Mutual Information Saddle Points in Channels of Exponential Family Type

Todd P. Coleman
ECE Department, University of Illinois
Urbana, IL 61801, USA
Email: colemant@illinois.edu

Maxim Raginsky
ECE Department, Duke University
Durham, NC 27708, USA
Email: m.raginsky@duke.edu

***Abstract*—This paper extends our prior work on "E-type" (exponential family type) channels. The channels considered here have transition kernels induced by an exponential family with a two-component sufficient statistic composed of an input-output distortion function and an output cost function. We demonstrate the existence of a mutual information saddle point in any E-type channel for which there exists a source distribution such that the induced output distribution is maximum-entropy under an output cost constraint. For additive-noise E-type channels, we provide necessary and sufficient conditions on the existence of saddle points which coincide with convolution divisibility of the additive noise law. This machinery generalizes many well-known saddle-point, capacity, and rate-distortion theorems, including those for the additive Gaussian and exponential-noise channels, and leads to a saddle point result on the non-additive exponential server timing channel, which appears to be new.**

I. Introduction

It is well known that the mutual information game

$$\max_X \min_W I(X; X+W) \tag{1}$$

where X and W are independent, variance-constrained random variables, has a Gaussian saddle point (cf. [1, Problem 9.21]). This fact leads to simple and elegant derivations of the Gaussian channel capacity and rate-distortion function.

Recent information-theoretic studies of timing-based communication via queues [2]–[4] have uncovered some remarkable similarities between Gaussian channels and queuing timing channels. In particular, the *exponential server timing channel* (ESTC) – the 'simplest' queuing timing channel, with exponentially distributed service times – has infinite memory, yet its capacity expression is remarkably similar to that of the Gaussian channel. Both have optimal input and output distributions that are maximum-entropy; for the ESTC, they correspond to Poisson processes. Just like the Gaussian channel, the ESTC is an "entropy-increasing operator" [5, Sec. 2], [6, Thm 1]. Moreover, Verdú [3] (cf. also [2, Thm. 3]) has shown that the game (1), where X and W are independent *nonnegative*, *mean-constrained* random variables, also has a saddle point, where the worst-case noise is exponential, while the best-case input is a mixture of a point mass and an exponential. This fact plays a prominent role in the seminal work of Anantharam and Verdú [2] on the ESTC.

These similarities prompted us to attempt a general description of source-channel pairs for which saddle-point, capacity, and rate-distortion results emerge in a coherent framework. In our earlier work [7], we described a class of additive-noise "E-type" channels, where the noise law is a member of an exponential family whose sufficient statistic is a difference distortion function. In this paper, we generalize the work in [7] to include channel laws induced by an exponential family with a two-component sufficient statistic consisting of a general (non-difference) *input-output* distortion function and an *output* cost function. A key benefit of this framework is that it covers the additive case, yet can also describe non-additive channels, such as the ESTC, when the latter is taken from a point process viewpoint [4]. Our main result states that a given E-type channel admits a mutual information saddle point if there exists a source distribution such that the induced output distribution is maximum-entropy with respect to the output cost function. Specializing this general result to additive-noise E-type channels, we provide *necessary and sufficient conditions* on the existence of saddle points, which coincide with convolution divisibility of the noise law. We demonstrate how this machinery subsumes known saddle-point, capacity, and rate-distortion theorems, including additive Gaussian- and exponential-noise channels. We also give a saddle-point result for the ESTC in terms of point process distributions, which, to the best of our knowledge, appears to be new.

II. Notation and preliminaries

Throughout this paper, we consider an input measurable space $(\mathsf{X}, \mathcal{F}_\mathsf{X})$, an output measurable space $(\mathsf{Y}, \mathcal{F}_\mathsf{Y})$, and a fixed σ-finite *reference measure* μ on $(\mathsf{Y}, \mathcal{F}_\mathsf{Y})$. The space of probability measures on, say, $(\mathsf{X}, \mathcal{F}_\mathsf{X})$ is denoted by $\mathcal{M}(\mathsf{X})$.

A *stochastic (also: Markov, transition) kernel* on $\mathcal{F}_\mathsf{Y} \times \mathsf{X}$ is a mapping $P_{Y|X}(\cdot|\cdot) : \mathcal{F}_\mathsf{Y} \times \mathsf{X} \to [0,1]$, such that $P_{Y|X}(\cdot|x) \in \mathcal{M}(\mathsf{Y})$ for every $x \in \mathsf{X}$ and $P_{Y|X}(A|\cdot)$ is $\mathcal{B}(\mathbb{R})/\mathcal{F}_\mathsf{X}$-measurable for every $A \in \mathcal{F}_\mathsf{Y}$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on the reals. We denote by $\mathcal{K}(\mathsf{X} \to \mathsf{Y})$ the space of all such stochastic kernels. Given any $P_X \in \mathcal{M}(\mathsf{X})$ and $P_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y})$, we will let $P_X \otimes P_{Y|X}$ denote the induced joint distribution on $(\mathsf{X} \times \mathsf{Y}, \mathcal{F}_\mathsf{X} \times \mathcal{F}_\mathsf{Y})$, defined by

