Math 626: High Dimensional Probability

Taught by Mark Rudelson
Lecture notes by Sayantan Khan

January 23, 2021

Contents

1 Introduction

High dimensional probability is the study of random variables taking values in $\mathbb{R}^n$ for large but fixed values of n. While this area has always been studied by classical probability theorists, it has also attracted attention from computer scientists, especially since the design and analysis of fast probabilistic algorithms requires tools and theorems from this field. A classical result that arises from pure mathematics, but has several real life applications is Dvoretzky's theorem, whose statement does not involve any probability at all, and yet the proof uses concentration inequalities.

Theorem 1.1 (Dvoretzky's theorem)**.** *Let X be an n-dimensional normed vector space. Then for any $\varepsilon > 0$, there exists a constant $c(\varepsilon) > 0$, and a linear subspace E of X such that, $\dim(E) \geq c(\varepsilon)\log(n)$, and for all $v \in E$, the following inequality relating the ambient norm and the Euclidean norm on E.*

$$(1-\varepsilon)\|v\|_2 \leq \frac{\|v\|_X}{M(X)} \leq (1+\varepsilon)\|x\|_2$$

Where $M(X)$ is a scaling factor that depends only on X, and $\|\cdot\|_X$ and $\|\cdot\|_2$ are the ambient and Euclidean norm on E.

Idea of proof. Pick a random subspace, and then show that with very high probability, the given inequality holds. We will prove the result in full detail later in the course. □

Remark 1.2. When the norm on X is the ℓ^1 norm, the lower bound on the dimension of E can be improved to be linear in $c(\varepsilon)n$. ◇

A computer science application of Dvoretzky's theorem Consider a subset S of ℓ_2^N, given by inequalities involving the norms of elements in ℓ_2^N. Suppose that we are required to optimize a linear function f on the set S. Since S is given by inequalities involve the ℓ^2-norm, it will be and intersection of interiors of ellipsoids, and consequently, optimizing f will be computationally expensive. But we can get around the computational expense by embedding ℓ_2^N into ℓ_1^M, where M is $O(N)$, by Remark 1.2. Since this embedding does not distort distances too much, we can replace S with a nearby polytope, given by inequalities involving the ℓ^1 norm instead. Optimizing a linear function on a polytope is computationally much easier, thanks to linear programming.

Empirical covariance estimation Let X be an $\mathbb{R}^n$-valued random variable, and let $\mathbb{E}(X) = 0$. The covariance of X is $\mathbb{E}(X^\top X)$, and will be denoted by A. Let $\{X_1, X_2, \ldots, X_m\}$ be i.i.d. samples of X. We define the sample covariance A_m in the following manner.

$$A_n = \frac{\sum_{i=1}^m X_i^\top X_i}{m}$$

As m tends to infinity, the sample covariance A_m will approach the true covariance, as we would expect the law of large numbers to predict. A harder, and more interesting question is to determine how many samples do we need to take to be within some threshold of the true covariance with high probability.

Just like in the scalar setting, one answers the question by proving appropriate concentration inequalities for matrix valued random variables. Here is the most general set up: Let A_m and A be matrices mapping some normed space X to some other normed space Y. We define the distance between A_m and A using the operator norm.

$$\begin{aligned}
\|A_m - A\|_{\text{op}} &= \max_{\|v\|_x \leq 1} \|(A_m - A)v\|_Y \\
&= \max_{\|v\|_x \leq 1} \max_{\|w\|_{Y^*} \leq 1} w\left((A_m - A)v\right) \\
&= \max_{\substack{(v,w) \in X \times Y^* \\ \|v\| \leq 1 \\ \|w\| \leq 1}} w((A_m - A)v)
\end{aligned}$$

We can consider $w((A_m - A)v)$ to be a family of scalar random variables parameterized by points in $X \times Y^*$, i.e. a random process V_u parameterized by $u \in X \times Y^*$. This leads to the following two questions.

(i) How to bound $\mathbb{E} \max V_u$.

(ii) How to bound $\mathbb{P}(|\max V_u - \mathbb{E} \max V_u| \geq t)$.

It turns out one can often answer (ii) without answering (i)m which may seem surprising given that the most elementary concentration inequalities involve the expectation (i.e. Markov's inequality). The starting point in answering (ii) is understanding *concentration of measure*.

2 Concentration of measure

Concentration of measure was originally observed Lévy, but first used by Milman in the early 70s. Roughly speaking, concentration of measure is the following phenomenon: suppose

$(X, d, \mathbb{P})$ is a metric space endowed with a probability measure and $f : X \to \mathbb{R}$ is a "nice" function. Then the value of f is essentially constant, i.e. there exists some constant $M(f)$ such that for small enough ε, the following probability bound holds.

$$\mathbb{P}(|f(x) - M(f)| < \varepsilon) \ll 1$$

Usually by nice, we will mean 1-Lipschitz, although similar results hold for a more general class of functions like convex or quasi-convex functions. We begin with the simplest version of measure concentration.

