

THE EXPONENTIAL DISTRIBUTION IN INFORMATION THEORY

S. Verdú

UDC 621.391.1:519.28

It is shown that the exponential distribution leads to information theoretic formulas which are strikingly similar to their Gaussian counterparts:

- *A saddle-point property satisfied by the mutual information between a random variable and its sum with an exponential random variable.*
- *Rate-distortion function of the Poisson process.*
- *Capacity of single-user and multiaccess channels with additive exponential noise.*
- *Capacity of Controlled Markov Processes.*

1. Introduction

Much of the prominent role that the Gaussian distribution plays in information theory can be attributed to the closed-form expressions it leads to, and to its extremal properties. Divergence, Mutual Information and Differential Entropy have simple expressions for Gaussian random variables as well as Gaussian random processes. Among random variables with constrained variance, the Gaussian distribution maximizes differential entropy. More generally, the game between variance-constrained independent distributions

$$\max_X \min_N I(X; X + N)$$

has a Gaussian saddle-point. This leads to simple and appealing expressions for the Gaussian channel capacity and the rate-distortion function and to the least-favorability of Gaussian noise.

The purpose of this paper is to demonstrate that the exponential distribution enjoys analogous properties and closed-form expressions. Exponential distributions are fundamental in continuous-time Markov processes, since the holding times of a Markov process are exponential (and independent). Examples of information-theoretic applications where the exponential distribution is encountered are

- The digital encoding of a Poisson stream.
- The capacity of the telephone-signaling channel.
- The capacity of an exponential queueing system.
- The transmission of information via an observed Markov process.

Let us denote the exponential probability density function (pdf) with mean equal to m by

$$e_m(t) = \frac{1}{m} e^{-t/m} u(t), \quad (1)$$

where $u(t)$ denotes the unit step function. The differential entropy of the exponential pdf with mean m is equal to $\log(me)$. As is well known, no other nonnegative random variable with mean m can achieve higher differential entropy [1]. As in the Gaussian case, this fact follows directly from the nonnegativity of

Translated from *Problemy Peredachi Informatsii*, Vol. 32, No. 1, pp. 100–111, January–March, 1996.

divergence. In fact, any result that can be proved for the differential entropy can be proved, in perhaps a more elegant form, for divergences and mutual information. Following our own preferences, differential entropy will not appear in the remainder of the paper.

Section 2 uses the divergence between arbitrary exponential distributions to obtain an unexpectedly Gaussian-looking mutual information involving an exponential random variable. This result turns out to be the gateway to a mutual-information saddle-point result which is completely parallel (in fact, even more general) to the corresponding well-known Gaussian result. Section 3 contains this saddle-point result which generalizes slightly a recent result in [2]. The rest of the results in this paper are new. Section 4 finds the rate-distortion function of the Poisson process for a new fidelity criterion which leads to a solution which is very similar to the familiar Gaussian rate-distortion function with mean-squared error criterion. This result is generalized to continuous-time Markov processes. The capacity of additive exponential-noise channels is considered in Sec. 5 both in the single-user case and for the multiple-access channel, and again the solutions turns out to mimic the capacity of Gaussian channels. Finally, Sec. 6 considers a new problem where both transmitter and receiver observe the evolution of a Markov process and the transmitter has the ability to transmit information by freezing the evolution of the process. The capacity turns out to be related to (but is somewhat more complex than) the water-filling solution of parallel independent Gaussian channels.

2. Preliminaries

Fact 1. *The divergence between the distributions of $d_1 + X_1$ and $d_2 + X_2$, where X_1 and X_2 are exponential with means a_1 and a_2 respectively, is equal to ∞ if $d_1 < d_2$; otherwise, it is given by*

$$D(e_{a_1}(\cdot - d_1) || e_{a_2}(\cdot - d_2)) = \log \frac{a_2}{a_1} + \left[\frac{d_1 + a_1 - d_2 - a_2}{a_2} \right] \log e. \quad (2)$$

