Fifty-second Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 1 - 3, 2014

Ergodic Capacity Under Channel Distribution Uncertainty

Sergey Loyka, Charalambos D. Charalambous and Ioanna Ioannou

Abstract—**The impact of channel distribution uncertainty on the performance of fading channels is studied. The compound capacity of a class of ergodic fading channels subject to channel distribution uncertainty is obtained, for arbitrary noise and nominal channel distribution. The saddle-point property is established, so that the compound capacity equals to the worst-case channel capacity, which is characterized as 1-D convex optimization problem. The properties of worst-case mutual information and channel distribution are studied. Closed-form solutions are obtained in the asymptotic regimes of small and large uncertainty, and an error floor effect is established in the latter case. The known results for the ergodic capacity of the Gaussian MIMO channel under i.i.d. Rayleigh fading are shown to hold under the channel distribution uncertainty as well.**

I. Introduction

CHANNEL state information (CSI) has a significant impact on channel performance as well as code design to achieve that performance. This effect is especially pronounced for wireless channels, due to their dynamic nature, limitations of a feedback link, channel estimation errors etc. [1][2].

When only incomplete or inaccurate CSI is available, performance analysis and coding techniques have to be modified properly. The impact of channel uncertainty has been extensively studied since late 1950s [3]-[5]; see [2] for an extensive literature review up to late 1990s. Since channel estimation is done at the receiver (Rx) and then transmitted to the transmitter (Tx) via a limited (if any) feedback link, most studies concentrate on limited CSI available at the Tx end assuming full CSI at the Rx end.

There are several typical approaches to this problem. In the compound channel model, the channel is unknown to the Tx but is known to belong to a certain class of channels. A member of the channel uncertainty class is selected at the beginning and held constant during the entire transmission, thus modeling a scenario with little dynamics (channel coherence time significantly exceeds the codeword duration [1][6]). A more dynamic approach is that of the arbitrary-varying channel, where the channel is allowed to vary from symbol to symbol being unknown to the Tx [2].

Incomplete CSI at the Tx end can be addressed by assuming that the channel is not known but its distribution is known to the Tx, the so-called channel distribution information (CDI) [1][6]. However, complete knowledge of CDI can be questioned on the same grounds as complete CSI: when only a limited sample set is available (always a practicality), channel distribution can be obtained with limited accuracy only (especially at the distribution tails); limited feedback link dictates quantization of the estimated CDI before transmission, thus introducing the quantization noise; presence of noise and channel dynamics makes any estimate inaccurate to a certain degree. This motivates us to study the impact of inaccurate channel distribution information on system performance.

In the context of non-ergodic fading channels such study has been reported in [7], where the main performance metric was the outage probability. It was demonstrated that inaccurate CDI limits the achievable outage probability: increasing the SNR over a certain threshold does not reduce the outage probability, i.e. an error floor effect. The key parameter characterizing the error floor effect is the distance between the nominal (estimated) and true distributions as measured by the relative entropy, regardless of any other channel specifics (e.g. nominal CDI, noise distribution etc.).

In the present paper, we carry out a similar investigation for ergodic settings, i.e. assuming that the channel is subject to an ergodic fading process so that the main performance metric is ergodic capacity [1]. However, since incomplete (inaccurate) CDI is assumed, the standard results on ergodic capacity [1][6] do not apply as certain achievable performance have to be demonstrated for the whole class of distributions, not just for a single one. We accomplish this using the standard compound channel approach [1][2] - properly extended to the ergodic setting. This allows us to establish the operational meaning of the max-min ergodic mutual information (MI), where min is over all channel distributions in the uncertainty class and max is over all feasible input distributions, as the largest achievable rate under the CDI uncertainty.

First, the worst-case ergodic MI is characterized as a 1-D convex optimization problem; its properties are studied and asymptotic analytical solutions (small/large uncertainty regimes) are obtained in closed forms. An error floor effect is established in the large-uncertainty regime: the worst-case MI and thus the compound capacity cannot be increased by increasing SNR but rather more accurate channel estimation is required to accomplish this. Our analysis of the small-uncertainty regime answers quantitatively the question ”how accurate is the perfect CDI?”.

