# Homework #0

Spring 2020, CSE 546: Machine Learning
**Roman Levin**
**1721898**

Collaborators: compared answers with Tyler Chen, Diya Sashidhar, Katherine Owens

# Problems A

## Probability and Statistics

A.1 *[2 points]* (Bayes Rule, from Murphy exercise 2.4.) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you dont have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

---

**Solution:**

Denote "+" the event of a positive test, denote "-" the event of a negative test, denote "disease" the event of having the disease, and denote "no disease" the event of not having the disease. Then, note that:

- $\mathbb{P}(+|\text{disease}) = 0.99$
- $\mathbb{P}(-|\text{no disease}) = 0.99 \Rightarrow \mathbb{P}(+|\text{no disease}) = 1 - \mathbb{P}(-|\text{no disease}) = 0.01$
- $\mathbb{P}(\text{disease}) = 10^{-4} \Rightarrow \mathbb{P}(\text{no disease}) = 1 - \mathbb{P}(\text{disease}) = 0.9999$

By the total probability law, we can find:

$$\mathbb{P}(+) = \mathbb{P}(+|\text{disease})\mathbb{P}(\text{disease}) + \mathbb{P}(+|\text{no disease})\mathbb{P}(\text{disease}) \tag{1}$$

Then, using the equation (1) and Bayes rule, we obtain:

$$\boxed{\begin{aligned}&\mathbb{P}(\text{disease}|+) = \frac{\mathbb{P}(+|\text{disease})\mathbb{P}(\text{disease})}{\mathbb{P}(+)} = \frac{\mathbb{P}(+|\text{disease})\mathbb{P}(\text{disease})}{\mathbb{P}(+|\text{disease})\mathbb{P}(\text{disease}) + \mathbb{P}(+|\text{no disease})\mathbb{P}(\text{disease})} = \\ &= \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.01 \cdot 0.9999} = \frac{1}{102}\end{aligned}} \tag{2}$$

---

A.2 For any two random variables $X, Y$ the *covariance* is defined as $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. You may assume $X$ and $Y$ take on a discrete values if you find that is easier to work with.

a. *[1 points]* If $\mathbb{E}[Y|X = x] = x$ show that $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

---

**Solution:**

Probably, a better way to solve this would be to use iterated expectation and conditioning on $X$ as an RV (not on the event $X = x$), but I am not sure if $\mathbb{E}[Y|X] = X$ follows from the problem statement, so I will just assume $X$ and $Y$ are discrete RVs. Then:

- By linearity of expectation: $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X\mathbb{E}[Y]] - \mathbb{E}[\mathbb{E}[X]Y] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- By the law of total expectation: $\mathbb{E}[XY] = \sum_x \mathbb{P}(X=x)\mathbb{E}[XY|X=x] =$ $= \sum_x \mathbb{P}(X=x)x\mathbb{E}[Y|X=x] = \sum_x \mathbb{P}(X=x)x^2 = \mathbb{E}[X^2]$
- By the law of total expectation: $\mathbb{E}[Y] = \sum_x \mathbb{P}(X=x)\mathbb{E}[Y|X=x] = \sum_x \mathbb{P}(X=x)x = \mathbb{E}[X]$

Combining the above, we get:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[(X-\mathbb{E}[X])^2] \qquad \square$$

---

b. *[1 points]* If $X, Y$ are independent show that $\mathrm{Cov}(X,Y) = 0$.

---

**Solution:**

- In a. we have shown: $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
- Since $X$ and $Y$ are independent: $\mathbb{E}[XY] = \sum_{x,y} \mathbb{P}(X=x, Y=y)xy = \sum_{x,y} \mathbb{P}(X=x)\mathbb{P}(Y=y)xy = \left(\sum_y \mathbb{P}(X=x)x\right)\left(\sum_y \mathbb{P}(Y=y)y\right) = \mathbb{E}[X]\mathbb{E}[Y]$

Combining the above, we get

$$\mathrm{Cov}(X,Y) = 0 \qquad \square$$

---

A.3 Let $X$ and $Y$ be independent random variables with PDFs given by $f$ and $g$, respectively. Let $h$ be the PDF of the random variable $Z = X + Y$.

a. *[2 points]* Show that $h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$. (If you are more comfortable with discrete probabilities, you can instead derive an analogous expression for the discrete case, and then you should give a one sentence explanation as to why your expression is analogous to the continuous case.).