Now, consider the following simple problem, which will be important in the sequel. Suppose that \bar{X} and \bar{N} are independent random variables with expected values equal to a and b respectively. What is the distribution of \bar{X} that makes $\bar{X} + \bar{N}$ exponential, if \bar{N} is exponential? A good way to solve this problem is by using the fact that the Laplace transform of the exponential pdf (1) is equal to

$$\int_0^{\infty} e^{-st} e_m(t) dt = \frac{1}{1 + ms}.$$

Thus, the Laplace transform of \bar{X} must be

$$\mathbf{E} \left[e^{-s\bar{X}} \right] = \frac{1 + bs}{1 + (a+b)s} = \frac{b}{a+b} + \frac{a}{a+b} \frac{1}{1 + (a+b)s}$$

which means that \bar{X} is equal to 0 with probability $\frac{b}{a+b}$ and is exponentially distributed with mean $a+b$ conditioned on it being positive.

The mutual information between \bar{X} and $\bar{X} + \bar{N}$ is given by

$$\begin{aligned} I(\bar{X}; \bar{X} + \bar{N}) &= \mathbf{E} \left[D(e_b(\cdot - \bar{X}) || e_{a+b}) \right] \\ &= \log \left(1 + \frac{a}{b} \right) + \mathbf{E} \left[\frac{\bar{X} - a}{a+b} \right] \log e \end{aligned} \quad (3)$$

$$= \log \left(1 + \frac{a}{b} \right), \quad (4)$$

where we used (2) to write (3). Note the striking similarity of the mutual information in (4) with that of Gaussian random variables \bar{X} and \bar{N} with variances a and b , respectively.

The Gaussian counterpart of the mutual information formula (4) can be generalized to give the mutual information between jointly Gaussian random variables [3]. For example, in the special case of X and Y jointly Gaussian with correlation coefficient ρ ,

$$I(X; Y) = \frac{1}{2} \log \left(\frac{1}{1 - \rho^2} \right).$$

If X and Y have a joint distribution given by the bivariate exponential distribution of Marshall and Olkin [4], (which satisfies a certain memoryless property), then it can be shown that

$$I(X; Y) = \infty.$$

3. Mutual information saddle-point

Theorem 1. Fix nonnegative scalars a and b . Let \bar{N} be exponentially distributed with mean b . As in Sec. 2, define \bar{X} as a nonnegative random variable independent of \bar{N} and with mean a , by the following mixture of a point mass and an exponential distribution:

$$\begin{aligned} P[\bar{X} = 0] &= \frac{b}{a + b}, \\ P[\bar{X} > x | \bar{X} > 0] &= e^{-x/(a+b)}. \end{aligned}$$

Then,

(a)

$$I(\bar{X}; \bar{X} + \bar{N}) = \log \left(1 + \frac{a}{b} \right)$$

(b) For any nonnegative random variable X , independent of \bar{N} , with mean a ,

$$I(X; X + \bar{N}) \leq I(\bar{X}; \bar{X} + \bar{N}),$$

with equality only if $X = \bar{X}$.

(c) For any nonnegative random variable N (possibly dependent on \bar{X}) with mean b

$$I(\bar{X}; \bar{X} + \bar{N}) \leq I(\bar{X}; \bar{X} + N),$$

with equality only if N is exponential independent of \bar{X} .

(d) For any independent nonnegative random variables X and N with means a and b , respectively,

$$I(X; X + N) = \log \left(1 + \frac{a}{b} \right) + D(P_N || P_{\bar{N}}) - D(P_{X+N} || P_{\bar{X} + \bar{N}}).$$

PROOF. (a) This was shown in Sec. 2, using the general expression for divergence between exponential distributions.

