Homework #0

Spring 2020, CSE 446/546: Machine Learning
Dino Bektesevic

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 don't 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.)

Denoting $X = 1$ as test result being positive and $Y = 1$ as actually having the disease, following Bayes rule we have

$$\begin{aligned} P(Y=1|X=1) &= \frac{P(X=1|Y=1)P(Y=1)}{P(X=1|Y=1)+P(X=1|Y=0)P(Y=0)} \\ &= \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.01 * 0.9999} \\ &= 0.0098 \approx 1\% \end{aligned}$$

since we are given that $P(X = 1|Y = 1) = P(X = 0|Y = 0) = 0.99$ and $P(Y = 1) = 1/10000 = 0.0001$ from which it follows that $P(Y = 0) = 1 - P(Y = 1) = 0.9999$ and $P(X = 1|Y = 0) = 1 - P(X = 0|Y = 0) = 0.01$

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 $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. Simplify the right hand side of the expression for covariance (marking expectation values that are scalars by replacing them with μ to show they can be taken out of expectation value operator):

$$\begin{aligned} E[(X - E(X))^2] &= E[X^2 - 2XE(X) + E(X)^2] = E[X^2 - 2\mu_x X + 2\mu_x^2] \\ &= E[X^2] - 2\mu_x E[X] + \mu_x^2 == E[X^2] - 2\mu_x^2 + \mu_x^2 \\ &= E[X^2] - \mathbb{E}[X]^2 = \text{Var}(X) \end{aligned}$$

It's clear now that we are actually looking to show that $\text{Cov}(X, Y) = \text{Var}(X)$. Apply what is given in the problem to law of iterative expectations:

$$E[Y] = \int yp(y)dy = \int E(Y|X = x)p(x)dx = \int xp(x)dx = E[X]$$

Finally, write out the expression for covariance of two rvs:

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = E[XY] - E[X]^2 = \text{Var}(X)$$

since we can write the joint probability $E[XY] = \int\int xp(x)yp_{Y|X=x}(y)dxdy = \int xp(x)xdx = \int x^2p(x)dx = E[X^2]$

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

$$\text{Cov}(X, Y) = E(X, Y) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

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.).
Random variables are independent when $p(x, y) = f(x)g(y)$. Since X and Y are independent and $Z = X + Y$ the domain of Z is given as a union of the domains of X and Y, i.e. in terms of values of rvs $z = x + y$. Picking a particular z and x limits, by the law of total probability, the allowable values event y can take. This can be expressed as the constraint $z = x + y \rightarrow y = z - x \rightarrow x = z - y$. So given a joint probability:

$$p(Z = z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} p(x, y)dxdy$$

and given that X and Y are independent:

$$p(Z = z) = \int_{x \in X} f(x)g(z-x)dx = \int_{y \in Y} f(z-y)g(y)dy$$

If we then further define $f(x) = 0; x \notin X$:

$$p(Z = z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$$

This is exactly the definition of convolution $(f * g)(t) \triangleq \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\,d\tau$

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 $Z = X + Y$?

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx = \int_{0}^{1} g(z-x)dx$$

Solutions are possible only for $0 \leq z - x \leq 1 \rightarrow z - 1 \leq x \leq z$. So we have 3 cases:
First case: $0 \leq z \leq 1$, *sa*:

$$h(z) = \int_{0}^{z} dx = z$$

Second case $1 < z \leq 2$:

$$h(z) = \int_{z-1}^{1} dx = 2 - z$$

and $h(z) = 0$ otherwise.

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

$$\begin{aligned} E[Y] &= aE[X] + b = 0 \\ \text{Var}[Y] &= a^2\text{Var}[X] = 1 \\ a\mu + b &= 0 \\ a^2\sigma^2 &= 1 \\ a &= 1/\sigma \\ b &= -\mu/\sigma \end{aligned}$$

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 $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and variance σ^2. If we define $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, what is the mean and variance of $\sqrt{n}(\widehat{\mu}_n - \mu)$?

