CSE 546 HW #0

Sam Kowash

October 4, 2018

(1) Analysis

1. A set $A \subseteq \mathbb{R}^n$ is convex if $\lambda x + (1-\lambda)y \in A$ for all $x, y \in A$. A norm $\|\cdot\|$ on $\mathbb{R}^n$ is non-negative, absolutely scalable, and satisfies the triangle inequality.

   (a) Let $A = \{x \in \mathbb{R}^n : \|x\| \leqslant 1\}$ for some norm $\|\cdot\|$, take $x, y \in A$, and take $\lambda \in [0,1]$. By the triangle inequality,

   $$\|\lambda x + (1-\lambda)y\| \leqslant \|\lambda x\| + \|(1-\lambda)y\|.$$

   Scalability then tells us

   $$\|\lambda x\| + \|(1-\lambda)y\| = |\lambda|\|x\| + |1-\lambda|\|y\|,$$

   and since $0 \leqslant \|x\|, \|y\| \leqslant 1$,

   $$|\lambda|\|x\| + |1-\lambda|\|y\| \leqslant |\lambda| + |1-\lambda|.$$

   Lastly, with $\lambda \in [0,1]$, $|\lambda| = \lambda$ and $|1-\lambda| = 1-\lambda$, so by applying transitivity through the chain of (in)equalities, we find

   $$\|\lambda x + (1-\lambda)y\| \leqslant 1,$$

   so $\lambda x + (1-\lambda)y \in A$ and A is convex, regardless of the norm used.

   (b) Consider the function $f(x) = \left(\sum_{i=1}^n |x_i|^{\frac{1}{2}}\right)^2$ for $x \in \mathbb{R}^n$. It is clearly non-negative for all inputs since it is a square of a sum of real numbers. Further, it is absolutely scalable, since a factor λ will come out of the sum as $\sqrt{|\lambda|}$, then be squared to give an overall scaling $|\lambda|$. However, consider $x = (1,0,0,\dots,0)$ and $y = (0,1,0,\dots,0)$. We have $f(x) = f(y) = 1$, but $f(x+y) = (\sqrt{1}+\sqrt{1})^2 = 4$, so $f(x+y) > f(x) + f(y)$ meaning that this function does not satisfy the triangle inequality and *cannot be a norm.* (We can also see this by plotting a level set of this function and noting that the region bounded by it is non-convex, in contradiction to the previous result.)

2. For $x \in \mathbb{R}^n$, define the norms

   $$\|x\|_1 = \sum_{i=1}^n |x_i|, \qquad \|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}, \qquad \|x\|_\infty = \max_{i=1,\dots,n} |x_i|.$$

   Each norm must obey the triangle inequality, so in particular (if we apply the inequality repeatedly),

   $$\|x\| = \left\|\sum_{i=1}^n x_i e_i\right\| \leqslant \sum_{i=1}^n \|x_i e_i\| = \sum_{i=1}^n |x_i| = \|x\|_1,$$

where e_i are unit basis vectors with respect to the given norm and the second equality comes by absolute scalability. Thus, we see that no norm exceeds the Manhattan norm. Further,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \geqslant \sqrt{(\max_{i=1,\ldots,n} |x_i|)^2} = \max_{i=1,\ldots,n} |x_i| = \|x\|_\infty,$$

so the infinity norm cannot exceed the Euclidean norm. This gives the ranking $\|x\|_\infty \leqslant \|x\|_2 \leqslant \|x\|_1$, as desired.

3. For $A, B \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}$, and $f(x,y) = x^T Ax + y^T Bx + c$ acting on $x, y \in \mathbb{R}^n$, we define the derivative

$$\nabla_z f(x,y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial z_1} & \frac{\partial f(x,y)}{\partial z_2} & \cdots & \frac{\partial f(x,y)}{\partial z_n} \end{bmatrix}^T.$$