(b) For any nonnegative random variable X , independent of \bar{N} with mean a ,

$$\begin{aligned} I(X; X + \bar{N}) &= D(P_{X+\bar{N}|X} || P_{X+\bar{N}} | P_X) \\ &= D(P_{X+\bar{N}|X} || Q | P_X) - D(P_{X+\bar{N}} || Q) \\ &\leq D(P_{X+\bar{N}|X} || Q | P_X), \end{aligned} \tag{5}$$

where Q is any distribution such that $P_{X+\bar{N}} \ll Q$. In this case, we choose Q to be exponential with mean $a + b$. Then, the right-hand side of (5) becomes

$$D(P_{X+\bar{N}|X} || Q | P_X) = \log \left(1 + \frac{a}{b} \right) - \frac{\log e}{b} \mathbf{E}[\bar{N}] + \frac{\log e}{a+b} \mathbf{E}[X + \bar{N}] = \log \left(1 + \frac{a}{b} \right),$$

regardless of the choice of X . This proof shows that if $P_X \neq P_{\bar{X}}$, then the inequality in (b) is strict, since in that case

$$D(P_{\bar{X}+\bar{N}} \| Q) > 0.$$

(c) Choose any nonnegative random variable N with mean b , and let $\bar{Y} = \bar{X} + \bar{N}$. Then,

$$\begin{aligned} I(\bar{X}; \bar{X} + N) &= D(P_{\bar{X}+N|\bar{X}} \| P_{\bar{X}+\bar{N}|\bar{X}} | P_{\bar{X}}) - D(P_{\bar{X}+N} \| P_{\bar{X}+\bar{N}}) \\ &+ \mathbf{E} \left[\log \frac{P_{\bar{Y}|\bar{X}}(\bar{X} + N | \bar{X})}{P_{\bar{Y}}(\bar{X} + N)} \right] \\ &\geq \mathbf{E} \left[\log \frac{P_{\bar{Y}|\bar{X}}(\bar{X} + N | \bar{X})}{P_{\bar{Y}}(\bar{X} + N)} \right] \\ &= \log \left(1 + \frac{a}{b} \right) - \frac{\log e}{b} \mathbf{E}[N] + \frac{\log e}{a+b} \mathbf{E}[\bar{X} + N] = \log \left(1 + \frac{a}{b} \right), \end{aligned} \quad (6)$$

regardless of the choice of N , where the inequality follows from the fact that conditioning increases divergence [5]. In order to check that (6) holds with equality only if $P_N = P_{\bar{N}}$, notice that

$$\begin{aligned} D(P_{\bar{X}+N|\bar{X}} \| P_{\bar{X}+\bar{N}|\bar{X}} | P_{\bar{X}}) &= D(P_{N|\bar{X}} \| P_{\bar{N}} | P_{\bar{X}}) = D(P_N \| P_{\bar{N}}) + I(\bar{X}; N) \\ &\geq D(P_N \| P_{\bar{N}}) \geq D(P_{\bar{X}+N} \| P_{\bar{X}+\bar{N}}), \end{aligned}$$

where the inequalities hold with equality only if N is independent of \bar{X} and has the same distribution as \bar{N} , respectively.

(d) Let $Y = X + N$. We can decompose the mutual information between X and Y as

$$I(X; Y) = \mathbf{E} \left[\log \frac{P_{Y|X}(Y | X)}{P_{\bar{Y}|\bar{X}}(Y | X)} \right] + \mathbf{E} \left[\log \frac{P_{\bar{Y}|\bar{X}}(Y | X)}{P_{\bar{Y}}(Y)} \right] - \mathbf{E} \left[\log \frac{P_Y(Y)}{P_{\bar{Y}}(Y)} \right], \quad (7)$$

where the expectations are with respect to the joint distribution of X and Y . It is easy to see that if we condition on X in the first expectation in the right side of (7) we obtain the constant $D(P_N \| P_{\bar{N}})$. The second expectation depends on X and Y only through their respective means, so it is equal to $I(\bar{X}; \bar{X} + \bar{N})$. Finally, the third term in (7) is equal to $D(P_Y \| P_{\bar{Y}})$. \triangle