Then, an operational meaning of the worst-case MI as the largest achievable rate for a given input distribution is established and the corresponding compound channel capacity is shown to be the max-min MI (where the min in over class of

S. Loyka is with the School of Electrical Engineering and Computer Science, University of Ottawa, Ontario, Canada, K1N 6N5, e-mail: sergey.loyka@ieee.org

I. Ioannou and C.D. Charalambous are with the ECE Department, University of Cyprus, 75 Kallipoleos Avenue, P.O. Box 20537, Nicosia, 1678, Cyprus, e-mail: aioannak@yahoo.gr, chadcha@ucy.ac.cy

978-1-4799-8009-3/14/$31.00 ©2014 IEEE

channels and max is over the input distribution); a number of its properties are also established. The saddle-point property is shown to hold, so that the compound channel capacity equals to the worst-case channel capacity, from which its game-theoretic interpretation follows.

These results are further extended to continuous fading distributions and an AWGN MIMO channel is considered subject to any unitary-invariant fading (of which i.i.d. Rayleigh fading is a special case). The optimal signaling is shown to be isotropic Gaussian, thus extending the corresponding result in [11] in several directions (from i.i.d. Rayleigh to any unitary-invariant fading; from a single fading channel to a compound channel setting to accommodate channel distribution uncertainty; the same optimal signaling is shown to hold under the total as well as per-antenna power constraints, thus demonstrating that no advantage is gained by trading off the power among the Tx antennas).

II. Channel Model

Let x and y be the channel input and output respectively, and h be the channel state (all can be sequences). Assume that the full channel state information (CSI) is available at the receiver but not the transmitter (see e.g. [1][6] for a detailed motivation of this assumption) and that the channel input x and state h are independent of each other. For any channel state and input distribution $p(x)$, the channel is characterized by its (instantaneous) mutual information (MI) $I(x;y,h) = I(x;y|h)$, where, following [1][6], we have augmented the output with the channel state (since it is known at the Rx) and have used the independence of x and h. We will further assume that the channel is subject to an ergodic fading characterized by its probability distribution f.

For a finite-state channel, $h \in \{h_1, ..., h_m\}$, f_i is the probability of $h = h_i$, and $I_i = I(x;y|h_i)$ is the (instantaneous) mutual information supported by channel realization h_i under given input distribution $p(x)$; without loss of generality, assume decreasing ordering $I_1 \geq I_2 \geq ... \geq I_m$ (unless otherwise indicated, we assume that not all I_i are the same). The ergodic mutual information supported by this channel is

$$I(x;y|f) = \sum_i f_i I_i \tag{1}$$

which is also a function of $f = \{f_1...f_m\}$. When f is known to the Tx, this is also the largest achievable rate for a given input distribution $p(x)$ [1].

Ergodic channel model is suitable in scenarios with significant channel dynamics so that a single codeword spans many different channel realizations and an encoder can take advantage of it [1][6]. However, in many practical scenarios, complete knowledge of channel distribution f may be not available at the transmitter, due to e.g.

- inaccuracy in estimating f at the receiver (due to finite sample size or estimation noise);
- limited (quantized) feedback link (quantization noise);
- outdated estimate,

so that the true channel distribution f differs from its estimate f_0 available at the transmitter.

To model this CDI uncertainty (inaccuracy), consider the scenario where the transmitter has only partial CDI. Namely, it knows that the true f is within a certain distance of the nominal (estimated) known f_0. We use the relative entropy as a measure of the distance between two distributions, so that all feasible distributions f satisfy the following inequality:

$$f = \{f_1...f_m\} : D(f||f_0) = \sum_i f_i \ln \frac{f_i}{f_{0i}} \leq d, \tag{2}$$

where $f_0 = \{f_{01}...f_{0m}\}$ is a nominal (known) distribution and $d \geq 0$ determines the size of the distribution uncertainty set.

Similar approach has been adopted in [7] to characterize the impact of channel distribution uncertainty on the performance of non-ergodic (quasi-static) fading channels, where the main performance metrics are outage probability (for a given target rate) or outage capacity (for a given outage probability). While the value of relative entropy as a measure of distance between two distributions is well-known [12], it will become clear from the present study that d is a critical parameter that characterizes the loss in performance due to the channel distribution uncertainty as well.

