IEMS 402: Statistical Learning **Winter 2024-2025**

Lecture 7: Concentration

Lecturer: Yiping Lu *Scribes: Matt Lu*

7.1 The High-dimensional Setting

So far we focused on $X_i \in \mathbb{R}^k, \theta \in \mathbb{R}^d$. We observe $X_1, \ldots, X_n \overset{i.i.d}{\sim} P_\theta$, with k, d fixed, and let $n \to \infty$. This is the traditional setting for **Asymptotic statistics**.

Under nice conditions, all the quantities we cared about (e.g. the MLE) were in linear functions of $n \to \infty$ samples in fixed dimensions. So the CLT applied to give us precise confidence intervals, and we are very pleased about that.

This framework is useful if k, d is small relative to n. Back in the day, this was generally true because humans took measurements, and models were analyzed by humans as well. How many measurements per specimen can one scientist take? Only $k = O(1)$ or so.

However, this is not the case in modern days. Nowadays many measurements are taken by a computer. Consider the following YouTube example:

Let X_i be the YouTube video watch count of user i, represented in $\mathbb{R}^k$ where each coordinate represents a video. The internet claims $K \approx 800$ million videos. The number of daily active YouTube users is ≈ 122 million (not important here), and the total number of users n is between 1 and 2 billion. So $\frac{n}{k} \leqslant 3$. Asymptotic theory will not apply here.

Under this high-dimensional setting, a different framework may be useful. We need a different setting, the **high-dimensional setting**. And we will focus on this in the following weeks.

High-dimensional statistics studies what happens when $d, n, k \to \infty$ together (i.e. k, d are no longer fixed). The focus is less on the limiting law (e.g. CLT) and more on getting rough confidence intervals that hold for all d, n, k (large).

Here are two case studies to offer us more sense of what's going on under the high dimensional setting.

7.2 Concentration

Now we are going to build tools for how well random variables concentrate. We will first look at some basic concentration inequalities, starting with Markov's inequality.

Theorem 7.1 (Markov's inequality) *Let X be a non-negative random variable. Then*

$$\mathbb{P}(x \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}$$

Proof: Using the law of total expectation, we write:

$$\mathbb{E}[X] = \mathbb{E}[X \mid X < a]\mathbb{P}(X < a) + \mathbb{E}[X \mid X \geqslant a]\mathbb{P}(X \geqslant a).$$

Here, note that $\mathbb{E}[X \mid X \geqslant a] \geqslant a$, because $X \geqslant a$ in this range by definition. Substituting this bound into the expression for $\mathbb{E}[X]$, we get:

$$\mathbb{E}[X] \geqslant a \cdot \mathbb{P}(X \geqslant a).$$

Rearranging this inequality gives:

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}.$$

■

Another basic concentration inequality is Chebyshev's inequality. It provides an upper bound on the probability that a random variable deviates from its mean by more than k standard deviations. This is particularly useful because it applies to any distribution (as long as the mean and variance exist), making it a very general result.

Theorem 7.2 (Chebyshev's inequality) *Let X be a random variable with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = Var(X)$. Then for any $k > 0$:*

$$\mathbb{P}(|X - \mu| \geqslant k) \leqslant \frac{\sigma}{k^2}.$$

Proof: Observe that $|X - \mu| \geqslant t \Leftrightarrow (X - \mu)^2 \geqslant t^2$. Applying Markov's, we get the desired result. ■

Before we continue, let's first take a look at moment generating functions (MGF). The MGF of a random variable provides a convenient way to summarize its entire probability distribution. It is defined as:

$$M_X(t) = \mathbb{E}[e^{tX}].$$

The name *moment generating function* comes from the fact that differentiating $M_X(t)$ and evaluating at $t = 0$ recovers the moments of X, such as the mean and variance.