4. Rate-distortion function of the Poisson process

The homogeneous rate- λ Poisson process is a point process whose interarrival times are independent with identical distribution equal to an exponential distribution with mean $1/\lambda$. Suppose it is desired to encode the times of arrival of a Poisson process with a finite number of bits per second. Encoding the arrival times is equivalent to encoding a memoryless source of exponential interarrival times. Once a fidelity criterion is chosen, the rate-distortion function reflects the optimal trade-off between encoding rate and achievable fidelity. This problem has been considered before in [6] with a distortion measure equal to the normalized absolute error between the true and reproduced interarrival times. However, that fidelity criterion does not result in closed-form results. An alternative criterion is presented in the next result which gives the same shape as the rate-distortion function of a Gaussian process with mean-squared error distortion. This result applies to situations where there is a hard constraint on how much a reproduced interarrival time can exceed the true interarrival time. This constraint, together with the constraint that the last arrival (sum of all n interarrival times) is declared no sooner than it occurs, translates into a nontrivial encoding of the Poisson process which becomes arbitrarily faithful as the permissible overestimation of each interarrival time diminishes to 0.

Theorem 2. *Consider a Poisson process with rate λ arrivals per second and the following lossy-source coding problem with fidelity parametrized by d seconds:*

- The interarrival times X_1, \dots, X_n are reproduced by $\hat{X}_1, \dots, \hat{X}_n$.

- $$\hat{X}_i \leq X_i + d. \quad (8)$$

- $$\sum_{i=1}^n \hat{X}_i \geq \sum_{i=1}^n X_i. \quad (9)$$

Let the rate-distortion function $R(d)$ be the minimum encoding rate such that the probability that the constraints (8) and (9) are not met vanishes with n . Then,

$$R(d) = \begin{cases} \lambda \log(1/\lambda d) \text{ bits per second} & \text{if } \lambda d \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

PROOF. In the case $\lambda d > 1$, the encoder need not transmit any information since the decoder can simply output

$$\hat{X}_i = d, \quad i = 1, \dots, n.$$

This ensures that the constraint (8) is satisfied. Moreover, the law of large numbers implies that the constraint (9) is violated with vanishing probability as $n \rightarrow \infty$. Henceforth, we will assume $\lambda d \leq 1$. The usual coding theorem methods [7] result in the expression for the rate-distortion function in bits per arrival:

$$\lambda^{-1} R(d) = \inf_{\substack{P_{\hat{X}|X} \\ \mathbf{E}[\hat{X}] \geq \lambda^{-1}, 0 \leq \hat{X} \leq X+d}} I(X; \hat{X}), \quad (10)$$

where X is exponential with mean λ^{-1} . Fix any choice of the joint distribution of X and \hat{X} compatible with that marginal of X and with the constraints in (10). Let us define the nonnegative random variable

$$N = X - \hat{X} + d.$$

Then,

$$I(X; \hat{X}) = I(X; X - N + d) = I(X; Y),$$

where we have used

$$Y = X - N.$$

Define the conditional distribution below

$$Q_{X|Y}(x|y) = e_d(x - y),$$

and consider the following decomposition of mutual information:

$$I(X; Y) = D(P_{X|Y} \| P_X | P_Y) = D(P_{X|Y} \| Q_{X|Y} | P_Y) + \mathbf{E} \left[\log \frac{Q_{X|Y}(X|Y)}{P_X(X)} \right], \quad (11)$$

where the expectation is with respect to P_{XY} . Using the expression for the exponential pdf, we obtain

$$\log \frac{Q_{X|Y}(X|Y)}{P_X(X)} = \log \left(\frac{1}{\lambda d} \right) + \lambda X \log e - \frac{X - Y}{d} \log e + \log u(X - Y) - \log u(X),$$

which, upon taking expectations with respect to P_{XY} , results in

$$\mathbf{E} \left[\log \frac{Q_{X|Y}(X|Y)}{P_X(X)} \right] = \log \left(\frac{1}{\lambda d} \right) + \frac{\mathbf{E}[\hat{X}] - \lambda^{-1}}{d} \log e \geq \log \left(\frac{1}{\lambda d} \right). \quad (12)$$

Since the conditional divergence in (11) is nonnegative, we can conclude from (12) that the infimum in (10) is lower bounded by $\log \left(\frac{1}{\lambda d} \right)$. We will now check that this lower bound is achieved by the following choice of the joint distribution of X and Y :

- Y has the Laplace transform:

$$\mathbf{E}[e^{-sY}] = \frac{sd+1}{\frac{s}{\lambda}+1} \quad (13)$$

- $P_{X|Y} = Q_{X|Y}$, which means that $X = Y + N$ with N a mean- d exponentially distributed random variable independent of Y .

