Mutual Information and Conditional Mean Estimation in Poisson Channels

Dongning Guo, *Member, IEEE,* Shlomo Shamai (Shitz), *Fellow, IEEE,* and Sergio Verdú, *Fellow, IEEE*

***Abstract*—Following the discovery of a fundamental connection between information measures and estimation measures in Gaussian channels, this paper explores the counterpart of those results in Poisson channels. In the continuous-time setting, the received signal is a doubly stochastic Poisson point process whose rate is equal to the input signal plus a dark current. It is found that, regardless of the statistics of the input, the derivative of the input–output mutual information with respect to the intensity of the additive dark current can be expressed as the expected difference between the logarithm of the input and the logarithm of its noncausal conditional mean estimate. The same holds for the derivative with respect to input scaling, but with the logarithmic function replaced by $x \log x$. Similar relationships hold for discrete-time versions of the channel where the outputs are Poisson random variables conditioned on the input symbols.**

Index Terms: Mutual information, nonlinear filtering, optimal estimation, point process, Poisson process, smoothing.

I. Introduction

Some fundamental relationships between input–output mutual information and conditional mean estimation have recently been discovered for additive Gaussian noise channels with arbitrary input [1]. In its simplest form, the derivative of the mutual information in nats as a function of the signal-to-noise ratio (SNR) is equal to half the minimum mean-square error (MMSE) regardless of the input statistics, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X; \sqrt{\gamma} X + N) = \frac{1}{2} \mathsf{E}\left\{ (X - \mathsf{E}\{X | \sqrt{\gamma} X + N\})^2 \right\} \quad (1)$$

for every P_X, where $N \sim \mathcal{N}(0,1)$ is standard Gaussian and $\gamma > 0$ stands for the SNR. Remarkably, the relationship also applies to continuous-time additive white Gaussian noise channels with arbitrary input process.

This paper develops parallel results for Poisson channels, the output of which are Poisson random variables or doubly stochastic Poisson point processes conditioned on the input. Such channels occur in direct-detection optical communication systems, in which incident radiation is intercepted by means of photon-sensitive devices to produce a point process, whose rate is typically the intensity of the incident radiation plus a (constant) "dark current".

Reference [2] presents a review of major developments of communication theory in the Poisson regime, of which we give a brief summary. Signal detection in Poisson channels has been studied since the 1960s and the general "Poisson matched filter" which yields optimal detection was found by 1969 [3]. Stochastic integration with Poisson point process observations was developed for various filtering problems in the 1970s (e.g., [4], [5]). In particular, the likelihood ratio for signal detection has been found as a stochastic integral. Using martingale theory, the likelihood ratio for detection based solely on the observation has been shown to admit an "estimator-correlator" type of formula (e.g., [6], cf. [7]). Furthermore, the mutual information can be expressed using the Liptser-Shiryaev formula (counterpart to Duncan's Gaussian noise formula [8]) as an integral of the expectation of the difference between a function ($x \log x$) of the input and the same function of the causal conditional mean estimate [4], [5]. The capacity of Poisson channels under peak- and average-power limits was found [9], [10], allowing infinite bandwidth. The reliability function at all rates below capacity is also known [11]. The only known closed-form expression for the rate-distortion function of the Poisson process was found in [12] under an appropriate distortion measure which finds a natural queueing interpretation [13]. Bounds on the capacity are found under bandwidth-like constraints [14]. More recently, the high signal-to-noise ratio asymptotic capacity of a peak and average power limited discrete-time Poisson channel is derived in [15] by observing that the entropy of the output is lower bounded by the differential entropy of the input. Poisson multiple-access channels, Poisson broadcast channels, Poisson multiple-input multiple-output channels, Poisson fading channels and Poisson arbitrarily varying channels are studied in references [16], [17], [18], [19] and [20] respectively. Also, a Poisson multiple-access channel where the electrical fields (instead of the energy) superpose is studied in [21] where it is found that time-division multiple access is optimal in terms of the cutoff rate [22].

Equipped with stochastic integration techniques, this paper studies the input–output mutual information of Poisson channels in discrete-time and continuous-time settings. A key result in this paper is that, regardless of the statistics of the input, the derivative of the input–output mutual information of a Poisson channel with respect to the intensity of the dark current is equal to the expected error between the logarithm of the actual input and the logarithm of its conditional mean estimate (noncausal in case of continuous-time). Equivalently,

Manuscript received ????; revised ????. This work was presented in part in the 2004 IEEE Information Theory Workshop, San Antonio, TX, USA. This work was partially supported by NSF Grants CCR-0312839 and CCF-0644344, DARPA Grant W911NF-07-1-0028, and the US-Israel Binational Science Foundation.

Dongning Guo is with the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL, 60208, USA.

Shlomo Shamai (Shitz) is with the Department of Electrical Engineering, Technion-Israel Institute of Technology, 32000 Haifa, Israel.

Sergio Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA.

the mutual information can be expressed as an integral of such an error as a function of the dark current. The derivative of the mutual information with respect to the scaling can also be expressed as a function of a certain error associated with the conditional mean estimate. In the continuous-time setting, together with the Liptser-Shiryaev formula [5], our results complete the triangle relationship of the mutual information and causal and noncausal conditional mean estimation errors.

The problem of Poisson channels studied in this paper is technically and conceptually more involved than its counterpart in the Gaussian regime. Some of the difficulties are inherent to Poisson channels: 1) The dark current and scaling cannot be consolidated into one parameter as in Gaussian channels; and 2) The channel conditioned on a degraded version of the output is no longer Poisson. Other difficulties are due to the fact that less is known about Poisson channels. For example, the hybrid continuous-discrete nature of the input–output pair appears harder to deal with; simple closed form expressions for conditional mean estimate and mutual information are known for fewer input distributions than in the Gaussian case; and little is known about "natural" metrics for measuring estimation errors.

In a wider context, this work reveals new connections between information theory and estimation theory. The results allow certain information measures to be expressed using solely estimation errors and vice versa. Since the work of [1] on Gaussian channels, such relationships have been developed not only for Poisson channels in [23] and this work, but also for a variety of other channels of interest, including additive non-Gaussian noise channels [24] and discrete memoryless channels [25] (see also [26]). Moreover, [25] obtained the derivatives of mutual information with respect to certain parameters of arbitrary random transformations. In all the above cases, the posterior distribution of the input given the observations plays an important role in the result.

The rest of the paper is organized as follows. Section II gives the necessary background on Poisson channels and conditional mean estimates. The main results of this work are presented in Section III, followed by some numerical examples in Section IV. Proofs of the results is relegated to Section V. Concluding remarks are given in Section VI.

II. Poisson Channels

A. Poisson Random Transformation

We start with a simple random transformation of the Poisson type that captures many of the properties of general Poisson channels. Let X and Y be a pair of random variables taking values in $[0,\infty)$ and the set of nonnegative integers respectively, where P_X denotes the distribution of X, and conditioned on $X=x$, the variable Y has Poisson distribution with mean equal to x:

$$P_{Y|X}(k|x) = \frac{1}{k!}x^k e^{-x}, \quad k=0,1,\dots \tag{2}$$

For convenience, we use the shorthand $\mathcal{P}(X)$ to denote an arbitrary Y related to X according to (2), i.e., $\mathcal{P}(X)$ is a conditionally Poisson random variable with its mean value equal to X. Note that $\mathcal{P}(X)$ can be regarded as a random transformation of X.

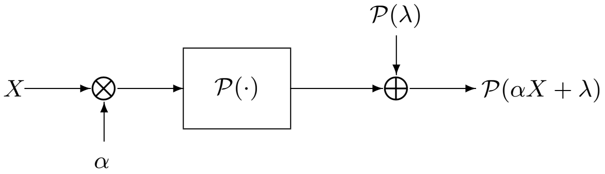

Fig. 1. Poisson random transformation with scaling α and dark current λ.

Given an arbitrary X and a conditionally Poisson variable $\mathcal{P}(X)$, consider the conditional mean estimate of X given $\mathcal{P}(X)$, which is aptly denoted using the angle bracket operator:

$$\langle X \rangle = \mathsf{E}\left\{ X \mid \mathcal{P}(X) \right\}. \tag{3}$$

Note that $\langle X \rangle$ is an implicit function of the conditionally Poisson variable $\mathcal{P}(X)$. Evidently, $\langle\cdot\rangle$ is a nonlinear operator: In general $\langle \alpha X+\lambda \rangle$ has a distribution different from that of $\alpha\langle X\rangle+\lambda$ for all $\alpha>0$, $\alpha\neq 1$ and/or $\lambda>0$.

B. Discrete-time Poisson Channels

Repeated independent use of the random transformation (2) defines a canonical discrete-time memoryless Poisson channel by regarding $P_{Y|X}$ as the input–output conditional distribution at each time instance. A general Poisson channel is defined by a transformation whose output, conditioned on the input $X=x$, is a Poisson random variable with its mean equal to $(\alpha x+\lambda)$. Here, $\alpha\geq 0$ is known as the scaling (factor) of the input, and $\lambda\geq 0$ (the intensity of) the "dark current". Figure 1 illustrates a construction of the general Poisson transformation using independent canonical ones. This setting has a direct counterpart in the Gaussian regime where α is the amplitude scaling and λ corresponds to the Gaussian noise level. Note that in the Gaussian case the scaling and the noise level consolidate to a single degree of freedom, the SNR, for all analysis purposes. This is not true in the Poisson case because, for one thing, $\mathcal{P}(\alpha X)$ and $\alpha\,\mathcal{P}(X)$ have different distributions unless $\alpha=0$ or $\alpha=1$.

More generally, the discrete-time input process $\{X_1, X_2, \dots\}$ to a discrete-time Poisson channel over time may have memory. The output is a discrete-time process $\{Y_n\}$ where $Y_n=\mathcal{P}(\alpha X_n+\lambda)$ are independent identically distributed (i.i.d.) conditioned on the input process.

C. Continuous-time Poisson Channels

The canonical continuous-time Poisson channel is the following. Let $\{X_t,\ 0\leq t\leq T\}$, or equivalently, X_0^T denote the input process, where X_t takes values in $[0,\infty)$. The output is a realization of a Poisson point process $\{Y_t\}$ whose time-varying expectation at any time t is equal to the integral of the "rate function" X_t. Precisely, for all $0\leq t<s\leq T$,

$$\mathsf{P}\left\{ Y_s - Y_t = k \mid X_0^T \right\} = \frac{1}{k!}\Lambda^k e^{-\Lambda}, \quad k=0,1,\dots \tag{4}$$

where

$$\Lambda = \int_t^s X_\xi \, \mathrm{d}\xi. \tag{5}$$

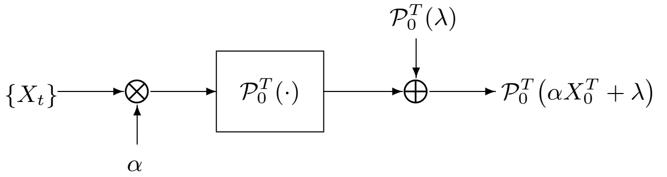
$\mathcal{P}_0^T(\lambda)$
$\{X_t\}$
$\mathcal{P}_0^T(\cdot)$
$\mathcal{P}_0^T\left(\alpha X_0^T+\lambda\right)$
α

Fig. 2. A continuous-time Poisson channel with scaling factor α and dark current level λ.

A general Poisson channel can be regarded as the canonical channel with its input replaced by the rate function $\alpha X_t+\lambda$, $0\le t\le T$. The output is known as a doubly stochastic Poisson process. Let the output at time t be denoted by $\mathcal{P}_t(\alpha X_0^t+\lambda)$, which depends on the input up to time t. Also let the output process in the interval $[r,s]$ by $\mathcal{P}_r^s(\alpha X_0^s+\lambda)$. In general, $\mathcal{P}_0^T\left(\alpha X_0^T+\lambda\right)$ can be regarded as the process $\mathcal{P}_0^T\left(\alpha X_0^T\right)$ superposed with an independent point process of constant rate λ as depicted in Figure 2. Moreover, we denote the conditional mean of the input at time given the Poisson channel output $\mathcal{P}_0^s(X_0^s)$ as

$$\langle X_t\rangle_s = \mathsf{E}\left\{X_t \mid \mathcal{P}_0^s(X_0^s)\right\}. \tag{6}$$

Note that the subscript s dictates the duration of the observation interval available to the conditional mean estimator. In particular, $\langle X_t\rangle_t$ is referred to as the causal (filtering) conditional mean estimate, and $\langle X_t\rangle_T$ the noncausal (smoothing) conditional mean estimate.