We will not assume any particular noise or channel distribution (except for examples) so that our results are general and apply to *any* such distribution.

III. Worst-Case Ergodic Mutual Information

Under a given $p(x)$, the worst-case ergodic mutual information for the CDI uncertainty set in (2) is given by

$$I_w = \min_{D(f||f_0) \leq d} I(x;y|f) \tag{3}$$

Its operational meaning will be established in the next section: when the nominal distribution f_0 and "radius" d are known at the transmitter, this is the largest achievable rate under the worst-case fading channel for a given $p(x)$ (and is a function of f_0 and d). The Theorem below gives its characterization as a 1-D convex optimization problem.

Theorem 1: For a given input distribution $p(x)$ and arbitrary nominal fading distribution f_0, the worst-case ergodic mutual information I_w in (3) can be expressed as a scalar convex optimization problem:

$$I_w = \max_{s \leq 0} s \Big(\ln \sum_i f_{0i} e^{I_i/s} + d \Big) \tag{4}$$

and the maximizing s^* can be found as a unique solution of the following equation

$$F(s) = \frac{\sum_i f_{0i} \frac{I_i}{s} e^{I_i/s}}{\sum_i f_{0i} e^{I_i/s}} - \ln \sum_i f_{0i} e^{I_i/s} = d \tag{5}$$

if $d \leq \ln \frac{1}{f_{0m}}$. The worst-case (minimizing) fading distribution f_i^* is

$$f_i^* = \frac{f_{0i} e^{I_i/s^*}}{\sum_i f_{0i} e^{I_i/s^*}}, \tag{6}$$

so that

$$I_w = \frac{\sum_i f_{0i} I_i e^{I_i/s^*}}{\sum_i f_{0i} e^{I_i/s^*}} \quad (7)$$

If $d \geq \ln \frac{1}{f_{0m}}$, $s^* = 0^-$ and

$$f_1^* ... f_{m-1}^* = 0, \ f_m^* = 1, \ I_w = I_m \quad (8)$$

i.e. all the probability mass is on the weakest channel and the worst-case ergodic MI equals to that of a weakest channel realization. If $d = 0$, then $f_i^* = f_{0i}$ and the corresponding worst-case MI is that under the nominal distribution: $I_w = I_0 = \sum_i f_{0i} I_i$, so that in general

$$I_m \leq I_w \leq I_0 \quad (9)$$

Proof: see Appendix. ■

We now proceed to establish a number of properties of $F(s)$ in (5), which reflect on corresponding solutions.

Proposition 1: The function $F(s)$ has the following properties:

1) $F(s)$ is increasing: $F'(s) \geq 0$, with strict inequality unless $s = -\infty$ or 0^- or all I_i are the same.
2) Its limiting values are $F(-\infty) = 0$, $F(0^-) = \ln \frac{1}{f_{0m}}$, so that
3) $0 \leq F(s) \leq \ln \frac{1}{f_{0m}}$ for $-\infty \leq s \leq 0^-$.

Note that $d = \ln \frac{1}{f_{0m}}$ is the threshold radius, beyond which the worst-case ergodic capacity equals to the point-wise (instantaneous) worst-case capacity and the worst-case fading distribution puts all the mass on the weakest channel realization.

A. Asymptotic regimes

Let us now study the worst-case MI in 2 asymptotic regimes, where more insights can be obtained.

Proposition 2: Consider the small uncertainty regime $d \to 0$. The worst-case ergodic MI can be approximated as follows:

$$I_w = I_0 - \sqrt{2d}\sigma_I + o(\sqrt{d}) \quad (10)$$

where $\sigma_I^2 = \sum_i f_{0i} I_i^2 - I_0^2$ is the variance of the instantaneous MI under the nominal fading distribution.

Proof: Based on the standard tools of asymptotic analysis [14]. ■

Note that, in this regime, the worst-case MI decreases proportionally to the standard deviation of the instantaneous MI (under the nominal fading distribution), the proportionality coefficient being $\sqrt{2d}$, and that increasing I_i results in smaller f_i^*., i.e. weaker channels get larger weights.

