CSE4088 - Introduction to Machine Learning
HW1 report

Zeynep Kumralbaş 150115051

October 18, 2019

1 Vectors and Matrices

1. $\begin{pmatrix} 9 & 8 \end{pmatrix} \cdot \begin{pmatrix} 7 \\ 6 \end{pmatrix} = 9 \times 7 + 8 \times 6 = 63 + 48 = 111$

2. $\begin{pmatrix} 9 & 8 \\ 7 & 6 \end{pmatrix} \cdot \begin{pmatrix} 9 \\ 8 \end{pmatrix} = \begin{pmatrix} 9 * 9 + 8 * 8 \\ 7 * 9 + 6 * 8 \end{pmatrix} = \begin{pmatrix} 81 + 64 \\ 63 + 48 \end{pmatrix} = \begin{pmatrix} 145 \\ 111 \end{pmatrix}$

3. $\left|X\right| = 9 \times 6 - 7 \times 8 = 54 - 56 =$ -2
   Since $\left|X\right|$ is not 0, the matrix X is invertible.
   $X^{-1} = \frac{1}{-2} \cdot \begin{pmatrix} 6 & -8 \\ -7 & 9 \end{pmatrix} = \begin{pmatrix} -3 & 4 \\ 7/2 & -9/2 \end{pmatrix}$

4. Reduced echolon form of $X = \begin{pmatrix} 1 & 8/9 \\ 7 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 8/9 \\ 0 & 370 \end{pmatrix}$
   Rank of X is the number of linearly independent column vectors.
   $\vec{r1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \vec{r2} = \begin{pmatrix} 8/9 \\ 370 \end{pmatrix}$
   $\vec{r1}$ and $\vec{r2}$ are linearly independent. Rank$(X) = 2$

2 Calculus

1. $\frac{dy}{dx} = 12x^2 - 2x$

2. $\frac{\partial y}{\partial x} = \tan(z)6zx^{6z-1}\text{-}\frac{[(7x+z)x^{-4}]'}{(7x+z)x^{-4}} = \tan(z)6zx^{6z-1}\text{-}\frac{7x^{-4}+(7x+z)-4x^{-5}}{(7x+z)x^{-4}}$
   $= \tan(z)6zx^{6z-1}\text{-}\frac{7x^{-4}-28x^{-4}-4x^{-5}z}{(7x+z)x^{-4}} = \tan(z)6zx^{6z-1}\text{-}\frac{-21-\frac{4z}{x}}{7x+z}$
   $= \tan(z)6zx^{6z-1} + \frac{21+\frac{4z}{x}}{7x+z}$

3 Probability and Statistics

1. $\bar{X} = \frac{0+1+1+0+0+1+1}{7} = \frac{4}{7} = 0.57$

2. $s^2 = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}$ $s^2 = \frac{(0-0.57)^2+(1-0.57)^2+(1-0.57)^2+(0-0.57)^2+(0-0.57)^2+(1-0.57)^2+(1-0.57)^2}{7-1}$
   $s^2 = \frac{1.7143}{6} = 0.2857$

3. $p(x=1) = 0.7$ $p(x=0) = 0.3$
   $p(S) = 0.3 \times 0.7 \times 0.7 \times 0.3 \times 0.3 \times 0.7 \times 0.7 = (0.3)^3 \times (0.7)^4 =$
   0.027×0.2401
   $p(S) = 0.0064827$

4. $p(x=1) = \theta$ $p(x=0) = 1-\theta$
   $p(S) = (1-\theta)^3 \times \theta^4$
   Maximize $p(S)$. $p'(S) = 0$
   $p'(S) = 3(1-\theta)^2(-1)\theta^4 + (1-\theta)^3 4\theta^3$
   $-3(1-\theta)^2\theta^4 + 4(1-\theta)^3\theta^3 = 0$
   $4(1-\theta)^3\theta^3 = 3(1-\theta)^2\theta^4$
   $4(1-\theta) = 3\theta$
   $4 - 4\theta = 3\theta$
   $\theta = \frac{4}{7}$
   If $p(x=1) = \frac{4}{7}$ and $p(x=0) = \frac{3}{7}$, the probability of sample S would be maximized.

