I. NOTES ON PROBABILITY AND RANDOM VARIABLES

A single-valued function $X(s)$ is called a *random variable* (actually a 'function' in traditional sense!) if it maps each outcome s of sample space $\mathcal{S}$ to a real number; two distinct outcomes may be mapped to same real number but not vice versa. One can then use the random variable to define events.

Cumulative distribution function (cdf) $F_X(x)$ of X is defined as

$$F_X(x) = P(X \leq x),\, x \in (-\infty, +\infty). \tag{1}$$

If a cdf changes values only in a countable number of jumps and is otherwise constant between two subsequent jumps, the corresponding X is a *discrete random variable*. Essentially, a discrete random variable is the one for which $\{X(s) : s \in \mathcal{S}\}$ is finite or countably infinite. On the other hand, the discrete random variable is *continuous* if $\{X(s) : s \in \mathcal{S}\}$ is either a single interval or a set of disjoint intervals of real line. Note that for a continuous random variable, $F_X(x)$ is continuous and its piecewise continuous derivative exists everywhere except maybe at a finite number of points. One can construct a *mixed* random variable that has a cdf with properties of both the discrete and the continuous random variables. The notion of *percentile* is defined via cdf: The u $(\in [0,1])$ percentile of X is the smallest real number x_u that satisfies $u = F_X(x_u)$.

For the discrete random variable, one can define *probability mass function* (pmf):

$$p_X(x) = P(X = x), \tag{2}$$

where

$$P(X = x_i) = F_X(x_i) - F_X(x_{i-1}). \tag{3}$$

Here subscript i denotes the points of jumps for the cdf.

Equivalently for the continuous random variable, one can define *probability density function* (pdf):

$$p_X(x) = \frac{d}{dx} F_X(x). \tag{4}$$

One can invert this to get,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} p_X(x) dx. \tag{5}$$

Although we are using the same symbol for pmf and pdf, it should not be source of a confusion as the context with make it clear.

It is straightforward to define *conditional cdf* of an event given some event E:

$$F(x|E) = P(X \leq x|E) = \frac{P(\{s : X(s) \leq x\} \cap E)}{P(E)}. \tag{6}$$

Using this definition, the corresponding definitions for *conditional pmf* or *conditional pdf* follow in analogy with what was done before for the unconditional cases.

An all important concept in the theory of random variable is that of the functions of random variables. Consider a function $f(x)$. Notationally,

$$Y = f(X) \tag{7}$$

defines a new random variable such that for any subset $\mathcal{I}$ with $f(x) \leq y$ and $\mathcal{I} \subseteq$ (range of X) $\subseteq \mathbb{R}$,

$$\{s : Y = f(X(s)) \leq y\} = \{s : X(s) \in \mathcal{I}\}. \tag{8}$$

Naturally, for continuous random variable,

$$F_Y(y) = P(Y \leq y) = P(X \in \mathcal{I}) = \int_{\mathcal{I}} p_X(x) dx\,. \tag{9}$$

The pdf of Y can then be obtained through,

$$p_Y(y) = \sum_i \frac{p_X(x_i)}{|f'(x_i)|}\,, \tag{10}$$

where x_i's are the real roots of $y = f(x)$.

The *expectation* of a function of a random variable is given by

$$E[Y] = E[f(X)] = \sum_i f(x_i) p_X(x_i) \text{ and } \int_{-\infty}^{+\infty} f(x) p_X(x) dx, \tag{11}$$

for discrete and continuous cases respectively. Expectations of the following three particular functions are useful:

- *Moment generating function*:

  $$M_X(t) \equiv E[e^{tX}],\, t \in \mathbb{R}. \tag{12}$$

  $M_X(t)$ need not exist as the integral or the sum may not converge for all t. The beauty of $M_X(t)$ is that nth *moment* m_n $(n \in \mathbb{N})$—defined as $E(X^n)$—can be obtained from it through

  $$m_n \equiv E[X^n] = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}. \tag{13}$$

  An important result is that if two random variables have same moment generating functions, then they necessarily have same probability distribution. Sometimes it is convenient to work with *central moments* μ_n $(n = 2, 3, \cdots)$ defined as

  $$\mu_n \equiv E[(X - m_1)^n], \tag{14}$$

  that measures deviation of the random variables from the *mean* m_1 (mostly commonly denoted by μ with no subscript; hence no confusion with the central moments). Mean measures the *central tendency* of a probability distribution, i.e., it is a measure that locates where the data are concentrated. Two other widely used measures of the central tendency are: *median* and *mode*. Median is value of the random variable such that the probability of getting smaller and greater values are equal; and mode is the most probable value of the probability distribution. Note that median is 0.5 percentile of X.