Note that this is indeed a valid choice, since the constraints in the feasible set of (10) are satisfied and the marginal pdf of X is given by the convolution of the pdf of Y and e_d , which has the Laplace transform:

$$\frac{sd+1}{\frac{s}{\lambda}+1} \frac{1}{sd+1} = \frac{1}{\frac{s}{\lambda}+1}$$

as desired.

Since in the present case $P_{X|Y} = Q_{X|Y}$, the conditional divergence in (11) is equal to 0 and $\mathbf{E}[\hat{X}] = \lambda^{-1}$. Therefore, the lower bound is attained with equality, and the infimum in (10) must be equal to $\log\left(\frac{1}{\lambda d}\right) \cdot \Delta$.

From the proof of Theorem 2 (and, in particular, the fact that the optimal output random variable is mixed-type (13)) we can see that an optimal encoding of the Poisson process with tolerance d (in the sense of (8) and (9)) assigns the value d to all sufficiently small interarrival times.

The result of Theorem 2 can be used to find the rate-distortion function of a continuous-time Markov process with a discrete state space \mathcal{S} . Let us assume that the process is ergodic; the transition rate from state i to j is denoted by λ_{ij} ; the stationary distribution (solution to the balance equations) is denoted by π_i , $i \in \mathcal{S}$, and the rate out of state i is denoted by

$$\lambda_i = \sum_{j \in \mathcal{S} \setminus \{i\}} \lambda_{ij}.$$

It is well known that given the state, the holding times are independent, exponential with means λ_i^{-1} . To simplify the problem, let us assume that the (discrete-time) sequence of states is to be encoded noiselessly. This takes a number of bits per transition which is equal to the entropy rate of the jump Markov chain. Now, the problem is to compress the holding times so that any decoded holding time at state i exceeds the true holding time by no more than d_i , and such that the sum of the decoded holding times for each state be not smaller than the sum of the actual holding times. Since both the encoder and decoder know the true sequence of states, the problem can be viewed as the decoupled encoding of sequences of independent identically distributed random variables with exponential distribution with mean λ_i^{-1} . Applying the result of Theorem 2, the required rate in bits per second is:

$$\sum_{i \in \mathcal{S}} \pi_i \lambda_i [\log 1/(\lambda_i d_i)]^+,$$

which in the special case where the tolerances are a fraction of the respective means $d_i = \delta/\lambda_i$ reduces to $\log \frac{1}{\delta}$ times $\sum_{i \in \mathcal{S}} \pi_i \lambda_i$.

Another way to pose the rate-distortion of the Markov process is to require $\hat{X}_i \leq X_i$ and

$$\frac{1}{n} \sum_{i=1}^n X_i - \hat{X}_i \leq d$$

with probability approaching 1 as $n \rightarrow \infty$. It is easy to see that this alternative form of the constraints leads to the same expression of the rate-distortion function of the Poisson process found in Theorem 2. The

advantage in the Markov process context is that we do not need to put separate constraints on the precisions with which holding times are encoded for each state; instead d_i is determined by the *water-flooding* solution:

$$d_i = \min\left(d^*, \frac{1}{\lambda_i}\right)$$

with the flood level d^* chosen so that $\sum_{i \in S} \pi_i d_i = d$.

5. Capacity of additive exponential-noise channels

Let us now consider the problem of finding the capacity of the following channel:

$$Y_i = X_i + N_i, \quad (14)$$

where N_i is an independent sequence of exponentially distributed random variables with mean b , and any codeword X_1, \dots, X_n is constrained to satisfy

$$X_i \geq 0 \quad (15)$$

and

$$\frac{1}{n} \sum_{i=1}^n X_i \leq a. \quad (16)$$

Examples of channels where this problem arises are the telephone signaling channel and the exponential queue with feedback [2]. Using the result in Theorem 1 it is straightforward to show the following result.