5. (a) The probability of A=0 and B=0 is $P(A = 0, B = 0) = 0.1$

   (b) $P(A = 1) = 0.2 + 0.3 = 0.5$

   (c) $P(A = 0|B = 1) = \frac{P(A=0,B=1)}{P(B=1} = \frac{0.4}{0.7} = \frac{4}{7}$

   (d) $P(A = 0 \vee B = 0) = P(A = 0)+P(B = 0)-P(A=0 \cap B=0) = 0.1 + 0.4 + 0.1 + 0.2 - 0.1 = 0.7$

4 Big-O Notation

1. $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \lim_{n\to\infty} \frac{n/2}{\log_2(n)} = \lim_{n\to\infty} \frac{n}{2} \frac{\ln 2}{\ln(n)} = \frac{\infty}{\infty}$
   $\lim_{n\to\infty} \frac{1}{\frac{1}{n}} = \lim_{n\to\infty} n = \infty$
   $g(n) \leq f(n)$ So, $g(n) = O(f(n))$ is true.

2. $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \lim_{n\to\infty} \frac{\ln n}{\log_2(n)} = \lim_{n\to\infty} \frac{\ln n)}{\ln n} \ln 2 = \lim_{n\to\infty} \ln 2 = \ln 2$
   $g(n) = O(f(n))$ and $f(n) = O(g(n))$ are true.

3. $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \lim_{n\to\infty} \frac{n^{100}}{100^n} = \frac{\infty}{\infty}$
   $\lim_{n\to\infty} \frac{100n^{99}}{\ln 100 \times 100^n} = \lim_{n\to\infty} \frac{n^{99}}{100^n} = 0$
   $f(n) \leq g(n)$ So, $f(n) = O(g(n))$ is true.

5 Algorithm

Assume the sorted array A=[-10 -10 ..... 0 ..... 10 10] with n integers. To find the location of 0, we can use binary search algorithm.
Steps of the algorithm:

1. Divide the array into 2 subarrays.

2. If $A[i] \geq 0$, continue with the subarray $A[0 : i - 1]$.

3. If $A[i] \leq 0$, continue with the subarray $A[i + 1 : n]$.

4. If $A[i] = 0$, the location of 0 is found, return i.

We search 0 by dividing the array into 2 subarrays and continue to search in one subarray. Therefore, the running time of the algorithm is $O(\log n)$.

6 Probability and Random Variables

6.1 Probability

1. Statement: $P(A|B)P(B) = P(B|A)P(A)$
   $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ $\quad$ $P(A|B)P(B) = P(B|A)P(A)$
   Statement is true.

2. Statement: $P(A \cup B) = P(A) + P(B) - P(A|B)$
   $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   $P(A|B) = \frac{P(A \cap B)}{P(B)}$ $\quad$ $P(A|B) \neq P(A \cap B)$
   Statement is false.

3. Statement: $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B \cup C)$
   $P(A|B \cup C)P(B \cup C) = P(A \cap (B \cup C))$
   $(B \cup C) = D \quad P(D) > 0$ Then, we are looking if $\frac{(A \cup D)}{P(D)} \geq P(A \cap D)$ is true.
   We know that $P(A \cap D) \geq P(A \cup D)$ from the properties of sets and $P(D) \leq 1$. Therefore, $\frac{(A \cup D)}{P(D)} \geq P(A \cap D)$ is true, meaning that the statement $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B \cup C)$ is true.

4. Statement: $P(B|A^C) + P(B|A) = 1$
   We know that $P(A) + P(A^C) = 1$. Since $P(B)>0$ and $P(A^C)>0$, $P(B|A^C) < P(A^C)$ and $P(B|A) < P(A)$.
   $P(B|A^C) + P(B|A) < P(A^C) + P(A)$ $\quad$ $P(B|A^C) + P(B|A) < 1$
   So, the statement is false.

5. Statement: If $P(\bigcap_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$, then $\{A_i\}_{i=1}^{n}$ are mutually independent.

If events A and B are mutually independent, $P(A \cap B) = P(A)P(B)$. If $P(\bigcap_{i=1}^{n} A_i = \prod_{i=1}^{n} P(A_i)$, then we can say $\{A_i\}_{i=1}^{n}$ are mutually independent. So, the statement is false.