- **Uniqueness**: If two random variables have the same MGF, they have the same distribution.
- **Moment Computation**: Taking derivatives of $M_X(t)$ at $t = 0$ gives the moments (mean, variance, etc.).
- **Simplifying Sums of Independent Variables**: For independent X_1, X_2, the MGF of their sum is simply the product of their MGFs:

  $$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t).$$

  This property is often used to derive the distribution of sums of independent variables.
- **Probability Bounds and Concentration Inequalities**: MGFs provide a powerful way to derive tail bounds, such as Chernoff bounds, which are useful in probability theory and statistical learning.

Think of the MGF as a transformation that encodes all the key statistical properties of a random variable in a single function. Since exponentials grow quickly, e^{tX} amplifies large values of X, making the MGF particularly useful in studying concentration of measure and large deviations.

In high-dimensional analysis (and beyond), an essential tool is concentration inequalities, often for sums of (nearly) independent R.V.s. Suppose $X_1, \ldots, X_n \in \mathbb{R}, i.i.d$ (we can have weaker assumptions later). We want to study how fast $\frac{1}{n}\sum_i X_i$ converges to $\mu = \mathbb{E}X_i$.

We will use this logic, often referring to "Chernoff bound":

$$\begin{aligned}\mathbb{P}[X - \mathbb{E}X > t] &= \mathbb{P}\left[e^{s(X-\mathbb{E}X)} > e^{st}\right] \text{ for any } t, s > 0 \\ &\leqslant \frac{\mathbb{E}e^{s(X-\mathbb{E}X)}}{e^{st}} \quad \text{(By Markov inequality).}\end{aligned}$$

By the arbitrariness of $s > 0$, we know $\mathbb{P}[X - \mathbb{E}X > t] \leqslant \inf_{s>0} \frac{\mathbb{E}e^{s(X-\mathbb{E}X)}}{e^{st}}$.

If $X = \sum_{i=1}^n X_i$ for X_i i.i.d, then $\mathbb{E}e^{sX} = \left(\mathbb{E}e^{sX_i}\right)^n$, which is "nice".

So understanding the MGF of X_i is enough to bound the above. It turns out that a condition weaker than independence suffices.

Definition 7.3 *We say a sequence $Y_1, Y_2, \ldots$ is a Martingale adapted to a filtration of σ-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ if for all $t \geqslant 1, \mathbb{E}|Y_t| < \infty$ and $\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] = Y_t$. ($\mathcal{F}_t$ represents "everything that has happened up to time t".)*

Example 1 *$Y_t = \sum_{i=1}^t (X_i - \mathbb{E}X_i)$ for X_i i.i.d bounded R.V.s.*

Example 2 *$Y_0 = 1$. $Y_{t+1} \sim \text{Unif}(\{0, 1, \ldots, 2Y_t\})$. $\mathbb{E}Y_t \leqslant 2^t < \infty$.*

Definition 7.4 *A R.V. X on $\mathbb{R}$ is called σ-subgaussian if $\mathbb{E}\exp(t(X - \mathbb{E}X)) \leqslant \exp\left(\sigma^2 t^2/2\right), \forall t \in \mathbb{R}$.*

This condition asks that the MGF of X is bounded by that of $N\left(0, \sigma^2\right)$, which is why we call it "subgaussian". It's very useful in getting bounds for concentration inequalities.

Claim: If $\text{supp}(X) \subseteq [a, b]$ then X is $\frac{(b-a)}{2}$-subgaussian.

Proof: See HW 6. ■

Theorem 7.5 (Azuma-Hoeffding inequality for subgaussian increments) *Suppose $Y_0, Y_1, Y_2, \ldots$ is a martingale sequence with conditional martingale difference $(Y_{i+1} - Y_i)|\mathcal{F}_i$ being σ_{i+1}-subgaussian almost surely. Then for all $t > 0$,*

$$\mathbb{P}\left[|Y_n - Y_0| > t\right] \leqslant 2 \cdot \exp\left(-t^2/(2\sum_{i=1}^n \sigma_i^2)\right).$$

It says $Y_n - Y_0$ behaves like gaussian. It makes sense since we can view it as the sum of i.i.d "Gaussian" R.V.s, then we get this analogy to the CLT.