Concentration for linear functions Let the metric space X in this setting be $\mathbb{R}^n$ for some fixed n, and let $\{X_1, \ldots, X_n\}$ be i.i.d scalar random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a linear function, given by taking inner product with some vector $\mathbf{a}$. We define Y to be $\sum_{i=1}^n a_i X_i$. We start by considering random variables X_i which have a subgaussian decay.

Definition 2.1 (Subgaussian decay)**.** A random variable X is said to be σ-subgaussian if there exists a constant $C > 0$, such that for all $t > 0$, the following inequality holds.

$$\mathbb{P}(|X| > t) \leq C \exp\left(-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right)$$

◇

Remark 2.2. The constant σ is often referred to as the variance proxy of the subgaussian random variable. ◇

Example 2.3. The following random variables are examples of subgaussian random variables.

(i) Normal random variables

(ii) Bounded random variables

◇

Lemma 2.4. *Let X be a random variable. Then the following are equivalent:*

(i) For all $t > 0$, $\mathbb{P}(|X| > t) \leq C \exp\left(-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right)$ (definition of subgaussian random variables).

(ii) There exists $a > 0$ such that $\mathbb{E}(\exp(aX^2)) < \infty$ (ψ_2 condition).

(iii) There exist C' and b such that for all $\lambda \in \mathbb{R}$, $\mathbb{E}(\exp(\lambda X)) \leq C' \exp(b\lambda^2)$ (Laplace transform condition).

(iv) There exists K such that for all $p \geq 1$, $\mathbb{E}(X^p)^{\frac{1}{p}} \leq K\sqrt{p}$ (moment condition).

Moreover, if $\mathbb{E}(X) = 0$, then the constant C' in the Laplace transform condition can be taken to be equal to 1.

Proof. $(i) \implies (ii)$ Using Fubini's theorem and a change of variables, we can express $\mathbb{E}(aX^2)$ as an integral involving tail bounds.

$$\begin{aligned}\mathbb{E}(aX^2) &= 1 + \int_0^\infty 2ate^{at^2} \cdot \mathbb{P}\left(|X| > t\right) \, \mathrm{d}t \\ &\leq 1 + \int_0^\infty 2Cat \exp\left(at^2 - \frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right) \, \mathrm{d}t\end{aligned}$$

Clearly, picking a value of a smaller than $\frac{1}{2\sigma^2}$ will make the integral converge.

$(ii) \implies (iii)$ Since we want to estimate the expectation of $\exp(\lambda X)$, we multiply and divide by $\exp(aX^2)$ and complete the square.

$$\exp(\lambda X) = \exp\left(aX^2\right) \cdot \exp\left(\frac{\lambda^2}{4a}\right) \cdot \exp\left(-\left(\sqrt{a}X + \frac{\lambda}{2\sqrt{a}}\right)^2\right)$$

Note that the third term in the product is always less than 1. We now take the expectation of the right hand side.

$$\begin{aligned}\mathbb{E}(\exp(\lambda X)) &= \mathbb{E}\left(\exp\left(aX^2\right) \cdot \exp\left(\frac{\lambda^2}{4a}\right) \cdot \exp\left(-\left(\sqrt{a}X + \frac{\lambda}{2\sqrt{a}}\right)^2\right)\right) \\ &\leq \exp\left(\frac{\lambda^2}{4a}\right) \mathbb{E}\left(\exp(aX^2)\right)\end{aligned}$$

Setting $b = \frac{1}{4a}$ and $C' = \mathbb{E}(\exp(aX^2))$, we get the result.

$(iii) \implies (iv)$ We begin by getting a crude estimate for $\mathbb{E}(X^p)$ using the infinite series for $\exp(\lambda X)$.

$$\begin{aligned}\mathbb{E}(X^p) &\leq \frac{p!}{\lambda^p}\mathbb{E}(\exp(\lambda X)) \\ &= \frac{C'p!\exp(b\lambda^2)}{\lambda^p}\end{aligned}$$

Note that this inequality works for all values of λ, but to get the best inequality, we minimize the right hand side by varying λ over $\mathbb{R}$. The minimum is attained when $\lambda = \frac{\sqrt{p}}{2b}$: plugging that into the right hand side, and taking p^{th} roots gives us the following.

$$\mathbb{E}(X^p)^{\frac{1}{p}} \leq C''\frac{(p!)^{\frac{1}{p}}}{\sqrt{p}}$$

Here, we have absorbed all the constants into C''. Using Stirling's approximation, the numerator is bounded above by p, giving us the inequality we want.

