We make a few more observations that will hint at things to come in subsequent topics.

### 3.1 Matrix form of \( k \)-means clustering

The \( k \)-means clustering objective can be written in matrix form. We will need a notion of matrix norm. A natural way to define a norm for matrices is to notice that an \( n \times m \) matrix \( A \) can be thought of as an \( nm \) vector, with one element for each entry of \( A \). Indeed, addition and scalar multiplication work exactly in the same way. Hence, we can define the 2-norm of a matrix in terms of the sum of its squared entries.

**Definition (Frobenius Norm):** The Frobenius norm of an \( n \times m \) matrix \( A \in \mathbb{R}^{n \times m} \) is defined as

\[
\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}.
\]

As we indicated before, for a collection of \( n \) data vectors \( x_1, \ldots, x_n \) in \( \mathbb{R}^d \), it is often convenient to stack them up into a matrix

\[
X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}.
\]
We can do the same with cluster representatives. Given \( \mu_1, \ldots, \mu_k \) also in \( \mathbb{R}^d \), we form the matrix

\[
U = \begin{bmatrix}
\mu_{11}^T & \mu_{12}^T & \cdots & \mu_{1d}^T \\
\mu_{21}^T & \mu_{22}^T & \cdots & \mu_{2d}^T \\
\vdots & \vdots & \ddots & \vdots \\
\mu_{k1}^T & \mu_{k2}^T & \cdots & \mu_{kd}^T
\end{bmatrix}.
\]

Perhaps less obviously, cluster assignments can also be encoded in matrix form. Recall that, given a partition \( C_1, \ldots, C_k \) of \( [n] \), we define \( c(i) = j \) if \( i \in C_j \). For \( i = 1, \ldots, n \) and \( j = 1, \ldots, k \), set \( z_{ij} = 1 \) if \( c(i) = j \) and 0 otherwise, and let \( Z \) be the \( n \times k \) matrix with entries \( z_{ij} \). That is, row \( i \) has exactly one entry with value 1, corresponding to the assigned cluster \( c(i) \) of data point \( x_i \), and all other entries 0.

With this notation, the representative of the cluster assigned to data point \( x_i \) is obtained through the matrix product

\[
\mu_{c(i)}^T = \sum_{j=1}^k z_{ij} \mu_j^T = (ZU)_{i,:}
\]

So

\[
G(C_1, \ldots, C_k; \mu_1, \ldots, \mu_k) = \sum_{i=1}^n \|x_i - \mu_{c(i)}\|^2
\]

\[
= \sum_{i=1}^n \sum_{\ell=1}^d (x_{i,\ell} - (ZU)_{i,\ell})^2
\]

\[
= \|X - ZU\|_F^2,
\]

where we used the definition of the Frobenius norm.

In other words, minimizing the \( k \)-means objective is equivalent to finding a matrix factorization of the form \( ZU \) that is a good fit to the data matrix \( X \) in Frobenius form. This formulation expresses in a more compact form the idea of representing \( X \) as a combination of a small number of representatives. Matrix factorization will come back repeatedly in this course.

**NUMERICAL CORNER** In Julia, the Frobenius norm of a matrix can be computed using the function `norm` (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.norm).

In [1]: # Julia version: 1.5.1
using Plots, LinearAlgebra, Statistics
To start explaining the quote above, we consider a simple experiment. Let $C = [-1/2, 1/2]^d$ be the $d$-cube with side lengths 1 centered at the origin and let $B = \{x \in \mathbb{R}^d : \|x\| \leq 1/2\}$ be the inscribed $d$-ball. In $d = 2$ dimensions:

$$\begin{align*}
\text{a - Side of square} \\
r - \text{Radius of circle}
\end{align*}$$

(Source: https://www.geeksforgeeks.org/program-to-calculate-area-of-an-circle-inscribed-in-a-square/)

Now pick a point $\mathbf{X}$ uniformly at random in $C$. What is the probability that it falls in $B$?

To generate $\mathbf{X}$, we pick $d$ independent random variables $X_1, \ldots, X_d \sim \text{U}[-1/2, 1/2]$, and form the vector $\mathbf{X} = (X_1, \ldots, X_d)$. Indeed, the PDF of $\mathbf{X}$ is then $f_{\mathbf{X}}(\mathbf{x}) = 1^d = 1$ if $\mathbf{x} \in C$ and 0 otherwise.