$$P_X \otimes P_{Y|X}(A \times B) \triangleq \int_A P_{Y|X}(B|x) P_X(dx)$$

for all $A \in \mathcal{F}_\mathsf{X}, B \in \mathcal{F}_\mathsf{Y}$, and by $P_X P_{Y|X}$ the marginal distribution $P_Y \in \mathcal{M}(\mathsf{Y})$, defined by

$$P_Y(B) \triangleq P_X \otimes P_{Y|X}(\mathsf{X} \times B) = \int_{\mathsf{X}} P_{Y|X}(B|x) P_X(dx).$$

Given two probability measures P, Q, we define the *divergence*

$$D\left(P \| Q\right) \triangleq \begin{cases} \mathbb{E}_P[\log(dP/dQ)], & \text{if } P \ll Q \\ +\infty, & \text{otherwise} \end{cases}$$

where $\ll$ denotes absolute continuity of measures; all logarithms are natural logarithms. Given two transition kernels $P_{Y|X}, P'_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y})$, along with a distribution $P_X \in \mathcal{M}(\mathsf{X})$, such that $P_{Y|X=x} \ll P'_{Y|X=x}$ P_X-a.s., where $P_{Y|X=x} \triangleq P_{Y|X}(\cdot|x)$ and $P'_{Y|X=x} \triangleq P'_{Y|X}(\cdot|x)$, the *conditional divergence* is given by

$$D\left(P_{Y|X} \| P'_{Y|X} | P_X\right) \triangleq \int_{\mathsf{X}} D\left(P_{Y|X=x} \| P'_{Y|X=x}\right) P_X(dx) \tag{2}$$

For $P_X \in \mathcal{M}(\mathsf{X}), P_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y})$, the *mutual information* is

$$I(P_X, P_{Y|X}) \triangleq D\left(P_{Y|X} \| P_Y | P_X\right). \tag{3}$$

We will also denote $I(P_X, P_{Y|X})$ by $I(X;Y)$ whenever P_X and $P_{Y|X}$ are clear from context. Using (3) and (2), we can express $I(P_X, P_{Y|X})$ as

$$I(X;Y) = D\left(P_{Y|X} \| Q | P_X\right) - D\left(P_Y \| Q\right) \tag{4}$$

for any $Q \in \mathcal{M}(\mathsf{Y})$ satisfying $P_Y \ll Q$. Another expression for the mutual information is

$$\begin{aligned} I(P_X, P_{Y|X}) &= D\left(P_{Y|X} \| P'_{Y|X} | P_X\right) \\ &\quad - D\left(P_{Y|X} P_X \| P'_{Y|X} P_X\right) \\ &\quad + \mathbb{E}_{P_X \otimes P_{Y|X}}\left[\log \frac{dP'_{Y|X}}{d(P'_{Y|X} P_X)}\right], \end{aligned} \tag{5}$$

which holds for any $P'_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y})$ satisfying $P_{Y|X=x} \ll P'_{Y|X=x}$ for P_X-almost all x.

III. Channels of exponential family type

We are interested in channels (stochastic kernels) that satisfy an extremal property called the *Gibbs variational principle (GVP)*. In statistical physics and theory of large deviations, GVP pertains to a duality between entropy and energy [8].

Fix two measurable maps $\rho_1 : \mathsf{X} \times \mathsf{Y} \to \overline{\mathbb{R}}_+$ and $\rho_2 : \mathsf{Y} \to \overline{\mathbb{R}}_+$, where $\overline{\mathbb{R}}_+ \triangleq \mathbb{R}_+ \cup \{+\infty\}$, and define $\rho : \mathsf{X} \times \mathsf{Y} \to \overline{\mathbb{R}}^2_+$ via $\rho(x,y) \triangleq [\rho_1(x,y), \rho_2(y)]^T$. Conceptually, $\rho_1(x,y)$ is an input-output *distortion function*, while $\rho_2(y)$ is an output *cost function*. For every $\beta = [\beta_1, \beta_2]^T \in \mathbb{R}^2$ and every $x \in \mathsf{X}$, define the *partition function*

$$Z_1(\beta|x) \triangleq \int_{\mathsf{Y}} e^{-\beta^T \rho(x,y)} \mu(dy).$$

We also assume that there exists a one-parameter curve $\beta = [\beta_1, \beta_2(\beta_1)]^T$ in $\mathbb{R}_+ \times \mathbb{R}$, where β_1 varies over an open set $T \subset \mathbb{R}_+$, such that $Z_1(\beta|x)$ is finite and *independent of* $x \in \mathsf{X}$ along this curve. From now on, we will assume that β is a point on this curve and write $Z_1(\beta)$ instead of $Z_1(\beta|x)$. Define $\overline{P}^{\beta}_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y})$ via the Radon–Nikodym derivative