Proof: Plugging into our earlier "Chernoff" reasoning:

$$\mathbb{P}\left[Y_n - Y_0 > t\right] \leqslant \inf_{s>0} \mathbb{E}\exp\left(s\left(Y_n - Y_0\right) - st\right).$$

Since:

$$\begin{aligned}
\mathbb{E}\exp\left(s\left(Y_n - Y_0\right)\right) &= \mathbb{E}\exp\left(s(Y_{n-1} - Y_0) + s(Y_n - Y_{n-1})\right) \\
\text{(By adaptivity)} \quad &= \mathbb{E}\left[\exp\left(s\left(Y_{n-1} - Y_0\right)\right) \cdot \mathbb{E}\left[\exp\left(s\left(Y_n - Y_{n-1}\right)\right) \mid \mathcal{F}_{n-1}\right]\right] \\
&\leqslant \mathbb{E}\left[\exp\left(s\left(Y_{n-1} - Y_0\right)\right)\right] \cdot \exp\left(\frac{s^2 \cdot \sigma_n^2}{2}\right) \\
\text{(By induction)} \quad &\leqslant \exp\left(s^2(\sum_{i=1}^{n}\sigma_i^2)/2\right).
\end{aligned}$$

Plug in for the RHS, we get:

$$\begin{aligned}
\inf_{s>0} \mathbb{E}\exp\left(s\left(Y_n - Y_0\right) - st\right) &\leqslant \inf_{s>0} \exp\left(\frac{s^2}{2}\left(\sum_{i=1}^{n}\sigma_i^2\right) - st\right) \\
\text{(By calculus, } s = t/\sum_{i=1}^{n}\sigma_i^2 \text{ is the minimizer)} \quad &= \exp\left(-t^2/(2\sum_{i=1}^{n}\sigma_i^2)\right).
\end{aligned}$$

Repeat the same procedure for $Y_0 - Y_n$ and then perform the union bound, and we complete the proof. ■

From the theorem, we get Gaussian-like tails for any sum of i.i.d random variables.

Example 3 (Hoeffding's inequality for general bounded random variables, see Theorem 2.2.6 in Vershynin)

Suppose $X_i \in [a,b]$ *i.i.d,* $\bar{X}_n = \frac{1}{n}\sum X_i$*, then we have* $\mathbb{P}\left[\left|\bar{X}_n - \mu\right| > t\right] \leqslant 2\exp\left(-\frac{t^2}{2}\frac{4n}{(b-a)^2}\right)$.

Proof: Since X_i is $\frac{(b-a)}{2}-$ subgaussian, we have $\bar{X}_n$ is $\sum_{i=1}^{n}\sigma_i^2 = \frac{(b-a)^2}{4n}-$ subgaussian.

So perform the change of variable $t = \frac{s}{\sqrt{n}}$ in Theorem 7.5, we get:

$$\mathbb{P}\left[\left|\bar{X}_n - \mu\right| > \frac{s}{\sqrt{n}}\right] \leqslant 2 \cdot \exp\left(-\frac{2s^2}{(b-a)^2}\right).$$

■

Keep in mind that $\mathrm{Var}\left(X_i\right) \leqslant \frac{(b-a)^2}{4}$ [see HW 6 P1(a)] so this is comparable to $\exp\left(-\frac{s^2}{2\,\mathrm{Var}(X_i)}\right)$, which is sort of like a "Quantitative CLT" guarantee for the average.

Note: The subgaussian condition is slightly too restrictive to capture e.g. $\chi^2 - R.V.s$. For that we introduce the following relaxed definition:

Definition 7.6 *A random variable* X *on* $\mathbb{R}$ *is* (σ,α)*-subexponential if* $\mathbb{E}\exp(t(X - \mathbb{E}X)) \leqslant \exp\left(\frac{t^2\sigma^2}{2}\right)$ *for all* $|t| < \frac{1}{\alpha}$.

This is a relaxation of subgaussian conditions.

