Lecture 1

Unexpected behaviors of high-dimensional spaces and introductory

Counter-intuition of high dimensional data

Vastness of hypersphere. Consider an inscribed hypersphere with radius r to a hypercube with edges of length $2r$ in d-dimensional Euclidean space,

$$V_{\text{hypersphere}} = \frac{2r^d \pi^{d/2}}{d\Gamma(d/2)}, \quad V_{\text{hypercube}} = (2r)^d,$$

where Γ is the gamma function. Then

$$\lim_{d\to\infty} \frac{V_{\text{hypersphere}}}{V_{\text{hypercube}}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} = 0, \tag{1.1}$$

which implies that *data points uniformly generated in a high-dimensional hypercube are concentrated in the corners.*

Concentration effect of L_p norms. For any fixed n, the difference between the minimum and that maximum distance under L_p norm between a random reference point Q and a list of n random data points $P_1, \dots, P_n$ become indiscernible compared to the minimum distance as

$$\lim_{d\to\infty} \mathbb{E}\left(\frac{\text{dist}_{\max}(d) - \text{dist}_{\min}(d)}{\text{dist}_{\min}(d)}\right) = 0, \tag{1.2}$$

where $\text{dist}_{\max}(d)$ and $\text{dist}_{\min}(d)$ denote the maximum and the minimum distance the reference point Q and n points $\{P_i\}_{i=1}^n$, respectively, in a d-dimensional space.

Concentration of Gaussian distribution. Let Z be a random vector in $\mathbb{R}^d$ with independent $\mathcal{N}(0,1)$ coordinates. Then

$$P(|\|Z\|_2 - \sqrt{d}| \geq t) \leq 2\exp(-ct^2), \tag{1.3}$$

where $c > 0$ is a constant, $t \geq 0$ and $\|\cdot\|$ is the Euclidean vector norm.

Almost orthogonality of independent vectors. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ be drawn at random with respect to the spherical Gaussian distribution with zero mean and unit variance. Then for every $\epsilon > 0$ and for all $d \geq 1$ the estimate

$$P\left[\left|\langle \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \rangle\right| \geq \epsilon\right] \leq \frac{2/\epsilon + 7}{\sqrt{d}}$$

holds.

From inside out

Non-asymptotic analysis

To illustrate the difference between asymptotic and non-asymptotic analysis, we recall the statement of the weak law of large numbers

Theorem 1.0.1: Weak law of large numbers (WLLN)

Let X be a real random variable with expectation $\mathbb{E}X = p$. Consider an iid sequence $(X_i : i \in \mathbb{N})$ of copies of X. From the running averages:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{for } n \in \mathbb{N}.$$

Then, for each $t > 0$, we have the limit

$$\mathbb{P}\{|\bar{X}_n - p| \geq t\} \to 0 \quad \text{as } n \to \infty.$$

The Weak Law of Large Numbers (WLLN) demonstrates that the sample average converges to the expectation of a random variable as the sample size increases, providing an asymptotic result. However, it does not address the question of how close the sample average is to the expectation for a fixed sample size n. That is precisely the focus of non-asymptotic analysis.

Goals of this course

Concentration Consider a fixed-size sample $(X_1, X_2, \ldots, X_n)$ generated from a distribution. For a measurable function $f : \mathbb{R}^n \to \mathbb{R}$, we define a random variable.

$$Z = f(X_1, X_2, \ldots, X_n).$$

Concentration inequalities provide an upper bound on the probability that the random variable Z deviates from its median $\mathbb{M}Z$ or expectation $\mathbb{E}Z$ by more than a given tolerance $t > 0$. These inequalities are in the forms of

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geq t\} \leq \Box,$$
$$\mathbb{P}\{|Z - \mathbb{M}Z| \geq t\} \leq \triangle.$$

We immediately recognize concentration inequalities tell stories in a non-asymptotic way.

For example, let f be the average function $Z = \frac{1}{n}\sum_{i=1}^{n} X_i$, the concentration inequality can tell us about the probability of the derivation of sample average from its expectation controlled by a given tolerance t, i.e.,

$$\mathbb{P}\{|Z - \mathbb{E}Z| \leq t\} = 1 - \mathbb{P}\{|Z - \mathbb{E}Z| \geq t\} \geq 1 - \square,$$

with fixed sample size.

Suprema The concentration inequalities do not offer any information on the value of $\mathbb{E}f(X_1, X_2, \ldots, X_n)$.. The estimation of this value depends on f. Here we analysis a specific but useful type of f, i.e., the max function.