Theorem 3. *The capacity of the additive exponential noise channel (14) under constraints (15) and (16) is equal to*

$$\log\left(1 + \frac{a}{b}\right).$$

Moreover, the capacity of any channel with nonnegative noise N with mean b satisfies (cf. [8])

$$\log\left(1 + \frac{a}{b}\right) \leq C \leq \log\left(1 + \frac{a}{b}\right) + D(P_N \| e_b).$$

In the Gaussian channel, a maximum likelihood decoder selects the closest codeword in Euclidean distance from the received word. In the present case, the maximum likelihood decoder can rule out any codeword for which there is a component that exceeds the observed value; among the surviving codewords, the most likely one is the one that corresponds to the smallest *sum* of noise components.

An extension of Theorem 3 considers the multiple-access channel:

$$Y_i = X_{1i} + X_{2i} + N_i,$$

where for $k = 1, 2$

$$X_{ki} \geq 0$$

and

$$\frac{1}{n} \sum_{i=1}^n X_{ki} \leq a_k.$$

The capacity region of this channel is given by the pentagon

$$\begin{aligned} 0 &\leq R_1 \leq \log\left(1 + \frac{a_1}{b}\right), \\ 0 &\leq R_2 \leq \log\left(1 + \frac{a_2}{b}\right), \\ R_1 + R_2 &\leq \log\left(1 + \frac{a_1 + a_2}{b}\right). \end{aligned}$$

This can be shown in an entirely analogous fashion to the Gaussian case: the converse follows immediately from the single-user result and the fact that all the components in the sum of both codewords are nonnegative and their sum does not exceed $a_1 + a_2$. The achievability of the corner point

$$(R_1, R_2) = \left(\log \left(1 + \frac{a_1}{b} \right), \log \left(1 + \frac{a_2}{a_1 + b} \right) \right) \quad (17)$$

can be argued from the fact that if \bar{X}_1 and \bar{X}_2 are defined as the independent mixed-type random variables below with means a_1 and a_2 , respectively

$$\begin{aligned} P[\bar{X}_1 = 0] &= \frac{b}{a_1 + b}, & P[\bar{X}_1 > x | \bar{X}_1 > 0] &= e^{-x/(a_1+b)}, \\ P[\bar{X}_2 = 0] &= \frac{a_1 + b}{a_1 + a_2 + b}, & P[\bar{X}_2 > x | \bar{X}_2 > 0] &= e^{-x/(a_1+a_2+b)}, \end{aligned}$$

then they achieve

$$\begin{aligned} I(\bar{X}_1; \bar{Y} | \bar{X}_2) &= \log \left(1 + \frac{a_1}{b} \right), \\ I(\bar{X}_2; \bar{Y}) &= \log \left(1 + \frac{a_2}{a_1 + b} \right). \end{aligned}$$

The usual symmetry and time-sharing arguments conclude the achievability proof.

Despite the stark similarity of the exponential multiaccess capacity region to the Gaussian case, it is interesting to note that a unique pair of distributions does not attain the whole capacity region. We have identified above the distribution pairs that attain the corner points (the symmetric point to (17) can be obtained by switching the roles of a_1 and a_2). As an alternative to time-sharing, we will show how to achieve any convex combination of corner points by using the rate-splitting method of [9]. This will show that the capacity region is not reduced in the absence of frame-synchronism. Any convex combination (\bar{R}_1, \bar{R}_2) of the corner points satisfies

$$\begin{aligned} \bar{R}_2 &= \log \left(1 + \frac{a_2}{b + \delta} \right), \\ \bar{R}_1 + \bar{R}_2 &= \log \left(1 + \frac{a_1 + a_2}{b} \right) \end{aligned}$$

with $0 \leq \delta \leq a_1$. In order to achieve this point, we will choose \bar{X}_2 with Laplace transform

$$L_2(s) = \frac{1 + (b + \delta)s}{1 + (a_2 + b + \delta)s}$$

and $\bar{X}_1 = \bar{X}_{10} + \bar{X}_{11}$, where \bar{X}_{10} and \bar{X}_{11} are independent with respective Laplace transforms

