1 Overview

In the last lecture we discussed LSH approach, and its runtime. In this lecture we will recall LSH
and introduce Large Deviation Inequalities for related matrices.

2 What is known about LSH for $\ell_p$-norms?

Recall that the runtimes that we could set all depends on $\rho := \frac{\log p_1}{\log p_2}$, where $p_1 > p_2$. For a good
LSH function, we want $\rho$ small.

Theorem 1 (See [1]). Let $p \in (0, 2], \delta, c \in (1, \infty)$, and $r \in \mathbb{R}^+$. There exists a LSH function $h:\mathbb{R}^D \rightarrow \mathbb{Z}$, w.r.t. $d(\vec{x}, \vec{y}) := ||\vec{x} - \vec{y}||_p$, with $\rho = \frac{\log p_1}{\log p_2} \leq \delta \cdot \max\{\frac{1}{c^p}, \frac{1}{c}\}$.

For $p = 2$ (Euclidean case), we showed how to do this with Gaussian random vectors.

Theorem 2 (See [2]). There exists an LSH function w.r.t. $l_2$-distance, and for all $r \in \mathbb{R}^+, c \in (1, \infty)$, that has $\rho = \frac{1}{c^2} + O\left(\frac{\log \log |X|}{\log \frac{1}{3}|X|}\right)$. (Here $X \subset \mathbb{R}^D$ is the arbitrary finite set we are hashing.)

Theorem 3 (See [3]). For large $D$ (i.e. in the limit), there exists $r, p_2$ for which $\rho \geq \frac{0.462}{c^p}$, for
any LSH funtion, w.r.t. any $l_p$-norm, for all $c, p \geq 1$.

3 Large Deviation Bounds Related to LSH

3.1 Problem

Given $\vec{g} \sim N(0, I_{D \times D})$, and $\vec{x} \in \mathbb{R}_D$, show that

$$\mathbb{P} \left[ |\vec{g}, \vec{x}|^2 - \|\vec{x}\|_2^2 \geq t\|\vec{x}\|_2^2 \right] \text{ is small in } t \right. \tag{1}$$

For LSH, we had computations involving $<\vec{g}, \vec{x}>$ for $\vec{x} \in \mathbb{R}_D, \vec{g} \in N(0, I_{D \times D})$, since $h(\vec{x}) = \lfloor \frac{<\vec{g}, \vec{x}> + u}{w} \rfloor$. LSH worked for $\ell_2$ exactly because this hash function sent vectors to buckets $\approx$ equal
to their length with high probability!
3.2 Discussion

Two very nice things happened that let us set our LSH function work for $\ell_2$:

1: $\langle \vec{x}, \vec{g} \rangle \sim N(0, ||\vec{x}||^2_2)$ because Gaussians are stable (i.e., when we add two Gaussians we get another one).

2: The bound (Eq. 1) held because the inner product was another Gaussian. This meant for LSH that vectors were hashed to $\approx$ their length (modulo $w$).

We are now going to generalize Equation 1 a little bit, and consider what happens if we take several gaussian measurements of a vector $\vec{x}$.

If $X \sim N(0,1)$, then $X^2 \sim \chi^2_1$ (chi-square r.v. with 1 degree of freedom).

Suppose that we have $D \chi^2_1$ (i.i.d.) $Y_1, \ldots, Y_D$, let $a \in \mathbb{R}^+$, $Z = \sum_{j=1}^D a Y_j$. Note that $Z \sim \chi^2_D$, with $D$ degrees of freedom. The moment generating function (MGF) for $Z$ is $\mathbb{E}[e^{uZ}] = (1 - 2u)^{-D/2}$, for all $u \in (-\infty, \frac{1}{2})$, and $\mathbb{E}[Z] = D$.

$$P[|Z - D| \geq \frac{t}{a}] = P[Z \geq D(1 + \frac{t}{Da})] + P[Z \leq D(1 - \frac{t}{Da})].$$

Note that,

$$P \left[ \left( 1 - \frac{t}{Da} \right) D \geq Z \right] = P \left[ e^{(1 - \frac{t}{Da})Du - uZ} \geq 1 \right]$$

$$\leq e^{(1 - \frac{t}{Da})Du} \mathbb{E} \left[ e^{-uZ} \right] \quad \text{(by the Markov Inequality)}$$

$$= e^{(1 - \frac{t}{Da})Du} (1 + 2u)^{-D/2}.$$

Similarly,

$$P[(1 + \frac{t}{Da})D \leq Z] \leq e^{-(1 + \frac{t}{Da})Du}(1 - 2u)^{-D/2},$$

So,

$$P[|Z - D| \geq t/a] \leq e^{-(1 + \frac{t}{Da})Du}(1 - 2u)^{-D/2} + e^{(1-t/Da)D\tilde{u}}(1 + 2\tilde{u})^{-D/2} \quad (2)$$

holds for any $u < 1/2$, and $\tilde{u} > -1/2$.

Define $f(u) := e^{-(1 + \frac{t}{Da})Du}(1 - 2u)^{-D/2}$, and $g(\tilde{u}) := e^{(1-t/Da)D\tilde{u}}(1 + 2\tilde{u})^{-D/2}$.

Optimize the choices of $u$ and $\tilde{u}$ by minimizing

$$\ln(f(u)) := - \left( 1 + \frac{t}{Da} \right) Du - \frac{D}{2} \ln(1 - 2u)$$

$$\ln(g(\tilde{u})) := \left( 1 - \frac{t}{Da} \right) Du - \frac{D}{2} \ln(1 + 2\tilde{u})$$

It is calculated that the following values minimize each of these:

$$u_{\text{min}} = \frac{t/(Da)}{2(1 + t/(Da))}, \quad \tilde{u}_{\text{min}} = \frac{t/(Da)}{2(1 - t/(Da))}, \quad (3)$$
Plugging these values of $u_{\min}$ and $\tilde{u}_{\min}$ back into (2) we see that

\[ P[|z - D| \geq t/a] \leq e^{-t^2/4Da^2} + e^{-3t^2 + 2t^3/(Da)} , \]

for all $t, a \in \mathbb{R}^+, D \in \mathbb{N}$.

We have basically proven the following,

**Lemma 1.** Let $G \in \mathbb{R}^{m \times D}$ be a random matrix with i.i.d. $N(0,1)$ entries, and $\vec{x} \in \mathbb{R}^D$, then

\[ P[|m^{-1}||G\vec{x}|^2 - ||\vec{x}||^2| \geq t||\vec{x}||^2] \leq e^{-t^2m/4} + e^{-3t^2 + 2t^3/m} , \]

Proof: $||G\vec{x}||^2 \sim ||\vec{x}||^2 \cdot \chi_m^2$, so that, $P[|m^{-1}||G\vec{x}||^2 - ||\vec{x}||^2| \geq t||\vec{x}||^2] = P[Z - m| \geq tm]$, where $Z \sim \chi_m^2$. The work above (see Equation (4)) now gives us the result when we set $a = 1/m$, $D = m$.

Note that $m = 1$ above is exactly the case of (1) related to LSH.

**References**