Given a discrete-time process $X_1,X_2,\dots$, an equivalent piecewise constant continuous-time process can be defined as $X_t=X_{\lceil t\rceil}$, $t>0$. Evidently, the output $\{Y_n\}$ of the discrete-time Poisson channel with $\{X_n\}$ as its input can be regarded as increment of the samples of the continuous-time doubly stochastic Poisson process $\{Y_t\}$ with input $\{X_t\}$.

III. Main Results

This section summarizes the fundamental relationships which relate derivatives of the mutual information to the conditional mean estimates described in Section II. For simplicity, we first present the results for the scalar Poisson random transformation. The results are then extended to general continuous-time and discrete-time Poisson channels. Proof of the main results are relegated to Section V.

A. Poisson Random Transformation

Theorem 1: For every $\lambda>0$ and positive random variable X with $\mathsf{E}\{X\log X\}<\infty$, the derivative of the input–output mutual information of the Poisson random transformation $X\mapsto\mathcal{P}(X+\lambda)$ with respect to the dark current is[1]

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} I\left(X;\mathcal{P}(X+\lambda)\right) = \mathsf{E}\left\{\log(X+\lambda)-\log\langle X+\lambda\rangle\right\} \tag{7}$$

$$= \mathsf{E}\left\{\log\frac{X+\lambda}{\langle X+\lambda\rangle}\right\}. \tag{8}$$

Evidently, the mutual information of the Poisson random transformation decreases as the dark current increases. Theorem 1 states that the rate of the decrease is equal to the mean difference between the logarithm of the actual input plus noise and that of its conditional mean estimate (or, the expected value of the logarithm of the ratio of the input plus noise and its estimate). Note that the mean difference in (7) is always negative due to Jensen's inequality. It is assumed that $\lambda>0$ in Theorem 1 because the derivative can be $-\infty$ at $\lambda=0$.

Scaling of the input to a Poisson channel cannot be absorbed into the additive dark current. Interestingly, the derivative of the mutual information with respect to the signal intensity also admits a formula in terms of the conditional mean estimate.

Theorem 2: For every $\alpha>0$, $\lambda\ge 0$ and positive X with $\mathsf{E}\{X\log X\}<\infty$,

$$\frac{\partial}{\partial\alpha} I(X;\mathcal{P}(\alpha X+\lambda)) = \mathsf{E}\left\{X\log\frac{\alpha X+\lambda}{\langle\alpha X+\lambda\rangle}\right\} \tag{9}$$

$$= \frac{1}{\alpha}\mathsf{E}\big\{\psi_\lambda(\alpha X+\lambda)-\psi_\lambda(\langle\alpha X+\lambda\rangle)\big\} \tag{10}$$

where $\psi_\lambda(t)=(t-\lambda)\log t$.

Theorems 1 and 2 are the Poisson counterpart of (1) for Gaussian channels, which relates the derivative of the mutual information to the MMSE achieved by conditional mean estimation.

The sufficient condition $\mathsf{E}\{X\log X\}<\infty$ in the theorems puts a constraint on the tail of the input distribution. Note that the condition also implies that $\mathsf{E}X$ exists, and so do $\mathsf{E}\{\langle X+\lambda\rangle\log\langle X+\lambda\rangle\}$ with $\lambda\ge 0$ and $\mathsf{E}\{\log\langle X+\lambda'\rangle\}$ with $\lambda'>0$ by Jensen's inequality.

Theorems 1 and 2 imply that the mutual information can be expressed as an integral of the estimation errors.

Corollary 1: If $\mathsf{E}\{X\log X\}<\infty$, then

$$I\left(X;\mathcal{P}(X)\right) = -\int_0^\infty \mathsf{E}\left\{\log\frac{\langle X+\lambda\rangle}{X+\lambda}\right\}\mathrm{d}\lambda \tag{11}$$

$$= \int_0^1 \mathsf{E}\left\{X\log\frac{\alpha X}{\langle\alpha X\rangle}\right\}\mathrm{d}\alpha. \tag{12}$$

Conditioned on $X=x$, the probability mass of $\frac{1}{\alpha}\mathcal{P}(\alpha X)$ concentrates at x as $\alpha\to\infty$ since its variance vanishes. In fact, the uncertainty of X given $\mathcal{P}(\alpha X)$ also vanishes as $\alpha\to\infty$. The following result is immediate in view of Corollary 1.

Corollary 2: For every positive discrete random variable X with $\mathsf{E}\{X\log X\}<\infty$,

$$H(X) = \int_0^\infty \mathsf{E}\left\{X\log\frac{\alpha X}{\langle\alpha X\rangle}\right\}\mathrm{d}\alpha. \tag{13}$$

In particular Corollary (2) implies that the right side of (13) is invariant to one-to-one transformations of X. We note that the conditional entropy $H(\mathcal{P}(\alpha X+\lambda)\,|X)$ is related to α, λ as well as the distribution of X. This is in contrast to the case in additive noise channels with noise density function, where the differential entropy $h(\alpha X+N|X)$ is unrelated to the input X and the channel gain α. This fact prevents us from obtaining a simple result for the derivatives of the entropy $H(\mathcal{P}(\alpha X+\lambda))$ using Theorems 1 and 2 like the one

[1] The unit of information measures is nats throughout the paper. All logarithms are natural. By convention, $0\log 0=0$.

in Gaussian channels. Neither can we find a counterpart to the De Bruijn identity in the Gaussian regime [1], [27]. In particular, the Fisher information is not defined for discrete random variables.[2] This is yet another indication that the mutual information–MMSE formula is more fundamental than de Bruijn's identity.

B. Continuous-time Poisson Channels

Consider the continuous-time Poisson channel depicted in Figure 2 where the input and output are X_0^T and $\mathcal{P}_0^T\big(\alpha X_0^T + \lambda\big)$ respectively.

Theorem 3: Suppose the input process satisfies

$$\mathsf{E}\int_0^T |X_t \log X_t| \,\mathrm{d}t < \infty, \tag{14}$$

then for every $\lambda > 0$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda} & I\left(X_0^T; \mathcal{P}_0^T\big(X_0^T + \lambda\big)\right) \\ &= \int_0^T \mathsf{E}\left\{\log\frac{X_t + \lambda}{\langle X_t + \lambda\rangle_T}\right\} \mathrm{d}t \qquad (15)\\ &= \int_0^T \mathsf{E}\left\{\log(X_t+\lambda) - \log\langle X_t+\lambda\rangle_T\right\}\ \mathrm{d}t. \qquad (16)\end{aligned}$$

Theorem 4: Suppose the input process satisfies (14), then

$$\begin{aligned}\frac{\partial}{\partial\alpha} & I\left(X_0^T; \mathcal{P}_0^T\big(\alpha X_0^T + \lambda\big)\right) \\ &= \int_0^T \mathsf{E}\left\{X_t \log\frac{\alpha X_t + \lambda}{\langle \alpha X_t + \lambda\rangle_T}\right\} \mathrm{d}t \qquad (17)\\ &= \frac{1}{\alpha}\int_0^T \mathsf{E}\big\{\psi_\lambda(\alpha X_t+\lambda) - \psi_\lambda\left(\langle \alpha X_t+\lambda\rangle_T\right)\big\}\,\mathrm{d}t \quad (18)\end{aligned}$$

for all $\alpha > 0$ and $\lambda \geq 0$, where $\psi_\lambda(t) = (t-\lambda)\log t$.

Theorems 3 and 4 are the Poisson counterpart of Theorem 6 in [1] for continuous-time Gaussian channels. In particular, the integrands in (16) and (18) are both average errors associated with the *noncausal* conditional mean estimate, which mirror the noncausal MMSE in Gaussian channels.

Theorems 3 and 4 complement the following relationship between the mutual information and the optimal *causal* estimate of the input, which takes a similar form as Duncan's result for Gaussian channels [8].

Theorem 5 (Liptser and Shiryaev [5]): Suppose the input process satisfies (14), then[3]

$$\begin{aligned} I\big(X_0^T;\mathcal{P}_0^T\big(X_0^T\big)\big) &= \mathsf{E}\int_0^T X_t \log X_t - X_t \log\langle X_t\rangle_t \,\mathrm{d}t \quad (19)\\ &= \mathsf{E}\int_0^T X_t\log X_t - \langle X_t\rangle_t \log\langle X_t\rangle_t\,\mathrm{d}t. \quad (20)\end{aligned}$$

Reference [4] states the theorem with a dark current of intensity λ in the Poisson channel, which is straightforward from (20) with X_t replaced by $X_t + \lambda$.

It is interesting to note from Theorems 3–5 that the causal and noncausal estimates are connected through the mutual information.

Corollary 3: For every input satisfying (14),

$$\begin{aligned}\int_0^T & \mathsf{E}\big\{X_t\log X_t - \langle X_t\rangle_t \log\langle X_t\rangle_t\,\big\}\ \mathrm{d}t \\ &= -\int_0^\infty\int_0^T \mathsf{E}\left\{\log\frac{X_t+\lambda}{\langle X_t+\lambda\rangle_T}\right\}\mathrm{d}t\ \mathrm{d}\lambda \qquad (21)\\ &= \int_0^1\int_0^T \mathsf{E}\left\{X_t\log\frac{\alpha X_t}{\langle\alpha X_t\rangle_T}\right\}\mathrm{d}t\,\mathrm{d}\alpha \qquad (22)\\ &= I\left(X_0^T;\mathcal{P}_0^T\big(X_0^T\big)\right). \qquad (23)\end{aligned}$$

Corollary 3 is a straightforward observation in light of the above theorems but it is not known how to establish equalities (21) and (22) from a purely estimation-theoretic viewpoint without resorting to the mutual information.

The mutual information $I\left(X_0^T;\mathcal{P}_0^T\big(\alpha X_0^T+\lambda\big)\right)$ can be regarded as a potential field on Quadrant I of a Cartesian plane, i.e., $\{(\alpha,\lambda)\,|\,\alpha,\lambda>0\}$. Theorems 3 and 4 give the two directional derivatives of the mutual information for all (α,λ), and hence its Taylor series expansion to the first order in scaling and dark current. It is clear that the mutual information vanishes as $\alpha\to 0$ or $\lambda\to\infty$. Thus the mutual information at any (α,λ) pair can be regarded as a path integral of some estimation errors from any $(0,\lambda_0)$ or (α_0,∞) to the point (α,λ), which is also evident from Corollary 3.

Suppose that the input $\{X_t\}$ is a stationary process, then the relationship between the causal and noncausal estimates in Corollary 3 reduces to the following.

Corollary 4: For every stationary input process $\{X_t\}$ with $\mathsf{E}\{X_t\log X_t\}<\infty$,

$$\begin{aligned}\mathsf{E}\big\{X_t\log & X_t - \langle X_t\rangle_t\log\langle X_t\rangle_t\,\big\}\\ &= \int_0^\infty \mathsf{E}\left\{\log\frac{\langle X_t+\lambda\rangle_\infty}{X_t+\lambda}\right\}\mathrm{d}\lambda \qquad (24)\\ &= \int_0^1 \mathsf{E}\left\{X_t\log\frac{\alpha X_t}{\langle\alpha X_t\rangle_\infty}\right\}\mathrm{d}\alpha \qquad (25)\end{aligned}$$

where $\langle X_t\rangle_s = \mathsf{E}\left\{X_t \mid \mathcal{P}_{-\infty}^s\big(X_{-\infty}^s\big)\right\}$ in this corollary.

Note that the random transformation described in Section II-A can be regarded as a special case of the continuous-time channel with a time-invariant input $X_t \equiv X$. It is easy to check that Y_T is a sufficient statistic of Y_0^T for X, so that $\langle X_t\rangle_T = \langle X\rangle$. Theorems 1 and 2 can thus be regarded as simple corollaries of Theorems 3 and 4.

C. Discrete-time Poisson Channels

Consider a discrete-time process $\{X_n, n=1,2,\dots,N\}$ (denoted by $\boldsymbol{X}^N$) and discrete-time doubly Poisson processes derived from it (denoted by $\mathcal{P}\big(\boldsymbol{X}^N\big)$ and the like) as described in Section II-B.

Corollary 5: If $\mathsf{E}\{X_n\log X_n\}<\infty$, $n=1,\dots,N$, then for all $\lambda>0$,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}I\Big(\boldsymbol{X}^N;\mathcal{P}\big(\boldsymbol{X}^N+\lambda\big)\Big) = \sum_{n=1}^N \mathsf{E}\left\{\log\frac{X_n+\lambda}{\langle X_n+\lambda\rangle_N}\right\} \tag{26}$$

[2] The reader is referred to a recent work [28] for a treatment of the *scaled Fisher information* related to Poisson statistics as an alternative.