$$\begin{aligned} L_{10}(s) &= \frac{1 + bs}{1 + (\delta + b)s}, \\ L_{11}(s) &= \frac{1 + (b + a_2 + \delta)s}{1 + (b + a_1 + a_2)s}, \end{aligned}$$

where note that $\mathbf{E}[\bar{X}_2] = a_2$, $\mathbf{E}[\bar{X}_{10}] = \delta$, and $\mathbf{E}[\bar{X}_{11}] = a_1 - \delta$. Consider a 3-user channel

$$\bar{Y} = \bar{X}_{10} + \bar{X}_{11} + \bar{X}_2 + N.$$

It is easy to check (multiplying Laplace transforms) that \bar{Y} is exponential with mean b . The achievable rates in this channel can be obtained as

$$R_{11} = I(\bar{X}_{11}; \bar{Y}) = \log \left(1 + \frac{a_1 - \delta}{b + a_2 + \delta} \right).$$

$$R_2 = I(\bar{X}_2; \bar{Y} | \bar{X}_{11}) = \log \left(1 + \frac{a_2}{b + \delta} \right),$$

$$R_{10} = I(\bar{X}_{10}; \bar{Y} | \bar{X}_{11}, \bar{X}_2) = \log \left(1 + \frac{\delta}{b} \right),$$

which satisfies $R_2 = \bar{R}_2$ and $R_{11} + R_{10} = \bar{R}_1$ as can be seen from

$$R_{11} + R_{10} + R_2 = \log \left(1 + \frac{a_1 + a_2}{b} \right).$$

6. Communication via observed Markov processes

The next obvious problem where a counterpart to the Gaussian case can be sought is the water-filling solution of the capacity of parallel Gaussian channels. Since that solution hinges exclusively on the functional form of the capacity of the individual Gaussian channel, which is identical to the exponential solution, the same water-filling solution is obtained for the capacity of parallel independent exponential-noise channels with different noise strengths and an overall constraint on the sum of the input symbols. The main usefulness of this solution in the Gaussian case is that its limiting form leads to the capacity of the non-white Gaussian channel. The counterpart of such a result could be sought by generalizing (14) to the case where the noise values are the interarrival times of a nonhomogeneous Poisson process. Instead of taking this direction, we will devote this section to a channel whose capacity has some connections to the water-filling interpretation.

Suppose that the receiver observes noiselessly the evolution of a continuous-time discrete-state Markov process with state space \mathcal{S} . The transmitter has the ability to freeze the process during any interval of time which may depend on the state of the process. We will assume that information can be transmitted only in the timing of those intervals chosen by the encoder where the process is frozen. In many cases of interest, information can also be transmitted by influencing the transition probabilities (controlled Markov chain); however, in this paper we will not consider this possibility in order to evaluate the information rate that can be encoded only in the timing. Naturally, the receiver does not observe the times at which the process has been frozen by the encoder, and can only obtain (a noisy version of) that information by observing the state of the process at every time. To motivate this problem, consider the following simple special case. Suppose that the transmitter sends strings of alternating symbols $\{\text{SHORT}, \text{LONG}\}$, which take an exponentially distributed amount of time to reach the receiver with means $1/\mu$ and $1/\lambda$ ($\lambda \leq \mu$) respectively. The transmitter knows when the symbols reach their destination and it cannot start sending the next symbol before the previous one has been received. Information can be transmitted by not sending the next symbol immediately upon the reception of the current one; instead a waiting time is inserted before sending each symbol, whose duration depends on the message to be transmitted. This corresponds to the special case of the model above, where $\mathcal{S} = \{\text{SHORT}, \text{LONG}\}$ and the transition rates from **SHORT** to **LONG**, and vice versa, are μ and λ , respectively. As a special case of the formula obtained below, the capacity in nats/sec is given by

$$C = \begin{cases} e^{-1} \sqrt{\lambda \mu}, & \text{if } e^{-2} \mu \leq \lambda \leq \mu, \\ \text{solution to } 1 + \frac{C}{\lambda} = \log \frac{\mu}{C}, & \text{if } \lambda < e^{-2} \mu. \end{cases}$$

Note that the dichotomy in this result is reminiscent of the water-filling solution for parallel Gaussian channels. If the **LONG** symbol is too long ($\lambda < e^{-2} \mu$), then its duration has too much uncertainty and it does not pay to delay its transmission (and thus pay a penalty in time). Otherwise, it does pay to encode timing information by delaying both symbols, and the capacity is proportional to the geometric mean of μ and λ in that case.