It will be cleaner to compute these derivatives with matrix operations expressed component-wise, so for the rest of this problem we adopt Einstein summation notation wherein repeated indices are summed over. We will also use ∂_{z_i} to denote $\frac{\partial}{\partial z_i}$. Then,

$$\begin{aligned}(\nabla_x f(x,y))_i &= \partial_{x_i} \left(A_{jk}x_j x_k + B_{jk}y_j x_k + c\right)\\ &= \delta_{ij}A_j k x_k + \delta_{ik}A_j k x_j + \delta_{ik}B_{jk}y_j\\ &= A_{ik}x_k + x_j A_{ji} + y_j B_{ji},\end{aligned}$$

so, returning to matrix notation,

$$\nabla_x f(x,y) = (A + A^T)x + B^T y. \tag{1.1}$$

Similarly,

$$\begin{aligned}(\nabla_y f(x,y))_i &= \partial_{y_i} \left(A_{jk}x_j x_k + B_{jk}y_j x_k + c\right)\\ &= \delta_{ij}B_{jk}x_k\\ &= B_{ik}x_k,\end{aligned}$$

so

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

4. Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices sharing eigenvectors $u_1, \ldots, u_n$ with corresponding eigenvalues $\alpha_1, \ldots, \alpha_n$ for A and $\beta_1, \ldots, \beta_n$ for B.
   (a) The matrix $C = A+B$ shares the eigenvectors $u_1, \ldots, u_n$, with eigenvalues $\alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n$.
   (b) The matrix $D = A-B$ shares the eigenvectors $u_1, \ldots, u_n$, with eigenvalues $\alpha_1 - \beta_1, \ldots, \alpha_n - \beta_n$.
   (c) The matrix $E = AB$ shares the eigenvectors $u_1, \ldots, u_n$, with eigenvalues $\alpha_1\beta_1, \ldots, \alpha_n\beta_n$.
   (d) Assuming that A is nonsingular, the matrix $F = A^{-1}B$ shares the eigenvectors $u_1, \ldots, u_n$, with eigenvalues $\frac{\beta_1}{\alpha_1}, \ldots, \frac{\beta_n}{\alpha_n}$.
5. A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive-semidefinite (PSD) if $x^T Ax \geqslant 0$ for all $x \in \mathbb{R}^n$.

(a) Let $y \in \mathbb{R}^n$. For any $x \in \mathbb{R}^n$, note that $x^T y = y^T x$, so

$$x^T y y^T x = (y^T x)^2 \geqslant 0,$$

so yy^T is PSD.

(b) Let X be a random vector in $\mathbb{R}^n$ with covariance matrix $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right]$. By the previous result, $(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T$ is PSD since it's the outer product of a vector with itself. Sums of PSD matrices must be PSD, so any linear combination of PSD matrices with non-negative weights is in turn PSD. The expected value is such a combination, with weights given by the PDF of X, so Σ is PSD.

(c) Take A to be a symmetric matrix so that (by the real spectral theorem) there exist an orthogonal matrix U and list of eigenvalues α satisfying $A = U \operatorname{diag}(\alpha) U^T$. The columns of U form basis for $\mathbb{R}^n$, so we can decompose any vector $x \in \mathbb{R}^n$ as $x = \sum_i c_i u_i$, where c_i are scalars and u_i are the columns of U. Because that basis is in fact orthonormal, we know that $U u_i = e_i$, so

$$\begin{aligned}
x^T A x &= \left(\sum_i c_i u_i^T\right) U \operatorname{diag}(\alpha) U^T \left(\sum_j c_j u_j\right) \\
x^T A x &= \left(\sum_i c_i e_i^T\right) \operatorname{diag}(\alpha) \left(\sum_j c_j e_j\right) \\
x^T A x &= \sum_i c_i^2 \alpha_i,
\end{aligned}$$

where orthonormality collapses one of the sums. The resulting sum can only be non-negative for all choices of components c_i if the eigenvalues α_i are each non-negative, which is to say that $\min_i \alpha_i \geqslant 0$.

6. Consider real independent variables X and Y with PDFs f and g, respectively. Let h be the PDF of $Z = X + Y$.

(a) The PDF for the sum of X and Y will be the convolution of their respective PDFs:

$$h(x) = \int_{-\infty}^{\infty} f(y) g(x - y) \mathrm{d}y.$$

We can understand this as accounting for the probabilities of all of the different ways that outcomes of X and Y can add to a particular outcome x of Z.