The average value of many random i.i.d. variables that have the same mean μ is the mean μ itself. So:

$$\hat{\mu}_n = \frac{1}{n}\sum X_i = \mu \rightarrow \sqrt{n}(\hat{\mu}_n - \mu) = 0$$

In a little more details:

$$\begin{aligned}
\mathbb{E}[\sqrt{n}(\hat{\mu}_n - \mu)] &= \sqrt{n}\mathbb{E}\left[\hat{\mu}_n - \mu\right] \\
&= \sqrt{n}\left(\mathbb{E}[\hat{\mu}_n] - \mathbb{E}[\mu]\right) = \sqrt{n}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - \mathbb{E}[\mu]\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] - \mu\right) = \sqrt{n}\left(\frac{1}{n}n\mu - \mu\right) \\
&= \sqrt{n}(\mu - \mu) = 0
\end{aligned}$$

The variance of many i.i.d. rvs can be written as:

$$\begin{aligned}
\mathrm{Var}[\sqrt{n}\hat{\mu}_n - \sqrt{n}\mu] &= n\mathrm{Var}[\hat{\mu}_n] + n\mathrm{Var}[\mu] \\
&= n\mathrm{Var}[\hat{\mu}_n] = \frac{n\sigma^2}{n} = \sigma^2
\end{aligned}$$

since $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$ and $\mathrm{Var}[\mu] = 0$ because all observed rvs have the same exact mean.

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 $g : \mathbb{R} \rightarrow \mathbb{R}$ 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, $\mathbf{1}\{x \leq a\}$ is 1 whenever $x \leq a$ and 0 whenever $x > a$. Note that $F(x) = \mathbb{E}[\mathbf{1}\{X \leq x\}]$. Let $X_1, \ldots, X_n$ be *independent and identically distributed* random variables with CDF $F(x)$. Define $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq x\}$. 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 $\mathbb{E}[\widehat{F}_n(x)]$?

$$\begin{aligned}
E[\hat{F}_n(x)] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \leq x\right\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[\mathbf{1}\left\{X_i \leq x\right\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} F(x) \\
&= \frac{1}{n}nF(x) = F(x)
\end{aligned}$$

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 $\mathrm{Variance}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$.

$$\begin{aligned}
Var\hat{F}_n(x) &= \mathrm{Var}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \leq x\right\} \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n}\mathbf{1}\left\{X_i \leq x\right\}\right)
\end{aligned}$$

Notice that $\mathbf{1}\left\{X_i \leq x\right\}$ takes values of 1 or 0, such that the elements of that sum follow a binomial distribution with n events, each having a probability $p = F(x)$ of being true. The variance of a binomial

distribution is given as $\text{Var}(X) = np(1-p)$ (Wikipedia). Substituting appropriately we write:

$$\begin{aligned}\text{Var}\hat{F}_n(x) &= \frac{1}{n^2}\text{Var}(\text{Bin}(n, F(x))) \\ &= \frac{1}{n^2} nF(x)\,(1-F(x)) \\ &= \frac{F(x)\,(1-F(x))}{n}\end{aligned}$$

c. *[1 points]* Using your answer to b, show that for all $x \in \mathbb{R}$, we have $\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \leq \frac{1}{4n}$. Maxima can usually be found by looking at the derivative and setting it to zero:

$$\begin{aligned}\frac{\partial}{\partial F(x)} E[(\widehat{F}_n(x) - F(x))^2] = \frac{\partial}{\partial F(x)} \text{Var}\widehat{F}_n(x) = \frac{\partial}{\partial F(x)} \frac{F(x)\,(1-F(x))}{n} = 0 \\ \frac{1-2F(x)}{n} = 0 \\ F(x) = \frac{1}{2}\end{aligned}$$

Plugging the maximum value of $F(x) = 1/2$ into the expression for variance:

$$\text{Var}\widehat{F}_n(x) = \frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{n} = \frac{1}{4n}$$

Which satisfies the given expression with the equality sign valid only at the maxima of the variance.

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

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?