Large uncertainty regime: this corresponds to $d \geq -\ln f_{0m}$, which is considered in Theorem 1 in (8). Note that in this regime further increase in d (beyond $-\ln f_{0m}$) does not result in any decrease in I_w, as the lower bound in (9) is already achieved. If $I_m = 0$ for any SNR (i.e. zero-gain channel realization) and $d \geq -\ln f_{0m}$, then $I_w = 0$ regardless of the SNR, so that the worst-case MI (and thus the compound channel capacity, which cannot exceed the worst-case MI under the optimal input distribution) cannot be increased by increasing the SNR in the large-uncertainty regime, i.e. there is an error floor effect induced by the channel distribution uncertainty. More accurate channel estimation (i.e. smaller d) is required to increase the worst-case MI in this case.

B. Properties of the worst-case channel distribution and MI

We study below the properties of the worst-case MI. Since the proofs follow mostly in a standard way from Theorem 1, they are omitted due to the page limit.

Proposition 3: The worst-case MI $I_w(d)$ as a function of "radius" d has the following properties:

1) $I_w(d)$ is a convex function of d, strictly so unless $d \geq \ln \frac{1}{f_{0m}}$.
2) $I_w(d)$ is a decreasing function of d, strictly so unless $d \geq \ln \frac{1}{f_{0m}}$,

$$I_w(d_1) > I_w(d_2) \quad \forall d_1 < d_2 < \ln \frac{1}{f_{0m}}. \quad (11)$$

3) Its boundary values are as follows:

$$I_w(0) = I_0, \quad I_w\Big(d \geq \ln \frac{1}{f_{0m}}\Big) = I_m. \quad (12)$$

Proposition 4: The worst-case MI is an increasing function of $I_i, i = 1...m$, strictly so if $d < -\ln f_{0m}$.

Proposition 5: Under the assumed instantaneous MI ordering $I_1 \geq I_2 \geq .. \geq I_m$, the normalized worst-case fading distribution $\alpha_i = f_i^*/f_{0i}$ is increasing in i: $\alpha_1 \leq \alpha_2 \leq .. \leq \alpha_m$. If $I_i < I_j$ and $d < -\ln f_{0m}$, then $\alpha_i > \alpha_j$.

Corollary 5.1: If the nominal fading distribution is uniform, $f_{01} = f_{02} = .. = f_{0m}$, the worst-case fading distribution is increasing in i: $f_1^* \leq f_2^* \leq .. \leq f_m^*$. If $I_i < I_j$ and $d < -\ln f_{0m}$, then $f_i^* > f_j^*$.

Corollary 5.2: If $f_{0i} = 0$, then $f_i^* = 0$. If $d < -\ln f_{0m}$, then $f_i^* = 0$ if and only if $f_{0i} = 0$.

IV. Optimizing Over the Input Distribution

The next step is to optimize the worst-case MI over the input distribution to obtain the compound channel capacity. The following Theorem establishes the operational meaning of this max-min MI. This corresponds to existence of a single code operating over the whole class of fading distributions.

Theorem 2: Consider an ergodic fading channel, whose distribution f is not known at the Tx, but is known to belong to a convex set $\mathcal{S}$ and assume that the set of all feasible input distributions $p(x)$ is convex. Its compound channel capacity C_c is the same as the worst-case channel capacity C_w,

$$C_c \overset{(a)}{=} \sup_{p(x)} \inf_{f \in \mathcal{S}} I(x;y|f) \overset{(b)}{=} \inf_{f \in \mathcal{S}} \sup_{p(x)} I(x;y|f) = C_w \quad (13)$$

Proof: The proof is done in 4 steps, as outlined below:

1) Assume first that $\mathcal{S}$ is of finite cardinality. In this case, (a) follows from Han's compound channel capacity theorem (see theorems 3.3.3 and 5 in [8]) by considering fading distribution f as a channel state.

2) When $\mathcal{S}$ is a convex polyhedron, (a) follows from 1) and the fact that any code that works for finite-cardinality set $\{f_i\}$ also works for its convex envelope $\sum_i \alpha_i f_i$.

