

# ENTROPY FOR MIXTURES OF DISCRETE AND CONTINUOUS VARIABLES

CHANDRA NAIR, BALAJI PRABHAKAR, AND DEVAVRAT SHAH

ABSTRACT. In this paper, we extend the notion of entropy in a natural manner for a mixed-pair random variable, a pair of random variables with one discrete and the other continuous. Our extensions are consistent in that there exist natural injections from discrete or continuous random variables into mixed-pair random variables such that their entropy remains the same. This extension of entropy allows us to obtain sufficient conditions for the entropy preservation under bijections between mixed-pair random variables.

The extended definition of entropy leads to an entropy rate for continuous time Markov chains. As applications of our results, we provide simpler proofs of some known probabilistic results. The framework developed in this paper is best suited for establishing probabilistic properties of complex processes, such as load balancing systems, queuing networks, caching algorithms, that have inherent discrete variables (choices made) and continuous variables (occurrence times).

## 1. INTRODUCTION

The notion of entropy for discrete random variables as well as continuous random variables is well defined. Entropy preservation of discrete random variable under bijection map is an extremely useful property. For example, Prabhakar and Gallager [PG03] used this entropy preservation property to obtain an alternate proof of the known result that Geometric processes are fixed points under certain queuing disciplines.

In many interesting situations, including Example 1.1 given below, the underlying random variables are mixtures of discrete and continuous random variables. Such systems exhibit natural bijective properties which allow one to obtain non-trivial properties of the system via *non-rigorous* “information preservation” arguments. In this paper we develop sufficient conditions to make such arguments rigorous.

We will extend the definition of entropy to random variables that form a mixed pair of discrete and continuous variables as well as obtain sufficient conditions for preservation of entropy. Subsequently, we will provide a rigorous justification of mathematical identities that follow in the example below.

**Example 1.1.** *Poisson Splitting:* Consider a Poisson Process,  $\mathcal{P}$ , of rate  $\lambda$ . Split the Poisson process into two baby-processes  $\mathcal{P}_1$  and  $\mathcal{P}_2$  as follows: for each point of  $\mathcal{P}$ , toss an independent coin of bias  $p$ ; if coin turns up heads then the point is assigned to  $\mathcal{P}_1$ , else to  $\mathcal{P}_2$ . It is well-known that  $\mathcal{P}_1$  and  $\mathcal{P}_2$  are independent Poisson processes with rates  $\lambda p$  and  $\lambda(1-p)$  respectively.

Entropy rate of a Poisson process with rate  $\mu$  is known to be  $\mu(1-\log \mu)$  nats per second. That is, entropy rates of  $\mathcal{P}$ ,  $\mathcal{P}_1$ , and  $\mathcal{P}_2$  are given by  $\lambda(1-\log \lambda)$ ,  $\lambda p(1-\log \lambda p)$  and  $\lambda(1-p)(1-\log \lambda(1-p))$  respectively. Further observe that the coin of bias  $p$  is tossed at a rate  $\lambda$  and each coin-toss has an entropy equal to  $-p \log p - (1-p) \log(1-p)$  nats.

It is clear that there is a bijection between the tuple  $(\mathcal{P}, \text{coin-toss process})$  and the tuple  $(\mathcal{P}_1, \mathcal{P}_2)$ . Observe that the joint entropy rate of the two independent baby-processes are given by their sum. This leads to the following “obvious” set of equalities.

$$\begin{aligned} H_{ER}(\mathcal{P}_1, \mathcal{P}_2) &= H_{ER}(\mathcal{P}_1) + H_{ER}(\mathcal{P}_2) \\ &= \lambda p(1-\log \lambda p) + \lambda(1-p)(1-\log \lambda(1-p)) \\ (1.1) \quad &= \lambda(1-\log \lambda) + \lambda(-p \log p - (1-p) \log(1-p)) \\ &= H_{ER}(\mathcal{P}) + \lambda(-p \log p - (1-p) \log(1-p)). \end{aligned}$$

The last sum can be identified as sum of the entropy rate of the original Poisson process and the entropy rate of the coin tosses. However the presence of differential entropy as well as discrete entropy prevents

this interpretation from being rigorous. In this paper, we shall provide rigorous justification to the above equalities.

## 2. DEFINITIONS AND SETUP

This section provides technical definitions and sets up the frame-work for this paper. First, we present some preliminaries.

**2.1. Preliminaries.** Consider a measure space  $(\Omega, \mathcal{F}, \mathbb{P})$ , with  $\mathbb{P}$  being a probability measure. Let  $(\mathbb{R}, B_{\mathbb{R}})$  denote the measurable space on  $\mathbb{R}$  with the Borel  $\sigma$ -algebra. A random variable  $X$  is a measurable mapping from  $\Omega$  to  $\mathbb{R}$ . Let  $\mu_X$  denote the induced probability measure on  $(\mathbb{R}, B_{\mathbb{R}})$  by  $X$ . We call  $X$  as *discrete random variable* if there is a countable subset  $\{x_1, x_2, \dots\}$  of  $\mathbb{R}$  that forms a support for the measure  $\mu_X$ . Let  $p_i = \mathbb{P}(X = x_i)$  and note that  $\sum_i p_i = 1$ .

The entropy of a discrete random variable is defined by the sum

$$H(X) = - \sum_i p_i \log p_i.$$

Note that this entropy is non-negative and has several well known properties. One natural interpretation of this number is in terms of the maximum compressibility (in bits per symbol) of an i.i.d. sequence of the random variables,  $X$  (cf. Shannon's data compression theorem [Sha48]).

A random variable  $Y$ , defined on  $(\Omega, \mathcal{F}, \mathbb{P})$ , is said to be a *continuous random variable* if the probability measure,  $\mu_Y$ , induced on  $(\mathbb{R}, B_{\mathbb{R}})$  is absolutely continuous with respect to the Lebesgue measure. These probability measures can be characterized by a non-negative density function  $f(x)$  that satisfies  $\int_{\mathbb{R}} f(x) dx = 1$ . The entropy (differential entropy) of a continuous random variable is defined by the integral

$$h(Y) = - \int_{\mathbb{R}} f(y) \log f(y) dy.$$

The entropy of a continuous random variable is not non-negative, though it satisfies several of the other properties of the discrete entropy function. Due to negativity, differential entropy clearly does not have interpretation of maximal compressibility. However, it does have the interpretation of being the limiting difference between the maximally compressed quantization of the random variable and an identical quantization of an independent  $U[0, 1]^*$  random variable [CT91] as the quantization resolution goes to zero. Hence the term differential entropy is usually preferred to entropy when describing this number.

**2.2. Our Setup.** In this paper, we are interested in a set of random variables that incorporate the aspects of both discrete and continuous random variables. Let  $Z = (X, Y)$  be a measurable mapping from the space  $(\Omega, \mathcal{F}, \mathbb{P})$  to the space  $(\mathbb{R} \times \mathbb{R}, B_{\mathbb{R}} \times B_{\mathbb{R}})$ . Observe that this mapping induces a probability measure  $\mu_Z$  on the space  $(\mathbb{R} \times \mathbb{R}, B_{\mathbb{R}} \times B_{\mathbb{R}})$  as well as two probability measures  $\mu_X$  and  $\mu_Y$  on  $(\mathbb{R}, B_{\mathbb{R}})$  obtained via the projection of the measure  $\mu_Z$ .