---

**Solution:**

- Joint PDF of $X, Y$: $X \perp\!\!\!\perp Y \Rightarrow f_{X,Y}(x,y) = f(x)g(y)$
- $F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(X+Y \leq z) = \iint_{x+y\leq z} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-x} f(x)g(y)dy\right)dx = \int_{-\infty}^{\infty} f(x)\left(\int_{-\infty}^{z-x} g(y)dy\right)dx = \int_{-\infty}^{\infty} f(x)F_Y(z-x)dx$
- $h(z) = \frac{d}{dz}F_Z(z) = \frac{d}{dz}\int_{-\infty}^{\infty} f(x)F_Y(z-x)dx = \int_{-\infty}^{\infty} f(x)\frac{d}{dz}F_Y(z-x)dx = \int_{-\infty}^{\infty} f(x)g(z-x)dx \qquad \square$

---

b. *[1 points]* If $X$ and $Y$ are both independent and uniformly distributed on $[0,1]$ (i.e. $f(x) = g(x) = 1$ for $x \in [0,1]$ and 0 otherwise) what is $h$, the PDF of $Z = X + Y$?

---

**Solution:**
Notation: $\mathbb{1}_{\{.\}}$ stands for an indicator function.

- $g(x) = f(x) = \mathbb{1}_{\{x\in[0,1]\}}$

- Plugging in to the result from a.:

$$\begin{aligned} h(z) &= \int_{-\infty}^{\infty} \mathbb{1}_{\{x\in[0,1]\}}\mathbb{1}_{\{z-x\in[0,1]\}}dx = \int_{-\infty}^{\infty} \mathbb{1}_{\{x\in[0,1]\}}(\mathbb{1}_{\{x\in(-\infty,z]\}} + \mathbb{1}_{\{x\in[z-1,\infty)\}})dx = \\ &= \int_0^1 \mathbb{1}_{\{x\in(-\infty,z]\}}dx + \int_0^1 \mathbb{1}_{\{x\in[z-1,\infty)\}}dx = z\mathbb{1}_{\{z\in[0,1]\}} + (2-z)\mathbb{1}_{\{z-1\in(0,1]\}} \end{aligned} \tag{3}$$

- This gives finally:

$$\boxed{h(z) = \begin{cases} z, & z\in[0,1] \\ 2-z, & z\in(1,2] \\ 0, & \text{o.w.} \end{cases}}$$

---

A.4 *[1 points]* A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is Gaussian distributed with mean $\mu$ and variance $\sigma^2$. Given that for any $a, b \in \mathbb{R}$, we have that $Y = aX + b$ is also Gaussian, find $a, b$ such that $Y \sim \mathcal{N}(0, 1)$.

---

**Solution:**
Consider

$$\boxed{a = \frac{1}{\sigma}, b = \frac{-\mu}{\sigma}}$$

Then, since multiplying an RV by a constant multiplies the mean by the constant and the variance by the squared constant and since adding a constant to an RV shifts the mean by that constant and does not affect the variance:

$$aX \sim \mathcal{N}(\frac{\mu}{\sigma}, 1) \text{ and } Y \sim \mathcal{N}(0,1) \qquad \square$$

---

A.5 *[2 points]* For a random variable $Z$, its mean and variance are defined as $\mathbb{E}[Z]$ and $\mathbb{E}[(Z - \mathbb{E}[Z])^2]$, respectively. Let $X_1, \dots, X_n$ be independent and identically distributed random variables, each with mean $\mu$ and variance $\sigma^2$. If we define $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, what is the mean and variance of $\sqrt{n}(\widehat{\mu}_n - \mu)$?