[3] Subtlety arises with the succinct notation of the form $\langle X\rangle = \mathsf{E}\{X \mid \mathcal{P}(X)\}$, which is an implicit function of the non-unique random transformation $\mathcal{P}(X)$. Naturally, it is understood that all occurrences of $\langle X_t\rangle_t$ are identical in (20). This convention is used throughout the paper.

and for all $\alpha > 0$ and $\lambda \geq 0$,

$$\frac{\partial}{\partial\alpha} I\Big(\boldsymbol{X}^N;\mathcal{P}\big(\alpha\boldsymbol{X}^N+\lambda\big)\Big) = \sum_{n=1}^{N} \mathsf{E}\left\{X_n \log\frac{\alpha X_n+\lambda}{\langle\alpha X_n+\lambda\rangle_N}\right\}. \tag{27}$$

Corollary 5 can be shown using Theorems 3 and 4 once we realize that the discrete-time samples of a continuous-time doubly stochastic Poisson process give sufficient statistics for the piecewise constant input process. In view of Theorem 5, we also have the following inequalities.

Corollary 6: If the input process satisfies $\mathsf{E}\{X_n \log X_n\} < \infty$ for all n, then

$$\begin{aligned}&\sum_{n=1}^{N} \mathsf{E}\left\{X_n \log X_n - \langle X_n\rangle_{n-1} \log \langle X_n\rangle_{n-1}\right\}\\ &\qquad\leq I\Big(\boldsymbol{X}^N;\mathcal{P}\big(\boldsymbol{X}^N\big)\Big) &(28)\\ &\qquad\leq \sum_{n=1}^{N} \mathsf{E}\left\{X_n \log X_n - \langle X_n\rangle_{n} \log \langle X_n\rangle_{n}\right\}. &(29)\end{aligned}$$

The inequalities are due to the discrepancy between $\langle X_t\rangle_{\lfloor t\rfloor}$, $\langle X_t\rangle_t$ and $\langle X_t\rangle_{\lceil t\rceil}$ in general. Interestingly, the input-output mutual information admits bounds based on the causal and one-step prediction estimates in this case.

D. Time Versus Scaling

An immediate consequence of Theorem 4 is the following small scaling expansion of mutual information to the first order.

Corollary 7: For every input process satisfying (14),

$$\begin{aligned}&I\left(X_0^T;\mathcal{P}_0^T\big(\alpha X_0^T\big)\right)\\ &\quad=\alpha\int_0^T \mathsf{E}\left\{X_t\log X_t\right\} - \mathsf{E}X_t\log\mathsf{E}X_t \ \mathrm{d}t + o(\alpha).\end{aligned} \tag{30}$$

Interestingly, the Liptser-Shiryaev formula (19) admits a new intuitive proof using Corollary 7 and the incremental channel idea. For simplicity, assume $\{X_t\}$ to be continuous with probability 1. The mutual information due to an infinitesimal extra observation time interval $(t, t+\alpha)$ is equal to the conditional mutual information of the same Poisson channel during the extra time interval given the past observation,

$$\begin{aligned}&I\left(X_0^{t+\alpha};\mathcal{P}_0^{t+\alpha}\big(X_0^{t+\alpha}\big)\right) - I\left(X_0^t;\mathcal{P}_0^t\big(X_0^t\big)\right)\\ &\qquad= I\left(X_t^{t+\alpha};\mathcal{P}_t^{t+\alpha}\big(X_0^{t+\alpha}\big)\,\middle|\,\mathcal{P}_0^t\big(X_0^t\big)\right)\end{aligned} \tag{31}$$

By expanding the small interval $(t, t+\alpha)$ to unit length, the conditional mutual information in (31) can be regarded as the mutual information of a channel with input attenuated by a factor of α, which, by Corollary 7, is obtained essentially as

$$\alpha\,\mathsf{E}\int_0^1 X_{t+\alpha s}\log X_{t+\alpha s} - \langle X_{t+\alpha s}\rangle_t \log\langle X_{t+\alpha s}\rangle_t \,\mathrm{d}s. \tag{32}$$

Theorem 5 is then established by continuity:

$$\lim_{\alpha\to 0}\langle X_{t+\alpha s}\rangle_t = \langle X_t\rangle_t\,. \tag{33}$$

Note that Theorems 1 and 2 for the Poisson random transformation can also be obtained by considering the special case of time-invariant input in the continuous-time setting. For example, the increase of the mutual information due to scaling $(1+\delta)$ is an outcome of the Liptser-Shiryaev formula. Let $X_t \equiv X$. By (19),

$$\frac{\mathrm{d}}{\mathrm{d}t} I(X;Y_0^t) = \mathsf{E}\left\{X\log X - \langle X\rangle_t \log\langle X\rangle_t\right\} \tag{34}$$

where $\langle X\rangle_t = \mathsf{E}\left\{X \mid Y_0^t\right\}$. Clearly,

$$\begin{aligned}&I\left(X;\mathcal{P}_0^1((1+\delta)X)\right) - I\left(X;\mathcal{P}_0^1(X)\right)\\ &\qquad= \delta\,\mathsf{E}\left\{X\log X - \langle X\rangle\log\langle X\rangle\right\} + o(\delta).\end{aligned} \tag{35}$$

The desired result is obtained once we note that increasing the observation time from 1 to $1+\delta$ is equivalent to keeping the observation time to $[0,1]$ but scaling the intensity of the input by $1+\delta$.

The above argument can be understood as a "time-scaling" transform, i.e., let $Z_t = X_{t/\alpha}$, $\forall t$, then $\mathcal{P}_0^{\alpha T}\big(Z_0^{\alpha T}\big)$ contains the same amount of information about X_0^T as $\mathcal{P}_0^T\big(\alpha X_0^T\big)$ does. Suppose the input is constant over time and there is no dark current. Increasing the scaling is equivalent to increasing the observation time and hence improves the mutual information. The same is true if a dark current is present, since dilating the input by slowing the time also reduces the effective dark current and hence further improves the mutual information.

IV. Numerical Illustration

A. Poisson Random Transformation

Theorem 1 is illustrated in Figure 3(a). The input is a binary random variable X equally likely to be 0 and $A = 2$. Both the input–output mutual information and the expected error in (8) are plotted against the dark current level λ. The mutual information at $\lambda = 0$ is 0.4858 nats [29]. The expected error at $\lambda = 0$ is infinite. Note that the capacity-achieving distribution for the Poisson channel with peak constraint A is binary as long as $A < 3.3679$, while the optimal allocation of probability mass onto the two points (0 and A) depends on the dark current [30].

Figure 3(b) illustrates (9) in Theorem 2 by showing the mutual information together with the estimation-theoretic quantity on the right hand side of (9) as functions of the scaling factor α, where $\lambda = 1$ and binary input equally likely to be 0 and 1 are assumed. It is interesting to note that as $\alpha \to \infty$ the mutual information exhibits the same asymptotic behavior as the mutual information of a Gaussian channel with the same input and SNR equal to α/λ.

B. Optimal Filtering and Optimal Smoothing

As mentioned in Section I, the problem of causal and noncausal estimation based on a doubly stochastic Poisson process observation has been studied since the 1970s (see e.g., [6], [31]–[33]). In [6], Snyder obtained a stochastic differential equation for the posterior probability density of the input process (hence its conditional mean estimate) which involves the Kolmogorov differential operator. In the case where the input process is Markov, the causal estimate can be obtained using Kalman filter type of formulas, since the future estimate is independent of the past observation conditioned on the current estimate. Explicit recursive formulas for obtaining the

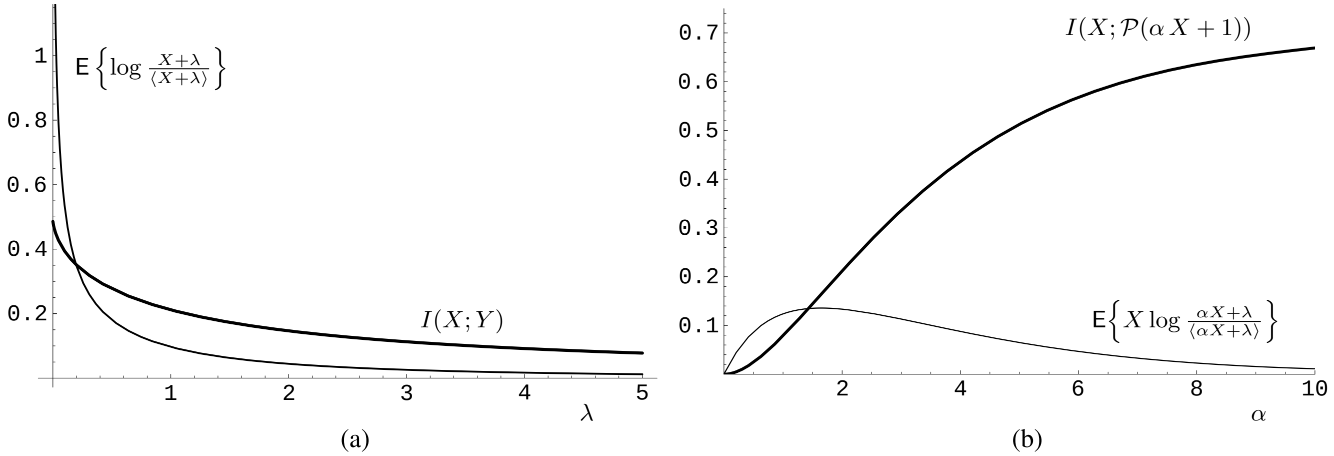

Fig. 3. Theorems 1 and 2 are verified for binary input. (a) The input is equally likely to be 0 and 2. The mutual information (in nats) and the expected error are plotted as a function of the dark current level. Unit scaling is assumed. (b) The mutual information (in nats) and the estimation error as functions of the scaling factor. The input X is binary and equally likely to be 0 and 1. Unit dark current is assumed.

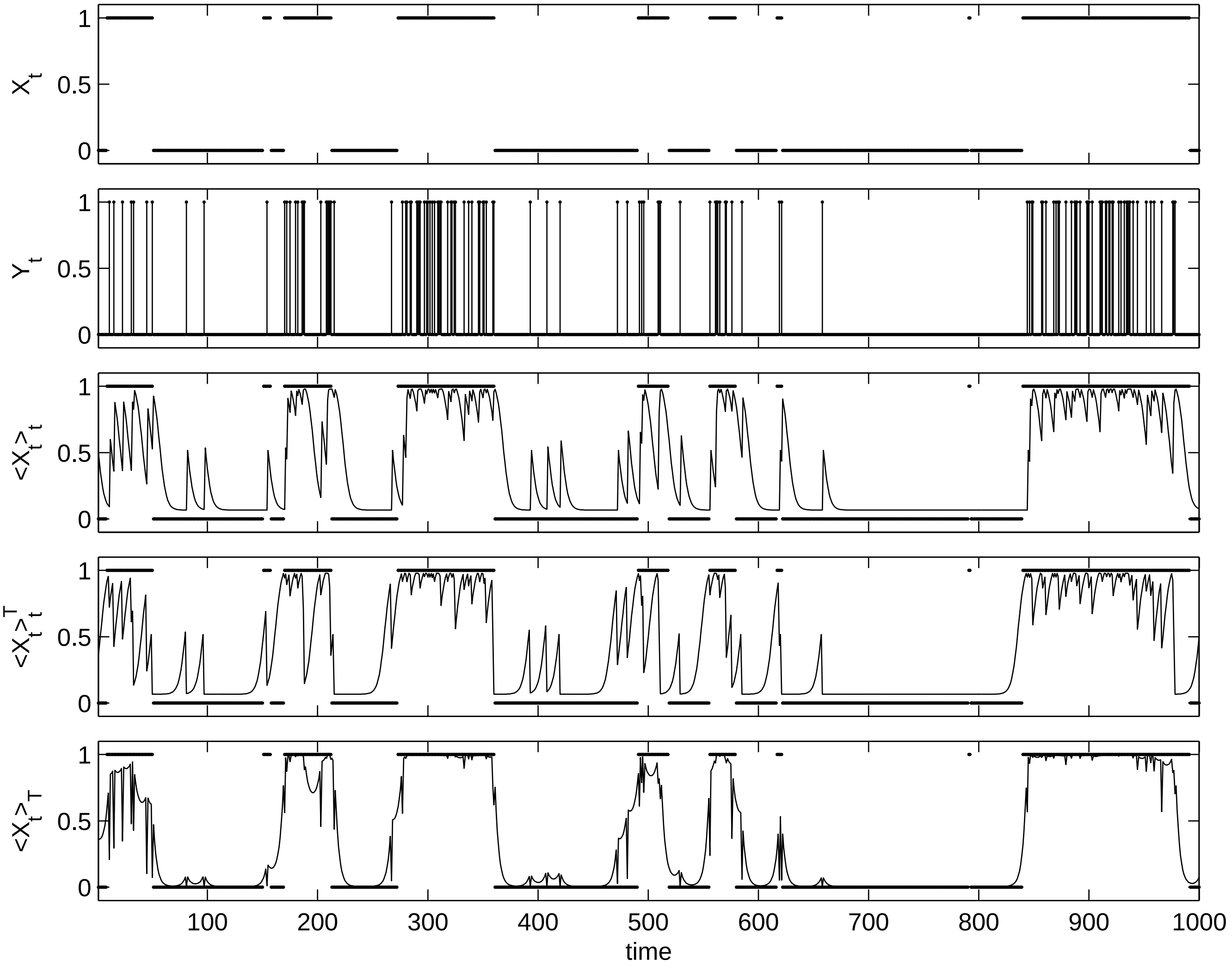

Fig. 4. Plots show the two-state input Markov rate process $\{X_t\}$, the output point process $\{Y_t\}$, as well as the causal, anticausal, and noncausal estimates of X_t based on $\{Y_t\}$, namely, $\langle X_t\rangle_t$, $\langle X_t\rangle_t^T$ and $\langle X_t\rangle_T$ respectively. The input process is also plotted along with each of the estimates for comparison.

causal estimate have also been found in the discrete-time setting [34], [35].