6.2 Discrete and Continuous Distributions

1. (a)Laplace - (h)
   (b)Multinomial - (i)
   (c)Poisson - (l)
   (d)Dirichlet - (k)
   (e)Gamma - (j)

6.3 Mean and Variance

1. (a) Mean of the random variable: np
   (b) Variance of the random variable: $np(1-p)$

2. (a) $\mathbb{E}[3X]$=$3\mathbb{E}[X] = 3 \times 1 = 3$
   (b) $\text{Var}(3X) = \mathbb{E}(3X - \mathbb{E}(3X))^2 = \mathbb{E}(9X^2 - 6X\mathbb{E}(3X) + \mathbb{E}(3X)\mathbb{E}(3X))$
   $= 9\mathbb{E}(X^2) - 6\mathbb{E}(X)\mathbb{E}(3X) + \mathbb{E}(3X)\mathbb{E}(3X)$
   $= 9\mathbb{E}(X^2) - 18\mathbb{E}(X)\mathbb{E}(X) + 9\mathbb{E}(X)\mathbb{E}(X)$
   $= 9\mathbb{E}(X^2) - 9(\mathbb{E}(X^2))^2 = 9((\mathbb{E}(X^2) - (\mathbb{E}(X))^2) = 9\text{Var}(X)$
   $= 9 \times 1 = 9$
   (c) $\text{Var}(X+3) = \mathbb{E}(X + 3 - \mathbb{E}(X+3))^2 = \mathbb{E}(X + 3 - \mathbb{E}(X) - 3)^2$
   $= \mathbb{E}(X - \mathbb{E}(X))^2 = \text{Var}(X) = 1$

6.4 Mutual and Conditional Independence

1. $\mathbb{E}[X] = \sum_{x \in X} xp(x)$
   $\mathbb{E}[XY] = \sum_{x,y \in X,Y} xyp(x,y)$
   If x and y are independent, $p(x,y) = p(x)p(y)$
   $\mathbb{E}[XY] = \sum_{x \in X} \sum_{y \in Y} xyp(x)p(y) = \sum_{x \in X} xp(x) \sum_{y \in Y} yp(y) = \mathbb{E}[X]\mathbb{E}[Y]$

2. $\mathrm{Var}(X) = \mathbb{E}(X - EX)^2$
   $\mathrm{Var}(X+Y) = \mathbb{E}(X+Y-E(X+Y))^2 = \mathbb{E}(X+Y-\mathbb{E}(X)-\mathbb{E}(Y))^2$
   $= \mathbb{E}((X+Y)-(\mathbb{E}(X)+\mathbb{E}(Y)))^2$
   $= \mathbb{E}((X+Y)^2 - 2(X+Y)(\mathbb{E}(X)+\mathbb{E}(Y)) + (\mathbb{E}(X)+\mathbb{E}(Y))^2)$
   $= \mathbb{E}((X+Y)^2) - 2(\mathbb{E}(X)+\mathbb{E}(Y))(\mathbb{E}(X)+\mathbb{E}(Y)) + \mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2$
   $= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - 2(EX)^2 - 4\mathbb{E}(X)\mathbb{E}(Y) - 2(EY)^2 + (EX)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + (EY)^2$
   $= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - 2(\mathbb{E}(X))^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - (\mathbb{E}(Y))^2$
   $= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 + \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 + 2[\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)]$
   $= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$
   $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
   We know that for independent random variables X and Y, $\mathbb{E}[X][Y] = \mathbb{E}[X]\mathbb{E}[Y]$ from the exercise 1. Then, $\mathrm{Cov}(X,Y) = 0$
   $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

3. In the first case, the result of the first die will not tell us something about the result of the second die.
   In the second case, the result of the second die is dependent of the first die. We know that the result of the first die is 1. The result of the second die should be an odd number(1, 3, 5) to make the sum of the two results even.