- *Characteristic function*:

$$\phi_X(t) \equiv E[e^{\sqrt{-1}tX}],\ t \in \mathbb{R}. \tag{15}$$

  Unlike $M_X(t)$, $\phi_X(t)$ always exists. One can prove that if two random variables have same caracteristic functions, then they necessarily have same probability distribution.

- *Cumulant generating function*: The cumulant generating function

$$K_X(t) \equiv \ln E[e^{tX}] = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!},\ t \in \mathbb{R}, \tag{16}$$

  yields *cumulants* κ_n through the following relation:

$$\kappa_n = \left.\frac{d^n}{dt^n} K_X(t)\right|_{t=0}. \tag{17}$$

  The cumulants are related to moments and central moments. In fact, the first cumulant κ_1 is the mean (μ); and the next two, κ_2 and κ_3, are respectively equal to the central moments μ_2 (usually denoted by σ^2, commonly known as the *variance*) and μ_3 (whose scaled form—μ_3/σ^3—is called *skewness*). Positive skewness means the probability distribution is skewed towards right and negative means the distribution is skewed towards left. The fourth cumulant κ_4 is not equal to any central moment but is related to μ_4 as follows: $\kappa_4 = \mu_4 - 3\sigma^4$. The scaled version of κ_4 is known as *kurtosis* $\kappa \equiv \kappa_4/\sigma^4$. A positive kurtosis makes the distribution *leptokurtic* meaning that the distribution's tail is heavier than in a *Gaussian* (also called *normal*) distribution $\mathcal{N}(\mu, \sigma^2)$ specified by

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{18}$$

  Whereas a negative value of kurtosis indicates that distribution is *platykurtic* meaning that the tail is lighter than in a Gaussian distribution. (Near zero value for kurtosis means that the distribution is *mesokurtic*. Interestingly, the "lightness" or "heaviness" in these Greek names do not refer to the tails; rather they refer to the region around the peak: Because of the normalization of probability distribution, while a heavier tail implies a lighter weight in the peak region, a lighter tail implies more weight around the peak region.)

Let us mention three extremely important practical results concerning sequences of *iid* (independent, identically distributed) random variables. While the mathematically precise meaning of the term iid is defined later but the idea should be clear from the following intention that we have: Our intention is to say something concrete about a sample that contains a sequence $X_1, X_2, \cdots, X_N$ (called *sample vector* or *random sample*; to be denoted as $\{X_i\}_{i=1}^{i=N}$ or simply $\{X_i\}$) of N observations for a particular experimental setup or phenomenon. Because we are interested in a particular fixed experimental setup, it is natural to assume that all the observations are modelled by identical random variables. Furthermore, assume that each observation is randomly generated independent of the other observations. One can define the *sample mean*, $\bar{X}_N \equiv (\sum_i^N X_i)/N$, for the sample. If many such samples are collected, what is the statistical behaviour of the sample mean?

- *(Strong) Law of large numbers*: If each of the iid random variables have a finite mean μ, then for any positive ϵ,

$$P\left(\lim_{N\to\infty} |\bar{X}_N - \mu| > \epsilon\right) = 0. \tag{19}$$

  It essentially says that the sequence of sample means converges to the mean of random variable X_i with full certainty. The weak law—$\lim_{N\to\infty} P\left(|\bar{X}_N - \mu| > \epsilon\right) = 0$—can be inferred from the strong law. If $E[X_i] = \infty$, the strong law does not hold but using the idea of conditionally convergent series, sometimes a finite expectation value may be defined and the weak law may be seen to be obeyed with convergence towards the redefined expectation value.

- *Central limit theorem* (CLT): If each of the iid random variables have a finite mean μ and a finite variance σ^2, then

$$\lim_{N\to\infty} \left(\frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}}\right) \xrightarrow{\text{distribution}} \mathcal{N}(0, 1). \tag{20}$$

  The CLT provides more detailed description about the sample mean than the laws of large numbers. It allows to approximately find a probability of $P(\bar{X}_N > x)$ because we know that we just have to use $\mathcal{N}(0, 1)$. However, such an approximation by the CLT in the tails—where x is quite distant from μ—will not be very accurate if N is not large enough. Moreover, the CLT does not furnish any detail about the convergence of the tail probabilities with the sample size. This is where a theorem from the theory of large deviations comes to the rescue.