---

**Solution:**

- $\mathbb{E}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mu = \mu n$ and $\mathbb{V}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \sigma^2 = \sigma^2 n$ since $X_i$ are iid.
- So $\mathbb{E}[\widehat{\mu}_n] = \mu$ and $\mathbb{V}[\widehat{\mu}_n] = \sigma^2/n$ (by the same argument with A.4).
- Then, again by the same argument with A.4:

$$\boxed{\mathbb{E}(\sqrt{n}(\widehat{\mu}_n - \mu)) = 0, \mathbb{V}(\sqrt{n}(\widehat{\mu}_n - \mu)) = (\sqrt{n})^2\sigma^2/n = \sigma^2}$$

---

A.6 If $f(x)$ is a PDF, the cumulative distribution function (CDF) is defined as $F(x) = \int_{-\infty}^{x} f(y)dy$. For any function $g : \mathbb{R} \to \mathbb{R}$ and random variable $X$ with PDF $f(x)$, recall that the expected value of $g(X)$ is defined as $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(y)f(y)dy$. For a boolean event $A$, define $\mathbf{1}\{A\}$ as 1 if $A$ is true, and 0 otherwise. Thus, $\mathbf{1}\{x \le a\}$ is 1 whenever $x \le a$ and 0 whenever $x > a$. Note that $F(x) = \mathbb{E}[\mathbf{1}\{X \le x\}]$. Let $X_1, \dots, X_n$ be *independent and identically distributed* random variables with CDF $F(x)$. Define $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le x\}$. Note, for every $x$, that $\widehat{F}_n(x)$ is an *empirical estimate* of $F(x)$. You may use your answers to the previous problem.

a. *[1 points]* For any $x$, what is $\mathbb{E}[\widehat{F}_n(x)]$?

**Solution:**

$$\boxed{\mathbb{E}[\widehat{F}_n(x)] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq x\}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{1}\{X_i \leq x\}] = \frac{1}{n}\sum_{i=1}^{n}F(x) = F(x)}$$

b. *[1 points]* For any $x$, the variance of $\widehat{F}_n(x)$ is $\mathbb{E}[(\widehat{F}_n(x) - F(x))^2]$. Show that Variance$(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$.

**Solution:**
Since $X_i$ are iid, then $\mathbf{1}\{X_i < x\}$ are also independent as functions of $X_i$, thus we can write (also observing that $(\mathbf{1}\{X_i \leq x\})^2 = \mathbf{1}\{X_i \leq x\}$):

$$\begin{aligned}\mathbb{V}[\widehat{F}_n(x)] = \mathbb{V}[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq x\}] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[\mathbf{1}\{X_i \leq x\}] = \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathbb{E}[(\mathbf{1}\{X_i \leq x\})^2] - \mathbb{E}[\mathbf{1}\{X_i \leq x\}]^2\right) = \\ = \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathbb{E}[\mathbf{1}\{X_i \leq x\}] - \mathbb{E}[\mathbf{1}\{X_i \leq x\}]^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}\left(F(x) - F(x)^2\right) = \frac{F(x)(1-F(x))}{n} \qquad \square\end{aligned} \tag{4}$$

c. *[1 points]* Using your answer to b, show that for all $x \in \mathbb{R}$, we have $\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \leq \frac{1}{4n}$.

**Solution:**

- Note that $\mathbb{E}[\widehat{F}_n(x) - F(x)] = 0$ (from a.).
- Therefore $\mathbb{V}[\widehat{F}_n(x) - F(x)] = \mathbb{E}[(\widehat{F}_n(x) - F(x))^2] = \frac{F(x)(1-F(x))}{n}$ (by b.)
- Since $F(x) \in [0,1] \forall x$, $\frac{F(x)(1-F(x))}{n}$ is maximized at $F(x) = \frac{1}{2}$ (it is just a parabola).
- Therefore
$$\boxed{\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] = \frac{F(x)(1-F(x))}{n} \leq \frac{0.5(1-0.5)}{n} = \frac{1}{4n} \qquad \square}$$

B.1 *[1 points]* Let $X_1, \ldots, X_n$ be $n$ independent and identically distributed random variables drawn unfiromly at random from $[0,1]$. If $Y = \max\{X_1, \ldots, X_n\}$ then find $\mathbb{E}[Y]$.

**Solution:**
Let

$$F_X(x) = \begin{cases} 1, & x > 1 \\ x, & x \in [0,1] \\ 0, & x < 0 \end{cases}$$

be the common CDF of $X_i$.