Rank is just the number of linearly independent columns of the matrix:

$$rank(A) = 2$$

$$rank(B) = 2$$

For the B matrix if we take away the last row from the first we are left with $[0, 1, 1]$ and if we take away second row from the last we are left again with $[0, 1, 1]$ so at least one of the rows is not linearly independent from the remaining two. For matrix A it's a bit harder to spot so we write:

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 2 \end{bmatrix} \equiv \begin{bmatrix} 1 & 2 & 1 \\ 0 & -2 & 2 \\ 1 & 1 & 2 \end{bmatrix} \equiv \begin{bmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \end{bmatrix} \equiv \begin{bmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

where we nullify the first column using first row, then divide second row by 2 and take away from the third.

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

Two columns of matrix A are linearly independent, so the basis of the image space is subset of $\mathbb{R}^2$ stretched by the set of column vectors:

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \right\}$$

Only two columns of matrix B are linearly independent. Starting from the note in previous problem:

$$A \equiv \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

from which we see that first column can be written as $c_1 = c_3 - c_2$. So the basis of the image space is a subset of $\mathbb{R}^2$ stretched by column vectors:

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right\}$$

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?
It's a matrix multiplication operation!

$$\begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2+4 \\ 2+4+2 \\ 3+3+1 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 7 \end{bmatrix}$$

b. *[2 points]* What is the solution to the linear system $Ax = b$? (Show your work).
Sequence of operations is as follows:

(a) switch first and second row and divide first row by 2 (sets (1,1) to 1) and zero out first column,

(b) divide second row by two and zero out second column,

(c) divide 3rd row by 4 and zero out 3rd column.

$$\begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ -4 \end{bmatrix}$$

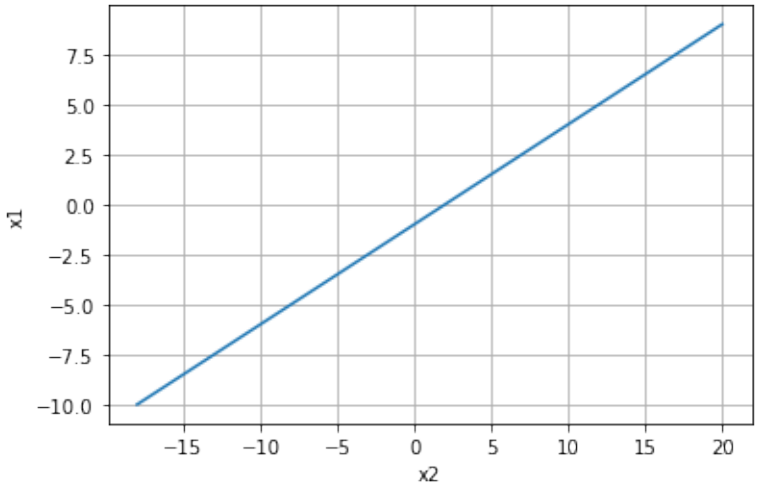
$$\left[\begin{array}{ccc|c} 0 & 2 & 4 & -2 \\ 2 & 4 & 2 & -2 \\ 3 & 3 & 1 & -4 \end{array}\right] \equiv \left[\begin{array}{ccc|c} 1 & 2 & 1 & -2 \\ 0 & 2 & 4 & -2 \\ 0 & -3 & -2 & -1 \end{array}\right] \equiv$$

$$\left[\begin{array}{ccc|c} 1 & 0 & 1 & -3 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 4 & -4 \end{array}\right] \equiv \left[\begin{array}{ccc|c} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right] \rightarrow x = \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}$$

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.
Effectively this gives us an equation of a line $-x_1 + 2x_2 + 2 = 0$ or $x_1 = 2x_2 + 2$. Plot line using following Python code:

```
import matplotlib.pyplot as plt
import numpy as np
x2 = np.arange(-10, 10, 1)
x1 = 2*x2+2
plt.plot(x1, x2)
plt.xlabel("x2")
plt.ylabel("x1")
plt.grid()
```