- *Cramér's theorem*: If each of the iid random variables have a finite mean μ and a cumulant generating function $K_{X_i}(t) = \ln E[e^{tX_i}]$, then $\bar{X}_N$ satisfies the *large deviation principle*, i.e.,

$$\lim_{N\to\infty} -\frac{1}{N} \ln P(\bar{X}_N \geq x) = I(x)\ \forall x > \mu, \tag{21}$$

  where $I(x)$—the Legendre–Fenchel transform of $K(t)$, i.e., $I(x) \equiv \sup_t(tx - K_{X_i}(t))$—is called *rate function* or *Cramér's function*. Sometimes, this is less rigorously presented as

$$P(\bar{X}_N \geq x) \approx e^{-NI(x)}, \tag{22}$$

where '≈' sign means that only the dominant exponential term has been kept and any sub-exponential terms have been ignored (hence the limit in the rigorous definition).

Readers must have already noted a technical point that $\bar{X}_N$ is defined on combined sample space—$\mathcal{S}^N \equiv \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_N$—where $\mathcal{S}_i$ is the sample space corresponding to the random variable X_i. Thus, by independent random variable we mean that $X_i(s_1, s_2, \cdots, s_N) = X_i(s_i)$, where $s_i \in \mathcal{S}_i$. Moreover, if $\mathcal{S}_i = \mathcal{S}_j = \mathcal{S}\ \forall i, j$ and $X_i(s_i) = X(s_i)\ \forall i$ (where X is some random variable defined on $\mathcal{S}$), then X_i's can be said to be identically distributed. This way of looking at multiple random variables begs the question what if two or more random variables are not independent or not identical; or more generally, how to generalize the concepts in single random variable theory to multiple random variables. While all the preceding ideas can be generalized to the case of multiple random variables, some newer concepts emerge. Let us list them. For convenience, we shall only work with the two random variables unless specified otherwise; further generalization is straightforward but notationally cumbersome.

First of all, we must realise that, for consistency of the concepts below, we should define a *bivariate* or *two-dimensional* random variable (X, Y) on the same sample space. (Y should not be confused as the function of X as used earlier in this section.) The *joint cumulative distribution function* can be defined as

$$F_{XY}(x, y) = P(X \leq x, Y \leq y);\ x, y \in (-\infty, +\infty). \quad (23)$$

The *joint probability mass function* is

$$p_{XY}(x, y) = P(X = x, Y = y), \quad (24)$$

and *joint probability density function* is

$$p_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y). \quad (25)$$

One defines *marginal cumulative distribution function* as

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y) = P(X \leq x, Y \leq \infty). \quad (26)$$

Furthermore, *conditional* pmf or pdf can now be defined as

$$p_{X|Y}(x|y) \equiv \frac{p_{XY}(x, y)}{p_Y(y)},\ p_Y(y) \neq 0. \quad (27)$$

Symmetry of $p_{XY}(x, y)$ in the arguments x and y, allows to write *Bayes' rule* for the random variables:

$$p_{X|Y}(x|y) \equiv \frac{p_{Y|X}(y|x) p_X(x)}{p_Y(y)},\ p_Y(y) \neq 0. \quad (28)$$

In the above notations, the random variables X and Y are defined to be independent if

$$F_{XY}(x, y) = F_X(x) F_Y(y), \quad (29)$$

and then, depending on whether we are dealing with discrete or continuous case, we have

$$p_{XY}(x, y) = p_X(x) p_Y(y), \quad (30)$$

$$p_{X|Y}(x|y) = p_X(x) \text{ and } p_{Y|X}(y|x) = p_Y(y). \quad (31)$$

Here, $p_X(x)$ and $p_Y(y)$ are the marginal versions obtained after summing or integrating over all possible values of y and x respectively.

One can furthermore define (i, j)th moment as

$$m_{ij} \equiv E[X^i Y^j]. \quad (32)$$

Similarly, one can define bivariate central moments and cumulants; and also different generating functions. Two concepts are of particular interest: *orthogonal* X and Y, and *uncorrelated* X and Y. X and Y are said to be orthogonal if $m_{11} = 0$; and they are said to be uncorrelated (otherwise, they are correlated) if their *covariance* $\text{Cov}(X, Y)$ (or σ_{XY}), given by

$$\sigma_{XY} \equiv E[(X - m_{10})(Y - m_{01})], \quad (33)$$

is zero. Note that since

$$\sigma_{XY} = E[XY] - E[X]E[Y], \quad (34)$$

independent random variables are uncorrelated but the converse need not be true.

