MATH-SHU 236
High Dimensional Geometry

Shuyang Ling

March 2, 2020

1 High dimensional data

Assume we have n points, denoted by $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, and each point is $\mathbb{R}^d$. In many applications, the dimension d of data is very high (not just in 2D or 3D).

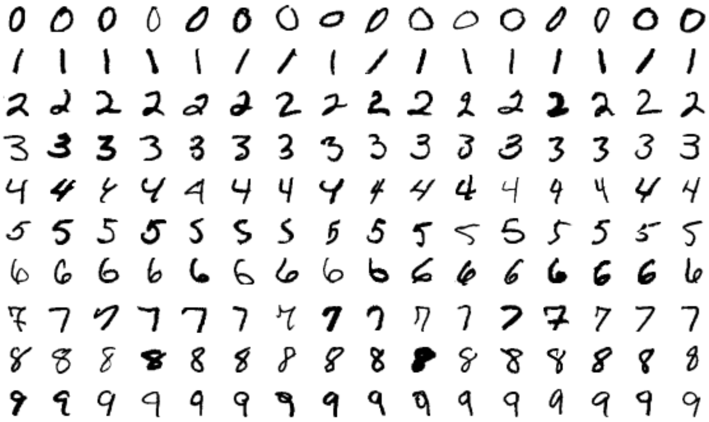
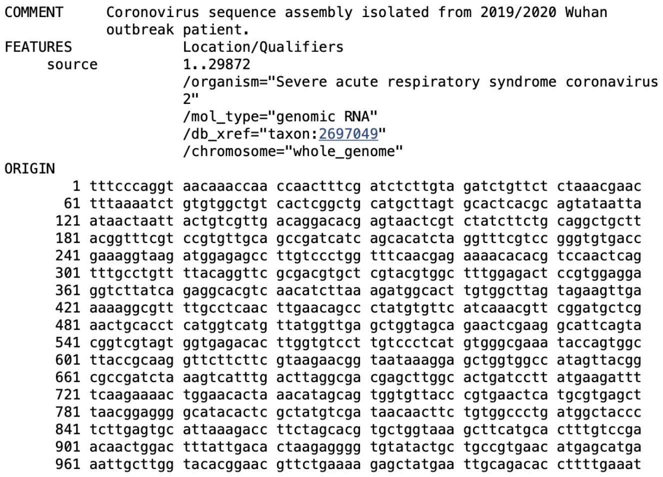


Figure 1: MNIST: Handwritten digits; each with size 28 × 28



```
COMMENT     Coronovirus sequence assembly isolated from 2019/2020 Wuhan
            outbreak patient.
FEATURES             Location/Qualifiers
     source          1..29872
                     /organism="Severe acute respiratory syndrome coronavirus
                     2"
                     /mol_type="genomic RNA"
                     /db_xref="taxon:2697049"
                     /chromosome="whole_genome"
ORIGIN
        1 tttcccaggt aacaaaccaa ccaactttcg atctcttgta gatctgttct ctaaacgaac
       61 tttaaaatct gtgtggctgt cactcggctg catgcttagt gcactcacgc agtataatta
      121 ataactaatt actgtcgttg acaggacacg agtaactcgt ctatcttctg caggctgctt
      181 acggtttcgt ccgtgttgca gccgatcatc agcacatcta ggtttcgtcc gggtgtgacc
      241 gaaaggtaag atggagagcc ttgtccctgg tttcaacgag aaaacacacg tccaactcag
      301 tttgcctgtt ttacaggttc gcgacgtgct cgtacgtggc tttggagact ccgtggagga
      361 ggtcttatca gaggcacgtc aacatcttaa agatggcact tgtggcttag tagaagttga
      421 aaaaggcgtt ttgcctcaac ttgaacagcc ctatgtgttc atcaaacgtt cggatgctcg
      481 aactgcacct catggtcatg ttatggttga gctggtagca gaactcgaag gcattcagta
      541 cggtcgtagt ggtgagacac ttggtgtcct tgtccctcat gtgggcgaaa taccagtggc
      601 ttaccgcaag gttcttcttc gtaagaacgg taataaagga gctggtggcc atagttacgg
      661 cgccgatcta aagtcatttg acttaggcga cgagcttggc actgatcctt atgaagattt
      721 tcaagaaaac tggaacacta aacatagcag tggtgttacc cgtgaactca tgcgtgagct
      781 taacggaggg gcatacactc gctatgtcga taacaacttc tgtggccctg atggctaccc
      841 tcttgagtgc attaaagacc ttctagcacg tgctggtaaa gcttcatgca ctttgtccga
      901 acaactggac tttattgaca ctaagagggg tgtatactgc tgccgtgaac atgagcatga
      961 aattgcttgg tacacggaac gttctgaaaa gagctatgaa ttgcagacac cttttgaaat
```