$$f^{\beta}_{Y|X}(y|x) \triangleq \frac{d\overline{P}^{\beta}_{Y|X=x}}{d\mu}(y) = \frac{e^{-\beta^T \rho(x,y)}}{Z_1(\beta)} \tag{6}$$

w.r.t. the reference measure μ. The GVP for $\overline{P}^{\beta}_{Y|X}$ can be phrased as follows: for each $x \in \mathsf{X}$, the probability measure $\overline{P}^{\beta}_{Y|X=x}$ minimizes the functional

$$F_{x,\beta}(P_Y) \triangleq \mathbb{E}_{P_Y}\left[\log \frac{dP_Y}{d\mu}\right] + \mathbb{E}_{P_Y}[\beta^T \rho(x,Y)]$$

over all $P_Y \in \mathcal{M}(\mathsf{Y})$. Indeed, a simple calculation shows that

$$F_{x,\beta}(P_Y) = D\left(P_Y \| \overline{P}^{\beta}_{Y|X=x}\right) - \log Z_1(\beta);$$

since the divergence is nonnegative, we see that $F_{x,\beta}(P_Y)$ is minimized by $\overline{P}^{\beta}_{Y|X=x}$. Note that $-\mathbb{E}_{P_Y}[\log(dP_Y/d\mu)]$ is the differential entropy $h(Y)$ of P_Y w.r.t. μ. Moreover, ρ_1 and ρ_2 can be viewed as *energy functions* (or potentials), where $\rho_1(x,\cdot)$ is that part of the energy cost which depends on the input $x \in \mathsf{X}$. Thus, $F_{x,\beta}(P_Y)$ quantifies the trade-off between the entropy and the (x-dependent) energy of P_Y, with β serving as a pair of Lagrange multipliers that control the relative weights of the entropy and the energy terms.

We will also be interested in situations when the *output* measure of a channel satisfies a Gibbs principle. To this end, for each scalar $\gamma \in \mathbb{R}$ define the second partition function as

$$Z_2(\gamma) \triangleq \int_{\mathsf{Y}} e^{-\gamma \rho_2(y)} \mu(dy).$$

We will deal only with those γ for which $Z_2(\gamma)$ is finite. Note that if $Z_1([0,\gamma]^T) < \infty$, then $Z_2(\gamma) = Z_1([0,\gamma]^T)$. Defining $\overline{P}^{\gamma}_Y \in \mathcal{M}(\mathsf{Y})$ via the density

$$f^{\gamma}_Y(y) \triangleq \frac{d\overline{P}^{\gamma}_Y}{d\mu}(y) = \frac{e^{-\gamma \rho_2(y)}}{Z_2(\gamma)}, \tag{7}$$

we can state the GVP: $\overline{P}^{\gamma}_Y$ minimizes

$$F_{\gamma}(P_Y) \triangleq \mathbb{E}\left[\log \frac{dP_Y}{d\mu}\right] + \gamma \mathbb{E}[\rho_2(Y)].$$

over all $P_Y \in \mathcal{M}(\mathsf{Y})$.

A. Additive-noise channels

Additive noise channels studied in our earlier paper [7] are a special case. To keep things simple, we will consider the case $(\mathsf{X}, \mathcal{F}_{\mathsf{X}}) = (\mathsf{Y}, \mathcal{F}_{\mathsf{Y}}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and μ the Lebesgue measure, although everything easily carries over to any Abelian topological group with a Polish topology and the associated Haar measure. We fix a measurable function $\rho_0 : \mathbb{R} \to \overline{\mathbb{R}}_+$, take

$$\rho_1(x,y) \triangleq \rho_0(y-x), \qquad \rho_2(y) \triangleq \rho_1(0,y) \equiv \rho_0(y)$$

and restrict the channel parameter β to have zero second coordinate: $\beta = [\beta_1, 0]^T$, where $\beta_1 \in \mathbb{R}_+$. From this point on, when dealing with additive channels, we will write β instead of β_1 (since β_2 is identically zero) and ρ instead of ρ_0. Then, from (6),

$$\frac{d\overline{P}^{\beta}_{Y|X=x}}{d\mu}(y) = \frac{e^{-\beta\rho(y-x)}}{Z_1(\beta)}. \tag{8}$$

In other words, $\overline{P}^{\beta}_{Y|X=x}$ is an additive-noise channel that effects a random transformation $Y = X + W$, where the additive noise W has law $\overline{P}_W \equiv \overline{P}^{\beta}_W$ described by the density

$$f^{\beta}_W(x) \triangleq \frac{d\overline{P}^{\beta}_W}{d\mu}(x) = \frac{e^{-\beta\rho(x)}}{Z_1(\beta)}, \tag{9}$$

and is statistically independent of X.

