part 2 from Lecture 12.

□

Definition 1. *A subgaussian random variable has parameter c if $\mathbb{E}\left[exp(\mathbb{X}\theta)\right] \leq exp(\theta^2 c)$, $\quad \forall \theta \in \mathbb{R}$.*

Lemma 1. *Let $\vec{Y}$ be a random vector with i.i.d. subgaussian entries, all with parameter c, mean 0, and variance 1. Then*
(i) $\mathbb{E}\left[|\left\langle \vec{Y}, \vec{x} \right\rangle|^2\right] = \|\vec{x}\|_2^2$, $\quad \forall \vec{x} \in \mathbb{R}^N$, and
(ii) $\left\langle \vec{Y}, \frac{\vec{x}}{\|\vec{x}\|} \right\rangle$ is a subgaussian random variable with $\mathbb{E}\left[exp(\theta \left\langle \vec{Y}, \frac{\vec{x}}{\|\vec{x}\|} \right\rangle)\right] \leq \exp(c\theta^2)$, $\forall \theta \in \mathbb{R}$.

Proof. Part (i) by a now-familiar calculation.
Part (ii) by Thm 1 above.

□

We can now prove the same type of concentration inequality for subgaussians that we had for gaussians (recall Lemma 1 in Lecture 10).

Theorem 2. *Let $A \in \mathbb{R}^{m\times N}$ be a matrix with i.i.d. mean zero, variance 1, subgaussian entries (each with parameter c). Then $\forall \vec{x} \in \mathbb{R}^N$ and $t \in (0,1)$*

$$\mathbb{P}\left[\left|\frac{1}{m}\|A\vec{x}\|_2^2 - \|\vec{x}\|_2^2\right| \geq t\|\vec{x}\|_2^2\right] \leq 2\exp(-\tilde{c}t^2 m)$$

where $\tilde{c} \in \mathbb{R}^+$ depends only on c. (e.g. $\tilde{c} = \frac{1}{(16c+1)8c}$ works)

Proof. Let $\vec{Y}_1, ..., \vec{Y}_m \in \mathbb{R}^N$ be the rows of $A \in \mathbb{R}^{m\times N}$, and set $Z_l := |\langle Y_l, \vec{x}\rangle|^2 - \|\vec{x}\|_2^2, \quad l \in [m]$.

- Note: $\mathbb{E}[Z_l] = 0$ by Lemma 1, (i).
- Also $\frac{1}{||\vec{x}||_2^2} Z_l$ is subexponential with $\beta = 2$, $K = \frac{1}{4c}$. Here's why:
  By Lemma 1, (ii), $\left\langle \vec{Y}, \frac{\vec{x}}{||\vec{x}||}\right\rangle$ is subgaussian with parameter c. Thus, Thm 1, part(2) from Lecture 12 implies that $\mathbb{E}[< \vec{Y}, \frac{\vec{x}}{||\vec{x}||} >] = 0$ and subgaussian with $\beta = 2$ and $\kappa = \frac{1}{4c}$.
  Thus, $\mathbb{P}[| < \vec{Y}_l, \frac{\vec{x}}{||\vec{x}||} > |^2 \geq r^2] \leq \beta e^{-kr^2}, \quad \forall r \in \mathbb{R}^+ \Rightarrow \frac{1}{||\vec{x}||_2^2} Z_l$ is subexponential with the same κ and β.

We are now able to see that the event we care about is

$$\begin{aligned}\frac{1}{||\vec{x}||_2^2}(m^{-1}||A\vec{x}||_2^2 - ||\vec{x}||_2^2) &= \frac{1}{m}\sum_{l=1}^{m} \frac{|< \vec{Y}_l, \vec{x} >|^2 - ||\vec{x}||_2^2}{||\vec{x}||_2^2} \\ &= \frac{1}{m}\sum_{l=1}^{m} \frac{Z_l}{||\vec{x}||_2^2}\end{aligned}$$

Furthermore, Bernstein's inequality for subexponential random variables (Thm 3, Lecture 11) implies that

$$\begin{aligned}\mathbb{P}\left[\frac{1}{m||\vec{x}||_2^2}\left|\sum_{l=1}^{m} Z_l\right| \geq t\right] &= \mathbb{P}\left[\left|\sum_{l=1}^{m}\frac{Z_l}{||\vec{x}||_2^2}\right| \geq mt\right] \\ &\leq 2\exp\left(\frac{-mt^2\kappa^2}{4\beta + 2kt}\right) \\ &\leq 2\exp\left(\frac{-\kappa^2}{4\beta + 2\kappa}(mt^2)\right) \quad \textit{since } t \in (0,1).\end{aligned}$$

□