We end this section by introducing a rather useful special multivariate random distribution: *n-variate Gaussian (normal) distribution*. Denoting $\boldsymbol{X}$ as a column vector $[X_1\, X_2\, \cdots\, X_n]^T$ and its particular realization as $\boldsymbol{x} \equiv [x_1\, x_2\, \cdots\, x_n]^T$, we define $\{X_i\}_{i=1}^{i=n}$ an n-variate Gaussian (normal) random variable if its joint pdf is

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathsf{C}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \mathsf{C}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}, \quad (35)$$

where $\boldsymbol{\mu}$ is a column vector $[E[X_1]\, E[X_2]\, \cdots\, E[X_n]]^T$ and $\mathsf{C}_{ij} \equiv \sigma_{X_i X_j}$ is the (i, j)-th element of the $n \times n$ matrix C—the *covariance matrix*—whose determinant has been denoted by $|\mathsf{C}|$ in the pdf. The set of such n variables are equivalently termed *jointly normal*, an apt term because the sum $\sum_{i=1}^{n} k_i x_i$ can be shown to be a normal random variable for any set of numbers, a_i's. We remark that if additionally the variables are uncorrelated, then they are independent as well.

II. NOTES ON THERMODYNAMICS

Thermodynamics is mainly the study of energy, work done, and heat exchanges in macroscopic systems where the microscopic details of the systems are ignored and a coarse-grained pragmatic description of them is adopted. A thermodynamic system can be *isolated*, *closed*, or *open* respectively meaning that the wall bounding the system is *adiabatic* (allows transfer of neither heat or matter), *diathermal* (allows transfer of only heat), or neither (i.e., allows transfer of both heat and matter). The state of such a system is characterized by thermodynamic variables/coordinates: *extensive* variables/coordinates (e.g., energy, volume, polarization, magnetization, and number of particles that are proportional to the size of the system) and *intensive* variables/coordinates (e.g, pressure, electric field, magnetic field and chemical potential that are independent of the size of the system). We shall consider only *homogeneous* system such that value of any of the thermodynamic variables is unchanged across the system. The *first law of thermodynamics*, which essentially is statement of the conservation of energy, states: $dE = đQ + đW$, where dE is the internal energy, $đQ$ is the heat supplied to the system, and $đW$ is the (generalized) work done on the system.The symbol 'd-cut' is used to denote inexact differential; the total heat transferred and the total work done depend on the details/path of the process.

There is a special thermodynamical state, called *equilibrium state*, such that its thermodynamic coordinates do not appear change over time scales of observation period that is much greater than microscopic time scales but naturally not infinitely long. When in *thermodynamic equilibrium*, a system is simultaneously in *thermal equilibrium* (uniform constant temperature), *mechanical equilibrium* (balanced mechanical forces), *chemical equilibrium* (constant chemical composition), and *phase equilibrium* (no phase changes like melting and evaporation). It may appear a bit of a circular argument that the concept of thermal equilibrium (defined using the concept of temperature) in the light of the *zeroth law of thermodynamics*, expounds the existence of temperature T as a state variable. The zeroth law states that the thermal equilibrium among thermodynamic systems is a transitive property; thus, two systems in thermal equilibrium have identical temperatures. In theoretical thermodynamics, the temperature is expressed in *thermodynamic scale* (in the unit of Kelvin) by choosing the triple point of water-ice-steam system to be 273.16 Kelvin; the thermodynamic temperature must always non-negative otherwise it can be shown to lead to the violation the *second law of thermodynamics* which we state below.

Suppose a *quasistatic* process on a system is performed meaning that the process is so slow that each step of the process the system is given enough time to settle down practically into an equilibrium state and consequently the specific values of the thermodynamic variables can define it unambiguously. In such a process, the infinitesimal work done, $đW$, due to change in the value of an extensive variable $\mathcal{D}_i$ under the influence of an intensive variable $\mathcal{F}_i$ (not constructed by dividing an extensive variable by system's volume or mass) can be written as $đW = \mathcal{F}_i d\mathcal{D}_i$; thus, $\mathcal{D}_i$ may be seen as a generalized displacement and $\mathcal{F}_i$ as its conjugate generalized force. However, what about the conjugate variable of temperature T, an intensive variable? We consider a further refinement of quasistatic process: *reversible process.* A reversible process is a quasistatic process that when run backward in time from the final state leads to the unchanged initial state. It is established by Clausius that for such a process $đQ/T$ is an exact differential, dS where S is an extensive variable and called it entropy; we specifically call it *Clausius entropy* in order to differentiate it from the other definitions of entropy. Therefore, the first law can be written as: $dE = TdS + \sum_i \mathcal{F}_i d\mathcal{D}_i$. For any arbitrary process (not necessarily reversible) leading to heat input in the system, $dS \geq đQ/T$; here, dS is the change in entropy assuming a virtual reversible process. It should be noted that the Clausius entropy is defined only for the equilibrium states.