- Since $X_i$ are iid: $F_Y(y) = \mathbb{P}(Y \leq y) = \prod_{i=1}^{n}\mathbb{P}(X_i \leq y) = \prod_i F_X(y) = F_X(y)^n$
- The PDF of Y:$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} ny^{n-1}, & y \in [0,1] \\ 0, & \text{o.w} \end{cases}$

- $$\boxed{\mathbb{E}[Y] = \int_0^1 yny^{n-1}dy = n\int_0^1 y^n dy = \frac{n}{n+1}}$$

## Linear Algebra and Vector Calculus

A.7 (Rank) Let $A = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}$. For each matrix $A$ and $B$,

a. *[2 points]* what is its rank?

**Solution:**
Let $a_i$ be the columns of A and $b_i$ be the columns of $B$.
For A, observe that $-3a_1 + a_2 + a_3 = 0$ but any two columns are obviously linearly independent, so

$$\boxed{\text{rank } A = 2}$$

For B, observe that $b_3 - b_2 = b_1$ but any two columns are obviously linearly independent, so

$$\boxed{\text{rank } B = 2}$$

b. *[2 points]* what is a (minimal size) basis for its column span?

**Solution:**
For $A$, since its rank is 2 and $a_1, a_2$ are linearly independent, we can take $\{a_1, a_2\}$ as the basis for its column space.
For $B$, since its rank is 2 and $b_1, b_2$ are linearly independent, we can take $\{b_1, b_2\}$ as the basis for its column space.

A.8 (Linear equations) Let $A = \begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix}$, $b = \begin{bmatrix} -2 & -2 & -4 \end{bmatrix}^T$, and $c = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$.

a. *[1 points]* What is $Ac$?

**Solution:**

$$\boxed{Ac = \begin{bmatrix} 6 & 8 & 7 \end{bmatrix}^T}$$

b. *[2 points]* What is the solution to the linear system $Ax = b$? (Show your work).

**Solution:**

$$\left[\begin{array}{ccc|c} 0 & 2 & 4 & -2 \\ 2 & 4 & 2 & -2 \\ 3 & 3 & 1 & -4 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 2 & 1 & -1 \\ 0 & 2 & 4 & -2 \\ 3 & 3 & 1 & -4 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 2 & 1 & -1 \\ 0 & 1 & 2 & -1 \\ 0 & -3 & -2 & -1 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -3 & 1 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 4 & -4 \end{array}\right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

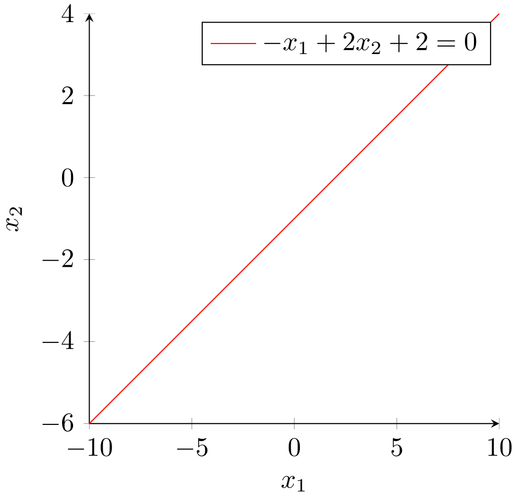
So

$$\boxed{x = [-2, 1, -1]^T}$$

A.9 (Hyperplanes) Assume $w$ is an $n$-dimensional vector and $b$ is a scalar. A hyperplane in $\mathbb{R}^n$ is the set $\{x : x \in \mathbb{R}^n,\ \text{s.t. } w^T x + b = 0\}$.

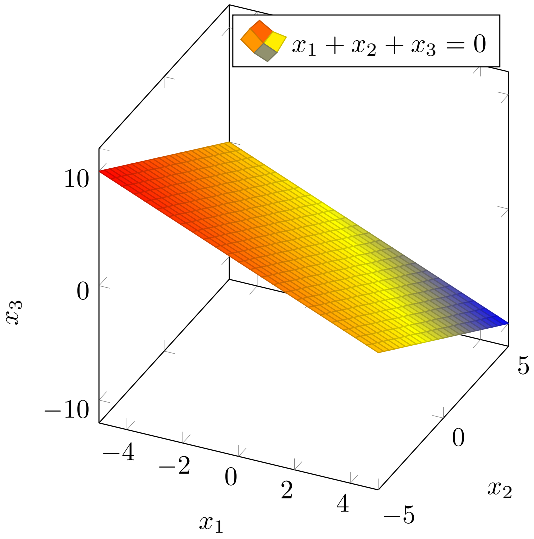
a. *[1 points]* ($n = 2$ example) Draw the hyperplane for $w = [-1, 2]^T$, $b = 2$? Label your axes.