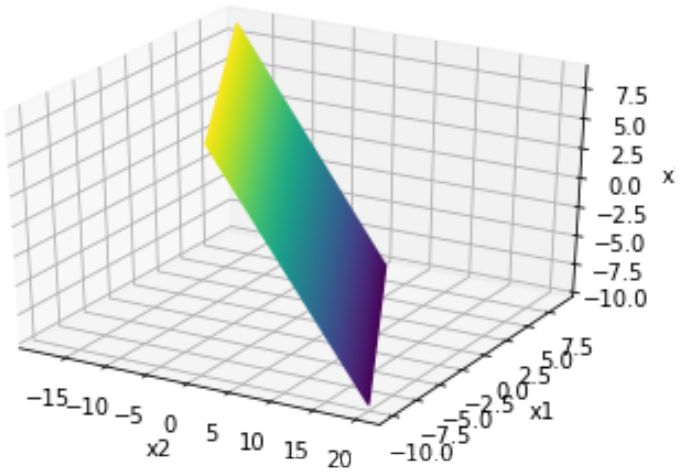


b. *[1 points]* ($n = 3$ example) Draw the hyperplane for $w = [1, 1, 1]^T$, $b = 0$? Label your axes. Following the same principles above $x + y + z = 0 \rightarrow z = -x - y$ we have:

```
import matplotlib.pyplot as plt
from mpl_toolkits import mplot3d
import numpy as np

x2 = np.arange(-10, 10, 1)
x3 = np.arange(-10, 10, 1)
X2, X3 = np.meshgrid(x2, x3)
X1 = -X2 -X3

fg = plt.figure()
ax = plt.axes(projection='3d')
ax.contour3D(X1, X2, X3, 150)
ax.set_zlabel("x3")
ax.set_xlabel("x2")
ax.set_ylabel("x1")
plt.grid()
```



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:

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

(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 $w^T\widetilde{x}_0$?)
We are looking to minimize the following function:

$$\begin{aligned} f(\vec{x}) &= \omega^T ||\vec{x} - \vec{x}_0||^2 + b = 0 \\ &= \omega^T ||\vec{x}^T\vec{x} - 2\vec{x}\cdot\vec{x}_0 + \vec{x}_0^T\vec{x}_0|| + b \\ &= \omega\vec{x}^T\vec{x} - 2\omega^T||\vec{x}\cdot\vec{x}_0|| + \omega^T\vec{x}_0^T\vec{x}_0 + b = 0 \end{aligned}$$

Note that values $\vec{x}^T\vec{x}$ and $\vec{x}_0^T\vec{x}_0$ are just scalars. Minimization can be performed via the method of Lagrange multipliers of the above equation given the constraint that $\vec{x}$ lies in the plane $\Phi = \omega^T\vec{x} + b = 0$:

$$\frac{\partial f(\vec{x})}{\partial x_i} - \lambda\frac{\partial \Phi(\vec{x})}{\partial x_i} = 0$$
$$-2\omega^T\frac{\partial||\vec{x}\cdot\vec{x}_0||}{\partial x_i} - \lambda\omega^T\frac{\partial\vec{x}}{\partial x_i} = 0$$
$$-2\omega^T\frac{\partial\vec{x}}{\partial x_i}\cdot\vec{x}_0 - \lambda\omega^T\frac{\partial\vec{x}}{\partial x_i} = 0$$

Following is speculative

$$\left(-2\omega^T\vec{x}_0 - \lambda\omega^T\right)\frac{\partial\vec{x}}{\partial x_i} = 0$$

Note that $\omega^T\vec{x}_0$ is a scalar, that x_0 denotes the magnitude of the radius vector of the point we're measuring distance to so it also is a scalar, we can

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.
Not sure what is meant by this but this is how the function would look like written out "explicitly":

$$f(x,y) = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} A_{11} & \dots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \dots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \begin{bmatrix} B_{11} & \dots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \dots & B_{nn} \end{bmatrix} \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} + c$$

Writing the above in terms of summations over indices we keep in mind the definition of matrix multiplication: $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$. Applied to Ax for example we have $Ax = \sum_{i=1}^{n} A_{1,i}x_i$. We write the following:

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

b. *[2 points]* What is $\nabla_x f(x,y)$ in terms of the summations over indices *and* vector notation?
In this context ∇_x represents the derivation operator with respect to elements of vector x. In summation form, if coordinates are orthogonal, it's defined as: $\nabla f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\frac{1}{h_i}\hat{e}_i$ where h are the Lame's coefficients (important for coordinate systems except for Cartesian where they're equal to 1). Effectively ∇_x turns a column vector to a row vector so in vector notation we can write $\nabla_x = \frac{\partial}{\partial x}^T$. Applied to $f(x,y)$:

$$\nabla_x f(x,y) = \nabla_x(x^TAx) + \nabla_x(y^TBx) = \frac{\partial}{\partial x_i}\sum_{j=1}^{n}\sum_{k=1}^{n} x_jA_{jk}x_k + \frac{\partial}{\partial x_i}\sum_{j=1}^{n}\sum_{k=1}^{n} y_jB_{jk}x_k$$

where $i \in \{1, ..., n\}$. Starting with the first term:

$$\begin{aligned} \frac{\partial}{\partial x_i}\sum_{j=1}^{n}\sum_{k=1}^{n} x_jA_{jk}x_k &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial x_j}{\partial x_i}A_{jk}x_k + \sum_{j=1}^{n}\sum_{k=1}^{n} x_jA_{jk}\frac{\partial x_k}{\partial x_i} \\ &= \sum_{k=1}^{n} A_{ik}x_k + \sum_{j=1}^{n} x_jA_{ji} = \sum_{k=1}^{n} A_{ik}x_k + \sum_{j=1}^{n} A_{ji}x_j \end{aligned}$$

In the first term the free index i iterates over the columns of A, and in the second term over the rows of A. Meanwhile, there is really no difference between indices k and j with respect to vector elements of x except to indicate the correct element of A within the sums themselves, additionally we know in general that matrix multiplication is distributive with respect to addition, so we can write the two individual sums over a single sum. :

$$\sum_{k=1}^{n} A_{ik}x_k + \sum_{j=1}^{n} A_{ji}x_j = \sum_{j=1}^{n}(A_{ij} + A_{ji})x_j$$

For a fixed i, we recognize the retrieved expression as the i-th element of a matrix expression $[(A+A^T)x]_i$ so finally we can write the result in vector form:

$$\nabla_x(x^TAx) = (A + A^T)x$$

For the second term we, similarly, write:

$$\begin{aligned} \frac{\partial}{\partial x_i}\sum_{j=1}^{n}\sum_{k=1}^{n} y_jB_{jk}x_k &= \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial y_j}{\partial x_i}B_{jk}x_k + \sum_{j=1}^{n}\sum_{k=1}^{n} y_jB_{jk}\frac{\partial x_k}{\partial x_i} \\ &= \sum_{j=1}^{n} y_jB_{ji} = \left[B^Ty\right]_i \end{aligned}$$

Which we recognize as the vector expression $\nabla_x(y^TBx) = B^Ty$. Put together we have that

$$\nabla_x f(x,y) = (A + A^T)x + B^Ty$$

c. *[2 points]* What is $\nabla_y f(x, y)$ in terms of the summations over indices *and* vector notation?
Similarly to the previous problem except that we only have to observe a single term that has a functional dependence on y:

$$\frac{\partial}{\partial y_i}\sum_{j=1}^{n}\sum_{k=1}^{n} y_j B_{jk} x_k = \sum_{j=1}^{n}\sum_{k=1}^{n}\frac{\partial y_j}{\partial y_i} B_{jk} x_k + \sum_{j=1}^{n}\sum_{k=1}^{n} y_j B_{jk}\frac{\partial x_k}{\partial y_i}$$
$$= \sum_{j=1}^{n} B_{ik} x_k = [Bx]_i$$

From which we recognize the vector expression $\nabla_y(y^T Bx) = Bx$. Since the remaining terms in the expression will evaluate to zero, as there is no dependence on y, we can immediately write:

$$\nabla_y f(x, y) = Bx$$

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 $Tr(AB) = Tr(BA)$.
It can be shown in general that trace is invariant under cyclic permutations, which for case of n=2 looks like commutation:

$$\text{tr}(AB) = \sum_{i}^{m}(ab)_{ii} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{m} b_{ji}a_{ij} = \sum_{i}^{n}(ab)_{jj} = \text{tr}(BA)$$

using the fact that product of $n \times m$ and $m \times n$ matrix is an $m \times m$ matrix and that the trace of a square matrix can be rewritten as:

$$\text{tr}(A^T B) = \sum_{i,j} A_{ij} B_{ij}$$

as per Wikipedia.

