Lemma 53. **??** *Let* $T:(\Omega_1,\mathcal{F}_1,\mathbf{P})\to(\Omega_2,\mathcal{F}_2,\mathbf{Q})$ *be measurable and* $\mathbf{Q}=\mathbf{P}T^{-1}$. *If* X *is an integrable r.v. on* Ω_2, *then* $X\circ T$ *is an integrable r.v. on* Ω_1 *and* $\mathbf{E}_{\mathbf{P}}[X\circ T]=\mathbf{E}_{\mathbf{Q}}[X]$.

Proof. For a simple r.v., $X=\sum_{i=1}^n c_i\mathbf{1}_{A_i}$, where $A_i\in\mathcal{F}_2$, it is easy to see that $X\circ T=\sum_{i=1}^n c_i\mathbf{1}_{T^{-1}A_i}$ and by definition $\mathbf{E}_{\mathbf{P}}[X\circ T]=\sum_{i=1}^n c_i\mathbf{P}\{T^{-1}A_i\}=\sum_{i=1}^n c_i\mathbf{Q}\{A_i\}$ which is precisely $\mathbf{E}_{\mathbf{Q}}[X]$. Use MCT to get to positive r.v.s and then to general integrable r.v.s. ∎

Corollary 54. *Let* X_i, $i\le n$, *be random variables on a common probability space. Then for any Borel measurable* $f:\mathbb{R}^n\to\mathbb{R}$, *the value of* $\mathbf{E}[f(X_1,\dots,X_n)]$ *(if it exists) depends only on the joint distribution of* $X_1,\dots X_n$.

Remark 55. The change of variable result shows the irrelevance of the underlying probability space to much of what we do. That is, in any particular situation, all our questions may be about a finite or infinite collection of random variables X_i. Then, the answers depend only on the joint distribution of these random variables and not any other details of the underlying probability space. For instance, we can unambiguously talk of the expected value of Exp(λ) distribution when we mean the expected value of a r.v having Exp(λ) distribution and defined on some probability space.

Density: Let ν be a measure on $(\Omega,\mathcal{F})$ and $X:\Omega\to[0,\infty]$ a r.v. Then set $\mu(A):=\int_A Xd\nu$. Clearly, μ is a measure, as countable additivity follows from MCT. Observe that $\mu\ll\nu$. If two given measures μ and ν are related in this way by a r.v. X, then we say that X is the *density* or *Radon-Nikodym derivative* of μ w.r.t ν and sometimes write $X=\frac{d\mu}{d\nu}$. If it exists Radon-Nikodym derivative is unique (up to sets of ν-measure zero). The Radon-Nikodym theorem asserts that whenever μ,ν are σ-finite measures with $\mu\ll\nu$, the Radon Nikodym derivative does exist. When μ is a p.m on $\mathbb{R}^d$ and $\nu=\mathbf{m}$, we just refer to X as the pdf (probability density function) of μ. We also abuse language to say that a r.v. has density if its distribution has density w.r.t Lebesgue measure.

Exercise 56. Let X be a non-negative r.v on $(\Omega,\mathcal{F},\mathbf{P})$ and let $\mathbf{Q}(A)=\frac{1}{\mathbf{E}_{\mathbf{P}}[X]}\int_A Xd\mathbf{P}$. Then, $\mathbf{Q}$ is a p.m and for any non-negative r.v. Y, we have $\mathbf{E}_{\mathbf{Q}}[Y]=\mathbf{E}_{\mathbf{P}}[XY]$. The same holds if Y is real valued and assumed to be integrable w.r.t $\mathbf{Q}$ (or YX is assumed to be integrable w.r.t $\mathbf{P}$).

It is useful to know how densities transform under a smooth change of variables. It is an easy corollary of the change of variable formula and well-known substitution rules for computing integrals.

