CS 395T: Sublinear Algorithms, Fall 2020 October 6th, 2020

Lecture 12: More subgamma properties; Coverings

Prof. Eric Price *Scribe: Lucas Gretta, Nikos Mouzakis*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the last lecture we introduced subgaussian and subgamma random variables.

In this lecture we prove certain properties and tools related to subgamma variables, and use these tools to prove Distributional Johnson–Lindenstrauss as well as introduce covering numbers.

2 Subgamma variables

2.1 Definition

Definition 1. *We say that X is subgamma(σ^2, c) if the following holds:*

1. $E[e^{\lambda(x-E[x])} \leq e^{\frac{\lambda^2\sigma^2}{2}}, \forall|\lambda| \leq \frac{1}{c}$

This implies the following properties:

2. $Pr[|x-E[x]| \geq t] \leq 2e^{-\min(\frac{t^2}{2\sigma^2}, \frac{t}{2c})}$
3. *With probability 1-δ:* $|x-E[x]| \leq \sigma\sqrt{2\ln\frac{2}{\delta}} + 2c\ln(\frac{2}{\delta})$

We can show that 2 and 3 imply 1, with $\sigma^2 \to O(\sigma^2+c^2)$ and $c \to O(c)$

2.2 Properties

We can show that the sum of independent subgamma variables is also subgamma.

Lemma 2. *If X, Y are independent subgamma variables with (σ_x^2, c_x) and (σ_y^2, c_y) respectively, then $X+Y$ is also subgamma($\sigma_x^2+\sigma_y^2, \max(c_x, c_y)$).*

We can also show that a scaling of a subgamma variable is also subgamma.

Lemma 3. *If X is subgamma(σ^2, c) and a is a constant, then aX is subgamma($a^2\sigma^2, ac$).*

2.3 Generating subgamma variables

An easy way to generate subgamma variable is by squaring Gaussians.

Claim 4. *If $X \sim N(0,1)$ then $X^2 \in subgamma(4,4)$*

By using Lemma 3 we can show that the following:

Claim 5. *If $X \sim N(0,\sigma^2)$ then $X^2 \in subgamma(4\sigma^4, 4\sigma^2)$*

This can be generalized to squaring subgaussians.

Lemma 6. *If $X \in subgaussian(\sigma^2)$ with mean 0, then $X^2 \sim subgamma(O(\sigma^4), O(\sigma^2))$*

2.4 Bernstein Bounds

Theorem 7. *If $|X - \mu| \leq n$ with probability 1 and $Var(X) = \sigma^2$ then $X \in subgamma(2\sigma^2, 2n)$*

Compare this with the guarantees given by Hoeffding bounds for the case of a Bernoulli variable in $X \in \{0,1\}$ and $Pr[x] = p$.

- *Hoeffding: $X \in subgamma(1/4, 0)$*
- *Bernstein: $X \in subgamma(2p, 2)$*

Example: For the sum of variables $X_i \in \{0,1\}, X = \sum X_i, Pr[X_i] = p_i$ we have that $E[X] = \sum p_i = \mu$ and $Var(X) = \sum p_i(1-p_i) \leq \mu$.

Using Hoeffding we get that $X \in subgaussian(\frac{n}{4})$ so we get that with probability 1-δ:

$$|X - \mu| \leq \sqrt{n \log(\frac{2}{\delta})}$$

However, if the p_i are small then (e.g. $n = 10^6$, $p_i = 10^{-4}$, $\mu = 100$) then we get a (1-δ)-guarantee of $100 \pm 1000\sqrt{\log(\frac{2}{\delta})}$, which is quite bad in terms of the constant.

On the other hand, if we used Chebyshev we would get an (1-δ)-guarantee of $\mu \pm \sqrt{\frac{Var(X)}{\delta}} = 100 \pm \frac{10}{\sqrt{\delta}}$. This has a much better constant, but the dependence on δ is much worse.

To get the best of both worlds we use a Bernstein type bound. Since each $X_i \in subgamma(2p_i, 2)$ then $X \in subgamma(2\mu, 2)$ by Lemma 2.

So by the third property of Definition 2.1, with probability $1 - \delta$:

$$|X - \mu| \leq 2\sqrt{\mu \ln(\frac{2}{\delta})} + 4\ln(\frac{2}{\delta}) = 20\sqrt{\ln(\frac{2}{\delta})} + 4\ln(\frac{2}{\delta})$$

, which has both a small constant and a good dependence on δ.

Equivalently, by the second property of Definition 2.1:

$$Pr[|X-\mu| \geq \epsilon\mu] \leq 2e^{-\min(\frac{\epsilon^2\mu^2}{4\mu},\frac{\epsilon\mu}{4})} = 2e^{-\frac{\mu}{4}\min(\epsilon^2,\epsilon)}$$

This is also known as a **multiplicative Chernoff bound**.

Essentially, it shows that X is concentrated like a Gaussian in $[0, 2\mu]$, as the Central Limit Theorem prescribes, but like an exponential for its tail (i.e. $x \geq 2\mu$).

3 Distributional Johnson-Lindenstrauss

We can use these tools to prove the Distributional Johnson-Lindenstrauss lemma.

Recall the Distributional Johnson-Lindenstrauss lemma.

Theorem 8. *If $A \in \mathbb{R}^{m\times n}$ with i.i.d. subgaussian($\frac{1}{m}$) entries with mean 0, then:*

$$\forall x, Pr[||Ax||_2^2 \notin (1\pm\epsilon)||x||_2^2] \leq e^{-\Omega(\epsilon^2 m)}$$

Proof. By the properties of subgaussians we have that $(Ax)_i$ is subgaussian($\frac{||x||_2^2}{m}$). Hence $(Ax)_i^2$ is subgamma($\frac{||x||_2^4}{m^2}, \frac{||x||_2^2}{m}$).