Figure 2: Whole genome of coronavirus found in Wuhan Seafood Market. Each sample is a character sequence of length around 30000 consisting of A,C,T, and G, the four bases in nucleic acid sequence.

1.1 What can go wrong when data are high dimensional?

Example 1.1 (Integration in high dimension)**.**

Suppose you have a function $f(\boldsymbol{x})$ over the hypercube $[0,1]^d$. Can you find the following integration

$$I = \int_{[0,1]^d} f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}?$$

We approximate the integration by Riemannian sum. Take samples of this function on a uniform grid:

$$\mathcal{U}_n^d = \left\{ \boldsymbol{z} \in \mathbb{R}^d : z_i \in \left\{ 0, \frac{1}{n-1}, \cdots, \frac{n-2}{n-1}, 1 \right\} \right\}$$

Then

$$I \approx \sum_{\boldsymbol{z} \in \mathcal{U}_n^d} f(\boldsymbol{z}) \cdot \frac{1}{(n-1)^d}.$$

Question What is wrong with this approach?

The number of samples is n^d. As the dimension increases with n fixed, the sample number grows exponentially fast! This is also referred to as *the curse of dimensionality.*

Example 1.2 (The geometry of high dimensional ball)**.**

The geometry of ℓ_p norm becomes bizarre when the dimension increases. Define the ℓ_p ball with radius r in $\mathbb{R}^d$:

$$B_p(r) := \left\{ \boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^{d} |x_i|^p \le r^p \right\}.$$

In particular, if $r = 1$, it is the unit ball:

$$B_p(1) := \left\{ \boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^{d} |x_i|^p \le 1 \right\}.$$

All the 2D unit ℓ_p balls are illustrated in Figure 3.

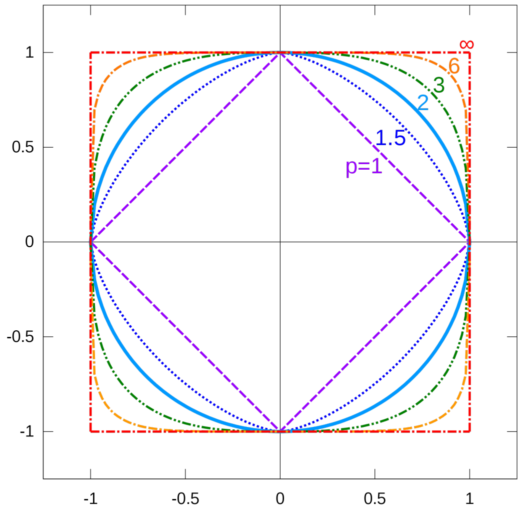


Figure 3: ℓ_p-ball in 2D

Now we will characterize the "shape" of ℓ_1 ball by studying the diameter, volume, and the radius of inscribed ball of unit ℓ_1-norm ball. Here, diameter of a set is defined as the maximal distance of two points in this set.