**Definition 2.1** (Mixed-Pair). Consider a random variables  $Z = (X, Y)$ . We call  $Z^\dagger$  a mixed-pair if  $X$  is a discrete random variable while  $Y$  is a continuous random variable. That is, the support of  $\mu_Z$  is on the product space  $\mathbb{S} \times \mathbb{R}$ , with  $\mathbb{S} = \{x_1, x_2, \dots\}$  is a countable subset of  $\mathbb{R}$ . That is  $\mathbb{S}$  forms a support for  $\mu_X$  while  $\mu_Y$  is absolutely continuous with respect to the Lebesgue measure.

Observe that  $Z = (X, Y)$  induces measures  $\{\mu_1, \mu_2, \dots\}$  that are absolutely continuous with respect to the Lebesgue measure, where  $\mu_i(A) = \mathbb{P}(X = x_i, Y \in A)$ , for every  $A \in B_{\mathbb{R}}$ . Associated with these measures  $\mu_i$ , there are non-negative density functions  $g_i(y)$  that satisfy

$$\sum_i \int_{\mathbb{R}} g_i(y) dy = 1.$$

---

\*  $U[0,1]$  represents a random variable that is uniformly distributed on the interval  $[0,1]$

†For the rest of the paper we shall adopt the notation that random variables  $X_i$  represent discrete random variables,  $Y_i$  represent continuous random variables and  $Z_i$  represent mixed-pair of random variables.

Let us define  $p_i = \int_{\mathbb{R}} g_i(y) dy$ . Observe that  $p_i$ 's are non-negative numbers that satisfy  $\sum_i p_i = 1$  and corresponds to the probability measure  $\mu_X$ . Further  $g(y) = \sum_i g_i(y)$  corresponds to the probability measure  $\mu_Y$ . Let

$$\tilde{g}_i(y) \triangleq \frac{1}{p_i} g_i(y)$$

be the probability density function of  $Y$  conditioned on  $X = x_i$ .

The following non-negative sequence is well defined for every  $y \in \mathbb{R}$  for which  $g(y) > 0$ ,

$$p_i(y) = \frac{g_i(y)}{g(y)}, \quad i \geq 1.$$

Now  $g(y)$  is finite except possibly on a set,  $A$ , of measure zero. For  $y \in A^c$ , we have that  $\sum_i p_i(y) = 1$ ;  $p_i(y)$  corresponds to the probability that  $X = x_i$  conditioned on  $Y = y$ . It follows from definitions of  $p_i$  and  $p_i(y)$  that

$$p_i = \int_{\mathbb{R}} p_i(y) g(y) dy.$$

**Definition 2.2** (Good Mixed-Pair ). A mixed-pair random variable  $Z = (X, Y)$  is called *good* if the following condition is satisfied:

$$(2.1) \quad \sum_i \int_{\mathbb{R}} |g_i(y) \log g_i(y)| dy < \infty.$$

Essentially, the good mixed-pair random variables possess the property that when restricted to any of the  $X$  values, the conditional differential entropy of  $Y$  is well-defined. The following lemma provides a simple sufficient conditions for ensuring that a mixed-pair variable is good.

**Lemma 2.1.** *The following conditions are sufficient for a mixed-pair random variable to be a good pair:*

- (a) *Random variable  $Y$  possess a finite  $\epsilon^{\text{th}}$  moment for some  $\epsilon > 0$ , i.e.*

$$M_\epsilon = \int_{\mathbb{R}} |y|^\epsilon g(y) dy < \infty.$$

- (b) *There exists  $\delta > 0$  such that  $g(y)$  satisfies*

$$\int_{\mathbb{R}} g(y)^{1+\delta} dy < \infty.$$

- (c) *The discrete random variable  $X$  has finite entropy, i.e.  $-\sum_i p_i \log p_i < \infty$ .*

*Proof.* The proof is presented in the appendix. □

**Definition 2.3** (Entropy of a mixed-pair). The entropy of a good mixed-pair random variable is defined by

$$(2.2) \quad \mathbb{H}(Z) = - \sum_i \int_{\mathbb{R}} g_i(y) \log g_i(y) dy.$$

**Definition 2.4** (Vector of Mixed-Pairs). Consider a random vector  $(Z_1, \dots, Z_d) = \{(X_1, Y_1), \dots, (X_d, Y_d)\}$ . We call  $(Z_1, \dots, Z_d)$  a vector of mixed-pairs if the support of  $\mu_{(Z_1, \dots, Z_d)}$  is on the product space  $\mathbb{S}^d \times \mathbb{R}^d$ , where  $\mathbb{S}^d \subset \mathbb{R}^d$  is a countable set. That is,  $\mathbb{S}^d$  forms the support for the probability measure  $\mu_{(X_1, \dots, X_d)}$  while the measure  $\mu_{(Y_1, \dots, Y_d)}$  is absolutely continuous with respect to the Lebesgue measure on  $\mathbb{R}^d$ .

**Definition 2.5** (Good Mixed-Pair Vector ). A vector of mixed-pair random variables  $(Z_1, \dots, Z_d)$  is called good if the following condition is satisfied:

$$(2.3) \quad \sum_{\mathbf{x} \in \mathbb{S}^d} \int_{\mathbf{y} \in \mathbb{R}^d} |g_{\mathbf{x}}(\mathbf{y}) \log g_{\mathbf{x}}(\mathbf{y})| d\mathbf{y} < \infty,$$

where  $g_{\mathbf{x}}(\mathbf{y})$  is the density of the continuous random vector  $Y^d$  conditioned on the event that  $X^d = \mathbf{x}$ .

Analogous to Lemma 2.1, the following conditions guarantee that a vector of mixed-pair random variables is good.

**Lemma 2.2.** *The following conditions are sufficient for a mixed-pair random variable to be a good pair:*

(a) *Random variable  $Y^d$  possess a finite  $\epsilon^{\text{th}}$  moment for some  $\epsilon > 0$ , i.e.*

$$M_\epsilon = \int_{\mathbb{R}^d} \|\mathbf{y}\|^\epsilon g(\mathbf{y}) d\mathbf{y} < \infty.$$

(b) *There exists  $\delta > 0$  such that  $g(\mathbf{y})$  satisfies*

$$\int_{\mathbb{R}^d} g(\mathbf{y})^{1+\delta} d\mathbf{y} < \infty.$$

(c) *The discrete random variable  $X^d$  has finite entropy, i.e.  $-\sum_{\mathbf{x} \in \mathbb{S}^d} p_{\mathbf{x}} \log p_{\mathbf{x}} < \infty$ .*

*Proof.* The proof is similar to that of Lemma 2.1 and is omitted.  $\square$

In rest of the paper, all mixed-pair variables and vectors are assumed to be *good*, i.e. assumed to satisfy the condition (2.1).