Corollary 57. *Suppose* $X=(X_1,\dots,X_n)$ *has density* $f(\mathbf{x})$ *on* $\mathbb{R}^n$. *Let* $T:\mathbb{R}^n\to\mathbb{R}^n$ *be injective and continuously differentiable. Write* $U=T\circ X$. *Then,* U *has density* g *which is given by* $g(\mathbf{u})=f(T^{-1}\mathbf{u})|\det(J[T^{-1}](\mathbf{u}))|$, *where* $[JT^{-1}]$ *is the Jacobian of the inverse map* T^{-1}.

More generally, if we can write $\mathbb{R}^n=A_0\cup\ldots\cup A_n$, *where* A_i *are pairwise disjoint,* $\mathbf{P}(X\in A_0)=0$ *and such that* $T_i:=T|_{A_i}$ *are one-one for* $i=1,2,\dots n$, *then,* $g(\mathbf{u})=\sum_{i=1}^n f(T_i^{-1}\mathbf{u})|\det(J[T_i^{-1}](\mathbf{u}))|$ *where the* i^{th} *summand is understood to vanish if* $\mathbf{u}$ *is not in the range of* T_i.

Proof. **Step 1** Change of Lebesgue measure under T: For $A\in\mathcal{B}(\mathbb{R}^n)$, let $\mu(A):=\int_A|\det(J[T^{-1}](\mathbf{u}))|dm(u)$. Then, as remarked earlier, μ is a Borel measure. For sufficiently nice sets, like rectangles $[a_1,b_1]\times\ldots\times[a_n,b_n]$, we know from Calculus class that $\mu(A)=\mathbf{m}(T^{-1}(A))$. Since rectangles generate the Borel sigma-algebra, and μ and $\mathbf{m}\circ T^{-1}$ agree on rectangles, by the $\pi-\lambda$ theorem we get $\mathbf{m}\circ T^{-1}=\mu$. Thus, $\mathbf{m}\circ T^{-1}$ is a measure with density given by $|\det(J[T^{-1}](\cdot))|$.

Step 2 Let B be a Borel set and consider

$$\begin{aligned}\mathbf{P}(U\in B) &= \mathbf{P}(X\in T^{-1}B)=\int f(x)\mathbf{1}_{T^{-1}B}(x)d\mathbf{m}(x)\\ &= \int f(T^{-1}u)\mathbf{1}_{T^{-1}B}(T^{-1}u)d\mu(u)=\int f(T^{-1}u)\mathbf{1}_B(u)d\mu(u)\end{aligned}$$

where the first equality on the second line is by the change of variable formula of Lemma **??**. Apply exercise 56 and recall that μ has density $|\det(J[T^{-1}](\mathbf{u}))|dm(u)$ to get

$$\mathbf{P}(U\in B)=\int f(T^{-1}u)\mathbf{1}_B(u)d\mu(u)=\int f(T^{-1}u)\mathbf{1}_B(u)|\det(J[T^{-1}](\mathbf{u}))|dm(u)$$

which shows that U has density $f(T^{-1}u)|\det(J[T^{-1}](\mathbf{u}))|$.

To prove the second part, we do the same, except that in the first step, (using $\mathbf{P}(X\in A_0)=0$, since $\mathbf{m}(A_0)=0$ and X has density) $\mathbf{P}(U\in B)=\cup_{i=1}^n\mathbf{P}(X\in T_i^{-1}B)=\sum_{i=1}^n\int f(x)\mathbf{1}_{T_i^{-1}B}(x)d\mathbf{m}(x)$. The rest follows as before. ∎

Suppose we know the joint distribution of $X = (X_1, \dots, X_n)$. Then we can find the distribution of any function of X because $\mathbf{P}(f(X) \in A) = \mathbf{P}(X \in f^{-1}(A))$. When X has a density, one can get simple formulas for the density of the sum, product etc., that are quite useful.

In the examples that follow, let us assume that the density is continuous. This is only for convenience, and so that we can invoke theorems like $\iint f = \int \left(\int f(x,y)dy\right)dx$. Analogous theorems for Lebesgue integral will come later (*Fubini's theorem*)...