Let us now consider a general Markov process and let us use the notation introduced in Sec. 4. In particular, λ_i denotes the sum of the transition rates out of state i . We will denote the stationary distribution

of the jump (discrete-time) Markov chain by p_i . It can be seen that the problem decouples into separate channels of the type (14), where $N_{i,j}$ are independent exponentially distributed with mean λ_j^{-1} . The average time by which the process is frozen when it is in state j is a degree of freedom which is denoted by α_j . Those averages are chosen to maximize the transmitted information rate:

$$C = \max_{\alpha_i, i \in \mathcal{S}} \frac{\sum_{i \in \mathcal{S}} p_i \log(1 + \alpha_i \lambda_i)}{\sum_{i \in \mathcal{S}} p_i (\alpha_i + \lambda_i^{-1})}, \quad (18)$$

which shows the trade-off inherent in the obvious fact that longer freezing times are more informative: both the numerator and denominator are increasing in α_i . Note that because of the property we showed in Theorem 1, when the optimal input distribution is used the overall holding time seen by the receiver is still exponential.

The solution to the optimization problem in (18) is readily shown:

$$\bar{\alpha}_i = \left[\frac{1}{C} - \frac{1}{\lambda_i} \right]^+, \quad (19)$$

which has a familiar flavor, except that in this case one cannot build a parametric solution with the water-level as an independent parameter since in this case it is the reciprocal of the capacity itself. Let us consider separately the case where

$$\min_{i \in \mathcal{S}} \lambda_i \geq e^{-1} \exp \left(\sum_{i \in \mathcal{S}} p_i \log \lambda_i \right).$$

In this case, all $\bar{\alpha}_i > 0$ and the capacity is given by

$$C = e^{-1} \exp \left(\sum_{i \in \mathcal{S}} p_i \log \lambda_i \right).$$

Otherwise, capacity is obtained by partitioning the state space into “fast” states \mathcal{F} and “slow” states $\mathcal{S} \setminus \mathcal{F}$, in such a way that if $\lambda_i < \lambda_j$ and $i \in \mathcal{F}$, then $j \in \mathcal{F}$. For every such partition, we can find the solution $C_{\mathcal{F}}$ to the equation

$$\sum_{i \in \mathcal{F}} p_i + \sum_{i \notin \mathcal{F}} p_i \frac{C_{\mathcal{F}}}{\lambda_i} = \sum_{i \in \mathcal{S}} p_i \left[\log \frac{\lambda_i}{C_{\mathcal{F}}} \right]^+.$$

Then, the capacity C in nats/sec is simply the largest $C_{\mathcal{F}}$ over all partitions of fast-slow states.

REFERENCES

1. T. M. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York (1991).
2. V. Anantharam and S. Verdú, “Bits through queues,” *IEEE Trans. Inf. Theory*, **42**, No. 1 (1996).
3. M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden Day, San Francisco (1964).
4. A. W. Marshall and I. Olkin, “A multivariate exponential distribution,” *J. American Statist. Assoc.*, **62**, 30–44 (1967).
5. I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York (1981).
6. I. Rubin, “Information rates and data-compression schemes for Poisson processes,” *IEEE Trans. Inf. Theory*, **20**, No. 2, 200–210 (1974).
7. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, New Jersey (1971).
8. S. Ihara, “On the capacity of channels with additive non-Gaussian noise,” *Inform. Control*, **37**, 34–39 (1978).
9. B. Rimoldi and R. Urbanke, “A rate-splitting approach to the Gaussian multiple-access channel,” *IEEE Trans. Inf. Theory* (to appear).