Example 4 *If $w = Z^2,\ Z \sim N(0,1)$, then w is $(2,4)-$subexponential. But it's not subgaussian.*

Proof:

$$\begin{aligned}\mathbb{E}\exp\left(t\left(w-1\right)\right) &= \frac{1}{\sqrt{2\pi}}\int \exp\left(t\left(z^2-1\right)\right)e^{-z^2/2}dz \\ &= \frac{1}{\sqrt{2\pi}}e^{-t}\cdot\int e^{-z^2\frac{(1-2t)}{2}}dz \\ &= \frac{1}{\sqrt{1-2t}}e^{-t}. \leftarrow \text{ not finite if } t \geqslant \frac{1}{2}, \text{ so not subgaussian}\end{aligned}$$

By calculus, we can get $\frac{1}{\sqrt{1-2t}}e^{-t} \leqslant e^{2t^2}$ for $|t| \leqslant \frac{1}{4}$, hence w is $(2,4)-$subexponential. ■

Theorem 7.7 (Azuma-Hoeffding (for subexponential increments)) *If $Y_0, Y_1, Y_2, \ldots$ is a Martingale sequence with $(Y_{i+1}-Y_i)|\mathcal{F}_i$ being $(\sigma_{i+1}, \alpha_{i+1})$-subexponential almost surely, then $Y_n - Y_0$ is $\left(\sqrt{\sum_{i=1}^n \sigma_i^2}, \max\limits_{i\leqslant n}\alpha_i\right)$-subexponential and $\forall t > 0$,*

$$\mathbb{P}\left[|Y_n - Y_0| \geqslant t\right] \leqslant 2\cdot\exp\left(-\min\left(\frac{t^2}{2\sum_{i=1}^n\sigma_i^2}, \frac{t}{2\max\limits_{i\leqslant n}\alpha_i}\right)\right).$$

Proof: The proof is almost the same as before, except that now we have a different MGF:

$$\mathbb{P}\left[Y_n - Y_0 > t\right] \leqslant \inf_{s>0}\mathbb{E}\exp\left(s\left(Y_n - Y_0\right) - ts\right)$$

$$\begin{aligned}\mathbb{E}\exp\left(s\left(Y_n - Y_0\right)\right) &= \mathbb{E}\exp\left(s(Y_{n-1} - Y_0) + s(Y_n - Y_{n-1})\right) \\ \text{(By adaptivity)} \quad &= \mathbb{E}\left[\exp\left(s\left(Y_{n-1}-Y_0\right)\right)\cdot\mathbb{E}\left[\exp\left(s\left(Y_n - Y_{n-1}\right)\right)\mid\mathcal{F}_{n-1}\right]\right] \\ &\leqslant \mathbb{E}\left[\exp\left(s\left(Y_{n-1}-Y_0\right)\right)\right]\cdot\exp\left(\frac{s^2\cdot\sigma_n^2}{2}\right) \\ \text{(By induction)} \quad &\leqslant \exp\left(s^2(\sum_{i=1}^n\sigma_i^2)/2\right)\end{aligned}$$

$$\text{where } s \leqslant \frac{1}{\alpha_{\max}} \quad \text{(Here, we denote } \alpha_{\max} = \max_{i\leqslant n}\alpha_i\text{)}.$$

Hence we have

$$\mathbb{P}\left[Y_n - Y_0 > t\right] \leqslant \inf_{s>0}\exp\left(\frac{s^2}{2}\sum_{i=1}^n\sigma_i^2 - ts\right) \text{ for } s \leqslant \frac{1}{\alpha_{\max}}.$$

Now the inf over s is achieved at $s = \frac{t}{\sum\sigma_i^2}$ only when $\frac{t}{\sum\sigma_i^2} \leqslant \frac{1}{\alpha_{\max}}$. In this case, the tail bound is $\exp(-\frac{t^2}{2\sum\limits_{i=1}^n\alpha_i^2})$.

On the other hand, if $t > \frac{\sum \sigma_i^2}{\alpha_{\max}}$, then the inf over s will be achieved at $s = \frac{1}{\alpha_{\max}}$, which gives the tail bound $\exp\left(-\frac{t}{2\alpha_{\max}}\right)$.