**Solution:**



b. *[1 points]* ($n = 3$ example) Draw the hyperplane for $w = [1, 1, 1]^T$, $b = 0$? Label your axes.

**Solution:**



c. *[2 points]* Given some $x_0 \in \mathbb{R}^n$, find the *squared distance* to the hyperplane defined by $w^T x + b = 0$. In other words, solve the following optimization problem:

$$\begin{aligned} &\min_x \|x_0 - x\|^2 \\ &\text{s.t. } w^T x + b = 0 \end{aligned}$$

(Hint: if $\widetilde{x}_0$ is the minimizer of the above problem, note that $\|x_0 - \widetilde{x}_0\| = |\frac{w^T(x_0-\widetilde{x}_0)}{\|w\|}|$. What is $w^T\widetilde{x}_0$?)

**Solution:**
Let $x^*$ be the minimizer of the optimization problem, that is, let $x^*$ be the projection of $x_0$ onto the hyperplane. Then, the distance from $x_0$ to the hyperplane would be given (as the hint states) by the difference of the lengths of the projections of $x_0$ and $x^*$ onto the normal vector $w$:

$$\|x_0 - x^*\| = \left| \frac{w^T x_0}{\|w\|} - \frac{w^T x^*}{\|w\|} \right| = \left| \frac{w^T(x_0 - x^*)}{\|w\|} \right|$$

Since $x^*$ lies in the hyperplane, we have $w^T x^* + b = 0$. Thus:

$$\boxed{\|x_0 - x^*\|^2 = \left| \frac{w^T x_0 + b}{\|w\|} \right|^2}$$

A.10 For possibly non-symmetric $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}$, let $f(x,y) = x^T\boldsymbol{A}x + y^T\boldsymbol{B}x + c$. Define $\nabla_z f(x,y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial z_1} & \frac{\partial f(x,y)}{\partial z_2} & \dots & \frac{\partial f(x,y)}{\partial z_n} \end{bmatrix}^T$.

a. *[2 points]* Explicitly write out the function $f(x,y)$ in terms of the components $A_{i,j}$ and $B_{i,j}$ using appropriate summations over the indices.

**Solution:**

$$\boxed{f(x,y) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j + \sum_{i=1}^n \sum_{j=1}^n y_i B_{ij} x_j + c}$$

b. *[2 points]* What is $\nabla_x f(x,y)$ in terms of the summations over indices *and* vector notation?

**Solution:**

$$\boxed{(\nabla_x f(x,y))_k = \frac{\partial f(x,y)}{\partial x_k} = \sum_{\substack{j=1\\ j\neq k}}^n A_{kj}x_j + \sum_{\substack{i=1\\ i\neq k}}^n A_{ik}x_i + 2A_{kk}x_k + \sum_{i=1}^n y_i B_{ik} = \sum_{j=1}^n A_{kj}x_j + \sum_{i=1}^n A_{ik}x_i + \sum_{i=1}^n y_i B_{ik}}$$

Expressing the above in the vector form:

$$\boxed{\nabla_x f(x,y) = (A + A^T)x + B^T y}$$

c. *[2 points]* What is $\nabla_y f(x,y)$ in terms of the summations over indices *and* vector notation?

**Solution:**

$$\boxed{(\nabla_y f(x,y))_k = \frac{\partial f(x,y)}{\partial y_k} = \sum_{j=1}^{n} B_{kj} x_j}$$

Epressing the above in the vector form:

$$\boxed{\nabla_y f(x,y) = Bx}$$

---

B.2 *[1 points]* The *trace* of a matrix is the sum of the diagonal entries; $Tr(A) = \sum_i A_{ii}$. If $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$, show that $Tr(AB) = Tr(BA)$.