**Definition 2.6** (Entropy of a mixed-pair vector). The entropy of a good mixed-pair vector of random variables is defined by

$$(2.4) \quad \mathbb{H}(Z) = - \sum_{\mathbf{x} \in \mathbb{S}^d} \int_{\mathbb{R}^d} g_{\mathbf{x}}(\mathbf{y}) \log g_{\mathbf{x}}(\mathbf{y}) d\mathbf{y}.$$

**Definition 2.7** (Conditional entropy). Given a pair of random variables  $(Z_1, Z_2)$ , the conditional entropy is defined as follows

$$\mathbb{H}(Z_1|Z_2) = \mathbb{H}(Z_1, Z_2) - \mathbb{H}(Z_2).$$

It is not hard to see that  $\mathbb{H}(Z_1|Z_2)$  evaluates to

$$- \sum_{x_1, x_2} \int_{\mathbb{R}^2} g_{x_1, x_2}(y_1, y_2) \log \frac{g_{x_1, x_2}(y_1, y_2)}{g_{x_2}(y_2)} dy_1 dy_2.$$

**Definition 2.8** (Mutual Information). Given a pair of random variables  $(Z_1, Z_2)$ , the mutual information is defined as follows

$$\mathbb{I}(Z_1; Z_2) = \mathbb{H}(Z_1) + \mathbb{H}(Z_2) - \mathbb{H}(Z_1, Z_2).$$

The mutual information evaluates to

$$\sum_{x_1, x_2} \int_{\mathbb{R}^2} g_{x_1, x_2}(y_1, y_2) \log \frac{g_{x_1, x_2}(y_1, y_2)}{g_{x_1}(y_1)g_{x_2}(y_2)} dy_1 dy_2.$$

Using the fact that  $1 + \log x < x$  for  $x > 0$  it can be shown that  $\mathbb{I}(Z_1; Z_2)$  is non-negative.

**2.3. Old Definitions Still Work.** We will now present injections from the space of discrete (or continuous) random variables into the space of mixed-pair random variable so that the entropy of the mixed-pair random variable is the same as the discrete (or continuous) entropy.

*Injection: Discrete into Mixed-Pair.* Let  $X$  be a discrete random variable with finite entropy. Let  $\{p_1, p_2, \dots\}$  denote the probability measure associated with  $X$ . Consider the mapping  $\sigma_d : X \rightarrow Z \equiv (X, U)$  where  $U$  is an independent continuous random variable distributed uniformly on the interval  $[0, 1]$ . For  $Z$ , we have  $g_i(y) = p_i$  for  $y \in [0, 1]$ . Therefore

$$\begin{aligned} \mathbb{H}(Z) &= - \sum_i \int_{\mathbb{R}} g_i(y) \log g_i(y) dy = \sum_i \int_0^1 -p_i \log p_i dy \\ &= - \sum_i p_i \log p_i = H(X) < \infty. \end{aligned}$$

Therefore we see that  $\mathbb{H}(Z) = H(X)$ .

*Injection: Continuous into Mixed-Pair:* Let  $Y$  be a continuous random variable with a density function  $g(y)$  that satisfies

$$\int_{\mathbb{R}} g(y) |\log g(y)| dy < \infty.$$

Consider the mapping  $\sigma_c : Y \rightarrow Z \equiv (X_0, Y)$  where  $X_0$  is the constant random variable, say  $\mathbb{P}(X_0 = 1) = 1$ . Observe that  $g(y) = g_1(y)$  and that the pair  $Z \equiv (X_0, Y)$  is a good mixed-pair that satisfies  $\mathbb{H}(Z) = h(Y)$ .

Thus  $\sigma_d$  and  $\sigma_c$  are injections from the space of continuous and discrete random variables into the space of good mixed-pairs that preserve the entropy function.

**2.4. Discrete-Continuous Variable as Mixed-Pair.** Consider a random variable<sup>‡</sup>  $V$  whose support is combination of both discrete and continuous. That is, it satisfies the following properties: (i) There is a countable set (possibly finite)  $S = \{x_1, x_2, \dots\}$  such that  $\mu_V(x_i) = p_i > 0$ ; (ii) measure  $\tilde{\mu}_V$  with an associated non-negative function  $\tilde{g}(y)$  (absolutely continuous w.r.t. the Lebesgue measure), and (iii) the following holds:

$$\int_{\mathbb{R}} \tilde{g}(y) dy + \sum_i p_i = 1.$$

Thus, the random variable  $V$  either takes discrete values  $x_1, x_2, \dots$  with probabilities  $p_1, p_2, \dots$  or else it is distributed according to the density function  $\frac{1}{1-p}\tilde{g}(y)$ ; where  $p = \sum_i p_i$ . Observe that  $V$  has neither a countable support nor is its measure absolutely continuous with respect to Lebesgue measure. Therefore, though such random variables are encountered neither the discrete entropy nor the continuous entropy is appropriate.

To overcome this difficulty, we will treat such variables as mixed-pair variables by appropriate injection of such variables into mixed-pair variables. Subsequently, we will be able to use the definition of entropy for mixed-pair variables.

*Injection: Discrete-Continuous into Mixed-Pair:* Let  $V$  be a discrete-continuous variable as considered above. Let the following two conditions be satisfied:

$$-\sum_i p_i \log p_i < \infty \text{ and } \int_{\mathbb{R}} \tilde{g}(y) |\log \tilde{g}(y)| dy < \infty.$$

Consider the mapping  $\sigma_m : V \rightarrow Z \equiv (X, Y)$  described as follows: When  $V$  takes a discrete value  $x_i$ , it is mapped on to the pair  $(x_i, u_i)$  where  $u_i$  is chosen independently and uniformly at random in  $[0, 1]$ . When  $V$  does not take a discrete value and say takes value  $y$ , it gets mapped to the pair  $(x_0, y)$  where  $x_0 \neq x_i, \forall i$ . One can think of  $x_0$  as an indicator value that  $V$  takes when it is not discrete. The mixed-pair variable  $Z$  has its associated functions  $\{g_0(y), g_1(y), \dots\}$  where  $g_i(y) = p_i, y \in [0, 1], i \geq 1$  and  $g_0(y) = \tilde{g}(y)$ . The entropy of  $Z$  as defined earlier is

$$\begin{aligned} \mathbb{H}(Z) &= -\sum_i \int_{\mathbb{R}} g_i(y) \log g_i(y) dy \\ &= -\sum_i p_i \log p_i - \int_{\mathbb{R}} \tilde{g}(y) \log \tilde{g}(y) dy. \end{aligned}$$

*Remark 2.1.* In the rest of the paper we will treat every random variable that is encountered as a mixed-pair random variable. That is, a discrete variable or a continuous variable would be assumed to be injected into the space of mixed-pairs using the map  $\sigma_d$  or  $\sigma_c$ , respectively.

### 3. BIJECTIONS AND ENTROPY PRESERVATION

In this section we will consider bijections between mixed-pair random variables and establish sufficient conditions under which the entropy is preserved. We first consider the case of mixed-pair random variables and then extend this to vectors of mixed-pair random variables.

---

<sup>‡</sup>Normally such random variables are referred to as mixed random variables.