$(iv) \implies (i)$ We rewrite the event $|X| > t$ in the following manner.

$$\begin{aligned}\mathbb{P}(|X| > t) &= \mathbb{P}(\exp(\lambda X^2) > \exp(\lambda t^2)) \\ &\leq \exp(-\lambda t^2)\mathbb{E}(\exp(\lambda X^2))\end{aligned}$$

Here, λ is a positive real number that we will specify later, and the inequality comes from Markov's inequality. Of course, we do not a priori know that $\mathbb{E}(\lambda X^2)$ is finite, but we will pick a λ small enough such that it is. Using Fubini's theorem, we can express $\mathbb{E}(\exp(\lambda X^2))$ in the following manner.

$$\mathbb{E}(\exp(\lambda X^2)) = 1 + \frac{\lambda \mathbb{E}(X^2)}{1!} + \frac{\lambda^2 \mathbb{E}(X^4)}{2!} + \cdots$$

Using the bound on the moments of X and Stirling's approximation, we get the following inequality.

$$\begin{aligned}\mathbb{E}(\exp(\lambda X^2)) &\leq \sum_{p=0}^{\infty} \frac{(2\lambda K^2 p)^p}{p!} \\ &\leq \sum_{p=0}^{\infty} \frac{(2\lambda e K^2 p)^p}{\sqrt{2\pi p} p^p}\end{aligned}$$

If we pick λ to be small enough that $2e\lambda K^2$ is much smaller than 1, then the infinite sum converges, and the expectation is finite. Setting $\frac{1}{2\sigma^2}$ to be equal to λ gives us the result.

We now show that the constant in the Laplace transform condition can be set to be 1 when $\mathbb{E}(X) = 0$. To do so, we recall the ψ_2 and the Laplace transform condition, i.e. there exist constants a, C' and b such that the following two inequalities hold for all $\lambda \in \mathbb{R}$.

$$\mathbb{E}(\exp(aX^2)) < \infty \tag{2.5}$$

$$\mathbb{E}(\exp(\lambda X)) \leq C \exp(b\lambda^2) \tag{2.6}$$

Suppose now that $\lambda^2 > 2a$. By the Laplace transform condition, we have the following inequality.

$$\begin{aligned}\mathbb{E}(\exp(2aX)) &\leq C' \exp(4ba^2) \\ &= \exp(4a^2 b')\end{aligned}$$

Where b' is $b + \frac{\log(C')}{4a^2}$. Since b' decreases as a increases, for any $\lambda^2 > 2a$, $\mathbb{E}(\exp(aX^2))$ will be less than $\exp(b'\lambda^2)$.

Now suppose that $\lambda^2 < 2a$. We begin by considering the special case where X is a symmetric random variable. By symmetry of X, we have the following inequality.

$$\begin{aligned}\exp(\lambda X) &= \frac{\exp(\lambda X) + \exp(-\lambda X)}{2} \\ &\leq \exp\left(\frac{\lambda^2 X^2}{2}\right)\end{aligned}$$

Taking expectations on both sides gives us the following.

$$\mathbb{E}(\exp(\lambda X)) \leq \mathbb{E}\left(\exp\left(\frac{\lambda^2}{2} X^2\right)\right)$$

Since $\lambda^2 < 2a$, $\frac{2a}{\lambda^2}$ is greater than 1, and we can use Hölder's inequality to bound the right hand term.

$$\mathbb{E}\left(\exp\left(\frac{\lambda^2}{2}X^2\right)\right) \leq \left(\mathbb{E}\left(\exp\left(aX^2\right)\right)\right)^{\frac{\lambda^2}{2a}}$$

Since $\mathbb{E}(\exp(aX^2))$ is a finite constant, the right hand side is $\exp(b''\lambda^2)$ for some constant b''.

Now suppose X is not symmetric. Let X' be an identical independent copy of X. Since $\mathbb{E}(X')$ is 0, we have the following equality.

$$\mathbb{E}(\exp(\lambda X)) = \mathbb{E}(\exp(\lambda(X - \mathbb{E}(X'))))$$

Since exp is a convex function, we can pull out the inner expectation, using Jensen's inequality.

$$\mathbb{E}(\exp(\lambda X)) \leq \exp(\lambda(X - X'))$$

Since $X - X'$ is symmetric, the result follows from the previous part, and the proof is complete. □

We now explain why care so much about the constant in the Laplace transform condition.

Theorem 2.7 (Hoeffding-Chernoff-Azuma inequality)**.** *Let $\{X_1, \ldots, X_n\}$ be i.i.d. subgaussian random variables with mean 0. Then for any $(a_1, \ldots, a_n) \in \mathbb{R}^n$ and any $t > 0$, the following probability bound holds.*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i X_i\right| > t\right) \leq C \exp\left(-c\frac{t^2}{\|\mathbf{a}\|_2}\right)$$

Where C and c are some absolute constants.

Proof. Without loss of generality, we can assume $\|\mathbf{a}\|_2 = 1$. It will suffice to show that the sum of subgaussian random variables is subgaussian. We will show it verifying the Laplace transform condition. Let $\lambda \in \mathbb{R}$, and let $Y = \sum_{i=1}^{n} a_i X_i$. We compute $\mathbb{E}(\exp(\lambda Y))$.

$$\begin{aligned}
\mathbb{E}(\exp(\lambda Y)) &= \prod_{i=1}^{n} \mathbb{E}(\exp(\lambda a_i X_i)) \\
&\leq \prod_{i=1}^{n} \mathbb{E}(\exp(b\lambda^2 a_i^2)) \\
&= \mathbb{E}(\exp(b\lambda^2))
\end{aligned}$$

This proves the result. Note that having the Laplace transform coefficient equal to 1 helped, because if the coefficient C was greater than 1, then we would pick up a constant of C^n, which would be very large for large values of n. □