3) When $\mathcal{S}$ is an arbitrary convex set, evoke 2) and use a sequence of increasingly finer inner/outer polyhedral approximations as in e.g. [13].

4) (b) follows from Von Neumann mini-max Theorem [9][10].

■

Applying this theorem to the setting in the previous section, one obtains the following.

Theorem 3: Consider the compound ergodic fading channel in (2) when the transmitter knows f_0 and d but not f, and the receiver has full CSI. Assume that the set of feasible input distributions $p(x)$ is convex and compact (e.g. average or maximum power constraint). The compound ergodic channel capacity in this setting is given by

$$C = \max_{p(x)} \min_{D(f||f_0)\le d} I(x;y|f) = \min_{D(f||f_0)\le d} \max_{p(x)} I(x;y|f) = C_w \tag{14}$$

i.e. the compound capacity equals to the worst-case channel capacity C_w and the saddle-point property holds for any feasible $p(x)$ and f,

$$I(x;y|f^*) \le C = I(x^*;y|f^*) \le I(x^*;y|f) \tag{15}$$

where x^* denotes the input under its optimal distribution $p^*(x)$ and (p^*, f^*) is a saddle point. ■

The inequalities in (15) have a well-known game-theoretic interpretation: the Tx chooses $p^*(x)$ and the adversary (nature) chooses f^*; neither player can deviate from this optimal strategy without incurring a penalty.

We are now in a position to obtain the compound channel capacity in the asymptotic regimes.

Proposition 6: Consider the large-uncertainty regime $d \ge -\ln f_{0m}$. The compound channel capacity in this regime is given by

$$C = \max_{p(x)} I_m \tag{16}$$

i.e. designing a single code for the whole class of fading channels is equivalent to designing a code for a weakest channel realization in this regime.

Proof: Follows from (14) and (8). ■

Proposition 7: Consider the small-uncertainty regime as in Proposition 2. The compound channel capacity in this regime is given by

$$C = \max_{p(x)} \{I_0 - \sqrt{2d}\sigma_I\} + o(\sqrt{d}) \tag{17}$$

$$\approx \max_{p(x)} I_0 \tag{18}$$

where 2nd approximation holds when

$$d \ll \frac{1}{2}\left(\frac{I_0}{\sigma_I}\right)^2. \tag{19}$$

Proof: Follows from (10) and (14). ■

In fact, (19) answers the question "how accurate is the perfect CDI?": when (19) holds, the CDI uncertainty is negligible and thus the CDI can be considered "perfect". Note that optimizing (designing a code for) the nominal MI I_0 is not optimal in general (as it does not necessarily optimize σ_I), but is optimal when uncertainty is negligible as in (19), so that one can "recycle" known optimal distributions (codes) in this small-uncertainty regime. On the other hand, one can "recycle" known distributions (codes) for a weakest channel realization in the large uncertainty regime.

Using the general inequality $I_m \le I_w \le I_0$, one obtains the general bounds on the compound ergodic capacity.

Proposition 8: The compound ergodic capacity of a finite-state fading channel can be bounded as follows

$$\max_{p(x)} I_m \le C \le \max_{p(x)} I_0 \tag{20}$$

and the bounds are tight: the lower bound is attained in the large uncertainty regime $d \ge -\ln f_{0m}$, and the upper bound is attained in the small-uncertainty regime $d \ll \frac{1}{2}(I_0/\sigma_I)^2$.

We would like to point out that the above results are general enough to apply to arbitrary nominal fading distribution and arbitrary noise (not necessarily Gaussian).

V. Continuous Fading Distributions

Here we consider a continuous fading distribution. The results follow from the finite-state case by using integrals instead of the sums (and calculus of variations to establish optimality).

In particular, the worst-case MI can be characterized as in Theorem 1 with integrals instead of the sums and a number of its properties mimic those for the finite-state channels.

In the asymptotic regimes, one obtains the following.

Proposition 9: Consider the small uncertainty regime $d \to 0$. When all moments of $I(h)$ are bounded, the worst-case ergodic MI can be approximated as follows:

$$I_w = I_0 - \sqrt{2d}\sigma_I + o(\sqrt{d}) \tag{21}$$

where $\sigma_I^2 = \int f_0(h) I^2(h) dh - I_0^2$ is the variance of the instantaneous MI under the nominal fading distribution.