---

**Solution:**

- Note that $(AB)_{ii} = \sum_{j=1}^{m} A_{ij}B_{ji}, \quad (BA)_{jj} = \sum_{i=1}^{n} B_{ji}A_{ij}$
- Then:

$$Tr(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}B_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} B_{ji}A_{ij} = \sum_{j=1}^{m}(BA)_{jj} = Tr(BA) \qquad \square$$

---

B.3 *[1 points]* Let $v_1, \ldots, v_n$ be a set of non-zero vectors in $\mathbb{R}^d$. Let $V = [v_1, \ldots, v_n]$ be the vectors concatenated.

a. What is the minimum and maximum rank of $\sum_{i=1}^{n} v_i v_i^T$?

---

**Solution:**
Note that $\sum_{i=1}^{n} v_i v_i^T = VV^T$. Since the $v_i$ are nonzero, the minimum rank cannot be zero (since only zero matrix is of zero rank), so we have:

$$\boxed{1 \leq \operatorname{rank} VV^T \leq \operatorname{rank} V \leq \min(n,d)}$$

An example of the minimum rank 1 would be to take $V$ with all the columns equal to the same vector, then both $V$ and $VV^T$ would be rank 1. An example of the maximum rank would be

- For $n >= d$: let $V$ be a matrix with $n$ orthonormal columns, then $VV^T$ and $V$ would be full rank.
- For $d >= n$: let $V$ be a matrix with $d$ orthonormal rows, then $VV^T$ and $V$ would be full rank.

(In fact, it can be shown that $\operatorname{rank} VV^T = \operatorname{rank} V$ by taking the reduced SVD of V and obtaining the SVD of $VV^T$ with the same number of nonzero singular values which is equal to rank.

$$V = \hat{U}\hat{\Sigma}\hat{V}^T \Rightarrow VV^T = \hat{U}\hat{\Sigma}\hat{V}^T\hat{V}\hat{\Sigma}\hat{U}^T = \hat{U}\hat{\Sigma}^2\hat{U}^T$$

Now, $\hat{\Sigma}$ and $\hat{\Sigma}^2$ have the same number of nonzero elements, so $\operatorname{rank} VV^T = \operatorname{rank} V \qquad \square$)

---

b. What is the minimum and maximum rank of $V$?

---

**Solution:**
Again, since $v_i$ are nonzero, we cannot have zero rank. The full rank would be $\min(d,n)$ (by definition). Examples are given in a.

$$\boxed{1 \leq \operatorname{rank} V \leq \min(d,n)}$$

---

c. Let $A \in \mathbb{R}^{D \times d}$ for $D > d$. What is the minimum and maximum rank of $\sum_{i=1}^n (Av_i)(Av_i)^T$?

**Solution:**
Note that $\sum_{i=1}^n (Av_i)(Av_i)^T = (AV)(AV)^T$. Then, since $D > d$:

$$\boxed{0 \leq \operatorname{rank}(AV)(AV)^T \leq \operatorname{rank}(AV) \leq \min(\operatorname{rank} A, \operatorname{rank} V) \leq \min(n, d)}$$

For rank 0 take zero $A$, for rank $\min(n, d)$ take full-rank $V$ and full-rank $A$ (because in the latter case $AV$ is full rank and so is $(AV)(AV)^T$ by a.).

d. What is the minimum and maximum rank of $AV$? What if $V$ is rank $d$?

**Solution:**

- $$\boxed{0 \leq \operatorname{rank}(AV) \leq \min(\operatorname{rank} A, \operatorname{rank} V) \leq \min(n, d)}$$

  For rank 0 take zero $A$, for rank $\min(n, d)$ take full-rank $V$ and full-rank $A$ (so that $AV$ is full-rank).
- $V$ of rank $d$ implies $n >= d$ and thus (since $D > d$):

  $$\boxed{0 \leq \operatorname{rank}(AV) \leq \min(\operatorname{rank} A, \operatorname{rank} V) \leq d}$$

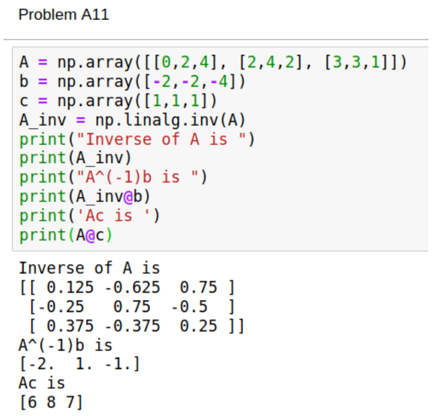
## Programming

A.11 For the $A, b, c$ as defined in Problem 8, use NumPy to compute (take a screen shot of your answer):

a. *[2 points]* What is $A^{-1}$?

b. *[1 points]* What is $A^{-1}b$? What is $Ac$?