(b) In particular we take X and Y both uniformly distributed on $[0, 1]$. It is easy to think about the convolution geometrically as sliding one unit square past the other and measuring the area of overlap. This falls to zero outside of $(0, 2)$ and rises to a maximum of 1 at $x = 1$, with linear change in between, so we can write

$$h(x) = \begin{cases} x & : 0 \leqslant x \leqslant 1 \\ 2 - x & : 1 < x \leqslant 2 \\ 0 & : \text{else} \end{cases}.$$

(c) We now wish to find $\mathbb{P}(X \leqslant 1/2 \mid X + Y \geqslant 5/4)$ for these distributions. This, too, we can think of geometrically: if we put values of X on the x-axis and those of Y on the

y-axis, the joint distribution has support on the unit square in the first quadrant. The event $X+Y \geqslant 5/4$ has probability equal to the area of the upper triangle bounded by $y = 5/4 - x$, and $(X \leqslant 1/2) \cap (X+Y \geqslant 5/4)$ is the portion of that triangle in the left half of the unit square. The diagonal boundary line for the triangle meets the edges of the square at $(1, 1/4)$ and $(1/4, 1)$, so the triangle has base and height $3/4$ and so area $9/32$. The smaller triangle has base and height $1/4$ and area $1/32$, so the conditional probability is $1/9$. See Fig. '1.1 for clarification.

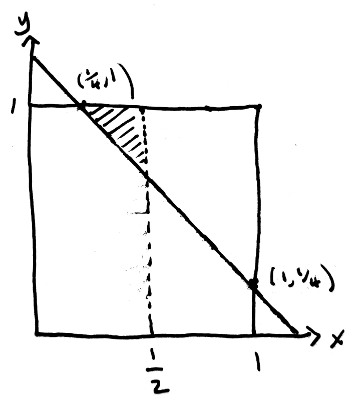

Figure 1.1: Geometry of conditional probability calculation

7. Given a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we wish to define $Y = aX + b$ such that $Y \sim \mathcal{N}(0,1)$. We can do this just by inspecting the differential probability

$$\mathrm{d}\mathbb{P}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \mathrm{d}x.$$

If we define $x = \frac{y-b}{a}$, we see

$$\mathrm{d}\mathbb{P}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}\right] \frac{\mathrm{d}y}{a},$$

which we can rearrange into

$$\mathrm{d}\mathbb{P}(y) = \frac{1}{\sqrt{2\pi(a\sigma)^2}} \exp\left[-\frac{(y-(b+a\mu))^2}{2(a\sigma)^2}\right] \mathrm{d}y.$$

It is then clear that $a = 1/\sigma$ and $b = -\frac{\mu}{\sigma}$ gives a Gaussian distribution with zero mean and unit variance, as desired.

8. Let $X_1, \ldots, X_n$ be i.i.d. random variables with CDF $F(x)$. We define $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leqslant x\}$, where $\mathbf{1}(A)$ is the indicator function for a Boolean event A, which is 1 when A is true and 0 otherwise.

(a) We know that for any random variable X, $\mathbb{E}[\mathbf{1}\{X \leqslant x\}]$ is equal to the CDF of X. Since all of our n variables have the same CDF,

$$\mathbb{E}[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n} F(x) = \frac{nF(x)}{n} = F(x).$$

(b) We now wish to find $\mathbb{E}[(\hat{F}_n(x) - F(x))^2]$, which by the previous result we recognize is equal to $\mathbb{E}[\hat{F}_n(x)^2] - F(x)^2$, so we should focus on computing $\mathbb{E}[\hat{F}_n^2]$.

$$\left(\hat{F}_n(x)\right)^2 = \frac{1}{n^2}\sum_{i,j=1}^{n} \mathbf{1}\{X_i \leqslant x\}\mathbf{1}\{X_j \leqslant x\}$$

This double sum has two types of terms. There are n diagonal terms where $i = j$ and we get a square of an indicator function, which is just the same indicator function back again and has expectation $F(x)$. There are also $n^2 - n$ off-diagonal terms where $i \neq j$ and we should think of the product as the indicator function for the event $(X_i \leqslant x) \cap (X_j \leqslant x)$ and its expected value as the joint CDF for X_i and X_j. Because all of our variables are independent, that joint CDF is just the product of CDFs for the variables, which are here the same, so these off-diagonal terms give $(F(x))^2$. Thus,

$$\mathbb{E}[(\hat{F}_n(x) - F(x))^2] = \frac{1}{n^2}\left[nF(x) + (n^2 - n)\left(F(x)\right)^2\right] - \left(F(x)\right)^2$$
$$\mathbb{E}[(\hat{F}_n(x) - F(x))^2] = \frac{F(x)\left(1 - F(x)\right)}{n}.$$