Then repeat the above procedure with $Y_0 - Y_n$ and apply union bound, we complete the proof. ■

Note: a special case is when $n = 1$, this shows us why these are called "subexponential": if $t < \frac{\sigma^2}{\alpha}$, we have subgaussian tails, but if $t > \frac{\sigma^2}{\alpha}$ they start to decay like exponential tails (a "phase transition"). And the t at which tails switch is $\frac{\sum \sigma_i^2}{\max \alpha_i}$.

Corollary 7.8 *If $Z \sim N(0, I_d)$, $||Z|| = \sqrt{d} \cdot (1 \pm o(1))$ with high probability.*

Proof: We know $\sum_{i=1}^{d} \left(Z_i^2 - 1\right)$ is martingale (index is d).

From Example 4, Z_i^2 is $(2, 4)$-subexponential. Apply this into the setting of Theorem 7.7, and we get $\sum_{i=1}^{d} \sigma_i^2 = 4d$ (since $\sigma_i = 2$, $i = 1, \ldots, d$). Now, from Theorem 7.7

$$\begin{aligned}\mathbb{P}\left[\left|\|Z\|^2 - d\right| > t\right] &\leqslant 2\exp\left(-\min\left(\frac{t^2}{2 \cdot 4d}, \frac{t}{2 \cdot 4}\right)\right) \\ &\leqslant \varepsilon \quad \text{if} \quad t = \Omega\left(\sqrt{d}\log\frac{1}{\varepsilon}\right).\end{aligned}$$

Now we complete the proof. ■

Note: This gives us the claim from earlier about the lower bound on n when Θ is finite and like $\mathbb{S}^{d-1}$ (Case B in Case Study **??**).

7.3 Johnson–Lindenstrauss lemma

In this section, we will see an application of concentration inequalities via the Johnson-Lindenstrauss lemma. The Johnson-Lindenstrauss (JL) lemma provides a powerful result in high-dimensional geometry: it states that a set of points in high-dimensional space can be embedded into a much lower-dimensional space, while approximately preserving the pairwise distances between points. This is achieved through a random linear projection. The key insight is that high-dimensional data often contains redundant information, and projecting the data onto a lower-dimensional subspace can preserve its essential geometric structure with high probability.

Motivation: The JL lemma is particularly useful in applications where:

- The data lies in a very high-dimensional space, making computations expensive.
- Approximate preservation of distances (e.g., Euclidean) is sufficient for tasks like clustering, nearest neighbor search, or dimensionality reduction.

The lemma ensures that we can reduce the dimensionality of the data significantly, without introducing much distortion, making it both computationally efficient and theoretically robust.

Key Idea: By using a random linear projection, the geometry of the points is preserved with high probability due to the concentration of measure phenomenon in high dimensions. This randomness is crucial, as deterministic methods often fail to provide similar guarantees.

Lemma 7.9 *For any $0 < \epsilon < 1$ and any integer n, let k be a positive integer such that*

$$k \geqslant \frac{24}{3\epsilon^2 - 2\epsilon^3} \log n$$

then for any set A of n points $\in \mathbb{R}^d$. there exists a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $x_i, x_j \in A$

$$(1-\epsilon)||x_i - x_j||^2 \leqslant ||f(x_i) - f(x_j)||^2 \leqslant (1+\epsilon)||x_i - x_j||^2$$

Proof:

Step 1: Define the Random Projection

Let R be a random matrix of size $k \times d$ where each entry $R_{ij} \sim \mathcal{N}(0,1)$ i.i.d. For any vector $u \in \mathbb{R}^d$, define $v = \frac{1}{\sqrt{k}} Ru$. Then $v \in \mathbb{R}^k$, and its entries v_i are given by:

$$v_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{d} R_{ij} u_j.$$

Step 2: Expected Norm Preservation

We can now show that the expected value of the Euclidean distance of the random projection is equal to the Euclidean distance of the original subspace. The squared norm of v is:

$$\|v\|_2^2 = \frac{1}{k} \sum_{i=1}^{k} \left(\sum_{j=1}^{d} R_{ij} u_j \right)^2.$$