In Figure 4 we plot the conditional mean estimates of a discrete-time random telegraph input process, which takes the value of 0 or 1, and at any time, the probability of a transition is given by a constant p. It is assumed that at each discrete sampling point, one and only one jump may occur with a probability that is linear in the input plus a dark current. This is asymptotically equivalent to observing a Poisson point process in the limit of vanishing sampling interval. The recursive formula for the causal estimate is found in [31]. The anticausal estimate is obtained similarly by simply reversing the time axis. The Markov property of the input process allows the noncausal likelihood ratio to be factored as a product of the causal and anticausal ones:

$$\begin{aligned}&\frac{\mathsf{P}\left\{X_t=1|Y_0^T\right\}}{\mathsf{P}\left\{X_t=0|Y_0^T\right\}}\\&\quad=\frac{\mathsf{P}\left\{X_t=1|Y_0^t\right\}}{\mathsf{P}\left\{X_t=0|Y_0^t\right\}}\frac{\mathsf{P}\left\{X_t=1|Y_t^T\right\}}{\mathsf{P}\left\{X_t=0|Y_t^T\right\}}\frac{\mathsf{P}\left\{X_t=0\right\}}{\mathsf{P}\left\{X_t=1\right\}}.\end{aligned} \tag{36}$$

Therefore, the noncausal conditional mean estimate can be obtained from the causal one and the anticausal one. In the special case of equal prior,

$$\langle X_t\rangle_T=\frac{\langle X_t\rangle_t\langle X_t\rangle_t^T}{(1/2)-\langle X_t\rangle_t-\langle X_t\rangle_t^T+4\langle X_t\rangle_t\langle X_t\rangle_t^T} \tag{37}$$

where $\langle X_t\rangle_t^T$ denotes the anticausal estimate of X_t given Y_t^T.

V. Proof

The proof of the theorems and corollaries given in Section III is essentially a task of estimating the change of the input–output mutual information due to an infinitesimal change in the quality of the Poisson channel. Although the results for the Poisson random transformation are simple corollaries of Theorems 3 and 4, they can be proved directly by working with random variables without recourse to deep results on continuous-time Poisson point processes. Such an exercise elucidates the essence of the proof technique and is included in the following on its own merit before more sophisticated undertakings on the point processes.

A. Poisson Random Transformation: Proof of Theorem 1

It suffices to establish (8) at $\lambda=0$ under the additional assumption that $X>\mu$ for some $\mu>0$, i.e.,

$$I\left(X;\mathcal{P}(X+\lambda)\right)-I\left(X;\mathcal{P}(X)\right)=\lambda\,\mathsf{E}\,\log\frac{X}{\langle X\rangle}+o(\lambda). \tag{38}$$

Equation (38) implies (8) for every $\lambda>0$ because one can always treat $X+\lambda$ as the input, which is bounded away from 0.

Let $N_\lambda=\mathcal{P}(\lambda)$ be independent of both X and $Y_0=\mathcal{P}(X)$. Let $Y_\lambda=Y_0+N_\lambda$. Clearly, Y_λ is a version of $\mathcal{P}(X+\lambda)$. By definition of mutual information,

$$I(X;Y_0)-I(X;Y_\lambda)=\mathsf{E}\left\{L(X,Y_0,Y_\lambda)\right\} \tag{39}$$

where the expectation is over the joint probability distribution of (X,Y_0,Y_λ), and the log-likelihood ratio is

$$L(x,k,l)=\log\frac{P_{Y_0|X}(k|x)}{P_{Y_0}(k)}-\log\frac{P_{Y_\lambda|X}(l|x)}{P_{Y_\lambda}(l)}. \tag{40}$$

Here the conditional Poisson distribution $P_{Y_0|X}$ is given by (2) and its marginal is

$$P_{Y_0}(k)=\frac{1}{k!}\mathsf{E}\left\{X^ke^{-X}\right\},\quad k=0,1,2,\ldots. \tag{41}$$

Also, $P_{Y_\lambda|X}(k|x)$ and $P_{Y_\lambda}(k)$ are similarly defined with x and X replaced by $x+\lambda$ and $X+\lambda$ respectively in (2) and (41). Clearly, the log-likelihood ratio can be written as

$$L(X,Y_0,Y_\lambda)=Y_0\log X-Y_\lambda\log(X+\lambda)+U \tag{42}$$

where

$$U=\log\frac{\mathsf{E}\left\{(X'+\lambda)^{Y_\lambda}e^{-X'}\,\middle|\,Y_\lambda\right\}}{\mathsf{E}\left\{(X')^{Y_0}e^{-X'}\mid Y_0\right\}} \tag{43}$$

where X' is identically distributed as X but independent of Y_0 and Y_λ, i.e., X' and X are i.i.d. given (Y_0,Y_λ). Taking expectation in (42),

$$\mathsf{E}L=\mathsf{E}\left\{X\log X-(X+\lambda)\log(X+\lambda)\right\}+\mathsf{E}U \tag{44}$$

where we replace Y_λ by Y_0+N_λ and write

$$\mathsf{E}U=\mathsf{E}\left\{\log\frac{\mathsf{E}\left\{(X'+\lambda)^{(Y_0+N_\lambda)}e^{-X'}\,\middle|\,Y_0,N_\lambda\right\}}{\mathsf{E}\left\{(X')^{Y_0}e^{-X'}\mid Y_0\right\}}\right\}. \tag{45}$$

Since N_λ is Poisson with mean λ and independent of Y_0,

$$\mathsf{E}U=\mathsf{E}\left\{u_0(Y_0,\lambda)\right\}+\mathsf{E}\left\{u_1(Y_0,\lambda)\right\}+\mathsf{E}\left\{u_2^+(Y_0,\lambda)\right\} \tag{46}$$

where

$$u_n(k,\lambda)=\frac{\lambda^ne^{-\lambda}}{n!}\log\frac{\mathsf{E}\left\{(X'+\lambda)^{k+n}e^{-X'}\right\}}{\mathsf{E}\left\{(X')^ke^{-X'}\right\}} \tag{47}$$

and

$$u_2^+(k,\lambda)=\sum_{n=2}^\infty u_n(k,\lambda). \tag{48}$$

The expectations $\mathsf{E}\left\{u_i(Y_0,\lambda)\right\}$ can be estimated as $\lambda\to0$:

$$\mathsf{E}\left\{u_0(Y_0,\lambda)\right\}=\lambda+o(\lambda) \tag{49}$$

$$\mathsf{E}\left\{u_1(Y_0,\lambda)\right\}=\lambda\mathsf{E}\log\langle X\rangle+o(\lambda) \tag{50}$$

$$\mathsf{E}\left\{u_2^+(Y_0,\lambda)\right\}=o(\lambda) \tag{51}$$

which will be justified shortly. Equation (38) is established using (39), (44), (46), and the estimates (49)–(51),

$$\begin{aligned}&I\big(X;\mathcal{P}(X+\lambda)\big)-I\left(X;\mathcal{P}(X)\right)\\&=\mathsf{E}\left\{(X+\lambda)\log(X+\lambda)-X\log X\right\}\\&\qquad\qquad-\lambda-\lambda\mathsf{E}\log\langle X\rangle+o(\lambda) &(52)\\&=\lambda\mathsf{E}\log\frac{X}{\langle X\rangle}+\mathsf{E}\left\{(X+\lambda)\log\frac{X+\lambda}{X}\right\}-\lambda+o(\lambda) &(53)\\&=\lambda\mathsf{E}\log\frac{X}{\langle X\rangle}+o(\lambda) &(54)\end{aligned}$$

where the final step is due to $X>\mu$ as well as the following.

Lemma 1: For every random variable $X \geq 0$,

$$\lim_{\lambda\to 0} \mathsf{E}\left\{\frac{X}{\lambda}\log\left(1+\frac{\lambda}{X}\right)\right\} = \mathsf{P}\left\{X>0\right\}. \tag{55}$$

Proof: Define $f_\lambda(x) = (x/\lambda)\log(1+\lambda/x)$, $x \in [0,\infty)$, where $f_\lambda(0)=0$. The function is concave and increasing in x, and is dominated by 1 because $\lim_{x\to\infty} f_\lambda(x) = 1$. Indeed,

$$\lim_{\lambda\to 0} f_\lambda(x) = 1_{\{x>0\}}, \quad \forall x \in [0,\infty). \tag{56}$$

Lemma 1 is evident by Lebesgue's Dominated Convergence Theorem [36]. ■

All that remains is to show (49)–(51).

1) Proof of (49)*:* Using the fact that the denominator in (47) is proportional to the unconditional distribution P_{Y_0} given by (41), one proceeds as

$$\mathsf{E}\left\{u_0(Y_0,\lambda)\right\} = \mathsf{E}\left\{\log\mathsf{E}\left\{\left(1+\frac{\lambda}{X}\right)^{Y_0}\,\middle|\,Y_0\right\}\right\} \tag{57}$$

$$\leq \log\mathsf{E}\left\{\left(1+\frac{\lambda}{X}\right)^{Y_0}\right\} \tag{58}$$

$$= \lambda \tag{59}$$

where (58) is due to Jensen's inequality. On the other hand,

$$\mathsf{E}\left\{u_0(Y_0,\lambda)\right\} \geq \mathsf{E}\left\{\log\left(1+\frac{\lambda}{X}\right)^{Y_0}\right\} \tag{60}$$

$$= \mathsf{E}\left\{X\log\left(1+\frac{\lambda}{X}\right)\right\} \tag{61}$$

$$= \lambda + o(\lambda) \tag{62}$$

where (60) follows from Jensen's inequality and (62) is by Lemma 1.

2) Proof of (50)*:* We establish the equivalent result that

$$\frac{e^\lambda}{\lambda}\mathsf{E}\left\{u_1(Y_0,\lambda)\right\} - \mathsf{E}\log\langle X\rangle = \mathsf{E}\left\{g_\lambda(Y_0+1)\right\} \tag{63}$$

vanishes as $\lambda \to 0$ where

$$g_\lambda(y) = \log\frac{\mathsf{E}\left\{(X'+\lambda)^y e^{-X'}\right\}}{\mathsf{E}\left\{X'^y e^{-X'}\right\}}. \tag{64}$$

For every y,

$$0 < g_\lambda(y) = \log\mathsf{E}\left\{\left(1+\frac{\lambda}{X}\right)^y\,\middle|\,Y_0=y\right\} \tag{65}$$

$$< \log\left(1+\frac{\lambda}{\mu}\right)^y. \tag{66}$$

Hence

$$\mathsf{E}\left\{g_\lambda(Y_0+1)\right\} < \mathsf{E}\left\{Y_0+1\right\}\log\left(1+\frac{\lambda}{\mu}\right) \tag{67}$$

which vanishes as $\lambda \to 0$.