IV. Saddle point theorems

The extremal properties detailed in the preceding section lead us to consider source-channel pairs $(\overline{P}_X, \overline{P}_{Y|X})$, such that $\overline{P}_{Y|X}$ is a maximum (conditional) entropy kernel under a constraint on ρ_1, while the induced output measure $\overline{P}_Y = \overline{P}_X\overline{P}_{Y|X}$ is a maximum-entropy measure under a constraint on ρ_2. We formalize this notion in the following definition:

Definition 1. *A source $\overline{P}_X \in \mathcal{M}(\mathsf{X})$ and a channel $\overline{P}_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y})$ are (ρ_1, ρ_2, μ)* saddle point admissible *if*

a) $\overline{P}_{Y|X} = \overline{P}^{\beta}_{Y|X}$ *for some* $\beta \in \mathbb{R}^2$.
b) $\overline{P}_Y \triangleq \overline{P}_X\overline{P}_{Y|X} = \overline{P}^{\gamma}_Y$ *for some* $\gamma \in \mathbb{R}$.
c) $I(\overline{P}_X, \overline{P}_{Y|X}) < +\infty$.

Define the functions $E_1(\beta)$ and $E_2(\gamma)$ via

$$E_1(\beta) \triangleq \mathbb{E}\left[\rho_1(X,Y)\right], \qquad E_2(\gamma) \triangleq \mathbb{E}\left[\rho_2(Y)\right]$$

where the expectation is taken w.r.t. $\overline{P}_X \otimes \overline{P}_{Y|X}$. From the definitions we see that, if $\overline{P}_X$ and $\overline{P}_{Y|X}$ are a saddle point admissible pair, then the mutual information is given by

$$\begin{aligned} I(\overline{P}_X, \overline{P}_{Y|X}) &= \mathbb{E}\left[\log \frac{f^{\beta}_{Y|X}(Y|X)}{f^{\gamma}_Y(Y)}\right] \\ &= \log \frac{Z_2(\gamma)}{Z_1(\beta)} + (\gamma - \beta_2)E_2(\gamma) - \beta_1 E_1(\beta). \end{aligned} \tag{10}$$

Next, we define the following sets:

$$\mathcal{S}(\overline{P}_X, \overline{P}_{Y|X}) \triangleq \Big\{ P_X \in \mathcal{M}(\mathsf{X}) :$$
$$\overline{P}_{Y|X=x} \ll \overline{P}_Y \;\; P_X\text{-a.s.}; \tag{11a}$$
$$\mathbb{E}_{P_X \otimes \overline{P}_{Y|X}}\left[\rho_1(X,Y)\right] = E_1(\beta); \tag{11b}$$
$$\mathbb{E}_{P_X \otimes \overline{P}_{Y|X}}\left[\rho_2(Y)\right] = E_2(\gamma) \Big\} \tag{11c}$$

and

$$\mathcal{N}(\overline{P}_X, \overline{P}_{Y|X}) \triangleq \Big\{ P_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y}) :$$
$$P_{Y|X=x} \ll \overline{P}_{Y|X=x} \;\; \overline{P}_X\text{-a.s.}; \tag{12a}$$
$$\mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\rho_1(X,Y)\right] \le E_1(\beta); \tag{12b}$$
$$\mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\rho_2(Y)\right] = E_2(\gamma) \Big\}. \tag{12c}$$

Then we have the following main theorem:

Theorem 1. *Let $\overline{P}_X \in \mathcal{M}(\mathsf{X})$ and $\overline{P}_{Y|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{Y})$ be a (ρ_1, ρ_2, μ) saddle point admissible pair. Then for any $P_X \in \mathcal{S}(\overline{P}_X, \overline{P}_{Y|X})$ and any $P_{Y|X} \in \mathcal{N}(\overline{P}_X, \overline{P}_{Y|X})$:*

$$I(P_X, \overline{P}_{Y|X}) \overset{\text{(a)}}{\le} I(\overline{P}_X, \overline{P}_{Y|X}) \overset{\text{(b)}}{\le} I(\overline{P}_X, P_{Y|X}). \tag{13}$$

Proof: Intuitively, (a) follows from the fact that, according to Definition 1b, $\overline{P}_Y$ is a maximum-entropy distribution over all $P_Y \in \mathcal{M}(\mathsf{Y})$ that satisfy (11c). Since $I(P_X, \overline{P}_{Y|X}) = h(Y) - h(Y|X)$, and since $h(Y|X)$ is constant by virtue of (11b), we obtain inequality (a) in (13). To be more precise, we proceed as follows. Because of (11a), $P_X\overline{P}_{Y|X} \ll \overline{P}_Y = \overline{P}^{\gamma}_Y$, so we can choose $Q = \overline{P}^{\gamma}_Y$ in (4):