**Solution:**
See Figure 1.



Problem A11

```
A = np.array([[0,2,4], [2,4,2], [3,3,1]])
b = np.array([-2,-2,-4])
c = np.array([1,1,1])
A_inv = np.linalg.inv(A)
print("Inverse of A is ")
print(A_inv)
print("A^(-1)b is ")
print(A_inv@b)
print('Ac is ')
print(A@c)
```

```
Inverse of A is
[[ 0.125 -0.625  0.75 ]
 [-0.25   0.75  -0.5  ]
 [ 0.375 -0.375  0.25 ]]
A^(-1)b is
[-2.  1. -1.]
Ac is
[6 8 7]
```

Figure 1: Screenshot of the solution for A11

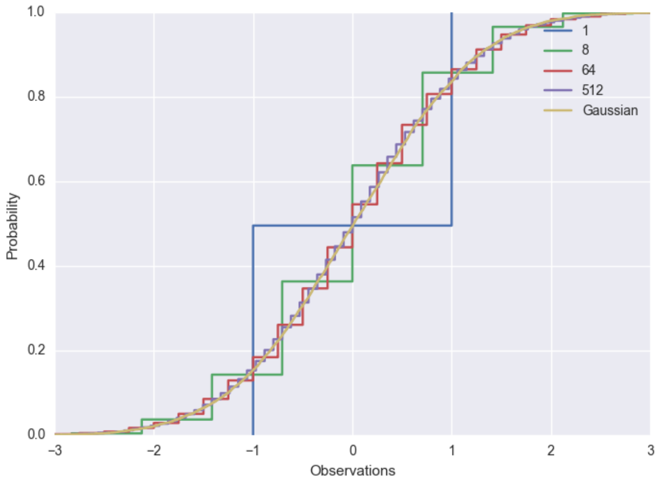
```
#Code for A11
import numpy as np
A = np.array([[0,2,4], [2,4,2], [3,3,1]])
b = np.array([-2,-2,-4])
c = np.array([1,1,1])
A_inv = np.linalg.inv(A)
print("Inverse of A is ")
print(A_inv)
print("A^(-1)b is ")
print(A_inv@b)
print('Ac is ')
print(A@c)
```

A.12 *[4 points]* Two random variables $X$ and $Y$ have equal distributions if their CDFs, $F_X$ and $F_Y$, respectively, are equal, i.e. for all $x$, $|F_X(x) - F_Y(x)| = 0$. The central limit theorem says that the sum of $k$ independent, zero-mean, variance-$1/k$ random variables converges to a (standard) Normal distribution as $k$ goes off to infinity. We will study this phenomenon empirically (you will use the Python packages Numpy and Matplotlib). Define $Y^{(k)} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} B_i$ where each $B_i$ is equal to $-1$ and $1$ with equal probability. From your solution to problem 5, we know that $\frac{1}{\sqrt{k}} B_i$ is zero-mean and has variance $1/k$.

a. For $i = 1, \ldots, n$ let $Z_i \sim \mathcal{N}(0,1)$. If $F(x)$ is the true CDF from which each $Z_i$ is drawn (i.e., Gaussian) and $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Z_i \leq x\}$, use the answer to problem 1.5 above to choose $n$ large enough such that, for all $x \in \mathbb{R}$, $\sqrt{\mathbb{E}[(\widehat{F}_n(x) - F(x))^2]} \leq 0.0025$, and plot $\widehat{F}_n(x)$ from $-3$ to $3$.
(Hint: use `Z=numpy.random.randn(n)` to generate the random variables, and `import matplotlib.pyplot as plt`;
`plt.step(sorted(Z), np.arange(1,n+1)/float(n))` to plot).

b. For each $k \in \{1, 8, 64, 512\}$ generate $n$ independent copies $Y^{(k)}$ and plot their empirical CDF on the same plot as part a.
(Hint: `np.sum(np.sign(np.random.randn(n, k))*np.sqrt(1./k), axis=1)` generates $n$ of the $Y^{(k)}$ random variables.)