3) Proof of (51)*:* Let us define

$$h_\lambda(y) = \mathsf{E}\left\{(X'+\lambda)^y e^{-X'}\right\}. \tag{68}$$

Then

$$\mathsf{E}u_2^+(Y_0,\lambda) = \sum_{n=2}^{\infty}\frac{\lambda^n e^{-\lambda}}{n!}\mathsf{E}\left\{\log\frac{h_\lambda(Y_0+n)}{h_0(Y_0)}\right\}. \tag{69}$$

Note that the term $h_0(Y_0)$ in (69) is not dependent on n and can be ignored because it contributes $o(\lambda)$ to the sum. In order to establish (51), we first show the finiteness of $\mathsf{E}\left\{\log h_\lambda(Y_0+n)\right\}$ for all $n = 0,1,\ldots$ and $\lambda \geq 0$. Since $X' > \mu$,

$$\mathsf{E}\{\log h_\lambda(Y_0+n)\} \geq \mathsf{E}\left\{\log\mathsf{E}\left\{\mu^{Y_0+n}e^{-X'}\,\middle|\,Y_0\right\}\right\} \tag{70}$$

$$= \mathsf{E}\left\{(Y_0+n)\log\mu + \log\mathsf{E}e^{-X'}\right\} \tag{71}$$

$$\geq (\mathsf{E}X+n)\log\mu - \mathsf{E}X. \tag{72}$$

Meanwhile, it is enough to consider $\lambda < \mu$ so that $X'+\lambda \leq 2X' \leq X'e$, and

$$\mathsf{E}\{\,\log h_\lambda(Y_0+n)\}$$

$$\leq \mathsf{E}\left\{\log\mathsf{E}\left\{(X'e)^{Y_0+n}e^{-X'}\,\middle|\,Y_0\right\}\right\} \tag{73}$$

$$= \mathsf{E}\left\{(Y_0+n) + \log\mathsf{E}\left\{(X')^{Y_0+n}e^{-X'}\,\middle|\,Y_0\right\}\right\} \tag{74}$$

$$\leq \mathsf{E}\left\{\log(n+Y_0)^{(n+Y_0)}\right\} \tag{75}$$

where the final step is due to the fact that $x^m e^{-x}$ achieves its maximum at $x=m$. The right hand side of (75) is finite because of the following auxiliary results.

Lemma 2: Let $N_\lambda = \mathcal{P}(\lambda)$. Then $\mathsf{E}\left\{N_\lambda\log N_\lambda\right\} < \lambda\log(1+\lambda)$.

Proof: Using the distribution of N_λ, we have

$$\mathsf{E}\left\{N_\lambda\log N_\lambda\right\} = \sum_{n=2}^{\infty}\frac{\lambda^n e^{-\lambda}}{n!}n\log n \tag{76}$$

$$= \lambda\mathsf{E}\left\{\log(1+N_\lambda)\right\} \tag{77}$$

$$< \lambda\log(1+\lambda) \tag{78}$$

by Jensen's inequality, ■

Corollary 8: Let $Y = \mathcal{P}(X)$. If $\mathsf{E}\left\{X\log X\right\} < \infty$, then $\mathsf{E}\left\{Y\log Y\right\} < \infty$.

By concavity of $(y+n)\log(y+n) - y\log y$, (75) can be upper bounded

$$\begin{aligned}&\mathsf{E}\{(n+Y_0)\log(n+Y_0)\} - \mathsf{E}\left\{Y_0\log Y_0\right\}\\ &\qquad\leq n\log(n+\mathsf{E}X) + \mathsf{E}X\log\left(1+\frac{n}{\mathsf{E}X}\right).\end{aligned} \tag{79}$$

Therefore,

$$\mathsf{E}\log h_\lambda(Y_0+n) \leq \mathsf{E}\left\{Y_0\log Y_0\right\} + n\log(n+\mathsf{E}X) + n. \tag{80}$$

It is clear from (72) and (80) that $\mathsf{E}\{\log h_\lambda(Y_0+n)\}$ is asymptotically upper bounded by $c_1 n\log n$ and lower bounded by $c_2 n$ where c_1 and c_2 are real-valued constants. Noticing that (see also (78))

$$\sum_{n=2}^{\infty}\frac{\lambda^n e^{-\lambda}}{n!}n \leq \sum_{n=2}^{\infty}\frac{\lambda^n e^{-\lambda}}{n!}n\log n = o(\lambda) \tag{81}$$

we obtain (51) from (69).

B. Poisson Random Transformation: Proof of Theorem 2

We first prove the following result, which is equivalent to Theorem 2 with $\lambda = 0$ and $\alpha = 1$.

Lemma 3: Suppose $\mathsf{E}\{X \log X\} < \infty$. As $\delta \to 0$,

$$\begin{aligned} I(X;&\mathcal{P}((1+\delta)X) - I(X;\mathcal{P}(X)) \\ &= \delta\,\mathsf{E}\{X \log X - \langle X \rangle \log \langle X \rangle\} + o(\delta). \end{aligned} \tag{82}$$

Proof: Consider first the case $\delta \to 0^+$. Let $Y = \mathcal{P}(X)$ and $Z = \mathcal{P}(\delta X)$ be independent conditioned on X. Let also $Y_\delta = Y + Z$. Then, the left hand side of (82) is

$$\begin{aligned} &I(X;Y_\delta) - I(X;Y) \\ &= \mathsf{E}\left\{\log \frac{p_{Y_\delta|X}(Y_\delta|X)}{p_{Y_\delta}(Y_\delta)} - \log \frac{p_{Y|X}(Y|X)}{p_Y(Y)}\right\} && (83) \\ &= \mathsf{E}\left\{\log \frac{(Y_\delta!)\,p_{Y_\delta|X}(Y_\delta|X)}{(Y!)\,p_{Y|X}(Y|X)} - \log \frac{(Y_\delta!)\,p_{Y_\delta}(Y_\delta)}{(Y!)\,p_Y(Y)}\right\} && (84) \\ &= \mathsf{E}\left\{Z \log X - \delta X - \log \frac{\mathsf{E}\{(X')^{Y_\delta} e^{-(1+\delta)X'} \mid Y_\delta\}}{\mathsf{E}\{(X')^{Y} e^{-X'} \mid Y\}}\right\} && (85) \end{aligned}$$

where X' takes the same distribution as X but independent of Y and Z, namely X' and X are i.i.d. conditioned on (Y, Z). Note that (85) remains true if Y_δ is replaced by $Y + Z$. Let us also introduce a random variable $\bar{X}$ which is i.i.d. with X and independent of all other variables conditioned on Y. The distribution of $\bar{X}$ given Y is $P_{X|Y}$. We rewrite (85) as

$$\begin{aligned} &I(X;Y_\delta) - I(X;Y) \\ &\quad = \delta\mathsf{E}\{X \log X - X\} - \mathsf{E}\left\{\log \mathsf{E}\{\bar{X}^Z e^{-\delta \bar{X}} | Y, Z\}\right\} \end{aligned} \tag{86}$$

where the change of variable X' to $\bar{X}$ uses the fact that the denominator in (86) is proportional to $P_Y(Y)$. Noting that Z is Poisson conditioned on X, the expectation over Z in (86) can be written as

$$\mathsf{E}\left\{\log \mathsf{E}\left\{\bar{X}^Z e^{-\delta \bar{X}} \,\middle|\, Y, Z\right\}\right\} = \delta \sum_{n=0}^{\infty} v_n(\delta) \tag{87}$$

where we define

$$\begin{aligned} v_n(\delta) &= \frac{1}{\delta}\mathsf{E}\left\{\frac{(\delta X)^n e^{-\delta X}}{n!} \log \mathsf{E}\left\{X^n e^{-\delta X} \,\middle|\, Y\right\}\right\} && (88) \\ &= \frac{\delta^{n-1}}{n!}\mathsf{E}\left\{\mathsf{E}\{X^n e^{-\delta X}|Y\} \log \mathsf{E}\left\{X^n e^{-\delta X}|Y\right\}\right\} && (89) \end{aligned}$$

where we have replaced $\bar{X}$ by X because they are i.i.d. conditioned on Y.

It is straightforward establish Lemma 3 if one can show

$$\lim_{\delta \to 0} v_0(\delta) \;=\; -\mathsf{E}X, \tag{90}$$

$$\lim_{\delta \to 0} v_1(\delta) \;=\; \mathsf{E}\{\langle X \rangle \log \langle X \rangle\}, \quad \text{and} \tag{91}$$

$$\textstyle\lim_{\delta \to 0} \sum\limits_{n=2}^{\infty} v_n(\delta) \;=\; 0. \tag{92}$$

1) Proof of (90)*:* For all $\delta > 0$,

$$\begin{aligned} v_0(\delta) &= (1/\delta)\,\mathsf{E}\left\{\mathsf{E}\left\{e^{-\delta X} \,\middle|\, Y\right\} \log \mathsf{E}\left\{e^{-\delta X} \,\middle|\, Y\right\}\right\} && (93) \\ &\geq (1/\delta)\,\mathsf{E}\left\{e^{-\delta X}\right\} \log \mathsf{E}\left\{e^{-\delta X}\right\} && (94) \\ &\geq (1/\delta)\,e^{-\delta \mathsf{E}X} \log e^{-\delta \mathsf{E}X} && (95) \\ &\geq -\mathsf{E}X && (96) \end{aligned}$$

where we have used Jensen's inequality repeatedly to arrive at the first two inequalities. Meanwhile, also by Jensen's inequality,

$$\begin{aligned} v_0(\delta) + \mathsf{E}X &\leq (1/\delta)\,\mathsf{E}\left\{e^{-\delta X} \log e^{-\delta X}\right\} + \mathsf{E}X && (97) \\ &= \mathsf{E}\left\{(1 - e^{-\delta X})X\right\} && (98) \end{aligned}$$

which vanishes with δ by the dominated convergence theorem.

2) Proof of (91)*:* For every $\delta > 0$,

$$v_1(\delta) \leq \mathsf{E}\{\psi_0(\langle X \rangle)\} < \infty \tag{99}$$

where $\psi_0(x) = x \log x$, because $\mathsf{E}\left\{X e^{-\delta X} \,\middle|\, Y\right\} \leq \mathsf{E}\{X \mid Y\}$ for every Y. By the Monotone Convergence Theorem [36],

$$\lim_{\delta \to 0} v_1(\delta) = \mathsf{E}\left\{\psi_0\left(\lim_{\delta \to 0} \mathsf{E}\left\{X e^{-\delta X} \,\middle|\, Y\right\}\right)\right\}. \tag{100}$$

A second use of the convergence theorem yields (91).

3) Proof of (92)*:* Using the fact that $t \log t \geq -1/e$ for all $t \geq 0$, one has

$$\begin{aligned} \lim_{\delta \to 0} \sum_{n=2}^{\infty} v_n(\delta) &\geq -\frac{1}{e\delta} \sum_{n=2}^{\infty} \frac{\delta^n}{n!} && (101) \\ &= -(e^\delta - 1 - \delta)/(e\delta) && (102) \end{aligned}$$

which vanishes as $\delta \to 0^+$. On the other hand,

$$\begin{aligned} &\lim_{\delta \to 0} \sum_{n=2}^{\infty} v_n(\delta) \\ &\leq \sum_{n=2}^{\infty} \frac{\delta^{n-1}}{n!}\mathsf{E}\left\{X^n e^{-\delta X} \log(X^n e^{-\delta X})\right\} && (103) \\ &= \mathsf{E}\left\{\left(X \log X \sum_{n=1}^{\infty} \frac{\delta^n X^n}{n!} - X \sum_{n=2}^{\infty} \frac{\delta^n X^n}{n!}\right) e^{-\delta X}\right\} && (104) \\ &= \mathsf{E}\left\{(X \log X - X)(1 - e^{-\delta X}) + \delta X^2 e^{-\delta X}\right\} && (105) \\ &\to 0 && (106) \end{aligned}$$

as $\delta \to 0^+$ by the dominated convergence theorem.

The case of $\delta \to 0^-$ can be similarly proved by letting $Y_\delta = \mathcal{P}((1+\delta)X)$, $Z = \mathcal{P}(-\delta X)$, $Y = Y_\delta + Z$, and essentially repeating the above. ■

Based on Theorem 1 and Lemma 3, we can obtain the following first order Taylor series expansion.