**3.1. Bijections between Mixed-Pairs.** Consider mixed-pair random variables  $Z_1 \equiv (X_1, Y_1)$  and  $Z_2 \equiv (X_2, Y_2)$ . Specifically, let  $\mathbb{S}_1 = \{x_{1i}\}$  and  $\mathbb{S}_2 = \{x_{2j}\}$  be the countable (possibly finite) supports of the discrete measures  $\mu_{X_1}$  and  $\mu_{X_2}$  such that  $\mu_{X_1}(x_{1i}) > 0$  and  $\mu_{X_2}(x_{2j}) > 0$  for all  $i \in \mathbb{S}_1$  and  $j \in \mathbb{S}_2$ . Therefore a bijection between mixed-pair variables  $Z_1$  and  $Z_2$  can be viewed as bijections between  $\mathbb{S}_1 \times \mathbb{R}$  and  $\mathbb{S}_2 \times \mathbb{R}$ .

Let  $F : \mathbb{S}_1 \times \mathbb{R} \rightarrow \mathbb{S}_2 \times \mathbb{R}$  be a bijection. Given  $Z_1$ , this bijection induces a mixed-pair random variable  $Z_2$ . We restrict our attention to the case when  $F$  is continuous and differentiable<sup>§</sup>. Let the induced projections be  $F_d : \mathbb{S}_1 \times \mathbb{R} \rightarrow \mathbb{S}_2$  and  $F_c : \mathbb{S}_1 \times \mathbb{R} \rightarrow \mathbb{R}$ . Let the associated projections of the inverse map  $F^{-1} : \mathbb{S}_2 \times \mathbb{R} \rightarrow \mathbb{S}_1 \times \mathbb{R}$  be  $F_d^{-1} : \mathbb{S}_2 \times \mathbb{R} \rightarrow \mathbb{S}_1$  and  $F_c^{-1} : \mathbb{S}_2 \times \mathbb{R} \rightarrow \mathbb{R}$  respectively.

As before, let  $\{g_i(y_1)\}$ ,  $\{h_j(y_2)\}$  denote the non-negative density functions associated with the mixed-pair random variables  $Z_1$  and  $Z_2$  respectively. Let  $(x_{2j}, y_2) = F(x_{1i}, y_1)$ , i.e.  $x_{2j} = F_d(x_{1i}, y_1)$  and  $y_2 = F_c(x_{1i}, y_1)$ . Now, consider a small neighborhood  $x_{1i} \times [y_1, y_1 + dy_1]$  of  $(x_{1i}, y_1)$ . From the continuity of  $F$ , for small enough  $dy_1$ , the neighborhood  $x_{1i} \times [y_1, y_1 + dy_1]$  is mapped to some small neighborhood of  $(x_{2j}, y_2)$ , say  $x_{2j} \times [y_2, y_2 + dy_2]$ . The measure of  $x_{1i} \times [y_1, y_1 + dy_1]$  is  $\approx g_i(y_1)|dy_1|$ , while measure of  $x_{2j} \times [y_2, y_2 + dy_2]$  is  $\approx h_j(y_2)|dy_2|$ . Since distribution of  $Z_2$  is induced by the bijection from  $Z_1$ , we obtain

$$(3.1) \quad g_i(y_1) \left| \frac{dy_1}{dy_2} \right| = h_j(y_2).$$

Further from  $y_2 = F_c(x_{1i}, y_1)$  we also have,

$$(3.2) \quad \frac{dy_2}{dy_1} = \frac{dF_c(x_{1i}, y_1)}{dy_1}.$$

These immediately imply a sufficient condition under which bijections between mixed-pair random variables imply that their entropies are preserved.

**Lemma 3.1.** *If  $\left| \frac{dF_c(x_{1i}, y_1)}{dy_1} \right| = 1$  for all points  $(x_{1i}, y_1) \in \mathbb{S}_1 \times \mathbb{R}$ , then  $\mathbb{H}(Z_1) = \mathbb{H}(Z_2)$ .*

*Proof.* This essentially follows from the change of variables and repeated use of Fubini's theorem (to interchange the sums and the integral). To apply Fubini's theorem, we use the assumption that mixed-pair random variables are *good*. Observe that,

$$(3.3) \quad \begin{aligned} \mathbb{H}(Z_1) &= - \sum_i \int_{\mathbb{R}} g_i(y_1) \log g_i(y_1) dy_1 \\ &\stackrel{(a)}{=} - \sum_j \int_{\mathbb{R}} h_j(y_2) \log \left( h_j(y_2) \left| \frac{dF_c(x_{1i}, y_1)}{dy_1} \right| \right) dy_2 \\ &\stackrel{(b)}{=} - \sum_j \int_{\mathbb{R}} h_j(y_2) \log h_j(y_2) dy_2 \\ &= \mathbb{H}(Z_2). \end{aligned}$$

Here (a) is obtained by repeated use of Fubini's theorem along with (3.1) and (b) follows from the assumption of the Lemma that  $\left| \frac{dF_c(x_{1i}, y_1)}{dy_1} \right| = 1$ .  $\square$

**3.2. Some Examples.** In this section, we present some examples to illustrate our definitions, setup and the entropy preservation Lemma.

**Example 3.1.** Let  $Y_1$  be a continuous random variable that is uniformly distributed in the interval  $[0, 2]$ . Let  $X_2$  be the discrete random variable that takes value 0 when  $Y_1 \in [0, 1]$  and 1 otherwise. Let  $Y_2 = Y_1 - X_2$ . Clearly  $Y_2 \in [0, 1]$ , is uniformly distributed and independent of  $X_2$ .

Let  $Z_1 \equiv (X_1, Y_1)$  be the natural injection,  $\sigma_c$  of  $Y_1$  (i.e.  $X_1$  is just the constant random variable.). Observe that the bijection between  $Z_1$  to the pair  $Z_2 \equiv (X_2, Y_2)$  that satisfies conditions of Lemma 3.1 and implies

$$\log 2 = \mathbb{H}(Z_1) = \mathbb{H}(Z_2).$$

<sup>§</sup>The continuity of mapping between two copies of product space  $\mathbb{S} \times \mathbb{R}$  essentially means that the mapping is continuous with respect to right (or  $Y$ ) co-ordinate for fixed  $x_i \in \mathbb{S}$ . Similarly, differentiability essentially means differentiability with respect to  $Y$  co-ordinate.

However, also observe that by plugging in the various definitions of entropy in the appropriate spaces,  $\mathbb{H}(Z_2) = \mathbb{H}(X_2, Y_2) = H(X_2) + h(Y_2) = \log 2 + 0$ , where the first term is the discrete entropy and the second term is the continuous entropy. In general it is not difficult to see that the two definitions of entropy (for discrete and continuous random variables) are compatible with each other if the random variables themselves are thought of as a mixed-pair.

**Example 3.2.** This example demonstrates that some care must be taken when considering discrete and continuous variables as mixed-pair random variables. Consider the following continuous random variable  $Y_1$  that is uniformly distributed in the interval  $[0, 2]$ . Now, consider the mixed random variable  $V_2$  that takes the value 2 with probability  $\frac{1}{2}$  and takes a value uniformly distributed in the interval  $[0, 1]$  with probability  $\frac{1}{2}$ .