We first answer the first question:

$$\mathrm{Vol}(B_1(1)) = \frac{2^d}{d!}, \quad \mathrm{Diam}(B_1(1)) = 2.$$

Proof: Let's compute its volume. Note that the set $\{\boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^d |x_i| \leq 1\}$ is symmetric: multiplying each x_i by either 1 or -1 still keeps $\boldsymbol{x}$ in the ℓ_1-norm ball.

$$\begin{aligned}
\int_{\sum_{i=1}^d |x_i| \leq 1} \mathrm{d}\boldsymbol{x} &= 2^d \int_{\sum_{i=1}^d x_i = 1,\ x_i \geq 0} \mathrm{d}\boldsymbol{x} \\
&= 2^d \int_{0 \leq \sum_{i=1}^{d-1} x_i \leq 1, x_i \geq 0} \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \int_0^{1-\sum_{i=1}^{d-1} x_i} \mathrm{d}x_d \\
&= 2^d \int_{0 \leq \sum_{i=1}^{d-1} x_i \leq 1, x_i \geq 0} \left(1 - \sum_{i=1}^{d-1} x_i\right) \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \\
&= 2^d \int_{0 \leq \sum_{i=1}^{d-2} x_i \leq 1, x_i \geq 0} \mathrm{d}x_1 \cdots \mathrm{d}x_{d-2} \int_0^{1-\sum_{i=1}^{d-2} x_i} \left(1 - \sum_{i=1}^{d-1} x_i\right) \mathrm{d}x_{d-1} \\
&= \frac{2^d}{2} \int_{0 \leq \sum_{i=1}^{d-2} x_i \leq 1, x_i \geq 0} \left(1 - \sum_{i=1}^{d-2} x_i\right)^2 \mathrm{d}x_1 \cdots \mathrm{d}x_{d-2} \\
&= \cdots \\
&= \frac{2^d}{(d-1)!} \int_{0 \leq x_1 \leq 1} (1 - x_1)^{d-1} \, \mathrm{d}x_1 = \frac{2^d}{d!}
\end{aligned}$$

For the diameter, we first pick $\boldsymbol{x}_0 = [1, 0, 0, \cdots, 0] \in \mathbb{R}^d$ and $\boldsymbol{y}_0 = [-1, 0, 0, \cdots, 0] \in \mathbb{R}^d$. This implies that

$$\|\boldsymbol{x}_0 - \boldsymbol{y}_0\| = 2, \quad \mathrm{Diam}(B_1(1)) \geq 2.$$

On the other hand, for any $\boldsymbol{x}$ and $\boldsymbol{y}$ in $B_1(r)$, we have

$$\|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_1 \leq \|\boldsymbol{x}\|_1 + \|\boldsymbol{y}\|_1 = 2.$$

Therefore, $\mathrm{Diam}(B_1(1)) = 2$ holds. □

Now we proceed to compute the radius of ℓ_2-norm inscribed ball in the unit ℓ_1-norm. The radius equals

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{x}\|_2 \text{ such that } \|\boldsymbol{x}\|_1 = 1.$$

The minimum is $1/\sqrt{d}$ which is achieved by $\boldsymbol{x} = \frac{1}{d}[1, \cdots, 1]^\top \in \mathbb{R}^d$. Why?

Proof: We first get a lower bound of $\|\boldsymbol{x}\|_2$ and then show this lower bound is attainable. By using Cauchy-Schwarz inequality,

$$1 = \|\boldsymbol{x}\|_1 = \sum_{i=1}^d |x_i| \leq \sqrt{\sum_{i=1}^d 1} \cdot \sqrt{\sum_{i=1}^d x_i^2} = \sqrt{d}\|\boldsymbol{x}\|_2.$$

As a result, we get

$$\|\boldsymbol{x}\|_2 \geq \frac{1}{\sqrt{d}}$$

and the equality is tight if $\boldsymbol{x} = \boldsymbol{x} = \frac{1}{d}[1, \cdots, 1]^\top$. □

What does it imply? The volume is very small for large d and goes to 0 as $d \to \infty$. However, the diameter remains unchanged. Moreover, the radius of inscribed ball is also small, i.e., of order $1/\sqrt{d}$. An illustration of unit ℓ_1-norm is given in Figure 4.