$$\begin{aligned} I(P_X, \overline{P}_{Y|X}) &= D\left(\overline{P}_{Y|X} \| \overline{P}_Y | P_X\right) - D\left(P_X\overline{P}_{Y|X} \| \overline{P}_Y\right) \\ &\le D\left(\overline{P}_{Y|X} \| \overline{P}_Y | P_X\right) \\ &= \mathbb{E}_{P_X \otimes \overline{P}_{Y|X}}\left[\log \frac{f^{\beta}_{Y|X}(Y|X)}{f^{\gamma}_Y(Y)}\right] \qquad (14) \\ &= \log \frac{Z_2(\gamma)}{Z_1(\beta)} + \mathbb{E}_{P_X\overline{P}_{Y|X}}\left[\gamma\rho_2(Y)\right] \\ &\qquad - \mathbb{E}_{P_X \otimes \overline{P}_{Y|X}}\left[\beta^T\rho(X,Y)\right] \qquad (15) \\ &= \log \frac{Z_2(\gamma)}{Z_1(\beta)} + (\gamma - \beta_2)E_2(\gamma) - \beta_1 E_1(\beta) \qquad (16) \end{aligned}$$

where (14) follows from (11a); (15) follows from (6) and (7); and (16) follows from (11b) and (11c).

Now we prove (b). Intuitively, it follows from the fact that, according to Definition 1a, $P_{Y|X}$ is a maximum conditional entropy kernel under the constraint (12b). More precisely, from (12a), we have $\overline{P}_X \otimes P_{Y|X} \ll \overline{P}_X \otimes \overline{P}_{Y|X}$ (and consequently $P_Y \ll \overline{P}_Y$). Thus, from (5)

$$\begin{aligned} I(\overline{P}_X, P_{Y|X}) &= D\left(P_{Y|X} \| \overline{P}_{Y|X} | \overline{P}_X\right) - D\left(\overline{P}_X P_{Y|X} \| \overline{P}_Y\right) \\ &\qquad + \mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\log \frac{d\overline{P}_{Y|X}}{d\overline{P}_Y}\right] \\ &\ge \mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\log \frac{d\overline{P}_{Y|X}}{d\overline{P}_Y}\right] \qquad (17) \\ &= \mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\log \frac{f^{\beta}_{Y|X}(Y|X)}{f^{\gamma}_Y(Y)}\right] \qquad (18) \\ &= \log \frac{Z_2(\gamma)}{Z_1(\beta)} + \mathbb{E}_{\overline{P}_X P_{Y|X}}\left[\gamma\rho_2(Y)\right] \\ &\qquad - \mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\beta^T\rho(X,Y)\right] \qquad (19) \\ &\ge \log \frac{Z_2(\gamma)}{Z_1(\beta)} + (\gamma - \beta_2)E_2(\gamma) - \beta_1 E_1(\beta) \qquad (20) \end{aligned}$$

where (17) follows because conditioning increases divergence [9, Ch. 5]; (18) follows from (12a); (19) follows from (6), (7); and (20) follows from (12b) and (12c). We also use the fact that $\beta_1, \rho_1(\cdot,\cdot) \ge 0$. ■

A. Additive-noise channels

In this section, we particularize Theorem 1 to additive-noise channels. We can phrase everything in terms of the input measure P_X and the noise law $\overline{P}_W \equiv \overline{P}_W^\beta$. Thus, we define $I(P_X, \overline{P}_W) \triangleq I(P_X, \overline{P}_{Y|X})$ where $\overline{P}_{Y|X}$ effects the random transformation $Y = X+W$, and revise the definition of saddle point admissibility as follows:

Definition 2. *Consider a source $\overline{P}_X \in \mathcal{M}(\mathsf{X})$ and a statistically independent additive noise $\overline{P}_W \in \mathcal{M}(\mathsf{X})$. Let $(\overline{X}, \overline{W})$ have joint law $\overline{P}_X \otimes \overline{P}_W$, and let $\overline{Y} = \overline{X} + \overline{W}$. Then $\overline{P}_X$ and $\overline{P}_W$ are (ρ, μ)* saddle point admissible *if*

a) $\overline{P}_W = \overline{P}_W^\beta$ *of the form* (9) *for some* β.
b) $\overline{Y} \sim \overline{P}_W^\gamma$ *for some* γ.
c) $I(\overline{P}_X, \overline{P}_W) \equiv I(\overline{X}; \overline{X} + \overline{W}) < +\infty$.

Remark 1. Saddle point admissibility for additive channels is equivalent to *convolution divisibility* of the exponential family (9). If $(\overline{P}_X, \overline{P}_W)$ are saddle point admissible, then $\overline{P}_W^\gamma = \overline{P}_X \star \overline{P}_W^\beta$ ($\star$ denotes the convolution). Conversely, given $\overline{P}_W^\gamma$ and $\overline{P}_W^\beta$, there exists a $\overline{P}_X$ such that $(\overline{P}_X, \overline{P}_W^\beta)$ is a saddle point admissible pair with $\overline{P}_Y = \overline{P}_W^\gamma$ only if $\overline{P}_W^\gamma = \overline{P}_W^\beta \star \overline{P}_X$.