The event we are interested in is $A = \{\|\mathbf{X}\|\leq 1/2\}$. The uniform distribution over the set $C$ has the property that $\Pr[A]$ is the volume of $A$ divided by the volume of $C$. In this case, the volume of $C$ is $1^d = 1$ and the volume of $A$ has an explicit formula (https://en.wikipedia.org/wiki/Volume_of_an_n-ball).
This leads to the following surprising fact:

**Theorem (High-dimensional Cube)** Let \( B = \{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1/2 \} \) and \( C = [-1/2, 1/2]^d \). Pick \( \mathbf{X} \sim \mathbf{U}[C] \). Then, as \( d \to +\infty \),
\[
\mathbb{P}[\mathbf{X} \in B] \to 0.
\]

In words, in high dimension if one picks a point at random from the cube, it is unlikely to be close to the origin. Instead it is likely to be in the corners. A geometric interpretation is that a high-dimensional cube is a bit like a spiky ball. A visualization of this theorem:

![Visualization](https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/)

We give a proof based on Chebyshev’s inequality. It has the advantage of providing some insight into this counter-intuitive phenomenon by linking it to the concentration of sums of independent random variables, in this case the squared norm of \( \mathbf{X} \).

**Proof idea:** We think of \( \|\mathbf{X}\|^2 \) as a sum of independent random variables and apply Chebyshev’s inequality. It implies that the norm of \( \mathbf{X} \) is concentrated around its mean, which grows like \( \sqrt{d} \). The latter is larger than \( 1/2 \) for \( d \) large.

**Proof:** Write out \( \|\mathbf{X}\|^2 = \sum_{i=1}^{d} X_i^2 \). Using linearity of expectation and the fact that the \( X_i \)'s are independent, we get
\[
\mathbb{E} \left[ \|\mathbf{X}\|^2 \right] = \sum_{i=1}^{d} \mathbb{E}[X_i^2] = d \mathbb{E}[X_1^2]
\]
and
\[
\text{Var} \left[ \|\mathbf{X}\|^2 \right] = \sum_{i=1}^{d} \text{Var}[X_i^2] = d \text{Var}[X_1^2].
\]
We bound the probability of interest as follows. We first square the norm and center around the mean:

$$\mathbb{P}[\|X\| \leq 1/2] = \mathbb{P}[\|X\|^2 \leq 1/4]$$

$$= \mathbb{P}[\|X\|^2 - \mathbb{E}[\|X\|^2] \leq 1/4 - d\mathbb{E}[X_1^2]].$$

Now notice that $\mathbb{E}[X_1^2] > 0$ does not depend on $d$. Take $d$ large enough that $d\mathbb{E}[X_1^2] > 1/4$. We then use the following fact: if $\alpha = d\mathbb{E}[X_1^2] - 1/4 > 0$ and $Z = \|X\|^2 - \mathbb{E}[\|X\|^2]$, we can write by monotonicity and the definition of the absolute value

$$\mathbb{P}[Z \leq -\alpha] \leq \mathbb{P}[Z \leq -\alpha \text{ or } Z \geq \alpha] = \mathbb{P}[|Z| \geq \alpha].$$

We arrive at

$$\mathbb{P}[\|X\| \leq 1/2] \leq \mathbb{P}[\|X\|^2 - \mathbb{E}[\|X\|^2] \geq d\mathbb{E}[X_1^2] - 1/4].$$

We can now apply Chebyshev’s inequality to the right-hand side, which gives

$$\mathbb{P}[\|X\| \leq 1/2] \leq \frac{\mathbb{V}ar[\|X\|^2]}{(d\mathbb{E}[X_1^2] - 1/4)^2}$$

$$= \frac{d\mathbb{V}ar[X_1^2]}{(d\mathbb{E}[X_1^2] - 1/4)^2}$$

$$= \frac{1}{d} \cdot \frac{\mathbb{V}ar[X_1^2]}{(\mathbb{E}[X_1^2] - 1/(4d))^2}. $$

Again, $\mathbb{V}ar[X_1^2]$ does not depend on $d$. So the right-hand side goes to 0 as $d \to +\infty$, as claimed. □

We will see later in the course that this high-dimensional phenomenon has implications for data science problems. It is behind what is referred to as the Curse of Dimensionality (https://en.wikipedia.org/wiki/Curse_of_dimensionality).