Lemma 4: For every P_X with $\mathsf{E}\{X \log X\} < \infty$ and every $\delta \to 0$ and $\lambda \to 0^+$,

$$\begin{aligned} &I(X;\mathcal{P}((1+\delta)X + \lambda)) \\ &\qquad = I(X;\mathcal{P}(X)) + \lambda\,\mathsf{E}\{\log X - \log \langle X \rangle\} + o(\lambda) \\ &\qquad\quad + \delta\,\mathsf{E}\{X \log X - \langle X \rangle \log \langle X \rangle\} + o(\delta). \end{aligned} \tag{107}$$

Proof: Applying (38) and Lemma 3, the change of the mutual information due to δ and λ can be written as

$$\begin{aligned} &I(X;\mathcal{P}((1+\delta)X + \lambda)) - I(X;\mathcal{P}(X)) \\ &\quad = I(X;\mathcal{P}((1+\delta)X + \lambda)) - I(X;\mathcal{P}((1+\delta)X)) \\ &\qquad + I(X;\mathcal{P}((1+\delta)X)) - I(X;\mathcal{P}(X)) && (108) \\ &\quad = \lambda\,\mathsf{E}\{\log((1+\delta)X) - \log \langle (1+\delta)X \rangle\} \\ &\qquad + \delta\,\mathsf{E}\{X \log X - \langle X \rangle \log \langle X \rangle\} + o(\lambda) + o(\delta). && (109) \end{aligned}$$

It remains to express $\mathsf{E}\log\langle(1+\delta)X\rangle$ in terms of $\mathsf{E}\log\langle X\rangle$ and the conditional mean estimates. Let $Y=\mathcal{P}(X)$ and $Z=\mathcal{P}(\delta X)$ be independent conditioned on X. Then

$$\frac{\langle(1+\delta)X\rangle}{(1+\delta)}=\frac{\mathsf{E}\left\{(X')^{Y+Z+1}e^{(1+\delta)X'}\,\middle|\,Y,Z\right\}}{\mathsf{E}\left\{(X')^{Y+Z}e^{(1+\delta)X'}\,\middle|\,Y,Z\right\}}. \tag{110}$$

Using similar techniques as in the proof of Theorem 2, it can be show that replacing Z in (110) by the indicator $1_{\{Z>0\}}$ introduces only second order difference. Some algebra yields

$$\begin{aligned}\mathsf{E}\{\log\langle(1+\delta)X\rangle\}&=\mathsf{E}\{\log((1+\delta)\langle X\rangle)\}\\&+\delta\mathsf{E}\left\{X\log\frac{\langle X^2\rangle}{\langle X\rangle^2}+X-\frac{\langle X^2\rangle}{\langle X\rangle}\right\}+o(\delta).\end{aligned}\tag{111}$$

Plugging (111) into (109) proves Lemma 4 since also $|\delta\lambda|\leq\max\left(|\delta|^2,|\lambda|^2\right)=o(\delta)+o(\lambda)$. ■

The proof for Theorem 2 can now be completed as follows.

Proof: Note that for every $\alpha>0$ and $\lambda\geq 0$,

$$\begin{aligned}&I(X;\mathcal{P}((\alpha+\epsilon)X+\lambda))-I(X;\mathcal{P}(\alpha X+\lambda))\\&=I\left(X;\mathcal{P}\left((1+\epsilon/\alpha)\left(\alpha X+\lambda\right)-\epsilon\lambda/\alpha\right)\right)\\&\quad-I(X;\mathcal{P}(\alpha X+\lambda)).\end{aligned}\tag{112}$$

In particular, the equality holds for $\lambda=0$ because of continuity of $I(X;\mathcal{P}(\alpha X+\lambda))$ at $\lambda=0^+$. For $\epsilon\to 0$, applying Lemma 4 to the right hand side of (112) establishes

$$\begin{aligned}\frac{\partial}{\partial\alpha}I(X;\mathcal{P}(\alpha X+\lambda))=\mathsf{E}\Big\{&X\log(\alpha X+\lambda)\\&-\frac{1}{\alpha}\left(\langle\alpha X+\lambda\rangle-\lambda\right)\log\langle\alpha X+\lambda\rangle\Big\}\end{aligned}\tag{113}$$

which is equivalent to (10). ■

The expectation in (10) is positive due to Jensen's inequality since $\psi_\lambda(t)=(t-\lambda)\log t$ is a convex function on $(0,\infty)$. Note that (9) does not apply directly to the special case of $\alpha=0$, which describes the mutual information that corresponds to a very small input. In particular, at $\alpha=0$ but $\lambda>0$, the derivative of $I(X;\mathcal{P}(\alpha X+\lambda))$ with respect to α is found to be 0 by taking the limit $\alpha\to 0^+$. At the point $\alpha=\lambda=0$, the derivative can be obtained from Theorem 2 by noting that

$$\lim_{\alpha\to 0}\langle\alpha X\rangle/\alpha=\mathsf{E}X. \tag{114}$$

Therefore,

Corollary 9: For every P_X satisfying $\mathsf{E}\{X\log X\}<\infty$,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}I\left(X;\mathcal{P}(\alpha X)\right)\Big|_{\alpha=0}=\mathsf{E}\{X\log X\}-\mathsf{E}X\log\mathsf{E}X. \tag{115}$$

C. Continuous-time Poisson Channels: Proof of Theorem 3

Theorem 3 can also be proved by examining the likelihood ratio as in the proof of Theorems 1 and 2. Note that since probability density functions are not defined for continuous-time processes, one has to resort to Radon-Nikodym derivatives.

Consider a continuous-time Poisson channel with dark current λ and scaling factor $\alpha=1$ as the one depicted by Figure 2. Let P^{ZY} denote the joint probability measure of the input $\{Z_t\}$ and the output $Y_0^T=\mathcal{P}_0^T\big(Z_0^T+\lambda\big)$. Let Q^{ZY} denote the product measure of $\{Z_t\}$ and an independent process $\mathcal{P}_0^T(\lambda)$, which corresponds to the output of the Poisson channel with zero input. The following Radon-Nikodym derivative between the two probability measures [4, p. 180], [5, p. 343],

$$\log\frac{\mathrm{d}P^{ZY}}{\mathrm{d}Q^{ZY}}=\int_0^T\log\left(1+\frac{Z_t}{\lambda}\right)\,\mathrm{d}Y_t-\int_0^T Z_t\,\mathrm{d}t \tag{116}$$

is the key to the information–estimation relationships as well as the hypothesis testing problem (of whether the input is $\{Z_t\}$ or zero). An illustration of the two probability measures corresponding two hypothesized inputs is shown in Figure 5. Note that the absolute continuity and existence of (116) is guaranteed under an even weaker condition than (14) (see [5] and references therein).

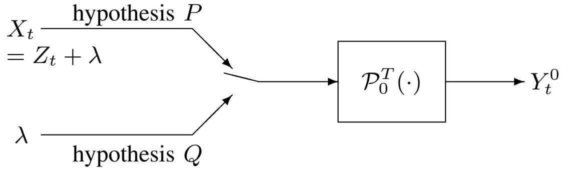

Fig. 5. Illustration of a continuous-time Poisson channel with two possible inputs corresponding to two hypotheses (or probability measures) P and Q. Under measure P, the output process Y^0 is caused by input process $X=Z+\lambda$. Under measure Q, Y^0 is caused by constant input λ and hence independent of X.

As a convention, restriction of a probability measure to the sub-σ-algebra generated by a process is denoted by superscript, e.g., P^Y denotes P restricted to the σ-algebra generated by $\{Y_t\}$. It is then clear that $P^Z=Q^Z$ and that independence implies that $Q^{YZ}=Q^Y\times Q^Z$. From (116) one can also derive the Radon-Nikodym derivative when only the observation $\{Y_t\}$ is accessible, which is reminiscent of the "estimator-correlator" principle found in Gaussian channels. That is, the resulting log-likelihood ratio $\log\mathrm{d}P^Y/\mathrm{d}Q^Y$ is given by (116) only with Z_t replaced by the causal estimate $\mathsf{E}\left\{Z_t\mid Y_0^t\right\}$ which denotes conditional expectation of Z_t with respect to measure P given Y_0^t.

Consider now a continuous-time Poisson channel with no dark current. Let P denote the probability measure under which $\{Y_t^0\}$ and $\{Y_t^\epsilon\}$ are conditional Poisson processes with intensity X_t and $X_t+\epsilon$ respectively. Here $Y_t^\epsilon=Y_t^0+N_t^\epsilon$ where $\{N_t^\epsilon\}$ is a point process with constant intensity ϵ independent of $\{X_t\}$ and $\{Y_t^0\}$. For ease of notation, let X, Y^0, Y^ϵ and N^ϵ denote the processes $\{X_t\}$, $\{Y_t^0\}$, $\{Y_t^\epsilon\}$ and $\{N_t^\epsilon\}$ in $[0,T]$ respectively. Let $\mathsf{E}\{\cdot\}$ denote expectation with respect to measure P.

Lemma 5: Let $\lambda>0$. For every input X_0^T satisfying (14) and $X_t>\lambda$, $\forall t\in[0,T]$,

$$I\left(X;Y^0\right)-I\left(X;Y^\epsilon\right)=\epsilon\int_0^T\mathsf{E}\log\frac{\langle X_t\rangle_T}{X_t}\,\mathrm{d}t+o(\epsilon). \tag{117}$$

Note that replacing X_t in (117) by $X_t+\lambda$ proves Theorem 3.

Proof: The mutual information is by definition

$$I\left(X;Y^0\right)=\mathsf{E}\left\{\log\frac{\mathrm{d}P^{XY^0}}{\mathrm{d}P^X\,\mathrm{d}P^{Y^0}}\right\}. \tag{118}$$

Hence,

$$I\left(X;Y^0\right) - I\left(X;Y^\epsilon\right) = \mathsf{E}\left\{\log \frac{\mathrm{d}P^{XY^0}\,\mathrm{d}P^{Y^\epsilon}}{\mathrm{d}P^{XY^\epsilon}\,\mathrm{d}P^{Y^0}}\right\}. \tag{119}$$

Let us introduce probability measure Q which differs with P in the following manner: The process $Y^0 = Y^\epsilon$ has constant intensity λ (instead of X_t or $X_t + \epsilon$) under Q. In particular, X, Y^0 and N^ϵ are independent under Q. Thus $Q^{XYN} = Q^X \times Q^Y \times Q^N$ and (119) can be rewritten as

$$\mathsf{E}\left\{\log \frac{\mathrm{d}P^{XY^0}\,\mathrm{d}Q^{XY^\epsilon}}{\mathrm{d}Q^{XY^0}\,\mathrm{d}P^{XY^\epsilon}}\right\} + \mathsf{E}\left\{\log \frac{\mathrm{d}P^{Y^\epsilon}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{Y^\epsilon}\,\mathrm{d}P^{Y^0}}\right\}. \tag{120}$$

In the following, we evaluate the two expectations in (120) separately.

In view of the description of probability measures leading to (116), Q and P can also be regarded as probability measures where the output Y is caused by zero input and input $X_t - \lambda$ respectively, with a dark current λ in both cases. Using formula (116) with $Z_t = X_t + \epsilon - \lambda$, the Radon-Nikodym derivative between the joint probability measures of (X, Y^ϵ) under hypotheses Q and P is

$$\log \frac{\mathrm{d}P^{XY^\epsilon}}{\mathrm{d}Q^{XY^\epsilon}} = \int_0^T \log \frac{X_t + \epsilon}{\lambda}\,\mathrm{d}Y_t^\epsilon - \int_0^T X_t - \lambda + \epsilon\,\mathrm{d}t. \tag{121}$$

By the same principle, equations (121) literally hold with ϵ replaced by 0. Note that by assumption $X_t > \lambda$, $\forall t$. Using (121) for all ϵ, the first expectation in (120) is evaluated as

$$\begin{aligned}
&\mathsf{E}\left\{\log \frac{\mathrm{d}P^{XY^0}\,\mathrm{d}Q^{XY^\epsilon}}{\mathrm{d}Q^{XY^0}\,\mathrm{d}P^{XY^\epsilon}}\right\} \\
&= \mathsf{E}\left\{\epsilon T - \int_0^T \log \frac{X_t + \epsilon}{\lambda}\,\mathrm{d}N_t^\epsilon - \int_0^T \log \frac{X_t + \epsilon}{X_t}\,\mathrm{d}Y_t^0\right\} && (122) \\
&= \mathsf{E}\left\{\epsilon T - \epsilon \int_0^T \log \frac{X_t + \epsilon}{\lambda}\,\mathrm{d}t - \int_0^T X_t \log \frac{X_t + \epsilon}{X_t}\,\mathrm{d}t\right\} && (123) \\
&= -\epsilon\,\mathsf{E}\left\{\int_0^T \log \frac{X_t}{\lambda}\,\mathrm{d}t\right\} + o(\epsilon) && (124)
\end{aligned}$$

where the final equality holds by the monotone convergence theorem.