Therefore, $||Ax||_2^2$ is subgamma($\frac{||x||_2^4}{m}, \frac{||x||_2^2}{m}$).

Also $E[(Ax)_i^2] = \frac{||x||_2^2}{m} \implies E[||Ax||_2^2] = ||x||_2^2$.

So by Property 2 of Def. 2.1:

$$Pr[|||Ax||_2^2 - ||x||_2^2| \geq \epsilon||x||_2^2] \leq 2e^{-\min(\frac{\epsilon^2||x||_2^4}{2||x||_2^4/m},\frac{\epsilon||x||_2^2}{2||x||_2^2/m})} = 2e^{-\frac{m}{2}\min(\epsilon^2,\epsilon)}$$

So $m = \frac{2}{\epsilon^2}\ln(\frac{2}{\delta})$ rows suffice for δ failure probability. □

4 Coverings

Given any set S, suppose we want $||Ax||_2^2 \in (1\pm\epsilon)||x||_2^2, \forall x \in S$. By the Johnson-Lindenstrauss lemma (which we proved in a previous class using the distributional version we showed in the previous section) it suffices to have $m = \frac{1}{\epsilon^2}\ln(\frac{2|S|}{\delta})$ rows.

However, this is problematic when the set S is infinite, for example if $S = \mathbb{R}^n$, since then $|S|$ is unbounded.

Idea: Use **coverings**. It suffices to consider the ball $B_2^n = \{x \in \mathbb{R}^n | ||x||_2 \leq 1\}$.

Definition 9. *A subset $T \subset B_2^n$ is an ϵ-cover of B_2^n if $\forall x \in B_2^n$ $\exists x' \in T$, such that $||x - x'|| \leq \epsilon$*

We can show that with $m = O(\frac{1}{\epsilon^2} \ln(\frac{|T|}{\delta}))$ we can have $||Ax||_2^2 \in [(1-\epsilon)||x||_2^2, (1+\epsilon)||x||_2^2], \forall x \in T$, using the JL lemma.

For the rest of $x \in B_2^n$, we use the following idea:

$$\forall x \in B_2^n, x = x_1 + \epsilon r_1, x_1 \in T, ||r_1|| \leq 1$$

$$\implies ||Ax||_2 \leq ||Ax_1||_2 + \epsilon||Ar_1|| \leq (1+\epsilon)||x_1|| + \epsilon||Ar_1||$$

We can repeat this k times to get:

$$x = x_1 + \epsilon x_2 + \ldots + \epsilon^k r_k, \forall i, x_i \in T \text{ and } ||r_k|| \leq 1$$

Similarly, we get that $||Ax||_2 \leq \sum_{i=1}^k (||Ax_i||_2 \epsilon^{i-1}) + \epsilon^k ||Ar_k||_2 \leq \frac{1+\epsilon}{1-\epsilon} + \epsilon = 1 + O(\epsilon)$

Similarly, for the lower bound we have:

$$||Ax||_2 \geq ||Ax_1|| - \epsilon||Ax_2|| - \epsilon^2||Ax_3|| - \ldots - \epsilon^k||Ar_k||_2$$

$$\geq 1 - \epsilon - (1+\epsilon)\epsilon - \ldots - \epsilon^k||Ar_k||_2 \geq 1 - \epsilon - \frac{\epsilon(1+\epsilon)}{1-\epsilon} - \epsilon^k||Ar_k||_2 = 1 - O(\epsilon)$$

The reason we can get rid of the $\epsilon^k||Ar_k||_2$ term as $k \to \infty$ is because for any fixed A, $||Ax||_2$ is bounded by the Frobenius norm of A, which gives us the following claim:

Claim 10. *For fixed A, $\lim_{k\to\infty} \epsilon^k ||Ar_k||_2 = 0$*

Proof. $||Ar_k||_2 \leq ||A||$ hence $\epsilon^k||Ar_k||_2 \leq \epsilon$ if $k > 1 + \log_\epsilon ||A||$ □

4.1 Metric Entropy

Hence we have shown that if T is an ϵ-cover of B_2^n and A preserves all $x' \in T$ up to $1 \pm \epsilon$, then A preserves all of $\mathbb{R}^n$, and in fact only $m = \frac{1}{\epsilon^2} \ln(\frac{2|T|}{\delta})$ rows suffice.

Question : How to actually find the minimum T that is an ϵ-cover?

Definition 11. *If $N(\epsilon, B_2, ||\cdot||_2)$ is the size of the minimum ϵ-cover of B_2, then $\log(N(\epsilon, B_2, ||\cdot||_2))$ is known as metric entropy.*

In the next class we will show more ways to compute this metric entropy quantity. In this case however we will try to upper bound it in a simple way.

Idea: Choose T greedily. Specifically, add any point in B_2^n that is not already covered (i.e. it is ϵ-far from all points already chosen). When the greedy algorithm terminates we have found an ϵ-cover.

Let us bound the number of points that this algorithm can add. Since all points are ϵ-far we have that the balls of radii $\epsilon/2$ that each point defines are all disjoint. Also all these balls must fit within a ball of radius $1+\epsilon/2$ around the origin. Hence, $|T| \leq \frac{Vol(B_2^n(1+\epsilon/2))}{Vol(B_2^n(\epsilon/2))} = (\frac{1+\epsilon/2}{\epsilon/2})^n = (1+2/\epsilon)^n$

Using this bound on the metric entropy of the ball, we get that we only need $m = O(\frac{1}{\epsilon^2} n \log(\frac{2}{\epsilon}))$ rows. It is possible to get rid of the $\log(\frac{2}{\epsilon})$ term, which we may show later.