**NUMERICAL CORNER** We can check the theorem in a simulation. Here we pick $n$ points uniformly at random in the $d$-cube $C$, for a range of dimensions $[d_{\min}, d_{\max}]$. We then plot the frequency of landing in the inscribed $d$-ball $B$ and see that it rapidly converges to 0. Alternatively, we could just plot the formula for the volume of $B$. But knowing how to do simulations is useful in situations where explicit formulas are unavailable or intractable.

```plaintext
In [4]: function highdim_cube(dmax, n)
  in_ball = zeros(Float64, dmax) # in-ball freq
  for d=1:dmax # for each dimension
    in_ball[d] = mean(((norm(rand(d) .- 1/2) < 1/2) for i=1:n))
  end
  plot(1:dmax, in_ball,
       legend=false, markershape=:diamond, xlabel="dim", ylabel="in-bal 1 freq")
end

Out[4]: highdim_cube (generic function with 1 method)
```
3.2.2 Gaussians in high dimension [optional]

In this optional section, we turn our attention to the Gaussian (or Normal) distribution (https://en.wikipedia.org/wiki/Normal_distribution) and its behavior in high dimension. Using Chebyshev’s inequality, we show that a standard Normal vector has the following counter-intuitive property in high dimension: a typical draw has 2-norm that is highly likely to be around $\sqrt{d}$. Visually, when $d$ is large, the joint PDF looks something like this:
This is unexpected because the joint PDF is maximized at $\mathbf{x} = \mathbf{0}$ for all $d$ (including $d = 1$ as can be seen in the figure above). But the rough intuition is the following: (1) there is only "one way" to obtain $\|\mathbf{X}\|^2 = 0$ -- every coordinate must be 0 by the point-separating property of the 2-norm; (2) on the other hand, there are "many ways" to obtain $\|\mathbf{X}\|^2 = \sqrt{d}$ -- and that compensates for the lower density.
Theorem (High-dimensional Gaussians) Let $\mathbf{X}$ be a standard Normal $d$-vector. Then, for any $\varepsilon > 0$,
\[
P \left[ \|\mathbf{X}\| \notin (\sqrt{d(1 - \varepsilon)}, \sqrt{d(1 + \varepsilon)}) \right] \to 0
\]
as $d \to +\infty$.

Proof idea: We apply Chebyshev’s inequality to the squared norm, which is a sum of independent random variables.

Proof: Let $Z = \|\mathbf{X}\|^2 = \sum_{i=1}^{d} X_i^2$ and notice that, by definition, it is a sum of independent random variables. Appealing to the expectation and variance formulas from the previous sections:
\[
\mathbb{E}[\|\mathbf{X}\|^2] = d \mathbb{E}[X_1^2] = d \text{Var}[X_1] = d
\]
and
\[
\text{Var}[\|\mathbf{X}\|^2] = d \text{Var}[X_1^2]
\]
where $\text{Var}[X_1^2]$ does not depend on $d$. By Chebyshev’s inequality
\[
P \left[ \|\mathbf{X}\| \notin (d(1 - \varepsilon), d(1 + \varepsilon)) \right] = P[\|\mathbf{X}\|^2 - d \geq \varepsilon d] \leq \frac{d \text{Var}[X_1^2]}{\varepsilon^2 d^2} = \frac{\text{Var}[X_1^2]}{\varepsilon^2}.
\]
Taking a square root inside the probability on the leftmost side and taking a limit as $d \to +\infty$ on the rightmost side gives the claim. $\square$

NUMERICAL CORNER We check our claim in a simulation. We generate standard Normal $d$-vectors using the `randn` function and plot the histogram of their 2-norm.

```julia
In [6]: function normal_shell(d, n)
    one_sample_norm = [norm(randn(d)) for i=1:n]
    histogram(one_sample_norm,
               legend=false, xlims=(0,maximum(one_sample_norm)), nbin=20)
end
Out[6]: normal_shell (generic function with 1 method)
```
In [7]: normal_shell(1, 10000)

Out[7]:

In higher dimension:

In [8]: normal_shell(100, 10000)

Out[8]:
Applying Chebyshev’s inequality to sums of independent random variables has useful statistical implications: it shows that, with a large enough number of samples \( n \), the sample mean is close to the population mean. Hence it allows us to infer properties of a population from samples. Interestingly, one can apply a similar argument to a different asymptotic regime: the limit of large dimension \( d \). But as we will see in this section, the statistical implications are quite different.

3.2.1 High-dimensional cube