The Radon-Nikodym derivative between the marginal probability measures of Y^ϵ under P and Q can be obtained as

$$\begin{aligned}
\frac{\mathrm{d}P^{Y^\epsilon}}{\mathrm{d}Q^{Y^\epsilon}} &= \mathsf{E}_Q\left\{\left.\frac{\mathrm{d}Q^{XY^\epsilon}}{\mathrm{d}P^{XY^\epsilon}}\,\right|\, Y^\epsilon\right\} && (125) \\
&= \mathsf{E}_Q\left\{\left.\frac{\mathrm{d}Q^{XY^\epsilon}}{\mathrm{d}P^{XY^\epsilon}}\,\right|\, Y^0, N^\epsilon\right\} && (126)
\end{aligned}$$

where Y^ϵ can be replaced by (Y^0, N^ϵ) in the final equality because X is independent of (Y^0, N^ϵ) under Q. Using (126),

$$\begin{aligned}
\frac{\mathrm{d}P^{Y^\epsilon}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{Y^\epsilon}\,\mathrm{d}P^{Y^0}} &= \mathsf{E}_Q\left\{\left.\frac{\mathrm{d}P^{XY^\epsilon}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{XY^\epsilon}\,\mathrm{d}P^{Y^0}}\,\right|\, Y^0, N^\epsilon\right\} && (127) \\
&= \mathsf{E}_Q\left\{\left.\frac{\mathrm{d}P^{XY^\epsilon}\,\mathrm{d}Q^{XY^0}\,\mathrm{d}P^{X|Y^0}}{\mathrm{d}Q^{XY^\epsilon}\,\mathrm{d}P^{XY^0}\,\mathrm{d}Q^X}\,\right|\, Y^0, N^\epsilon\right\} && (128) \\
&= \mathsf{E}\left\{\left.\frac{\mathrm{d}P^{XY^\epsilon}\,\mathrm{d}Q^{XY^0}}{\mathrm{d}Q^{XY^\epsilon}\,\mathrm{d}P^{XY^0}}\,\right|\, Y^0, N^\epsilon\right\} && (129)
\end{aligned}$$

where we have used $P^{XY} = P^Y \times P^{X|Y}$ and independence of X and N^ϵ in (128) and changed the underlying measure of the expectation in (129). Thus the likelihood ratios of the marginals have been expressed in terms of those of the joint measures given by (121). Plugging (121) into (129), the second expectation in (120) is expressed as

$$\mathsf{E}\left\{\log \frac{\mathrm{d}P^{Y^\epsilon}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{Y^\epsilon}\,\mathrm{d}P^{Y^0}}\right\} = \mathsf{E}\left\{\log u\left(Y^0, N^\epsilon\right)\right\} \tag{130}$$

where

$$\begin{aligned}
u\left(Y^0, N^\epsilon\right) = \mathsf{E}\Bigg\{\exp\Bigg[&\int_0^T \log \frac{X_t + \epsilon}{X_t}\,\mathrm{d}Y_t^0 - \epsilon T \\
&+ \int_0^T \log \frac{X_t + \epsilon}{\lambda}\,\mathrm{d}N_t^\epsilon\Bigg]\Bigg|\, Y^0, N^\epsilon\Bigg\}.
\end{aligned} \tag{131}$$

Note that N^ϵ is a Poisson process independent of X and Y^0. Moreover, N^ϵ remains all zero with probability $1 - \epsilon T + o(\epsilon)$, contains one jump with probability $\epsilon T + o(\epsilon)$, and two or more jumps with probability $o(\epsilon)$. Using similar techniques as in the proof of Theorem 1 in the Poisson random transformation case (see Section V-A), we can show that

$$\mathsf{E}\left\{\log u\left(Y^0, N^\epsilon\right) \,\middle|\, N_T^\epsilon \neq 1\right\} = o(\epsilon). \tag{132}$$

Therefore, as far as (130) is concerned, it suffices to evaluate the expectation conditioned on that N^ϵ contains one jump at a random time S, which is uniformly distributed in $[0, T]$. Evidently in this case,

$$\int_0^T \log \frac{X_t + \epsilon}{\lambda}\,\mathrm{d}N_t^\epsilon = \log \frac{X_S + \epsilon}{\lambda} \tag{133}$$

and thus $u\left(Y^0, N^\epsilon\right)$ is rewritten as

$$\mathsf{E}\left\{\frac{X_S + \epsilon}{\lambda}\,\exp\left[\int_0^T \log \frac{X_t + \epsilon}{X_t}\,\mathrm{d}Y_t^0 - \epsilon T\right]\Bigg|\, Y^0\right\}. \tag{134}$$

Using the dominated convergence theorem, we can show that

$$\begin{aligned}
&\mathsf{E}\left\{\log u\left(Y^0, N^\epsilon\right) \,\middle|\, N_T^\epsilon = 1\right\} \\
&\quad= \mathsf{E}\left\{\log \mathsf{E}\left\{\left.\frac{X_S + \epsilon}{\lambda}\,\right|\, Y^0\right\}\right\} + o(\epsilon)
\end{aligned} \tag{135}$$

By (130), (132) and (135),

$$\begin{aligned}
\mathsf{E}\left\{\log \frac{\mathrm{d}P^{Y^\epsilon}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{Y^\epsilon}\,\mathrm{d}P^{Y^0}}\right\} &= \epsilon T\,\mathsf{E}\left\{\log \frac{\langle X_S\rangle_T}{\lambda}\right\} + o(\epsilon) && (136) \\
&= \epsilon\,\mathsf{E}\int_0^T \log \frac{\langle X_t\rangle_T}{\lambda}\,\mathrm{d}t + o(\epsilon). && (137)
\end{aligned}$$

Lemma 5 is thus established by (120), (124) and (137). ■

D. Continuous-time Poisson Channels: Proof of Theorem 4

Let $\delta > 0$. Consider doubly Poisson point process $\left\{Y_t^0\right\}$ with rate X_t, and $Y_t^\delta = Y_t^0 + Z_t$, $\forall t$, where $\{Z_t\}$ is a point process conditionally independent of $\left\{Y_t^0\right\}$ with rate δX_t. Clearly, $\left\{Y_t^\delta\right\}$ is a point process with rate $(1 + \delta)X_t$. Let X, Z, Y^0, Y^δ denote the respective processes $\{X_t\}$, $\{Z_t\}$, $\left\{Y_t^0\right\}$ and $\left\{Y_t^\delta\right\}$ in $[0, T]$.

Lemma 6: For every input X_0^T satisfying (14),

$$\lim_{\delta\to 0^+} \frac{I\left(X;Y^0\right) - I\left(X;Y^\delta\right)}{\delta} = \mathsf{E}\int_0^T X_t \log\frac{\langle X_t\rangle_T}{X_t}\,\mathrm{d}t. \quad (138)$$

Proof: Assume for now that X_t is bounded away from 0, i.e., $X_t > \lambda$, $\forall t \in [0,T]$, for some $\lambda > 0$. This constraint will eventually be removed. Let P denote the underlying probability measure of (138), where X and Y are the input and output of the continuous-time Poisson channel. We also introduce a measure Q, under which $Y^0 = Y^\delta$ is a Poisson point process of fixed intensity λ (or, it is caused by zero input with dark current intensity λ). In view of the proof of Lemma 5, the mutual information difference in (138) admits literally the same expression (120) only with ϵ replaced by δ.

Formula (116) leads to

$$\log\frac{\mathrm{d}P^{XY^\delta}}{\mathrm{d}Q^{XY^\delta}} = \int_0^T \log\frac{(1+\delta)X_t}{\lambda}\,\mathrm{d}Y_t^\delta - \int_0^T (1+\delta)X_t - \lambda\;\mathrm{d}t \quad (139)$$

which also holds for $\delta = 0$. Clearly,

$$\begin{aligned}
&\mathsf{E}\left\{\log\frac{\mathrm{d}P^{XY^\delta}\,\mathrm{d}Q^{XY^0}}{\mathrm{d}Q^{XY^\delta}\,\mathrm{d}P^{XY^0}}\right\}\\
&= \mathsf{E}\Bigg\{\int_0^T \log\frac{X_t}{\lambda}\,\mathrm{d}\left(Y_t^\delta - Y_t^0\right)\\
&\qquad + \int_0^T \log(1+\delta)\,\mathrm{d}Y_t^\delta - \delta\int_0^T X_t\,\mathrm{d}t\Bigg\} \quad (140)\\
&= -\delta\,\mathsf{E}\left\{\int_0^T X_t\log\frac{X_t}{\lambda}\;\mathrm{d}t\right\} + o(\delta). \quad (141)
\end{aligned}$$

It is important to note that the small adjustment $o(\delta)$ in (141) does not depend on λ, or in other words, the convergence in δ is uniform for all λ.

Using the same techniques leading to (126), we express the likelihood ratio between the marginals as

$$\begin{aligned}
\frac{\mathrm{d}P^{Y^\delta}}{\mathrm{d}Q^{Y^\delta}} &= \mathsf{E}_Q\left\{\frac{\mathrm{d}Q^{XY^\delta}}{\mathrm{d}P^{XY^\delta}}\,\middle|\,Y^\delta\right\} \quad (142)\\
&= \mathsf{E}_Q\left\{\frac{\mathrm{d}Q^{XY^\delta}}{\mathrm{d}P^{XY^\delta}}\,\middle|\,Y^0, Z\right\} \quad (143)
\end{aligned}$$

where Y^δ can be replaced by $\left(Y^0, N^\epsilon\right)$ in (143) because X is independent of $\left(Y^0, Z\right)$ under Q. Furthermore, using similar techniques leading to (129), we obtain

$$\begin{aligned}
\frac{\mathrm{d}P^{Y^\delta}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{Y^\delta}\,\mathrm{d}P^{Y^0}} &= \mathsf{E}_Q\left\{\frac{\mathrm{d}P^{XY^\delta}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{XY^\delta}\,\mathrm{d}P^{Y^0}}\,\middle|\,Y^0, Z\right\} \quad (144)\\
&= \mathsf{E}_R\left\{\frac{\mathrm{d}P^{XY^\delta}\,\mathrm{d}Q^{XY^0}}{\mathrm{d}Q^{XY^\delta}\,\mathrm{d}P^{XY^0}}\,\middle|\,Y^0, Z\right\} \quad (145)
\end{aligned}$$

where R is a probability measure of X conditioned on the filtration of $\left(Y^0, Z\right)$ defined in the following manner: Under measure R, the distribution of X (conditioned on the filtration of Y^0) is identical to the conditional distribution of X under P (i.e., $R_{X|Y^0} = P_{X|Y^0}$), while X is conditionally independent of Z given Y^0. Note that, unlike in the proof of Lemma 5, the dependence of X and Z distinguishes the measure R from P. By (139) and (145), we have

$$\mathsf{E}\left\{\log\frac{\mathrm{d}P^{Y^\delta}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{Y^\delta}\,\mathrm{d}P^{Y^0}}\right\} = \mathsf{E}\Big\{\log v_\lambda\left(Y^0, Z\right)\Big\} \quad (146)$$

where

$$\begin{aligned}
v_\lambda\big(Y^0, Z\big) = \mathsf{E}_R\Bigg\{&\exp\Bigg[\int_0^T \log(1+\delta)\,\mathrm{d}Y_t^0\\
&- \delta\int_0^T X_t\,\mathrm{d}t + \int_0^T \log\frac{(1+\delta)X_t}{\lambda}\,\mathrm{d}Z_t\Bigg]\Bigg|\,Y^0, Z\Bigg\}. \quad (147)
\end{aligned}$$

In order to isolate λ, let $w\left(Y^0, Z\right) = v_\lambda\left(Y^0, Z\right)\lambda^{Z_T}$ which is not related to λ. Since X is conditionally independent of Z given Y^0, (146) can be rewritten as a sum over the all possible number of jumps contained in $\{Z_t\}$ in $[0,T]$:

$$\sum_{n=0}^{\infty}\mathsf{E}\left\{\mathsf{P}\left(Z_T = n|Y^0\right)\,\log w(Y^0, Z)\right\} - \mathsf{E}Z_T\log\lambda \quad (148)$$

where $\mathsf{P}\left(Z_T = n|Y^0\right) = q^n e^{-q}/n!$ with

$$q = \delta\int_0^T \mathsf{E}\left\{X_t \mid Y_0^T\right\}\;\mathrm{d}t. \quad (149)$$

Using similar bounding techniques as used in the proof of Theorem 2, it can be shown that

$$\begin{aligned}
\mathsf{E}\left\{\mathsf{P}\left(Z_T = 0|Y^0\right)\,\log w(Y^0, Z)\right\} &= o(\delta), \quad (150)\\
\sum_{n=2}^{\infty}\mathsf{E}\left\{\mathsf{P}\left(Z_T = n|Y^0\right)\,\log w(Y^0, Z)\right\} &= o(\delta). \quad (151)
\end{aligned}$$

Furthermore, consider the case where $\{Z_t\}$ contains a single jump at $S \in [0,T]$. Conditioned on Y^0, the density of S is

$$p_S(t) = \delta\,\langle X_t\rangle_T/q, \quad t \in [0,T]. \quad (152)$$

We can write

$$\begin{aligned}
&\mathsf{E}\left\{\mathsf{P}\left(Z_T = 1|Y^0\right)\,\log w(Y^0, Z)\right\}\\
&= \mathsf{E}\Bigg\{qe^{-q}\,\log\mathsf{E}_R\Bigg\{(1+\delta)X_S\\
&\qquad \times\exp\left[Y_T^0\log(1+\delta) - \delta\int_0^T X_t\,\mathrm{d}t\right]\Bigg\}\Bigg|\,Y, S\Bigg\} \quad (153)\\
&= \mathsf{E}\left\{qe^{-q}\,\log\mathsf{E}_R\left\{(1+\delta)X_S \mid Y, S\right\}\right\} + o(\delta) \quad (154)\\
&= \delta\,\mathsf{E}\left\{\int_0^T \langle X_t\rangle\log\langle X_t\rangle\;\mathrm{d}t\right\} + o(\delta) \quad (155)
\end{aligned}$$

where the bounding techniques for arriving at (154) follows the principles developed in the proof of Theorem 2. The final form (155) is obtained by writing the expectation over S in (154) as an integral over the density (152).