Clearly, there is a mapping that allows us to create  $V_2$  from  $Y_1$  by just mapping  $Y_1 \in (1, 2]$  to the value  $V_2 = 2$  and by setting  $Y_1 = V_2$  when  $Y_1 \in [0, 1]$ . However, given  $V_2 = 2$  we are not able to reconstruct  $Y_1$  exactly. Therefore, intuitively one expects that  $\mathbb{H}(Y_1) > \mathbb{H}(V_2)$ .

However, if you use the respective injections, say  $Y_1 \rightarrow Z_1$  and  $V_2 \rightarrow Z_2$ , to the space of mixed-pairs of random variables, we can see that

$$\mathbb{H}(Y_1) = \mathbb{H}(Z_1) = \log 2 = \mathbb{H}(Z_2).$$

This shows that if we think of  $\mathbb{H}(Z_2)$  as the entropy of the mixed random variable  $V_2$  we get an intuitively paradoxical result where  $\mathbb{H}(Y_1) = \mathbb{H}(V_2)$  where in reality one would expect  $\mathbb{H}(Y_1) > \mathbb{H}(V_2)$ .

The careful reader will be quick to point out that the injection from  $V_2$  to  $Z_2$  introduces a new continuous variable,  $Y_{22}$ , associated with the discrete value of 2, as well as a discrete value  $x_0$  associated with the continuous part of  $V_2$ . Indeed the "new" random variable  $Y_{22}$  allows us to precisely reconstruct  $Y_1$  from  $Z_2$  and thus complete the inverse mapping of the bijection.

*Remark 3.1.* The examples show that when one has mappings involving various types of random variables and one wishes to use bijections to compare their entropies; one can perform this comparison as long as the random variables are thought of as mixed-pairs.

**3.3. Vector of Mixed-Pair Random Variables.** Now, we derive sufficient conditions for entropy preservation under bijection between vectors of mixed-pair variables. To this end, let  $Z_1 = (Z_1^1, \dots, Z_1^d)$  and  $Z_2 = (Z_2^1, \dots, Z_2^d)$  be two vectors of mixed-pair random variables with their support on  $\mathbb{S}_1 \times \mathbb{R}^d$  and  $\mathbb{S}_2 \times \mathbb{R}^d$  respectively. (Here  $\mathbb{S}_1, \mathbb{S}_2$  are countable subsets of  $\mathbb{R}^d$ .) Let  $F : \mathbb{S}_1 \times \mathbb{R}^d \rightarrow \mathbb{S}_2 \times \mathbb{R}^d$  be a continuous and differentiable bijection that induces  $Z_2$  by its application on  $Z_1$ .

As before, let the projections of  $F$  be  $F_d : \mathbb{S}_1 \times \mathbb{R}^d \rightarrow \mathbb{S}_2$  and  $F_c : \mathbb{S}_1 \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ . We consider situation where  $F_c$  is differentiable. Let  $g_i(\mathbf{y}), \mathbf{y} \in \mathbb{R}^d$  for  $\mathbf{x}_i \in \mathbb{S}_1$  and  $h_j(\mathbf{y}), \mathbf{y} \in \mathbb{R}^d$  for  $\mathbf{w}_j \in \mathbb{S}_2$  be density functions as defined before. Let  $(\mathbf{x}_i, \mathbf{y}^1) \in \mathbb{S}_1 \times \mathbb{R}^d$  be mapped to  $(\mathbf{w}_j, \mathbf{y}^2) \in \mathbb{S}_2 \times \mathbb{R}^d$ . Then, consider  $d \times d$  Jacobian

$$J(\mathbf{x}_i, \mathbf{y}^1) \equiv \left[ \frac{\partial y_k^2}{\partial y_l^1} \right]_{1 \leq k, l \leq d},$$

where we have used notation  $\mathbf{y}^1 = (y_1^1, \dots, y_d^1)$  and  $\mathbf{y}^2 = (y_1^2, \dots, y_d^2)$ . Now, similar to Lemma 3.1 we obtain the following entropy preservation for bijection between vector of mixed-pair random variables.

**Lemma 3.2.** *If for all  $(\mathbf{x}_i, \mathbf{y}^1) \in \mathbb{S}_1 \times \mathbb{R}^d$ ,*

$$|\det(J(\mathbf{x}_i, \mathbf{y}^1))| = 1,$$

*then  $\mathbb{H}(Z^1) = \mathbb{H}(Z^2)$ . Here  $\det(J)$  denotes the determinant of matrix  $J$ .*

*Proof.* The main ingredients for the proof of Lemma 3.1 for the scalar case were the equalities (3.1) and (3.2). For a vector of mixed-pair variable we will obtain the following equivalent equalities: For change of  $d\mathbf{y}^1$  at  $(\mathbf{x}_i, \mathbf{y}^1)$ , let  $d\mathbf{y}^2$  be induced change at  $(\mathbf{x}_j, \mathbf{y}^2)$ . Let  $\text{vol}(d\mathbf{y})$  denote the volume of  $d$  dimensional rectangular region with sides given by components of  $d\mathbf{y}$  in  $\mathbb{R}^d$ . Then,

$$(3.4) \quad g_i(\mathbf{y}^1) \text{vol}(d\mathbf{y}^1) = h_j(\mathbf{y}^2) \text{vol}(d\mathbf{y}^2).$$

Further, at  $(\mathbf{x}_i, \mathbf{y}^1)$ ,

$$(3.5) \quad \text{vol}(d\mathbf{y}^2) = |\det(J(\mathbf{x}_i, \mathbf{y}^1))| \text{vol}(d\mathbf{y}^1).$$

Using exactly the same argument that is used in (3.3) (replacing  $dy_k$  by  $\text{vol}(d\mathbf{y}^k)$ ,  $k = 1, 2$ ), we obtain the desired result. This completes the proof of Lemma 3.2.  $\square$

*Remark 3.2.* Essentially, for every discrete choice  $X$ , if the mapping between the continuous vectors has one, then the bijection preserves entropy.

#### 4. ENTROPY RATE OF CONTINUOUS TIME MARKOV CHAINS

A continuous time Markov chain is composed of the point process that characterizes the time of transitions of the states as well as the discrete states between which the transition happens. Specifically, let  $x_i \in \mathbb{R}$  denote the time of  $i^{\text{th}}$  transition or jump with  $i \in \mathbb{Z}$ . Let  $V_i \in \mathbb{S}$  denote the state of the Markov chain after the jump at time  $x_i$ , where  $\mathbb{S}$  be some countable state space. For simplicity, we assume  $\mathbb{S} = \mathbb{N}$ . Let transition probabilities be  $p_{k\ell} = \mathbb{P}(V_i = \ell | V_{i-1} = k)$ ,  $k, \ell \in \mathbb{N}$  for all  $i$ .

We recall that the entropy rate of a point process  $\mathcal{P}$  was defined in section 13.5 of [DVJ88] according to the following: ‘‘Observation of process conveys information of two kinds: the actual number of points observed and the location of these points given their number.’’ This led them to define the entropy of a realization  $\{x_1, \dots, x_N\}$  as

$$\mathbb{H}(N) + \mathbb{H}(x_1, \dots, x_N | N)$$

The entropy rate of the point process  $\mathcal{P}$  is defined as follows: let  $N(T)$  be the number of points arrived in time interval  $(0, T]$  and the instances be  $\mathbf{x}(T) = (x_1, \dots, x_{N(T)})$ . Then, the entropy rate of the process is

$$\mathbb{H}_{ER}(\mathcal{P}) = \lim_{T \rightarrow \infty} \frac{1}{T} [\mathbb{H}(N(T)) + \mathbb{H}(\mathbf{x}(T) | N(T))],$$

if the above limit exists.