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 $\sum_{i=1}^{n} v_i v_i^T$?
Basis of an $\mathbb{R}^d$ is spanned by d linearly independent vectors. So the maximum rank of a matrix will be d. Minimal rank of a matrix would always be 0 in case of a zero matrix, but since the vectors v must be non-zero the minimal rank of V would be 1 in the case all v are the same unit vectors.

b. What is the minimum and maximum rank of V?
Same as a) (are these supposed to be like trick questions?)

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

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

```
In [51]: 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])

         print(np.linalg.inv(A))
         print()
         print(np.linalg.inv(A)@b)
         print()
         print(A@c)

         [[ 0.125 -0.625  0.75 ]
          [-0.25   0.75  -0.5  ]
          [ 0.375 -0.375  0.25 ]]

         [-2.  1. -1.]

         [6 8 7]
```

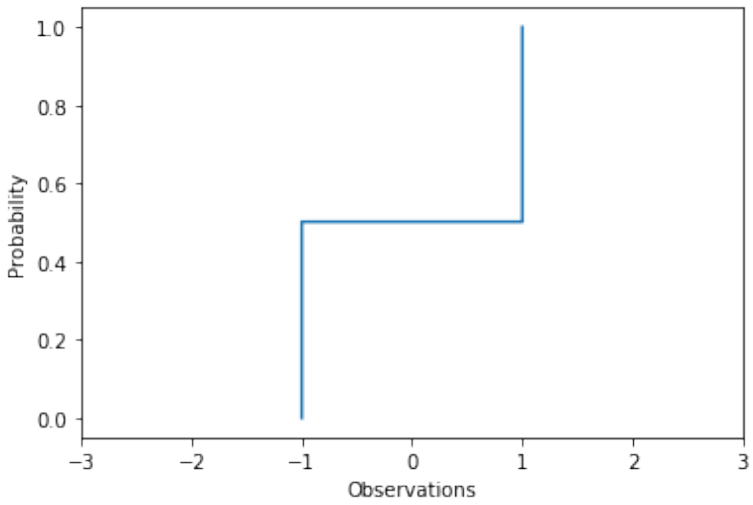
A.12 *[4 points]* Two random variables X and Y have equal distributions if their CDFs, F_X and F_Y, respectively, are equal, i.e. for all x, $|F_X(x) - F_Y(x)| = 0$. The central limit theorem says that the sum of k independent, zero-mean, variance-$1/k$ random variables converges to a (standard) Normal 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).

```
import matplotlib.pyplot as plt
import numpy as np

n = 20000
Z = np.sign(np.random.randn(n))
plt.step(sorted(Z), np.arange(1, n + 1) / float(n))

plt.xlim(-3, 3)
plt.xlabel("Observations")
plt.ylabel("Probability")
plt.show()
```



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.)

```
import matplotlib.pyplot as plt
import numpy as np

def Yk(n, k=1):
    """Returns and array of samples of function:

        Y^{(k)} = 1/sqrt(k) \sum_1^k B_i

    where B_i = +1 or +1 with equal probablility.

    Parameters
    ----------
    n : 'int'
        Number of samples of Y^(k).
    k : 'int'
        Upper bound of the Y^(k) sum
    """
    B = np.sign(np.random.randn(n, k))
    return np.sum(np.sqrt(1.0/k)*B, axis=1)


n = 20000
for k in [1, 8, 64, 512]:
    Y = Yk(n, k)
    plt.step(sorted(Y), np.arange(1, n + 1) / float(n), label='{0}'.format(k))
```

```
gaus = np.random.normal(size=n)
plt.step(sorted(gaus), np.arange(1, n + 1) / float(n), label='Gaussian')

plt.legend()
plt.xlim(-3, 3)
plt.xlabel("Observations")
plt.ylabel("Probability")
plt.show()
```