Be sure to always label your axes. Your plot should look something like the following (Tip: checkout `seaborn` for instantly better looking plots.)



**Solution:**
a. See Figure 2 for the plot of empirical CDF of the standard normal. Using the $\frac{1}{4n}$ bound in A.6, we find that $n = 40000$ is sufficient to guarantee the required error from the true CDF.

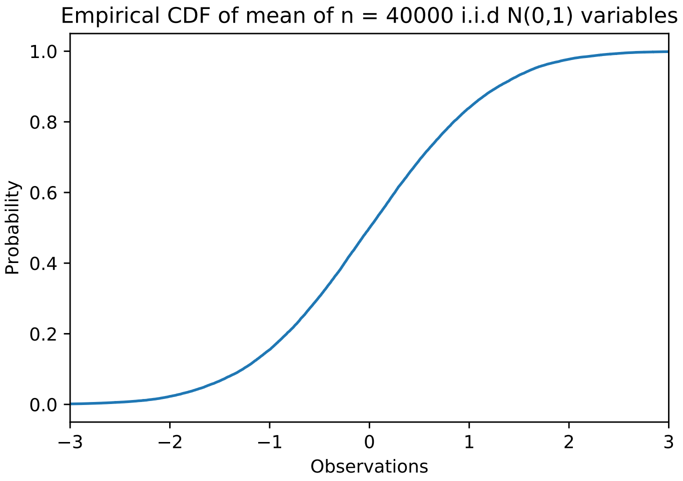


Figure 2: Empirical CDF for A12a

```python
#Code for A12.a
import numpy as np
import matplotlib.pyplot as plt
n = 40000
Z = np.random.randn(n)
ax = plt.gca()
plt.step(sorted(Z), np.arange(1,n+1)/float(n))
plt.title("Empirical CDF of mean of n = 40000 i.i.d N(0,1) variables")
ax.set_xlim((-3,3))
plt.xlabel(’Observations’)
plt.ylabel(’Probability’)
plt.savefig(’A12a_CDF.pdf’)
```

b. See Figure 3.

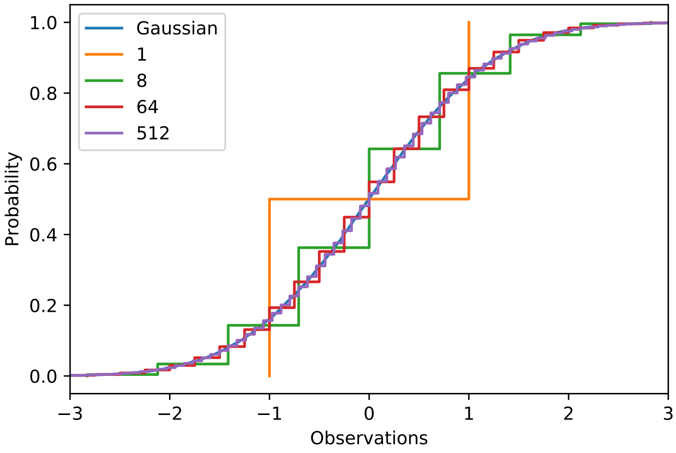


Figure 3: A12b

```
#Code for A12.b
import numpy as np
import matplotlib.pyplot as plt
n = 40000
Z = np.random.randn(n)
ax = plt.gca()
plt.step(sorted(Z), np.arange(1,n+1)/float(n), label = 'Gaussian')

for k in [1,8,64,512]:
    Z_k = np.sum(np.sign(np.random.randn(n, k))*np.sqrt(1./k), axis=1)
    plt.step(sorted(Z_k), np.arange(1,n+1)/float(n), label='%s' % k)

ax.set_xlim((-3,3))
plt.legend()
plt.xlabel('Observations')
plt.ylabel('Probability')
plt.savefig('A12b.pdf')
```