Taking the expectation:

$$\begin{aligned}
\mathbb{E}[||v||_2^2] &= \mathbb{E}\left[\sum_{i=1}^{k} v_i^2 \right] \\
&= \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}\left[\left(\sum_{j=1}^{d} R_{ij} u_j \right)^2 \right] \\
&= \frac{1}{k} \sum_{i=1}^{k} \sum_{j,k}^{d} u_j u_k \mathbb{E}[R_{ij} R_{ik}] \\
&= \frac{1}{k} \sum_{i=1}^{k} \sum_{j,k}^{d} u_j u_k \delta_{j,k} \\
&= \sum_{i \leqslant j \leqslant d} u_j^2 \\
&= ||u||_2^2
\end{aligned}$$

where $\delta_{j,k}$ is the Kronecker-δ function (i.e. $\delta_{j,k} = \{0 \text{ if } j \neq k; 1 \text{ if } j = k\}$).

Step 3: Concentration of Norms

Next, we will use concentration inequalities to show the probability that the variance of the Euclidean distance is greater than a specified error. First, let $X = \sum_{i=1}^{k} x_i$, where $x_i = R_i^\top \cdot u$ (1-dim projection). Then,

$$\begin{aligned}\mathbb{P}\left(||v||_2^2 \geqslant (1+\epsilon)||u||_2^2\right) &= \mathbb{P}\left(\frac{1}{K}||U||_2^2 X \geqslant (1+\epsilon)||u||_2^2\right) \\ &= \mathbb{P}(x \geqslant (1+\epsilon)k) \\ &= \mathbb{P}(x^{\lambda x} \geqslant e^{\lambda(1+\epsilon)k})\end{aligned}$$

By Markov's inequality,

$$\begin{aligned}\mathbb{P}(x^{\lambda x} \geqslant e^{\lambda(1+\epsilon)k}) &\leqslant \frac{\mathbb{E}[e^{\lambda x}]}{e^{\lambda(1+\epsilon)k}} \\ &= \prod_{i=1}^{k} \frac{\mathbb{E}[e^{\lambda x_i^2}]}{e^{\lambda(1+\epsilon)k}} \qquad \text{(as } x_i\text{'s are iid)} \\ &= \left[\frac{\mathbb{E}[e^{\lambda x_i^2}]}{e^{\lambda(1+\epsilon)k}}\right]^k \\ &\leqslant \left(\frac{1}{\sqrt{1-2\lambda}\cdot e^{\lambda(1+\epsilon)}}\right)^k\end{aligned}$$

where the last inequality follows using the mgf of χ^2. Now setting, $\lambda = \frac{\epsilon}{2(1+\epsilon)}$, we have $\leqslant [(1+\epsilon)e^{-\epsilon}]^{\frac{k}{2}}$. Finally, using the inequality $\log(1+x) < x - \frac{x^2}{2} + \frac{x^3}{3}$ and $k \geqslant \frac{24}{3\epsilon^2 - 2\epsilon^3}\log n$, we have

$$\begin{aligned}&\leqslant e^{-(\epsilon^2/2 - \epsilon^3/3)\frac{k}{2}} \\ &\leqslant e^{-2\log n} \\ &\leqslant n^{-2}\end{aligned}$$

We can conclude that $\mathbb{P}\left(||v||_2^2 \geqslant (1+\epsilon)||u||_2^2\right) > 1 - n^{-2}$. We can repeat the process above to show that $\mathbb{P}\left(||v||_2^2 \geqslant (1-\epsilon)||u||_2^2\right) > 1 - n^{-2}$.

Step 4: Union Bound for All Pairs

To ensure that the distance between all n points in A is preserved, apply the union bound over all $\binom{n}{2}$ pairs of points. Set $k \geqslant O(\log n/\epsilon^2)$ to ensure that the total failure probability is less than $1/n$. ■