By (146), (148), (150), (151) and (154),

$$\mathsf{E}\log\frac{\mathrm{d}P^{Y^\delta}\,\mathrm{d}Q^{Y^0}}{\mathrm{d}Q^{Y^\delta}\,\mathrm{d}P^{Y^0}} = \delta\,\mathsf{E}\int_0^T \langle X_t\rangle\log\frac{\langle X_t\rangle_T}{\lambda} + o(\delta). \quad (156)$$

Putting (141) and (156) together, we have shown the desired result (138). Finally, note that the small terms $o(\delta)$ in (150)–(154) are not dependent on λ. Hence λ, the lower bound for X_t, can be sent to 0. Lemma 6 holds as long as (14) is satisfied. Furthermore, the above arguments essentially also apply to the case $\delta \to 0^-$ by reversing the roles of Y^0 and Y^δ. ■

VI. Concluding Remarks

New relationships between the input–output mutual information and conditional mean estimation in Poisson channels have been identified in this paper. In particular, the derivatives of the mutual information with respect to the intensity of the dark current (resp. input scaling) is expressed in the expected difference in the logarithm function $\log x$ (resp. $x \log x$) evaluated at the actual input and the same function evaluated at its conditional mean estimate. The general relationships hold for the discrete-time and continuous-time Poisson channels as well as for the Poisson random transformation.

We expect that, by replacing the (nonlinear) conditional mean estimate with linear estimates in the information–estimation formulas, bounds can be developed for the mutual information, which is often hard to compute otherwise. Moreover, linear filtering of doubly Poisson and Gaussian processes are tightly connected (see e.g., [37]), which allows one to tap into the rich estimation theory in the Gaussian regime.

Underlying the analysis and results in both [1] and this paper are common properties of Gaussian and Poisson distributions, namely, 1) infinite divisibility of Gaussian and Poisson distributions; and 2) independent increments of Gaussian and Poisson processes. In fact, the entire class of processes with independent increments can be characterized by not much more than a mixture of Wiener and Poisson processes [38]. It is even speculated in [1] that information and estimation satisfy similar relationships as long as the output has independent increments conditioned on the input.

Acknowledgment

The authors are indebted to the anonymous referee for the valuable comments and meticulous scrutiny of the proof of the results, which improved the presentation significantly. D. Guo would also like to thank Jun Luo for useful discussion.

References

[1] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1261–1282, Apr. 2005.

[2] S. Verdú, "Poisson communication theory," *Invited talk in the International Technion Communication Day in honor of Israel Bar-David*, Mar. 1999. [Online] Available at: http://www.princeton.edu/∼verdu.

[3] I. Bar-David, "Communication under the Poisson regime," *IEEE Trans. Inform. Theory*, vol. 15, pp. 31–37, Jan. 1969.

[4] P. Brémaud, *Point Processes and Queues*. New York: Springer-Verlag, 1981.

[5] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes II: Applications*. Springer, 2nd ed., 2001.

[6] D. L. Snyder, "Filtering and detection for doubly stochastic Poisson processes," *IEEE Trans. Inform. Theory*, vol. 18, pp. 91–102, Jan. 1972.

[7] T. Kailath, "A general likelihood-ratio formula for random signals in Gaussian noise," *IEEE Trans. Inform. Theory*, vol. 15, pp. 350–361, May 1969.

[8] T. E. Duncan, "On the calculation of mutual information," *SIAM Journal of Applied Mathematics*, vol. 19, pp. 215–220, July 1970.

[9] Y. M. Kabanov, "The capacity of a channel of the Poisson type," *Theory of Probability and Its Applications*, vol. 23, pp. 143–147, 1978.

[10] M. H. A. Davis, "Capacity and cutoff rate for Poisson-type channels," *IEEE Trans. Inform. Theory*, vol. 26, pp. 710–715, Nov. 1980.

[11] A. D. Wyner, "Capacity and error exponent for the direct detection photon channel—Part I and Part II," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1449–1471, Nov. 1988.

[12] S. Verdú, "The exponential distribution in information theory," *Problems of Information Transmission*, vol. 32, pp. 86–95, 1996. Translated from Problemy Peredachi Informatsii, Vol. 32, No. 1, pp. 100–111, January–March, 1996.

[13] A. S. Bedekar, "On the information about message arrival times required for in-order decoding," in *Proc. IEEE Int. Symp. Inform. Theory*, p. 227, Washington, DC, USA, 2001.

[14] S. Shamai and A. Lapidoth, "Bounds on the capacity of a spectrally constrained Poisson channel," *IEEE Trans. Inform. Theory*, vol. 39, pp. 19–29, Jan. 1993.

[15] A. Lapidoth and S. M. Moser, "Bounds on the capacity of the discrete-time Poisson channel," in *Proceedings 41st Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Oct. 2003.

[16] A. Lapidoth and S. Shamai, "The Poisson multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 44, pp. 488–501, Mar. 1998.

[17] A. Lapidoth, I. E. Telatar, and R. Urbanke, "On wide-band broadcast channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 3250–3258, Dec. 2003.

[18] S. M. Haas and J. H. Shapiro, "Capacity of wireless optical communications," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 1346–1357, Oct. 2003.

[19] K. Chakraborty and P. Narayan, "The Poisson fading channel," *IEEE Trans. Inform. Theory*, vol. 53, pp. 2349–2364, July 2007.

[20] S. I. Bross and S. Shamai, "Capacity and decoding rules for the Poisson arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 49, pp. 3076–3093, Nov. 2003.

[21] P. Narayan and D. L. Snyder, "Signal set design for band-limited memoryless multiple-access channels with soft decision demodulation," *IEEE Trans. Inform. Theory*, vol. 33, pp. 539–556, July 1987.

[22] S. Verdú, "Multiple access channels with point-process observations: Optimum demodulation," *IEEE Trans. Inform. Theory*, vol. 32, pp. 642–651, Sept. 1986.

[23] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," in *Proc. IEEE Inform. Theory Workshop*, pp. 265–270, San Antonio, TX, USA, 2004.

[24] D. Guo, S. Shamai, and S. Verdú, "Additive non-Gaussian noise channels: Mutual information and conditional mean estimation," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sept. 2005.

[25] D. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Trans. Inform. Theory*, vol. 53, pp. 453–470, Feb. 2007.

[26] C. Méasson, R. Urbanke, A. Montanari, and T. Richardson, "Life above threshold: From list decoding to area theorem and MSE," in *Proc. IEEE Inform. Theory Workshop*, San Antonio, TX, USA, 2004.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2nd ed., 2006.

[28] M. Madiman, O. Johnson, and I. Kontoyiannis, "Fisher information, compound Poisson approximation, and the Poisson channel," in *Proc. IEEE Int. Symp. Inform. Theory*, Nice, France, 2007.

[29] G. L. Fillmore and G. Lachs, "Information rates for photocount detection systems," *IEEE Trans. Inform. Theory*, vol. 15, pp. 465–468, July 1969.

[30] S. Shamai, "Capacity of a pulse amplitude modulated direct detection photon channel," *IEE Proceedings*, vol. 137, pp. 424–430, Dec. 1990.

[31] A. Segall, "Recursive estimation from discrete-time point processes," *IEEE Trans. Inform. Theory*, vol. 22, pp. 422–431, July 1976.

[32] J. H. Van Schuppen, "Filtering, prediction and smoothing for counting process observations, a martingale approach," *SIAM J. Appl. Math.*, vol. 32, pp. 552–570, May 1977.

[33] R. K. Boel and V. E. Benes, "Recursive nonlinear estimation of a diffusion acting as the rate of an observed Poisson process," *IEEE Trans. Inform. Theory*, vol. 26, pp. 561–575, Sept. 1980.

[34] V. Krishnamurthy and R. J. Elliott, "Filters for estimating Markov modulated Poisson processes and image-based tracking," *Automatica*, vol. 33, no. 5, pp. 821–833, 1997.

[35] J. Evans and V. Krishnamurthy, "Exact filters for doubly stochastic AR models with conditionally Poisson observations," *IEEE Trans. Inform. Theory*, vol. 44, pp. 794–798, Apr. 1999.

[36] H. L. Royden, *Real Analysis*. Macmillan, 1988.

[37] D. L. Snyner and M. I. Miller, *Random Point Processes in Time and Space*. New York: Springer-Verlag, 2nd ed., 1991.

[38] J. Bertoin, *Lévy Processes*. Cambridge University Press, 1996.

PLACE
PHOTO
HERE

Dongning Guo (S'97-M'05) joined the Department of Electrical Engineering & Computer Science at Northwestern University an Assistant Professor in 2004. He received the Ph.D. and M.Sc. degrees from Princeton University, the M.Eng. degree from the National University of Singapore and the B.Eng. degree from University of Science & Technology of China. He was an R&D Engineer in the Centre for Wireless Communications (now the Institute for Infocom Research), Singapore from 1998 to 1999. He was a Visiting Professor at Norwegian University of Science and Technology in summer 2006. He received the Huber and Suhner Best Student Paper Award in the International Zurich Seminar on Broadband Communications in 2000 and the National Science Foundation Faculty Early Career Development (CAREER) Award in 2007. His research interests are in information theory, communications and networking.

PLACE
PHOTO
HERE

Shlomo Shamai (Shitz) (S'80-M'82-SM'89-F'94) received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 1975, 1981, and 1986, respectively. During 1975–1985, he was with the Communications Research Labs in the capacity of a Senior Research Engineer. Since 1986, he has been with the Department of Electrical Engineering, Technion-Israel Institute of Technology, where he is now the William Fondiller Professor of Telecommunications. His research interests cover a wide spectrum of topics in information theory and statistical communications. Dr. Shamai (Shitz) is a member of the Union Radio Scientifique Internationale (URSI). He is the recipient of the 1999 van der Pol Gold Medal of URSI, and a corecipient of the 2000 IEEE Donald G. Fink Prize Paper Award, the 2003 and 2004 joint IEEE IT/COM Societies Paper Award, and the 2007 IEEE Information Theory Society Paper Award. He is also the recipient of the 1985 Alon Grant for distinguished young scientists and the 2000 Technion Henry Taub Prize for Excellence in Research. He has served as Associate Editor for Shannon Theory for IEEE TRANSACTIONS ON INFORMATION THEORY, and also serves on the Board of Governors of the IEEE Information Theory Society.

PLACE
PHOTO
HERE

Sergio Verdú (S'80-M'84-SM'88-F'93) received the Telecommunications Engineering degree from the Universitat Politècnica de Barcelona in 1980, and the Ph.D. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1984. Since 1984 he has been a member of the faculty of Princeton University. Sergio Verdú is the recipient of the 2007 Claude E. Shannon Award, the 2008 Richard W. Hamming Medal, and a member of the National Academy of Engineering. He is a recipient of several paper awards from the IEEE: the 1992 Donald Fink Paper Award, the 1998 Information Theory Outstanding Paper Award, an Information Theory Golden Jubilee Paper Award, the 2002 Leonard Abraham Prize Award, and the 2006 Joint Communications/Information Theory Paper Award. In 1998, Cambridge University Press published his book Multiuser Detection, for which he received the 2000 Frederick E. Terman Award from the American Society for Engineering Education. In 2005, he received a Doctorate Honoris Causa from the Universitat Politècnica de Catalunya. Sergio Verdú served as President of the IEEE Information Theory Society in 1997. He is currently Editor-in-Chief of Foundations and Trends in Communications and Information Theory.