We extend the above definition to the case of Markov chain in a natural fashion. Observation of a continuous time Markov chain over a time interval  $(0, T]$  conveys information of three types: the number of points/jumps of the chain in the interval, the location of the points given the number as well as the value of the chain after each jump. Treating each random variable as a mixed-pair allows us to consider all the random variables in a single vector.

As before, let  $N(T)$  denote the number of points in an interval  $(0, T]$ . Let  $\mathbf{x}(T) = (x_1, \dots, x_{N(T)})$ ,  $\mathbf{V}(T) = (V_0, V_1, \dots, V_{N(T)})$  denote the locations of the jumps as well as the values of the chain after the jumps. This leads us to define the entropy of the process during the interval  $(0, T]$  as

$$(4.1) \quad \mathbb{H}_{(0,T]} = \mathbb{H}(N(T), \mathbf{V}(T), \mathbf{x}(T)).$$

Observe that the  $(N(T), \mathbf{V}(T), \mathbf{x}(T))$  is a random vector of mixed-pair variables.

For a single state Markov chain the above entropy is the same as that of the point process determine the jump/transition times. Similar to the development for point processes, we define the entropy rate of the Markov chain as

$$\mathbb{H}_{ER} = \lim_{T \rightarrow \infty} \frac{\mathbb{H}_{(0,T]}}{T}, \text{ if it exists.}$$

**Proposition 4.1.** *Consider a Markov chain with underlying Point process being Poisson of rate  $\lambda$ , its stationary distribution being  $\pi = (\pi(i))$  with transition probability matrix  $P = [p_{ij}]$ . Then, its entropy rate is well-defined and*

$$\mathbb{H}_{ER} = \lambda(1 - \log \lambda) + \lambda H_{MC},$$

where  $H_{MC} = -\sum_i \pi(i) \sum_j p_{ij} \log p_{ij}$ .

*Proof.* For Markov Chain as described in the statement of proposition, we wish to establish that

$$\lim_{T \rightarrow \infty} \frac{\mathbb{H}_{(0,T]}}{T} = \mathbb{H}_{ER},$$

as defined above. Now

$$\begin{aligned} \mathbb{H}_{(0,T]} &= \mathbb{H}(\mathbf{x}(T), N(T), \mathbf{V}(T)) \\ &= \mathbb{H}(\mathbf{x}(T), N(T)) + \mathbb{H}(\mathbf{V}(T) | N(T), \mathbf{x}(T)). \end{aligned}$$



Consider the term on the right hand side of the above equality. This corresponds to the points of a Poisson process of rate  $\lambda$ . It is well-known (cf. equation (13.5.10), pg. 565 [DVJ88]) that

$$(4.2) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{H}(\mathbf{x}(T), N(T)) = \lambda(1 - \log \lambda).$$

Now consider the term  $\mathbb{H}(\mathbf{V}(T)|\mathbf{x}(T), N(T))$ . Since  $\mathbf{V}(T)$  is independent of  $\mathbf{x}(T)$ , we get from the definition of conditional entropy that

$$(4.3) \quad \mathbb{H}(\mathbf{V}(T)|\mathbf{x}(T), N(T)) = \mathbb{H}(\mathbf{V}(T)|N(T)).$$

One can evaluate  $\mathbb{H}(\mathbf{V}(T)|N(T))$  as follows,

$$\mathbb{H}(\mathbf{V}(T)|N(T)) = \sum_k p_k \mathbb{H}(V_0, \dots, V_k),$$

where  $p_k$  is the probability that  $N(T) = k$ . The sequence of states  $V_0, \dots, V_k$  can be thought of as sequence of states of a discrete time Markov chain with transition matrix  $P$ . For a Markov chain, with stationary distribution  $\pi$  (i.e.  $P\pi = \pi$ ), it is well-known that

$$\begin{aligned} \lim_{k \rightarrow \infty} \frac{1}{k} \mathbb{H}(V_0, \dots, V_k) &= - \sum_i \pi(i) \sum_j p_{ij} \log p_{ij} \\ &= H_{\text{MC}}. \end{aligned}$$

Thus, for any  $\epsilon > 0$ , there exists  $k(\epsilon)$  large enough such that for  $k > k(\epsilon)$

$$\left| \frac{1}{k} \mathbb{H}(V_0, \dots, V_k) - H_{\text{MC}} \right| < \epsilon.$$

For  $T$  large enough, using tail-probability estimates of Poisson variable it can be shown that

$$\mathbb{P}(N(T) \leq k(\epsilon)) \leq \exp\left(-\frac{\lambda T}{8}\right).$$

Putting these together, we obtain that for given  $\epsilon$  there exists  $T(\epsilon)$  large enough such that for  $T \geq T(\epsilon)$

$$\begin{aligned} \frac{\mathbb{H}(\mathbf{V}(T)|N(T))}{T} &= \frac{1}{T} \left( \sum_k k p_k \frac{\mathbb{H}(V_0, \dots, V_k)}{k} \right) \\ &= \frac{\sum_{k \geq k(\epsilon)} k p_k (H_{\text{MC}} \pm \epsilon) + O(k(\epsilon))}{T} \\ &= (H_{\text{MC}} \pm \epsilon) \frac{\lambda T + O(k(\epsilon))}{T} \\ &= \lambda H_{\text{MC}} \pm 2\epsilon. \end{aligned}$$

That is

$$\lim_{T \rightarrow \infty} \frac{\mathbb{H}(\mathbf{V}(T)|N(T))}{T} = \lambda H_{\text{MC}}.$$

Combining (4.2), (4.3) and the above equation we complete the proof of the Proposition 4.1.  $\square$

*Fact 4.1* (cf. Ch. 13.5 [DVJ88]). Consider the set of stationary ergodic point processes with mean rate  $\lambda$ . Then the entropy of this collection is maximized by a Poisson Process with rate  $\lambda$ . That is, if  $\mathcal{P}$  is a stationary ergodic point process with rate  $\lambda$  then

$$\mathbb{H}_{ER}(\mathcal{P}) \leq \lambda(1 - \log \lambda).$$

## 5. APPLICATION

**5.1. Computation of continuous entropies.** In this section we show how our previous results aid in the computation of traditional continuous time entropies. Let  $(X_1, X_2)$  be two i.i.d. random variables whose distributions satisfy the conditions required of it to be a good-mixed pair. Let  $(Y_1, Y_2)$ ,  $Y_1 < Y_2$  be the ordering of  $(X_1, X_2)$ . Then the following holds:

**Lemma 5.1.**  $h(Y_1, Y_2) = h(X_1, X_2) - \log 2$ .

*Proof.* Let  $\mathbb{I}$  represent the indicator function such that  $\mathbb{I} = 0$  implies  $X_1 = Y_1$ , and  $\mathbb{I} = 1$  implies  $X_1 = Y_2$ . Clearly  $\mathbb{I}$  is independent of  $Y_1, Y_2$  and probability of  $\mathbb{I} = 1$  is  $\frac{1}{2}$ . It is further easy to see that  $(\mathbb{I}, Y_1, Y_2) \leftrightarrow (X_1, X_2)$  with the corresponding Jacobians evaluating to 1. Thus, viewed as mixed-pairs we can equate the entropies, yielding

$$\log 2 + h(Y_1, Y_2) = h(X_1, X_2).$$

This can be of course be shown using traditional methods but the proof here is an illustration of how our results can be used to obtain such results in an easier fashion.  $\square$

In a similar fashion if  $(Y_1, \dots, Y_n)$  is an ordering, in increasing order, of the i.i.d. random variables  $(X_1, \dots, X_n)$ , then

$$h(Y_1, \dots, Y_n) = h(X_1, \dots, X_n) - \log(n!)$$

**5.2. Poisson Splitting via Entropy Preservation.** In this section, we use the sufficient conditions developed in Lemma 3.2 to obtain proof of the following property.