Example 58. Suppose (X,Y) has density $f(x,y)$ on $\mathbb{R}^2$. What is the distribution of X? Of $X+Y$? Of X/Y? We leave you to see that X has density $g(x) = \int_{\mathbb{R}} f(x,y)dy$. Assume that f is continuous so that the integrals involved are also Riemann integrals and you may use well known facts like $\iint f = \int \left(\int f(x,y)dy\right)dx$. The condition of continuity is unnatural and the result is true if we only assume that $f \in L^1$ (w.r.t. Lebesgue measure on the plane). The right to write Lebesgue integrals in the plane as iterated integrals will be given to us by *Fubini's theorem* later.

Suppose (X,Y) has density $f(x,y)$ on $\mathbb{R}^2$.

(1) X has density $f_1(x) = \int_{\mathbb{R}} f(x,y)dy$ and Y has density $f_2(y) = \int_{\mathbb{R}} f(x,y)dx$. This is because, for any $a < b$, we have

$$\mathbf{P}(X \in [a,b]) = \mathbf{P}((X,Y) \in [a,b] \times \mathbb{R}) = \int\limits_{[a,b]\times\mathbb{R}} f(x,y)dxdy = \int\limits_{[a,b]} \left(\int\limits_{\mathbb{R}} f(x,y)dy \right) dx.$$

This shows that the density of X is indeed f_1.

(2) Density of X^2 is $(f_1(\sqrt{x}) + f_1(-\sqrt{x}))/2\sqrt{x}$ for $x > 0$. Here we notice that T is one-one on $\{x > 0\}$ and $\{x < 0\}$ (and $\{x = 0\}$ has zero measure under f), so the second statement in the proposition is used.

(3) The density of $X+Y$ is $g(t) = \int_{\mathbb{R}} f(t-v,v)dx$. To see this, let $U = X+Y$ and $V = Y$. Then the transformation is $T(x,y) = (x+y,y)$. Clearly $T^{-1}(u,v) = (u-v,v)$ whose Jacobian determinant is 1. Hence by corollary 57, we see that (U,V) has the density $g(u,v) = f(u-v,v)$. Now the density of U can be obtained like before as $h(u) = \int g(u,v)dv = \int f(u-v,v)dv$.

(4) To get the density of XY, we define $(U,V) = (XY,Y)$ so that for $v \neq 0$, we have $T^{-1}(u,v) = (u/v,v)$ which has Jacobian determinant v^{-1}.

We claim that $X+Y$ has the density $g(t) = \int_{\mathbb{R}} f(t-v,v)dx$.

Exercise 59. (1) Suppose (X,Y) has a continuous density $f(x,y)$. Find the density of X/Y. Apply to the case when (X,Y) has the *standard bivariate normal distribution* with density $f(x,y) = (2\pi)^{-1}\exp\{-\frac{x^2+y^2}{2}\}$.

(2) Find the distribution of $X+Y$ if (X,Y) has the standard bivariate normal distribution.

(3) Let $U = \min\{X,Y\}$ and $V = \max\{X,Y\}$. Find the density of (U,V).

Given a r.v. or a random vector, expectations of various functions of the r.v give a lot of information about the distribution of the r.v. For example,

Proposition 60. *The numbers* $\mathbf{E}[f(X)]$ *as* f *varies over* $C_b(\mathbb{R})$ *determine the distribution of* X.

Proof. Given any $x \in \mathbb{R}^n$, we can recover $F(x) = \mathbf{E}[\mathbf{1}_{A_x}]$, where $A_x = (-\infty, x_1] \times \ldots \times (-\infty, x_n]$ as follows. For any $\delta > 0$, let $f(y) = \min\{1, \delta^{-1} d(y, A^c_{x+\delta\mathbf{1}})\}$, where d is the L_∞ metric on $\mathbb{R}^n$. Then, $f \in C_b(\mathbb{R})$, $f(y) = 1$ if $y \in A_x$, $f(y) = 0$ if $y \notin A_{x+\delta\mathbf{1}}$ and $0 \le f \le 1$. Therefore, $F(x) \le \mathbf{E}[f \circ X] \le F(x + \delta\mathbf{1})$. Let $\delta \downarrow 0$, invoke right continuity of F to recover $F(x)$. ■

Much smaller sub-classes of functions are also sufficient to determine the distribution of X.