Specifically, let

$$Z = \sup_{t \in T} X_t,$$

Z is defined as the supremum of a random process $\{X_t\}_{t \in T}$, a family of random variables indexed by a set T. For example, let $X_i \sim \mathcal{N}(0, \sigma^2)$, given an indexed set $T = \{1, 2, \ldots, N\}$,

$$Z = \max_{i=1,2,\ldots,N} X_i.$$

The reason that suprema plays an important role in high-dimensional problem arises in twofold. First, a family of random variables may be interdependent and taking the supreme controls all of them simultaneously. Second, some quantities are naturally mathematically presented in suprema.

Example (Random matrices) Let $\boldsymbol{A} = (A_{ij})$ be a n-by-n random matrix with each of its element is an iid Gaussian random variable. Suppose we want to estimate the largest singular value of $\boldsymbol{A}$, i.e., the spectral norm of $\boldsymbol{A}$, which is mathematically defined as

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{u}, \boldsymbol{v} \in \mathcal{B}} \langle \boldsymbol{u}, \boldsymbol{A}\boldsymbol{v} \rangle,$$

where $\mathcal{B}$ denotes the Euclidean unit ball. Let $X_{\boldsymbol{u},\boldsymbol{v}} := \langle \boldsymbol{u}, \boldsymbol{A}\boldsymbol{v} \rangle$, $\|\boldsymbol{A}\|$ is the supreme of the random process $\{X_{\boldsymbol{u},\boldsymbol{v}}\}_{(\boldsymbol{u},\boldsymbol{v}) \in \mathcal{B} \times \mathcal{B}}$.

Example (Empirical risk minimization) The core issue in machine learning is computing

$$\arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \underbrace{\mathbb{E}[\ell(\boldsymbol{\theta}, X)]}_{\text{generalization error}}.$$

In practice, the distribution of X is unknown, and alternatively we collect an iid sample $(X_1, X_2, \ldots, X_n)$ from the distribution and minimize the empirical risk solving

$$\arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, X_i)}_{\text{empirical risk}},$$

with the hope that

$$\mathbb{E}[\ell(\boldsymbol{\theta}, X)] \approx \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, X_i).$$

Measuring how close is the empirical risk to the generalization error over the the parameter space $\mathbf{\Theta}$ leads the investigate the uniform derivation

$$\sup_{\boldsymbol{\theta}\in\mathbf{\Theta}} \left| \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, X_i) - \mathbb{E}[\ell(\boldsymbol{\theta}, X)] \right|.$$

We will not talk about **Universality** and **Phase transitions**, which you can learn from APC 550 Lecture Notes[1]

Review of Expectation and Variance

Expectation of a random variable X with density $p(x)$ is defined as

$$\mathbb{E}X = \int_{-\infty}^{\infty} xp(x)dx.$$

Generally,

$$\mathbb{E}f(X) = \int_{-\infty}^{\infty} f(x)p(x)dx.$$

If X is a random variable in $\mathbb{R}^n$,

$$\mathbb{E}f(X) = \int_{\mathbb{R}^n} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}.$$

Variance of a random variable is defined as

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2.$$

Properties of expectation and variance

- *Linearity of Expectation* Suppose there are a sequence of random variables $X_1, X_2, \cdots, X_n$, we have
$$\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n$$
- *Association of Expectation of Independent Random Variable* If X_1 and X_2 are independent,
$$\mathbb{E}\left[X_1 X_2\right] = \mathbb{E}X_1 \cdot \mathbb{E}X_2.$$
- *Linearity of Variance for Independent Random Variables* If $X_1, X_2, \cdots, X_n$ are independent,
$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n).$$

Lemma 1.0.2

Let X be a random variable and X' is an independent copy of X, i.e., X and X' are iid. Then we have

$$\mathrm{Var}(X) = \frac{1}{2}\mathbb{E}(X - X')^2.$$

[1] Ramon van Handel. Probability in High Dimension. https://web.math.princeton.edu/ rvan/APC550.pdf

Some classical inequalities

Jensen's inequality For any random variable and a convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X).$$

Lemma 1.0.3: Integral identity

Let X be a non-negative random variable, then

$$\mathbb{E}X = \int_0^\infty P\{X > t\}dt.$$

Proof. Any non-negative real number x can be expressed as

$$x = \int_0^x 1dt = \int_0^\infty \mathbf{1}_{\{x>t\}}dt.$$

By the definition of expectation

$$\begin{aligned} \mathbb{E}X &= \int_0^\infty \int_0^\infty p(x)\mathbf{1}_{\{x>t\}}dtdx \\ &= \int_0^\infty \int_0^\infty p(x)\mathbf{1}_{\{x>t\}}dxdt \\ &= \int_0^\infty \int_t^\infty p(x)dxdt \\ &= \int_0^\infty P\{X \geq t\}dx. \end{aligned}$$