Large uncertainty regime: this corresponds to $d \ge -\ln f_{0m}$ when there is a point mass f_{0m} at $h = h_m$ so that $I_w = I_m$ if $d \ge -\ln f_{0m}$. When there is no such mass, $I_w \to I_m$ as $d \to \infty$, which corresponds to $f_{0m} \to 0$.

A. An example: Gaussian MIMO channel

In this section, we consider an example of ergodic Gaussian MIMO fading channel when the nominal fading distribution is unitary-invariant. In the special case of i.i.d. Rayleigh fading, its capacity has been established in [11] and the optimal signalling is isotropic Gaussian. Our example extends this in two directions: (i) we consider a class of fading channels thus allowing channel distribution uncertainty, and (ii) we allow the nominal distribution to be any unitary-invariant one, of which i.i.d. Rayleigh fading is a special case. The key result is that the optimal signaling is still isotropic Gaussian, exactly as in [11]. The channel model is

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\xi} \tag{22}$$

where $\boldsymbol{x}, \boldsymbol{y}$ are the input and output signals, $\boldsymbol{\xi}$ is AWG noise, $\boldsymbol{\xi} \sim CN(\boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{H}$ is the channel matrix. Under given channel distribution $f(\boldsymbol{H})$, its ergodic capacity, under the total power constraint $tr\boldsymbol{R} \leq P_T$, is [11]

$$C(f) = \max_{tr\boldsymbol{R} \leq P_T} \int f(\boldsymbol{H}) \ln |\boldsymbol{I} + \boldsymbol{HRH}^+| d\boldsymbol{H} \quad (23)$$

where $\boldsymbol{R} = \overline{\boldsymbol{xx}^+}$ is the covariance of $\boldsymbol{x}$, $(\cdot)^+$ denotes Hermitian conjugation and $|\cdot|$ denotes determinant; we have also used the fact that Gaussian signaling is optimal since the noise is Gaussian.

When the channel distribution is uncertain and belongs to the class in (2), the compound capacity becomes

$$C = \max_{tr\boldsymbol{R} \leq P_T} \min_{D(f||f_0) \leq d} \int f(\boldsymbol{H}) \ln |\boldsymbol{I} + \boldsymbol{HRH}^+| d\boldsymbol{H} \quad (24)$$

The following Proposition characterizes it for a broad class of nominal fading distributions.

Proposition 10: Consider an ergodic-fading AWGN MIMO channel as in (22) whose fading distribution belongs to the class in (2) and assume that the nominal fading distribution $f_0(\boldsymbol{H})$ is right unitary invariant, i.e. $f_0(\boldsymbol{H}) = f_0(\boldsymbol{HU})$ for any unitary $\boldsymbol{U}$. The compound channel capacity is

$$C = \max_{s \leq 0} s \Big\{ \ln \int |\boldsymbol{I} + \gamma \boldsymbol{HH}^+|^{s^{-1}} f_0(\boldsymbol{H}) d\boldsymbol{H} + d \Big\} \quad (25)$$

where m is the number of transmit antennas, $\gamma = P_T/m$ is the per-antenna SNR, i.e. an optimal covariance $\boldsymbol{R}^* = \gamma \boldsymbol{I}$, so that isotropic Gaussian signaling is optimal. This holds under the total as well as per-antenna power constraints: $tr\boldsymbol{R} \leq P_T$ or $r_{ii} \leq P_T/m$, where r_{ii} is i-th diagonal entry of $\boldsymbol{R}$. ■

While this optimal signaling is the same as in the case of i.i.d. Rayleigh-fading channel in [11], the present result extends [11] in three directions:

• a class of fading distributions is considered, rather than a single one, thus allowing fading distribution uncertainty typical in wireless communications;

• i.i.d. Rayleigh fading in [11] is extended to any right-unitary-invariant distribution, of which any spherically-symmetric and thus i.i.d. Raleigh fadings are just special cases;

• the same optimal signaling and capacity are shown to hold under the total as well as the per-antenna power constraints; since the per-antenna power constraint $r_{ii} \leq P_T/m$ implies the total power constraint $tr\boldsymbol{R} \leq P_T$ but not vice-versa, this indicates that nothing is gained by allowing transmitters to trade-off the power under an ergodic, unitary-invariant fading. This may have important applications in multi-user systems.