6.5 Law of Large Numbers and the Central Limit Theorem

1. Outcomes of flipping two coins can be "hh","ht","tt","th"(h:head,t:tail). The probability of observing two heads is $P(hh) = 0.25$. From the Bernoulli distribution, $\mathbb{E}(\bar{X}_n) = \mu = 0.25$, $\sigma^2 = p(1-p) = 0.25 \times 0.75 = 0.1875$.
   $\sigma^2 = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{0.1875}{40000} = 0.0000046875 \qquad \sigma = 0.00217.$

$\mu - 3\sigma < \bar{X} < \mu + 3\sigma \quad 0.24349 < \bar{X} < 0.25651$

9739.6 <number of times the result was two heads< 10260.4

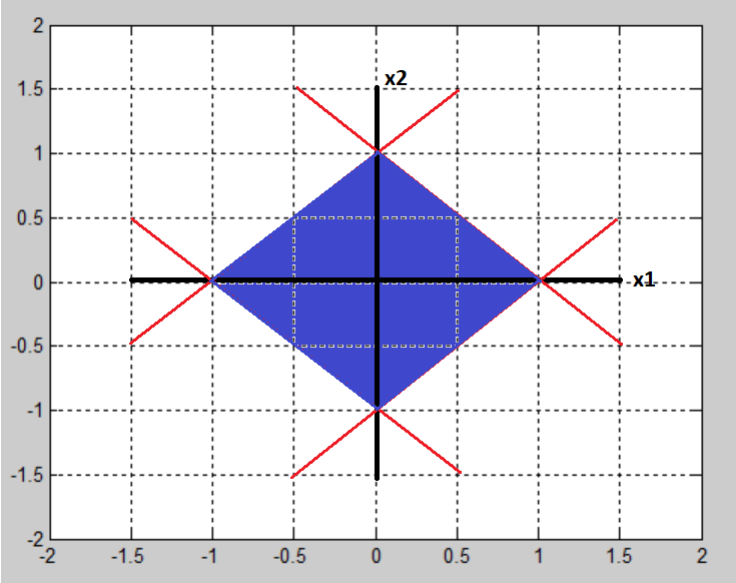
From the law of large numbers, as $n \to \infty, \mathrm{Var}(\bar{X}_n) = 0$.

2. X_i is a normal distribution with $\mu = 0$ and $\sigma^2 = 1$. From the law of large numbers, $\mathbb{E}(\sqrt{n}\bar{X}) = \mu = 0$. $\mathrm{Var}(\sqrt{\mathrm{n}}\bar{\mathrm{X}}) = (\sqrt{n})^2 \frac{\sigma^2}{n} = \sigma^2 = 1$. So, the distribution of $\bar{X}$ satisfies $\sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$.