For additive channels, $E_1(\cdot), E_2(\cdot)$ are both given by

$$E(\alpha) = \int_{\mathsf{X}} \frac{\rho(x)e^{-\alpha\rho(x)}}{Z(\alpha)} \mu(dx), \qquad \forall \alpha \in \mathbb{R} \text{ s.t. } Z(\alpha) < \infty.$$

The mutual information $I(\overline{P}_X, \overline{P}_W)$ simplifies to

$$I(\overline{P}_X, \overline{P}_W) = \log \frac{Z(\gamma)}{Z(\beta)} + (\gamma - \beta)E(\gamma) - \beta E(\beta). \tag{21}$$

We also redefine the sets $\mathcal{S}(\overline{P}_X, \overline{P}_{Y|X})$ and $\mathcal{N}(\overline{P}_X, \overline{P}_{Y|X})$ appropriately:

$$\mathcal{S}_{\text{add}}(\overline{P}_X, \overline{P}_W) \triangleq \Big\{ P_X \in \mathcal{M}(\mathsf{X}) :$$
$$\overline{P}_{X+W|X=x} \ll \overline{P}_{X+W} \;\; P_X\text{-a.s.}; \tag{22a}$$
$$\mathbb{E}_{P_X \otimes \overline{P}_W} \left[\rho(X+W)\right] = E(\gamma) \Big\} \tag{22b}$$

and

$$\mathcal{N}_{\text{add}}(\overline{P}_X, \overline{P}_W) \triangleq \Big\{ P_{W|X} \in \mathcal{K}(\mathsf{X} \to \mathsf{X}) :$$
$$P_{W|X=x} \ll \overline{P}_W, \;\; \overline{P}_X - \text{a.}s. \tag{23a}$$
$$\mathbb{E}_{\overline{P}_X \otimes P_{W|X}} \left[\rho(W)\right] \le E(\beta); \tag{23b}$$
$$\mathbb{E}_{\overline{P}_X \otimes P_{W|X}} \left[\rho(X+W)\right] = E(\gamma) \Big\}, \tag{23c}$$

Note that in (23) we are allowing *input-dependent* additive noise W. Also note that the counterpart to (11b) is absent in (22) because, under $\overline{P}_W$, W is independent of X, and so (11b) is always satisfied. Then we have the following saddle point theorem:

Theorem 2. *Let the source $\overline{P}_X$ and additive noise $\overline{P}_W$ be (ρ, μ) saddle point admissible. Then for any $P_X \in \mathcal{S}_{\text{add}}(\overline{P}_X, \overline{P}_W)$ and any $P_W \in \mathcal{N}_{\text{add}}(\overline{P}_X, \overline{P}_W)$:*

$$I(P_X, \overline{P}_W) \overset{\text{(a)}}{\le} I(\overline{P}_X, \overline{P}_W) \overset{\text{(b)}}{\le} I(\overline{P}_X, P_{W|X}),$$

where equality in (a) holds if and only if $P_X = \overline{P}_X$, and equality in (b) holds if and only if $P_{W|X} = \overline{P}_W$.

Proof: The proof of the inequalities follows directly from Theorem 1. Conditions for equality follow from the uniqueness property under convolution divisibility: if $\overline{P}_W^\gamma = \overline{P}_W^\beta \star P$ and $\overline{P}_W^\gamma = \overline{P}_W^\beta \star Q$, then $P \equiv Q$. ■

V. Examples

The Gaussian channel. Let $\rho(x) = x^2$. Then $\mathcal{N}(0, \sigma^2)$, the zero-mean normal distribution with variance σ^2, is of the form (9) with $\beta = 1/2\sigma^2$ and $Z(\beta) = \sqrt{\pi/\beta}$. Thus, the scalar Gaussian channel $Y = X + W$ with $W \sim \mathcal{N}(0, \sigma^2)$ is of the E-type form (8).

Let $\overline{P}_W = \mathcal{N}(0, \sigma^2)$ and $\overline{P}_X = \mathcal{N}(0, P)$. Then $\overline{P}_X$ and $\overline{P}_W$ are a saddle point admissible pair with $\gamma = \frac{1}{2(P+\sigma^2)}$. Moreover,

$$\mathcal{S}_{\text{add}} = \left\{ P_X : \mathbb{E}[X^2] = P \right\}, \tag{24}$$
$$\mathcal{N}_{\text{add}} = \left\{ P_{W|X} : \mathbb{E}_{\overline{P}_X \otimes P_{W|X}}[W^2] \le \sigma^2 \right\} \tag{25}$$

and

$$I(\overline{P}_X, \overline{P}_W) = C(P, \sigma^2) \triangleq \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2}\right).$$

Note that the strict equality can be replaced with $\le$ in (24) due to the monotonicity of $C(P, \sigma^2)$ with respect to P. Thus, we recover (i) the Gaussian saddle-point theorem[1], (ii) the classic result on the capacity of the Gaussian channel under the average power constraint, and (iii) the Gaussian squared-error rate-distortion theorem.