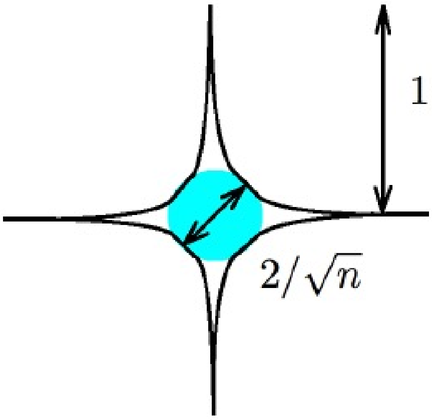


Figure 4: An illustration of ℓ_1-ball in high dimension [?], V. Milman's "hyperbolic" drawings of high dimensional convex sets. It is not accurate in the sense that the ℓ_1-ball is actually convex. The shape of ℓ_1 ball is quite spiky: with small inscribed ball and most volume contained in the spikes.

The volume of ℓ_2 ball in $\mathbb{R}^d$ is

$$V_d(r) = \frac{\pi^{d/2} r^d}{\Gamma(d/2+1)}$$

and the diameter is 2. Note that

$$\lim_{d\to\infty} V_d(r) = 0, \quad \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, \mathrm{d}x$$

for any fixed radius r. In particular, $\Gamma(n) = (n-1)!$ and $\Gamma(1/2) = \sqrt{\pi}$. You can find the proof on most calculus books as well as [?, Chapter 2].

Example 1.3 (Volumes concentrate around the surface)**.**

For any set A in $\mathbb{R}^d$, the volume of $(1-\epsilon)A$ is

$$\mathrm{Vol}((1-\epsilon)A) = (1-\epsilon)^d \,\mathrm{Vol}(A).$$

In other words, the volume between $(1-\epsilon)A$ and A is given by

$$\mathrm{Vol}(A - (1-\epsilon)A) = \mathrm{Vol}(A) - \mathrm{Vol}((1-\epsilon)A) = (1-(1-\epsilon)^d)\,\mathrm{Vol}(A)$$

Now, we can see that if we shrink the set by ϵ, the volume decreases by a factor of $(1-\epsilon)^d$. For large d, the volume around the surface takes the majority of volume in A.

Proof: Let A be any set in $\mathbb{R}^d$. Now we shrink this set by ϵ, i.e.,

$$(1-\epsilon)A = \{(1-\epsilon)\boldsymbol{z} : \boldsymbol{z} \in A \subset \mathbb{R}^d\}.$$

Now we compute the volume of $(1-\epsilon)A$:

$$\mathrm{Vol}((1-\epsilon)A) = \int_{(1-\epsilon)A} \mathrm{d}\boldsymbol{x} = \int_{(1-\epsilon)A} \mathrm{d}x_1 \cdots \mathrm{d}x_d$$

We do a change of variable: $\boldsymbol{z} = \frac{\boldsymbol{x}}{1-\epsilon}$. Then

$$\boldsymbol{x} \in (1-\epsilon)A \Longleftrightarrow \boldsymbol{z} \in A$$

and

$$\mathrm{d}\boldsymbol{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_d = (1-\epsilon)^d \,\mathrm{d}z_1 \cdots \mathrm{d}z_d = (1-\epsilon)^d \,\mathrm{d}\boldsymbol{z}.$$

As a result, it holds

$$\int_{(1-\epsilon)A} \mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{z}\in A} (1-\epsilon)^d \,\mathrm{d}\boldsymbol{z} = (1-\epsilon)^d \,\mathrm{Vol}(A)$$

□

2 Blessings of high dimensional data

However, not every aspect of high dimensional data is pessimistic. In fact, when data become high dimensional, additional "structure" and phenomena show up.

2.1 Concentration of measure

In the analysis of high dimensional data, probability theory is one powerful tool. One fundamental phenomena in high dimensional probability is called *concentration of measure* which has found numerous applications including learning theory, high dimensional statistical inference, and signal processing. Informally speaking, concentration of measure refers to "a random variable that depends (in a "smooth" way) on the influence of many independent variables (but not too much on any of them) is essentially constant", quoted from [**?**]. See [**?**, Chapter 2 and 3] and [**?**, Chapter 2 and 3] for more details on concentration of measure.