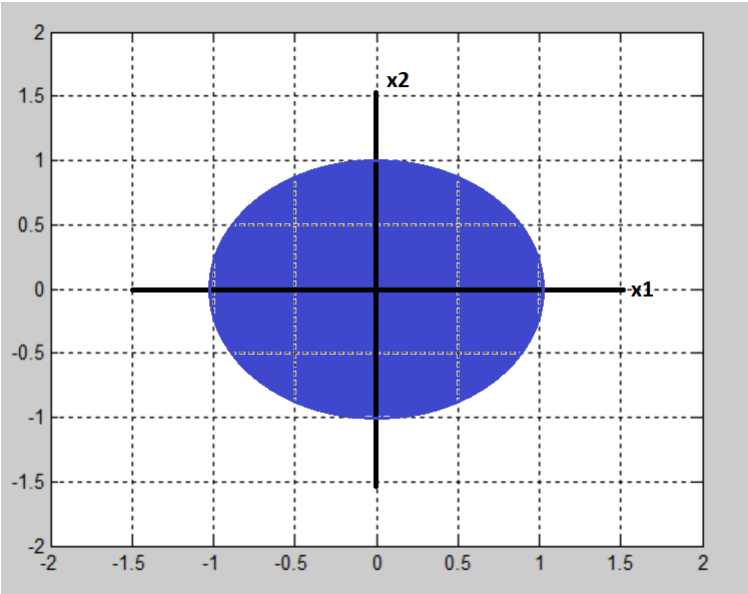
7 Linear Algebra

7.1 Norms

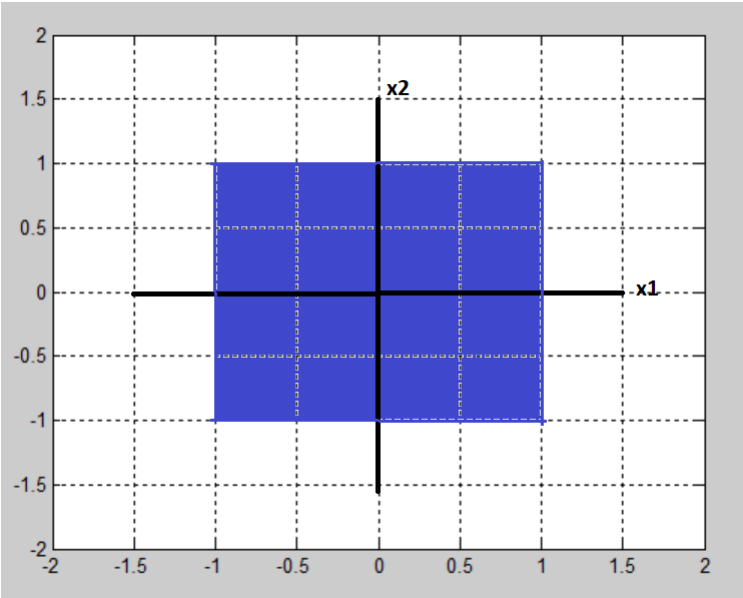
1. $||x||_1 \leq 1$



2. $||x||_2 \leq 1$



3. $||x||_\infty \leq 1$



7.2 Geometry

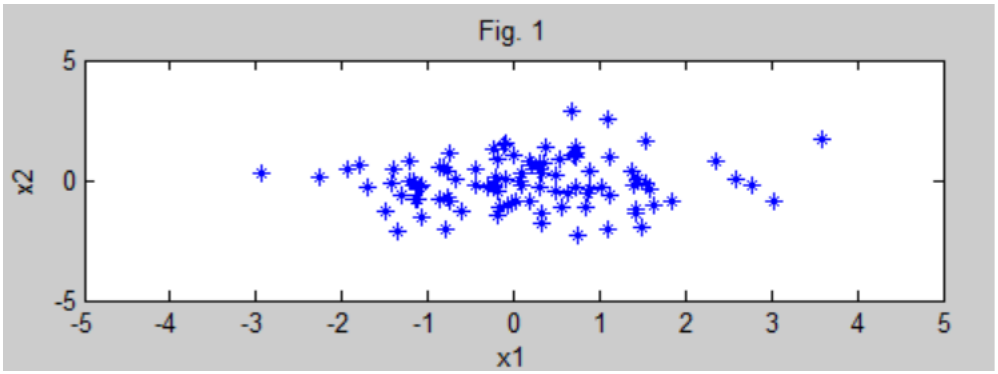
1. Let $|\vec{a}|$ is the smallest Euclidean distance from the origin to some point x in the hyperplane.
   $|\vec{a}| = |\vec{x}|\cos\alpha \qquad < \vec{w}, \vec{x} >= |\vec{w}||\vec{x}|\cos\alpha$
   $|\vec{x}|\cos\alpha = \frac{<\vec{w},\vec{x}>}{|\vec{w}|} = \frac{w^T x}{|\vec{w}|} = \frac{-b}{||\mathbf{w}||_2}$ The distance: $\frac{|b|}{||\mathbf{w}||_2}$
   Statement is true.

2. Smallest Euclidean distance from the origin to some point x in the hyperplane 1 is $\frac{-b_1}{||\mathbf{w}||_2}$, to some point x in the hyperplane 2 is $\frac{-b_2}{||\mathbf{w}||_2}$.
   Distance between two hyperplanes:
   $\frac{-b_1}{||\mathbf{w}||_2} - \frac{-b_2}{||\mathbf{w}||_2} = \frac{-b_1-(-b_2)}{||\mathbf{w}||_2} = \frac{-b_1+b_2}{||\mathbf{w}||_2} = \frac{|-b_1+b_2|}{||\mathbf{w}||_2}$ or