The futile endeavours of inventing perfect engine (which transforms heat completely to work) and perfect refrigerator (which transfers heat from cold body to hot body without any energy consumption) led to an empirical law—*the second law of thermodynamics*—that is often expressed using the Clausius entropy: Entropy of an isolated thermodynamic system can not decrease in any (spontaneous) process, i.e., $dS \geq 0$. The entropy of a system with its isolated subsystems at individual distinct equilibria, goes to an equilibrium state of higher entropy, following irreversible exchanges of heat inside it once all the adiabatic walls are removed between the subsystems. The intermediate non-equilibrium states are captured by the time-varying values of $\mathcal{D}_i$'s. Thus, stated another way, in the final equilibrium state, the entropy of an isolated system at the fixed total energy is maximized by the set of $\mathcal{D}_i$'s that take their equilibrium values in the final state. This is the *entropy maximum principle.*

In passing, we observe that the Clausius entropy is defined only up to an additive constant. In this context, we mention that there is one more law, the *third law of thermodynamics*, that states that the entropy of any system is a universal constant—conveniently taken as zero—at $T = 0$. It can be shown that the third law implies unattainability (in finite number of steps) of the absolute zero temperature, and also implies vanishing of thermal expansivities and heat capacities of the system at $T = 0$. However, since at very low temperature the classical mechanics is not strictly valid, the third law explicitly requires quantum mechanics for its justification.

The Clausius entropy for a system in equilibrium may be envisaged to be determinable from E and $\mathcal{D}_i$'s, i.e., $S = S(E, \{\mathcal{D}_i\})$—called *(entropic) fundamental relation*—whose specific form would depend on the system in hand. All the intensive variables can, thus, be obtained as first order partial derivatives of S with respect to $\mathcal{D}_i$'s. The relations relating the intensive with the extensive variables are called the *equation of states.* Furthermore, inverting the preceding equation to write the *(energetic) fundamental relation*, $E = E(S, \{\mathcal{D}_i\})$, we can mathematically write the assumed extensivity of energy as $E(\lambda S, \{\lambda \mathcal{D}_i\}) = \lambda E(S, \{\mathcal{D}_i\})$ ($\lambda \in \mathbb{R}_{\geq 0}$) which is valid in case there is short-range interactions between constituent particles of the system because then the energy is additive and hence extensive (converse may not always hold). Along with the first law, it implies the *Eu-*

ler relation, $E = TS + \sum_i \mathcal{F}_i \mathcal{D}_i$, that in turn leads to the *Gibbs–Duhem relation*: $SdT + \sum_i \mathcal{D}_i d\mathcal{F}_i$=0.

At this juncture, it is of use to comment (without proof) that the maximum entropy principle is equivalent to the *minimum energy principle*: In the final equilibrium state, the energy of a closed system at the fixed total entropy is minimized by the set of $\mathcal{D}_i$'s that take their equilibrium values in the final state. In practice, the tuneable thermodynamic variables may not be the extensive variables and so E or S may not be the right quantities for finding the equilibrium state variables. Therefore, with the Euler relation in the back of ones mind, one uses the Legendre transformation to define *enthalpy* $\mathcal{H} \equiv E - \sum_i \mathcal{F}_i \mathcal{D}_i$, *Helmholtz free energy* $F \equiv E - TS$, *Gibbs free energy* $G \equiv E - TS - \sum_i \mathcal{F}_i \mathcal{D}_i$, and *grand potential* $\mathcal{G} \equiv E - TS - \sum_i \mu_i N_i$. (Note here we are not considering the pairs (μ_i, N_i)—the chemical potential of a species and the corresponding number of particles—in the set of the pairs $(\mathcal{F}_i, \mathcal{D}_i)$; in other words, we are separating chemical work from mechanical work.) They are respectively minimized during adiabatic transformation with mechanical work at constant generalized forces, isothermal transformation in absence of mechanical work, isothermal transformation with mechanical work at constant generalized forces, and isothermal transformation with chemical work at constant chemical potentials. In view of the afore-discussed minimization schemes, E, $\mathcal{H}$, F, G, and $\mathcal{G}$ are appositely called *thermodynamic potentials* in analogy with the role of the potential energy in the Newtonian dynamics.

If we can find the fundamental relation in terms of the Clausius entropy (or, equivalently, in terms of the thermodynamic potentials), then all the macroscopic thermodynamic properties of the system can be evaluated.