Here we give an example. Let $\boldsymbol{x}$ be a vector in $\mathbb{R}^d$ whose each coordinate is an i.i.d. symmetric Bernoulli random variable $\mathbb{P}(x_i = 1) = \mathbb{P}(x_i = -1) = 1/2$. Now we consider the sample average of its coordinates, i.e.,

$$f(\boldsymbol{x}) := \frac{1}{d}\sum_{i=1}^{d} x_i = \frac{1}{d}\langle \boldsymbol{x}, \mathbf{1}_d\rangle = \left\langle \frac{1}{\sqrt{d}}\boldsymbol{x}, \frac{1}{\sqrt{d}}\mathbf{1}_d \right\rangle.$$

It is obvious that $f(\boldsymbol{x})$ is the cosine of angle between $\boldsymbol{x}$ and a constant vector.

Note that the max/min of $f(\boldsymbol{x})$ is 1 and -1 respectively. However, the probability of attaining the maximum and minimum is rather small:

$$\mathbb{P}(f(\boldsymbol{x}) = 1) = \mathbb{P}(f(\boldsymbol{x}) = -1) = 1/2^d$$

On the other hand, we will later show that with very large probability (close to 1), $f(\boldsymbol{x}) \approx 0$ holds by bounding the tail probability of $\mathbb{P}(|f(\boldsymbol{x})| \geq \epsilon)$. It means most of vectors are almost perpendicular to a constant vector. Also, we can show that any random vectors in high dimensional space are near-orthogonal!

2.2 Data are not as "dense" as you think

Recall the MNIST dataset. In the training set, there are 60000 pictures of size 28 × 28. Are these pictures saturated in the whole Euclidean space of $784 = 28 \times 28$? In fact, these pictures are intrinsically low-dimensional. For example, we apply 2D Fourier transform to each image by treat each image as a matrix. This is easily made possible by calling `fft2`.