The exponential channel. Let $\rho(x) = x$ if $x \ge 0$ and $+\infty$ otherwise. Then $\mathcal{E}(b)$, the exponential distribution with mean $b > 0$, is of the form (9) with $\beta = 1/a$ and $Z(\beta) = 1/\beta$.

Let $\overline{P}_W = \mathcal{E}(b)$. Given $a > 0$, let $\overline{P}_X$ be the mixture of a point mass at 0 and an exponential $\mathcal{E}(a + b)$:

$$\overline{P}_X(X = 0) = \frac{b}{a+b}, \qquad \overline{P}_X(X > x | X > 0) = e^{-\frac{x}{a+b}}.$$

Then, from [2, Thm. 3] and [3, Thm. 1] it follows that $\overline{P}_X$ and $\overline{P}_W$ are a saddle point admissible pair with $\gamma = \frac{1}{a+b}$. Moreover,

$$\mathcal{S}_{\text{add}}(\overline{P}_X, \overline{P}_W) = \left\{ P_X : X \ge 0 \text{ a.s., } \mathbb{E}[X] = a \right\}, \tag{26}$$
$$\mathcal{N}_{\text{add}}(\overline{P}_X, \overline{P}_W) = \left\{ P_{W|X} : W \ge 0 \text{ a.s.}; \mathbb{E}_{\overline{P}_X \otimes P_{W|X}}[W] \le b \right\}$$

and

$$I(\overline{P}_X, \overline{P}_W) = C(a, b) = \log \left(1 + \frac{a}{b}\right).$$

Again, the strict equality can be replaced with $\le$ in (26) due to the monotonicity of $C(a, b)$ with respect to a. Thus, we recover the saddle-point theorem for the exponential-noise channel [2, Thm 3], [3, Thm 1], the rate-distortion theorem for the Poisson process [3, Thm 2], and the capacity theorem for the exponential-noise channel [3, Thm 3].

The exponential server timing channel (ESTC). Finally, we consider an important E-type channel that does not have an

[1] In a more generalized form: here the statistics of the noise are even allowed to depend on the input X in (25).

additive structure: the *exponential server timing channel* [2], [4], [10], which models the input-output behavior of a $\cdot/M/1$ queue. Fix a $T \in (0,\infty)$. Following Sundaresan and Verdú [4], we take X to be the set of functions $x : [0,T] \to \mathbb{Z}_+ \cup \{\infty\}$ that are non-decreasing and right-continuous. In other words, X is the set of point processes on $[0,T]$. Let $\{\mathcal{F}_t^X : t \in [0,T]\}$ denote the filtration over X defined via $\mathcal{F}_t^X \triangleq \sigma\{x_s : s \in [0,t]\}$. We take $\mathcal{F}_{\mathsf{X}} = \mathcal{F}_T^X$. Also, let $\mathsf{Y} = \{y \in \mathsf{X} : \ y_0 = 0\}$ and take $\mathcal{F}_{\mathsf{Y}} = \mathcal{F}_T^Y$. Define the *queue state* $q_t = x_t - y_t$. Note that, with this framework, $q_0 = x_0$, and so x_0 determines the queue state at time 0.

Let $P^\lambda \in \mathcal{M}(\mathsf{Y})$ denote the distribution of a Poisson process of rate λ, and let the reference measure $\mu \equiv P^1$. Then the ESTC with the initially empty queue ($q_{0-} = 0$) and service rate ν can be modeled via the transition density of the form [4], [11]

$$\frac{d\overline{P}^\beta_{Y|X=x}}{d\mu}(y) = \nu^{y_T} e^T \exp\left\{-\nu \int_0^T \tilde{\rho}_1(x_t - y_t)dt\right\} = \frac{e^{-\beta^T \rho(x,y)}}{Z_1(\beta)}$$

where

$$\tilde{\rho}_1(q) \triangleq \begin{cases} 1 & q > 0 \\ 0 & q = 0 \\ \infty & q < 0 \end{cases}$$

$$\rho_1(x,y) \triangleq \int_0^T \tilde{\rho}_1(x_t - y_t)dt, \ \ \rho_2(y) \triangleq y_T$$

$$\beta_1 = \nu \qquad \beta_2 = -\log\nu \qquad Z_1(\beta) = e^{-T}$$

Thus, the ESTC is E-type, according to (6).