□

Theorem 1.0.4: Markov's inequality

For any non-negative random variable X and $t > 0$, we have

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}.$$

Proof.

$$\begin{aligned} \mathbb{E}X &= \int_0^\infty xp(x)dx = \int_0^t xp(x)dx + \int_t^\infty xp(x)dx \\ &\overset{x\geq 0}{\geq} \int_t^\infty xp(x)dx \overset{x\geq t}{\geq} t\int_t^\infty p(x)dx = t\mathbb{P}\{X \geq t\}. \end{aligned}$$

□

Theorem 1.0.5: Chebyshev's inequality

Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t}.$$

Proof. Denote $Z := (X - \mu)^2$. Since $Z \geq t$ is equivalent to $|X - \mu| \geq t$, we have

$$\mathbb{P}\{|X - \mu| \geq t\} = \mathbb{P}\{Z \geq t\} \leq \frac{\mathbb{E}Z}{t}.$$

The last inequality is a result of Markov's inequality.

Note that $\mathbb{E}Z = \mathbb{E}(X - \mathbb{E}X)^2 = \sigma^2$, we finish the proof. □

Lemma 1.0.6: Tower rule

Let X and Y be two random variables with distributions p_X and p_Y, respectively. Then, we have

$$\mathbb{E}_X[X] = \mathbb{E}_Y\left[\mathbb{E}_X[X|Y]\right].$$

Proof.

$$\begin{aligned}
\mathbb{E}_Y\left[\mathbb{E}_X[X|Y]\right] &= \int_Y p(y)\left[\int_X xp(x|y)\right]dxdy \\
&= \int_X x\left[\int_Y p(y)p(x|y)dy\right]dx \\
&= \int_X xp(x)dx = \mathbb{E}_X[X]
\end{aligned}$$

□

Integration in high-dimensional spaces

For simplicity, we take a bounded function $f : [0,1]^d \to \mathbb{R}$ as an example to calculate the integral

$$\int_{[0,1]^d} f(\boldsymbol{x})d\boldsymbol{x}.$$

Given resolution at $\epsilon > 0$, the grid method takes $(1/\epsilon)^d$ points over the d-dimensional space $[0,1]^d$, which suffers in high dimensionality.

Monte-Carlo's Method. Alternatively, we solve this problem in a probabilistic way. Define a random variable X over the d- dimensional space $[0,1]^d$, with density

$$p(\boldsymbol{x}) = \begin{cases} 1, \boldsymbol{x} \in [0,1]^d \\ 0, \boldsymbol{x} \notin [0,1]^d \end{cases}$$

Obviously, we have

$$\int_{[0,1]^d} f(\boldsymbol{x})d\boldsymbol{x} = \mathbb{E}f(X).$$

Draw a sequence of iid random variables $(X_1, X_2, \dots, X_n)$ from p with sample size n, and take the sample average as

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Hopefully, we want

$$\int_{[0,1]^d} f(\boldsymbol{x})d\boldsymbol{x} = \mathbb{E}f(X) \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Is it a good estimator? We measure the error by

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right)^2 \text{ (Q: why do we take the expectation?)}$$

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right)^2 \\
=&\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} (f(X_i) - \mathbb{E}f(X))\right)^2 \\
=&\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) \qquad \text{(def. of var.)} \\
=&\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(f(X_i)) \qquad (\text{iid of } X_i) \\
=&\frac{1}{n}\mathrm{Var}(f(X)) \\
=&\frac{1}{n}\mathbb{E}\left(f(X) - \mathbb{E}f(X)\right)^2 \leq \frac{1}{n}2M^2 \qquad (\text{suppose } |f| \leq M)
\end{aligned}$$

By concavity of $\sqrt{\cdot}$ and Jensen' inequality, we have

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right| \leq \sqrt{\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right)^2} \leq M\sqrt{\frac{2}{n}}.$$

Hence,

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right| \lesssim \frac{1}{\sqrt{n}},$$

where we use the symbol "$\lesssim$" to hide to quantities independent of n.

The error is **INDEPENDENT OF DIMENSION**. The result is not obtained for free. We have made compromises to derive an upper bound for the error, **ON AVERAGE**.