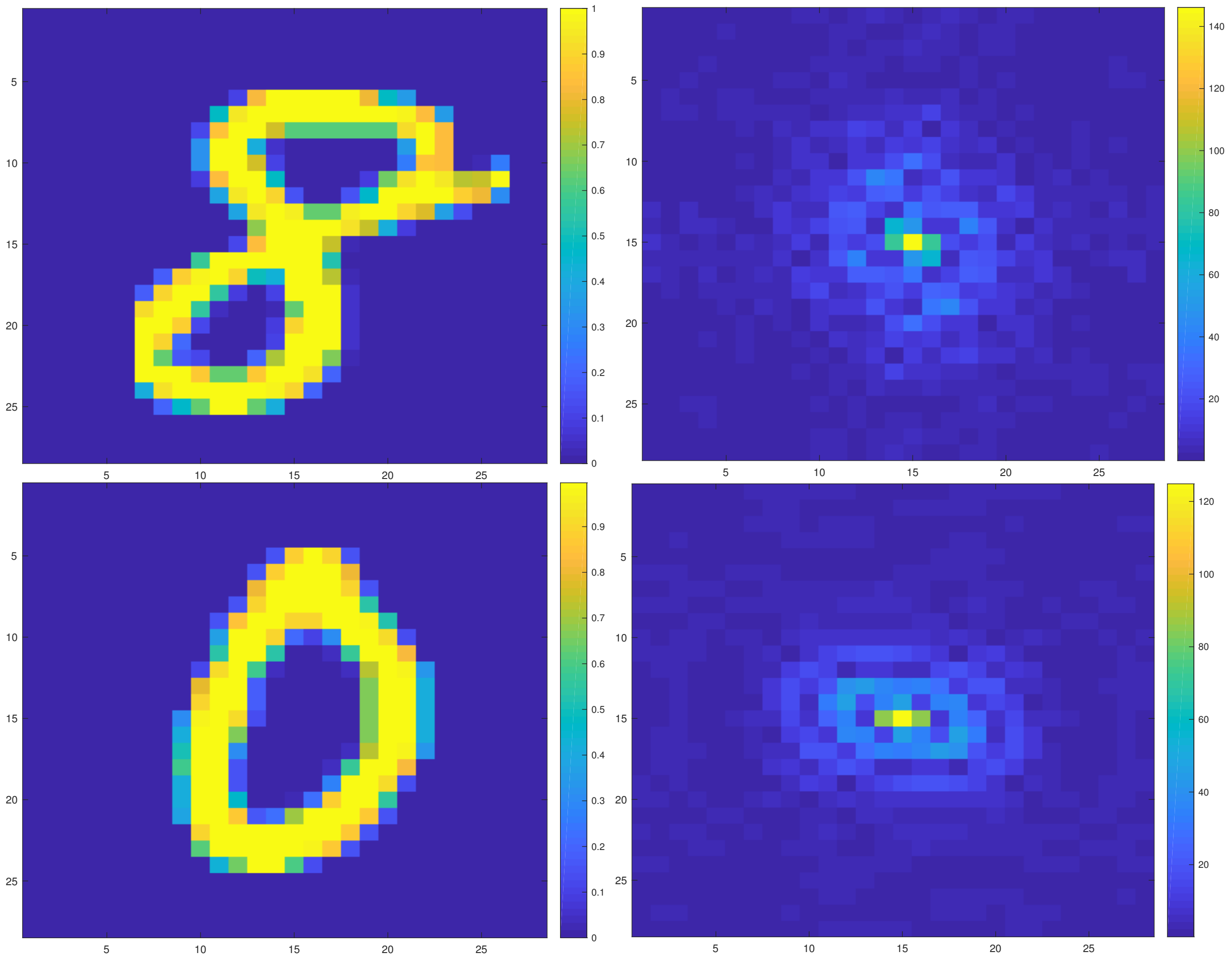


Figure 5: Two images from MNIST and their magnitude of Fourier transform.

From Figure 5, we can see that despite they are different images, most of their Fourier coefficients are small and thus most of information is contained in those large coefficients. In other words, these images could be well-approximated by a small amount of Fourier coefficients, i.e., a low dimensional representation of images.

In fact, Fourier transform is one important approach to find a concise representation of images. There are many other tools such as wavelets, curvelets, shearlets in computational harmonic analysis, see [**?**] for details.

Even for data without structure, the data are not "dense" in the whole space. How many vertices are there in ℓ_∞ unit ball, i.e., the hypercube in $\mathbb{R}^d$?

- $d = 1$: 1 and -1
- $d = 2$: $(1, 1)$, $(-1, 1)$, $(1, -1)$, and $(-1, -1)$
- $d = 3$: $(1, 1, 1)$, $(1, 1, -1)$, $(1, -1, 1)$,$(1, -1, -1)$, $(-1, 1, 1)$, $(-1, 1, -1)$, $(-1, -1, 1)$, and $(-1, -1, -1)$.

- For general d, 2^d vertices!

Suppose n data points are given, they only occupy at most $\log_2(n)$ vertices. Later, we will talk about Johnson-Lindenstrauss Lemma [**?**, Chapter 5.3] and [**?**, Chapter 2.7]. It says that there exists a mapping which maps any set of n points in $\mathbb{R}^d$ to $\mathbb{R}^k$ with $k = \Omega(\log n)$ and their pairwise ℓ_2 distance is well preserved. This mapping could be constructed via random projection.

References

[1] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science.* Cambridge University Press, 2020.

[2] S. Mallat. *A Wavelet Tour of Signal Processing.* Elsevier, 1999.

[3] M. Talagrand. A new look at independence. *The Annals of Probability*, pages 1–34, 1996.

[4] R. Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, A Renaissance*, pages 3–66. Springer, 2015.

[5] R. Vershynin. *High-dimensional Probability: An introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

[6] M. J. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.