**Lemma 5.2.** *Consider a Poisson process,  $\mathcal{P}$ , of rate  $\lambda$ . Split the process  $\mathcal{P}$  into two baby-processes  $\mathcal{P}_1$  and  $\mathcal{P}_2$  as follows: for each point of  $\mathcal{P}$ , toss an independent coin of bias  $p$ . Assign the point to  $\mathcal{P}_1$  if coin turns up head, else assign it to  $\mathcal{P}_2$ . Then, the baby-processes  $\mathcal{P}_1$  and  $\mathcal{P}_2$  have the same entropy rate as Poisson processes of rates  $\lambda p$  and  $\lambda(1-p)$  respectively.*

*Proof.* Consider a Poisson Process,  $\mathcal{P}$ , of rate  $\lambda$  in the interval  $[0, T]$ . Let  $N(T)$  be the number of points in this interval and let  $\mathbf{a}(T) = \{a_1, \dots, a_{N(T)}\}$  be their locations. Further, let  $\mathbf{C}(T) = \{C_1, \dots, C_{N(T)}\}$  be the outcomes of the coin-tosses and  $M(T)$  denote the number of heads among them. Denote  $\mathbf{r}(T) = \{R_1, \dots, R_{M(T)}\}$ ,  $\mathbf{b}(T) = \{B_1, \dots, B_{N(T)-M(T)}\}$  as the locations of the baby-processes  $\mathcal{P}_1, \mathcal{P}_2$  respectively.

It is easy to see that the following bijection holds:

$$(5.1) \quad \{\mathbf{a}(T), \mathbf{C}(T), N(T), M(T)\} \rightleftharpoons \{\mathbf{r}(T), \mathbf{b}(T), N(T) - M(T), M(T)\}.$$

Given the outcomes of the coin-tosses  $\mathbf{C}(T)$ ,  $\{\mathbf{r}(T), \mathbf{b}(T)\}$  is a permutation of  $\mathbf{a}(T)$ . Hence, the Jacobian corresponding to any realization of  $\{\mathbf{C}(T), N(T), M(T)\}$  that maps  $\mathbf{a}(T)$  to  $\{\mathbf{r}(T), \mathbf{b}(T)\}$  is a permutation matrix, i.e determinant is  $\pm 1$ .

Therefore, Lemma 3.2 implies that

$$(5.2) \quad \begin{aligned} & \mathbb{H}(\mathbf{a}(T), \mathbf{C}(T), N(T), M(T)) \\ &= \mathbb{H}(\mathbf{r}(T), \mathbf{b}(T), N(T) - M(T), M(T)) \\ &\leq \mathbb{H}(\mathbf{b}(T), N(T) - M(T)) + \mathbb{H}(\mathbf{r}(T), M(T)). \end{aligned}$$

$M(T)$  is completely determined by  $\mathbf{C}(T)$  and it is easy to deduce from the definitions that

$$\mathbb{H}(M(T)|\mathbf{a}(T), \mathbf{C}(T), N(T)) = 0.$$

Hence

$$(5.3) \quad \begin{aligned} & \mathbb{H}(\mathbf{a}(T), \mathbf{C}(T), N(T), M(T)) \\ &= \mathbb{H}(\mathbf{a}(T), \mathbf{C}(T), N(T)) + \mathbb{H}(M(T)|\mathbf{a}(T), \mathbf{C}(T), N(T)) \\ &= \mathbb{H}(\mathbf{a}(T), \mathbf{C}(T), N(T)). \end{aligned}$$

Since the outcome of the coin-tosses along with their locations form a continuous time Markov chain, using Proposition 4.1 we can see that

$$\begin{aligned}
(5.4) \quad & \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{H}(\mathbf{a}(T), \mathbf{C}(T), N(T), M(T)) \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{H}(\mathbf{a}(T), \mathbf{C}(T), N(T)) \\
&= \lambda(1 - \log \lambda) - \lambda(p \log p + (1 - p) \log(1 - p)) \\
&= \lambda p(1 - \log \lambda p) + \lambda(1 - p)(1 - \log \lambda(1 - p)).
\end{aligned}$$

It is well known that  $\mathcal{P}_1, \mathcal{P}_2$  are stationary ergodic processes of rates  $\lambda p, \lambda(1 - p)$  respectively. Hence from Fact 4.1 we have

$$\begin{aligned}
(5.5) \quad & \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{H}(\mathbf{r}(T), M(T)) \leq \lambda p(1 - \log \lambda p), \\
& \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{H}(\mathbf{b}(T), N(T) - M(T)) \leq \lambda(1 - p)(1 - \log \lambda(1 - p)).
\end{aligned}$$

Combining equations (5.2), (5.4), (5.5) we can obtain

$$\begin{aligned}
(5.6) \quad & \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{H}(\mathbf{r}(T), M(T)) = \lambda p(1 - \log \lambda p), \\
& \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{H}(\mathbf{b}(T), N(T) - M(T)) = \lambda(1 - p)(1 - \log \lambda(1 - p)).
\end{aligned}$$

Thus, the entropy rates of processes  $\mathcal{P}_1$  and  $\mathcal{P}_2$  are the same as that of Poisson processes of rates  $\lambda p$  and  $\lambda(1 - p)$  respectively. This completes the proof of Lemma 5.2.  $\square$

## 6. CONCLUSIONS

This paper deals with notions of entropy for random variables that are mixed-pair, i.e. pair of discrete and continuous random variables. Our definition of entropy is a natural extension of the known discrete and differential entropy. Situations where both continuous and discrete variables arise are common in the analysis of randomized algorithms that are often employed in networks of queues, load balancing systems, etc. We hope that the techniques developed here will be very useful for the analysis of such systems and for computing entropy rates for the processes encountered in these systems.

## REFERENCES

- [CT91] T Cover and J Thomas, Elements of information theory, Wiley Interscience, 1991.
- [DVJ88] D J Daley and D Vere-Jones, An introduction to the theory of point processes, Springer Verlag, 1988.
- [PG03] B Prabhakar and R Gallager, Entropy and the timing capacity of discrete queues, IEEE Trans. Info. Theory **IT-49** (February, 2003), 357–370.
- [Sha48] C E Shannon, A mathematical theory of communication, Bell System Technical Journal **27** (July and October, 1948), 379–423 and 623–656.

