Math 626: High Dimensional Probability

Taught by Mark Rudelson
Lecture notes by Sayantan Khan

January 21, 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'll 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 doesn't 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).