Now consider any $\lambda \in (0,\nu)$. Under the constraint $\mathbb{E}[\rho_2(Y)] = \mathbb{E}[Y_T] = \lambda T$, $P^\lambda \in \mathcal{M}(\mathsf{Y})$ is the Gibbs measure $\overline{P}^\gamma_Y$ with respect to ρ_2, as in (7):

$$f_Y(y) \triangleq \frac{d\overline{P}^\gamma_Y}{d\mu}(y) = \lambda^{y_T} e^{-(\lambda-1)T} = \frac{e^{-\gamma\rho_2(y)}}{Z_2(\gamma)}, \qquad (27)$$

with $\gamma = -\log\lambda$ and $Z_2(\gamma) = e^{(\lambda-1)T}$.[2] Let us specify an input distribution $\overline{P}_X$ as follows: (i) X_0 is geometric with parameter $\frac{\lambda}{\nu}$, and (ii) the point process $(X_t - X_0 : 0 \le t \le T) \in \mathsf{Y}$ is statistically independent of X_0, with distribution $\overline{P}^\gamma_Y$. Then the queue states Q_t form an ergodic Markov chain, in steady state at time 0, and so Burke's theorem [11] ensures that $Y \sim P^\lambda$. The resulting pair $(\overline{P}_X, \overline{P}^\beta_{Y|X})$ is saddle point admissible, and

$$\mathcal{S}(\overline{P}_X, \overline{P}_{Y|X}) \triangleq \Bigg\{ P_X : \\ \mathbb{E}_{P_X \otimes \overline{P}_{Y|X}}\left[\frac{1}{T}\int_0^T 1_{\{X_t > Y_t\}} dt\right] \le \frac{\lambda}{\nu}; \qquad (28) \\ \mathbb{E}_{P_X \otimes \overline{P}_{Y|X}}\left[\frac{Y_T}{T}\right] = \lambda \Bigg\} \qquad (29)$$

[2]Note that here, because $\lambda \in (0,\nu)$, $Z_2(\gamma) \neq Z_1([0,\gamma]^T)$, which is infinite.

$$\mathcal{N}(\overline{P}_X, \overline{P}_{Y|X}) \triangleq \Bigg\{ P_{Y|X} : \\ X_t \ge Y_t \ \forall t \in [0,T], \quad \overline{P}_X\text{-a.s.}; \qquad (30) \\ \mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\frac{1}{T}\int_0^T 1_{\{X_t > Y_t\}} dt\right] \le \frac{\lambda}{\nu}; \qquad (31) \\ \mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}\left[\frac{Y_T}{T}\right] = \lambda \Bigg\}. \qquad (32)$$

The mutual information (normalized by T) is

$$\frac{1}{T} I(\overline{P}_X, \overline{P}^\beta_{Y|X}) = C(\lambda,\nu) \triangleq \lambda \log\left(\frac{\nu}{\lambda}\right),$$

which is simultaneously the rate-constrained capacity of the ESTC [2], as well as the rate-distortion function of the poisson Process with distoriton measure $\rho_1(x,y)$ at $D = \lambda/\nu$ [10]. This example demonstrates the importance of the equality constraint on the expected output cost $\mathbb{E}_{\overline{P}_X \otimes P_{Y|X}}[\rho_2(Y)]$, stated here in (12c): For the ESTC, $C(\lambda,\nu)$ is *not* monotonic in λ (cf. [2, Fig 3]), and so turning the equality in (32) into $\le$ clearly violates the theorem. To the best of our knowledge, this mutual information saddle-point theorem for ESTC on the $[0,T]$ interval is new. The second part of the saddle point inequality closely resembles the optimality of the ESTC in the rate-distortion problem for a Poisson process (Y) with a 'queuing distortion measure' (encoded in (30) and (31)) [10], although there is a slight difference because the expectations in [10] should be taken with respect to $P_{X|Y} \otimes \overline{P}_Y$.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments. T.P. Coleman would also like to acknowledge support provided by the DARPA ITMANET program via US Army RDECOM contract W911NF-07-1-0029, NSF grant CNS 08-31488, and the AFOSR Complex Networks Program via award no FA9550-08-1-0079.

References

[1] T. M. Cover and J. T. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
[2] V. Anantharam and S. Verdú, "Bits through queues," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 4–18, 1996.
[3] S. Verdú, "The exponential distribution in information theory," *Problems Inform. Transmission*, vol. 32, no. 1, pp. 86–95, 1996.
[4] R. Sundaresan and S. Verdú, "Capacity of queues via point-process channels," *IEEE Trans. Inf. Theory*, June 2006.
[5] B. Prabhakar and N. Bambos, "The entropy and delay of traffic processes in ATM networks," in *Proceedings of the Conference on Information Science and Systems (CISS)*, Baltimore, Maryland, 1995, pp. 448–453.
[6] B. Prabhakar and R. Gallager, "Entropy and the timing capacity of discrete queues," *IEEE Trans. Inf. Theory*, vol. 49, 2003.
[7] M. Raginsky and T. P. Coleman, "Mutual information and posterior estimates in channels of exponential family type," in *Proc. IEEE Inform. Theory Workshop*, Taormina, Sicily, October 2009, pp. 399–403.
[8] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley, 1997.
[9] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
[10] T. P. Coleman, N. Kiyavash, and V. Subramanian, "The rate-distortion function of a Poisson process with a queuing distortion measure," *IEEE Trans. Inf. Theory*, submitted May 2008; revised Dec 2009.
[11] P. Brémaud, *Point Processes and Queues: Martingale Dynamics*. New York: Springer, 1981.