(c) Looking at the numerator of the previous result as a quadratic function of $F(x)$, we see that it is maximal when $F(x) = 1/2$, which gives the numerator a maximum value of $1/4$. This means that

$$\sup_{x\in\mathbb{R}} \mathbb{E}\left[\left(\hat{F}_n(x) - F(x)\right)^2\right] \leqslant \frac{1}{4n}.$$

(2) Programming

9. We investigate the central limit theorem numerically, using the family of random variables $Y^{(k)} = \frac{1}{\sqrt{k}}\sum_{i=1}^{k} B_i$, where each B_i takes values -1 and 1 with equal probability. The last result above tells us that

$$\sup_{x} \sqrt{\mathbb{E}[(\hat{F}_n(x) - F(x))^2]} \leqslant \frac{1}{2\sqrt{n}},$$

so for $n = 40000$ we can guarantee that the standard deviation of the empirical CDF $\hat{F}_n(x)$ is no greater than 0.0025.

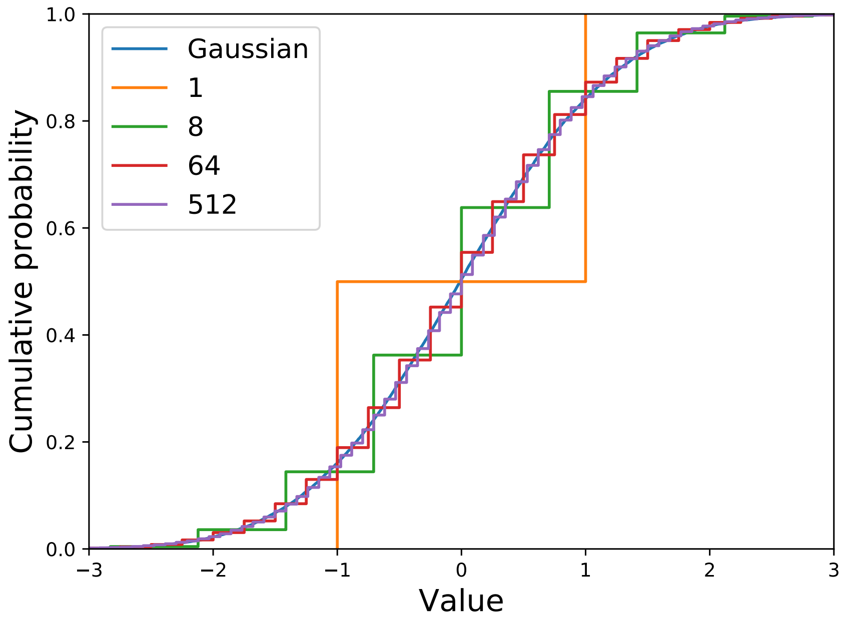

Figure 2.1: Empirical CDFs of Gaussian and of $Y^{(k)}$ for various k

Figure 2.1 plots the estimator $\hat{F}_n(x)$ for a Gaussian random variable and for our family $Y^{(k)}$ at various values of k. We can see that as k increases, $\hat{F}_n^{(k)}$ tends toward $\hat{F}_n$, justifying the claim of the central limit theorem that $Y^{(k)}$ tends toward a normal random variable as $k \to \infty$. Code follows on next page.

```python
#!/usr/bin/env python
import numpy as np
import matplotlib.pyplot as plt

n = 40000

#gaussian cdf
Z = np.random.randn(n)
plt.step(sorted(Z),np.arange(1,n+1)/n,label="Gaussian")


#y^(k) cdfs
for k in [1,8,64,512]:
    Yk = np.sum(np.sign(np.random.randn(n,k))*np.sqrt(1/k),axis=1)
    plt.step(sorted(Yk),np.arange(1,n+1)/n,label=k)




#plot details
plt.xlim((-3,3))
plt.ylim((0,1))
plt.xlabel("Value",fontsize=16)
plt.ylabel("Cumulative probability",fontsize=16)
plt.legend(fontsize=14)


#uncomment if testing
#plt.show()

#comment out if testing
plt.tight_layout()
plt.savefig('../figures/cdfs.pdf')
```