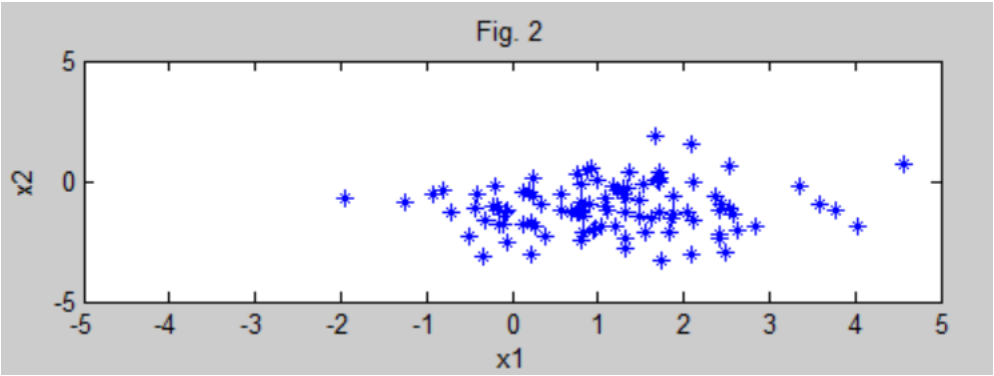
   $\frac{-b_2}{||\mathbf{w}||_2} - \frac{-b_1}{||\mathbf{w}||_2} = \frac{-b_2-(-b_1)}{||\mathbf{w}||_2} = \frac{-b_2+b_1}{||\mathbf{w}||_2} = \frac{b_1-b_2}{||\mathbf{w}||_2} = \frac{|b_1-b_2|}{||\mathbf{w}||_2}$
   Statement is true.

8 Programming Skills

1. Mean:$(0,0)^T$ Covariance Matrix: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

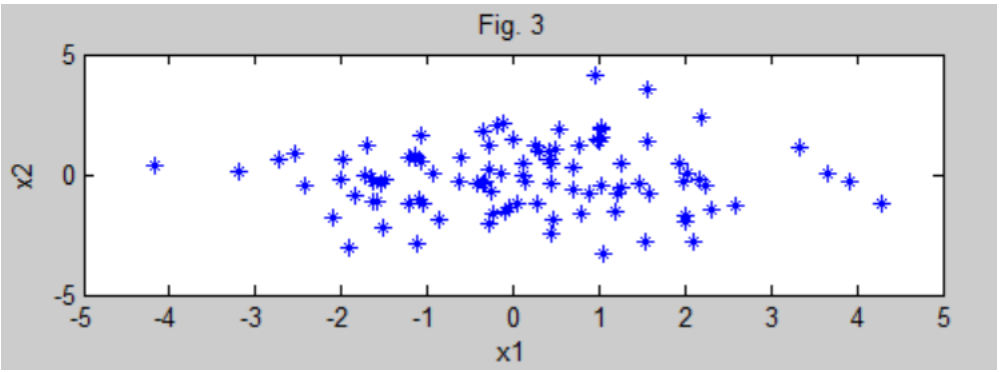


2. Mean:$(1,-1)^T$ Covariance Matrix: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$



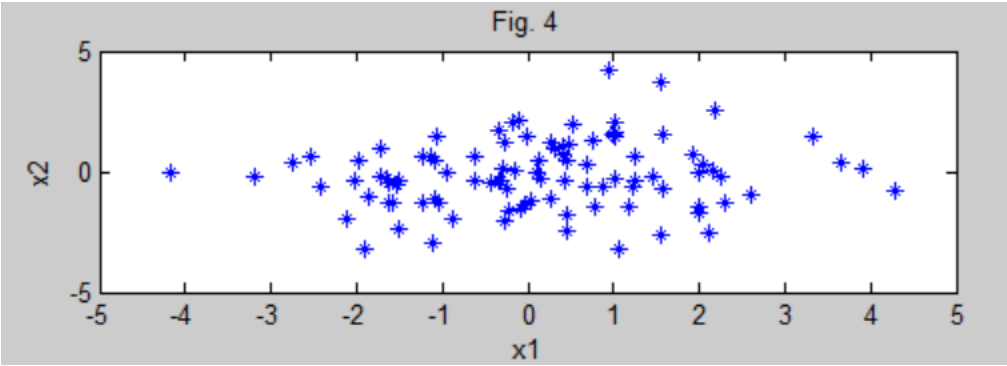
The sample points are shifted right by 1 unit in x1 axis, below 1 unit in x2 axis compared with the Fig. 1, because of the change in the mean value.

3. Mean:$(0,0)^T$ Covariance Matrix: $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$



The sample points are spread in x1 and x2 axis compared with the Fig. 1.

4. Mean:$(0,0)^T$ Covariance Matrix: $\begin{pmatrix} 2 & 0.2 \\ 0.2 & 2 \end{pmatrix}$



5. Mean:$(0,0)^T$ Covariance Matrix: $\begin{pmatrix} 2 & -0.2 \\ -0.2 & 2 \end{pmatrix}$