## APPENDIX

6.1. **Proof of Lemma 2.1.** We wish to establish that the conditions of Lemma 2.1 guarantee that

$$(6.1) \quad \sum_i \int g_i(y) |\log g_i(y)| dy < \infty.$$

Let  $(a)_+ = \max(a, 0)$  and  $(a)_- = \min(a, 0)$  for  $a \in \mathbb{R}$ . Then,

$$a = a_+ + a_-, \quad \text{and} \quad |a| = a_+ - a_-.$$

By definition  $g_i(y) \geq 0$ . Observe that

$$|\log g_i(y)| = 2(\log g_i(y))_+ - \log g_i(y).$$

Therefore to guarantee (6.1) it suffices to show the following two conditions:

$$(6.2) \quad \sum_i \int_{\mathbb{R}} g_i(y) (\log g_i(y))_+ dy < \infty,$$

$$(6.3) \quad \sum_i \left| \int_{\mathbb{R}} g_i(y) \log g_i(y) dy \right| < \infty.$$

The next two lemmas show that equations (6.2) and (6.3) are satisfied and hence completes the proof of Lemma 2.1.

**Lemma 6.1.** *Let  $Y$  be a continuous random variable with a density function  $g(y)$  such that for some  $\delta > 0$*

$$\int_{\mathbb{R}} g(y)^{1+\delta} dy < \infty.$$

*Further if  $g(y)$  can be written as sum of non-negative functions  $g_i(y)$ , the*

$$\sum_i \int_{\mathbb{R}} g_i(y) (\log g_i(y))_+ dy < \infty.$$

*Proof.* For given  $\delta$ , there exists finite  $B_\delta > 1$  such that for  $x \geq B_\delta$ ,  $\log x \leq x^\delta$ . Using this, we obtain

$$(6.4) \quad \begin{aligned} \int_{\mathbb{R}} g_i(y) (\log g_i(y))_+ dy &= \int_{g_i(y) \geq 1} g_i(y) \log g_i(y) dy \\ &= \int_{1 \leq g_i(y) < B_\delta} g_i(y) \log g_i(y) dy + \int_{B_\delta \leq g_i(y)} g_i(y) \log g_i(y) dy \\ &\leq \log B_\delta \int_{\mathbb{R}} g_i(y) dy + \int_{\mathbb{R}} g_i(y)^{1+\delta} dy \\ &= p_i \log B_\delta + \int_{\mathbb{R}} g_i(y)^{1+\delta} dy. \end{aligned}$$

Therefore,

$$\begin{aligned} \sum_i \int_{\mathbb{R}} g_i(y) (\log g_i(y))_+ dy &\leq \sum_i \left( p_i \log B_\delta + \int_{\mathbb{R}} g_i(y)^{1+\delta} dy \right) \\ &\stackrel{(a)}{=} \log B_\delta + \int_{\mathbb{R}} \sum_i g_i(y)^{1+\delta} dy \\ &\stackrel{(b)}{\leq} \log B_\delta + \int_{\mathbb{R}} g(y)^{1+\delta} dy < \infty. \end{aligned}$$

In (a) we use the fact that  $g_i(y)$  is positive to interchange the sum and the integral. In (b), we again use the fact that  $g_i(y) \geq 0$  to bound  $\sum_i g_i(y)^{1+\delta}$  with  $(\sum_i g_i(y))^{1+\delta}$ . □

**Lemma 6.2.** *In addition to the hypothesis of  $Y$  in Lemma 6.1 assume that  $Y$  has a finite  $\epsilon$  moment for some  $\epsilon > 0$ . Then the following holds:*

$$\sum_i \left| \int_{\mathbb{R}} g_i(y) \log g_i(y) dy \right| < \infty.$$

*Proof.* Let for some  $\epsilon > 0$ ,

$$M_\epsilon = \int_{\mathbb{R}} |y|^\epsilon g(y) dy < \infty.$$

Note that for any  $\epsilon > 0$ , there is a constant  $C_\epsilon > 0$ , such that  $\int_{\mathbb{R}} C_\epsilon e^{-|y|^\epsilon} dy = 1$ . Further, observe that the density  $\tilde{g}_i(y) = g_i(y)/p_i$  is absolutely continuous w.r.t. the density  $f(y) (\triangleq C_\epsilon e^{-|y|^\epsilon})$ . Thus from the fact

that the Kullback-Liebler distance  $D(\tilde{g}_i||f)$  is non-negative we have

$$\begin{aligned} 0 &\leq \int_{\mathbb{R}} g_i(y) \log \frac{g_i(y)}{p_i f(y)} dy = \int_{\mathbb{R}} g_i(y) \log \frac{g_i(y)}{p_i C_\epsilon e^{-|y|^\epsilon}} dy \\ &= \int_{\mathbb{R}} g_i(y) \log g_i(y) dy - p_i \log p_i - p_i \log C_\epsilon + \int_{\mathbb{R}} |y|^\epsilon g_i(y) dy. \end{aligned}$$

Therefore

$$(6.5) \quad - \int_{\mathbb{R}} g_i(y) \log g_i(y) dy \leq -p_i \log p_i + p_i |\log C_\epsilon| + \int_{\mathbb{R}} |y|^\epsilon g_i(y) dy.$$

From (6.4) we have

$$(6.6) \quad \int_{\mathbb{R}} g_i(y) \log g_i(y) dy \leq \int_{\mathbb{R}} g_i(y) (\log g_i(y))_+ dy \leq p_i \log B_\delta + \int_{\mathbb{R}} g_i(y)^{1+\delta} dy.$$

Combining equations (6.5) and (6.6), we obtain

$$(6.7) \quad \left| \int_{\mathbb{R}} g_i(y) \log g_i(y) dy \right| \leq -p_i \log p_i + p_i |\log C_\epsilon| + \int_{\mathbb{R}} |y|^\epsilon g_i(y) + p_i \log B_\delta + \int_{\mathbb{R}} g_i(y)^{1+\delta} dy.$$

Now using the facts

$$\begin{aligned} - \sum_i p_i \log p_i &< \infty, \quad \sum_i \int_{\mathbb{R}} |y|^\epsilon g_i(y) dy = \int_{\mathbb{R}} |y|^\epsilon g(y) dy < \infty, \\ \text{and } \sum_i \int_{\mathbb{R}} g_i(y)^{1+\delta} dy &< \int_{\mathbb{R}} g(y)^{1+\delta} dy < \infty, \end{aligned}$$

we obtain from (6.7) that

$$\sum_i \left| \int_{\mathbb{R}} g_i(y) \log g_i(y) dy \right| < \infty.$$

□

MICROSOFT RESEARCH, REDMOND, WA 98052  
*E-mail address:* [cnair@microsoft.com](mailto:cnair@microsoft.com)

DEPTS. OF EE AND CS, STANFORD UNIVERSITY, STANFORD, CA  
*E-mail address:* [balaji@ee.stanford.edu](mailto:balaji@ee.stanford.edu)

DEPT. OF EECS, MIT, CAMBRIDGE, MA  
*E-mail address:* [devavarat@mit.edu](mailto:devavarat@mit.edu)