Exercise 61. Show that the values $\mathbf{E}[f \circ X]$ as f varies over the class of all smooth (infinitely differentiable), compactly supported functions determine the distribution of X.

Expectations of certain functionals of random variables are important enough to have their own names.

Definition 62. Let X be a r.v. Then, $\mathbf{E}[X]$ (if it exists) is called the *mean* or *expected value* of X. $\mathrm{Var}(X) := \mathbf{E}\left[(X - \mathbf{E}X)^2\right]$ is called the *variance* of X, and its square root is called the *standard deviation* of X. The standard deviation measures the spread in the values of X or one way of measuring the uncertainty in predicting X. For any $p > 0$, if it exists, $\mathbf{E}[X^p]$ is called the p^{th}*-moment* of X. The function ψ defined as $\psi(\lambda) := \mathbf{E}[e^{\lambda X}]$ is called the *moment generating function* of X. Note that the m.g.f of a non-negative r.v. exists for all $\lambda < 0$. It may exist for some $\lambda > 0$ also. A similar looking object is the *characteristic function* of X, define by $\phi(\lambda) := \mathbf{E}[e^{i\lambda X}] := \mathbf{E}[\cos(\lambda X)] + i\mathbf{E}[\sin(\lambda X)]$. This exists for all $\lambda \in \mathbb{R}$.

For two random variables X, Y on the same probability space, we define their *covariance* to be $\mathrm{Cov}(X,Y) := \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$. The *correlation coefficient* is measured by $\frac{\mathrm{Cov(X,Y)}}{\sqrt{\mathrm{Var(X)Var(Y)}}}$. The correlation coefficient lies in $[-1, 1]$ and measures the association between X and Y. A correlation of 1 implies $X = Y$ a.s while a correlation of -1 implies $X = -Y$ a.s. Covariance and correlation depend only on the joint distribution of X and Y.

Exercise 63. (i) Express the mean, variance, moments of $aX + b$ in terms of the same quantities for X.

(ii) Show that $\mathrm{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$.

(iii) Compute mean, variance and moments of the Normal, exponential and other distributions defined in section 7.

Example 64 (The exponential distribution). Let $X \sim \mathrm{Exp}(\lambda)$. Then, $\mathbf{E}[X^k] = \int x^k d\mu(x)$ where μ is the p.m on $\mathbb{R}$ with density $\lambda e^{-\lambda x}$ (for $x > 0$). Thus, $\mathbf{E}[X^k] = \int x^k \lambda e^{-\lambda x} dx = \lambda^{-k} k!$. In particular, the mean is λ, the variance is $2\lambda^2 - (\lambda)^2 = \lambda^2$. In case of the normal distribution, check that the even moments are given by $\mathbf{E}[X^{2k}] = \prod_{j=1}^{k}(2j-1)$.

Remark 65 (Moment problem). Given a sequence of numbers $(\alpha_k)_{k \ge 0}$, is there a p.m μ on $\mathbb{R}$ whose k^{th} moment is α_k? If so, is it unique?

This is an extremely interesting question and its solution involves a rich interplay of several aspects of classical analysis (orthogonal polynomials, tridiagonal matrices, functional analysis, spectral theory etc). Note that there are are some non-trivial conditions for (α_k) to be the moment sequence of a p.m. μ. For example, $\alpha_0 = 1$, $\alpha_2 \ge \alpha_1^2$ etc. In the homework you were asked to show that $((\alpha_{i+j}))_{i,j \le n}$ should be a n.n.d. matrix for every n. The non-trivial answer is that these conditions are also sufficient!

Note that like proposition 60, the uniqueness question is asking whether $\mathbf{E}[f \circ X]$, as f varies over the space of polynomials, is sufficient to determine the distribution of X. However, uniqueness is not true in general. In other words, one can find two p.m μ and ν on $\mathbb{R}$ which have the same sequence of moments!