VI. Appendix: Proof of Theorem 1

The Lagrangian for the optimization problem in (3) is

$$L = \sum_i f_i I_i + \lambda \left(\sum_i f_i \ln \frac{f_i}{f_{0i}} - d \right) + \mu \left(\sum_i f_i - 1 \right) \quad (26)$$

and the corresponding KKT conditions are

$$\frac{\partial L}{\partial f_i} = I_i + \lambda \left(\ln \frac{f_i}{f_{0i}} + 1 \right) + \mu = 0, \quad (27)$$

$$\lambda \left(\sum_i f_i \ln \frac{f_i}{f_{0i}} - d \right) = 0, \ \lambda \geq 0, \ \sum_i f_i = 1. \quad (28)$$

It is straightforward to see that the problem is convex (since the objective $I(x;y|f)$ in (1) is linear in f_i and the constraint in (2) is convex) and the Slater's condition holds (for any $d > 0$), so that the KKT conditions are sufficient for optimality [9]. Combining (27) with the constraint $\sum_i f_i = 1$ one obtains, after some manipulations, the minimizing distribution

$$f_i^* = \frac{f_{0i} e^{-I_i/\lambda}}{\sum_i f_{0i} e^{-I_i/\lambda}}, \quad (29)$$

Using this in (26), one obtains, after some manipulations, the Lagrange dual function $L(\lambda)$:

$$L(\lambda) = -\lambda \left(\ln \sum_i f_{0i} e^{-I_i/\lambda} + d \right), \ \lambda \geq 0 \quad (30)$$

Since the duality gap is zero, the problem in (3) is equivalent to its dual,

$$I_w = \max_{\lambda \geq 0} L(\lambda) \quad (31)$$

Changing the dual variable $s = -\lambda$ results in (4).

To prove (5), let

$$Q(s) = s \Big(\ln \sum_i f_{0i} e^{I_i/s} + d \Big) \quad (32)$$

and observe that $F(s) = d - Q(s)'$. Furthermore,

$$Q(s)'' = -F(s)' \leq 0 \quad (33)$$

This clearly demonstrates that $Q(s)$ is concave and, thus, the problem in (4) is convex (strictly so, unless $s = 0^-$ or all I_i are the same, so that the solution is unique), and that $F(s)$ is increasing (unless $s = -\infty$ or 0^-), so that the equation in (5) has a unique solution if $d \leq -\ln f_{0m}$, which corresponds to the maximizer in (4) (note that $Q(s)' = 0 \leftrightarrow F(s) = d$) and can be easily found numerically using any suitable algorithm (e.g. bisection or Newton decent method [9]). (5) can also be obtained from complementary slackness in (28) when $\lambda > 0$.

If $d > -\ln f_{0m}$, then $Q(s)' = d - F(s) > 0$ (from Proposition 1) so that $s^* = 0^-$ and (8) follows. The same solution applies when $d = -\ln f_{0m}$.

References

[1] E. Biglieri, J. Proakis, and S. Shamai, "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Trans. Inform. Theory*, vol. 44, No. 6, pp. 2619-2692, Oct. 1998.

[2] A. Lapidoth and P. Narayan, "Reliable Communication Under Channel Uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, No. 6, Oct. 1998.

[3] L. Dobrushin, "Optimal information Transmission through a channel with unknown parameters," *Radiotekhnika i Electronika*, vol. 4, pp. 1951-1956, 1959.

[4] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels," *Ann. Math. Statist.*, vol. 30, pp. 1229-1241, December 1959.

[5] W. L. Root, P. P. Varaya, "Capacity of Classes of Gaussian Channels", *SIAM J. Appl. Math.*, vol. 16, no. 6, pp. 1350-1393, Nov. 1968.
[6] G. Caire and K. Kumar, "Information Theoretic Foundations of Adaptive Coded Modulation," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2274–2298, Dec. 2007.
[7] I. Ioannou, C.D. Charalambous, S. Loyka, "Outage Probability Under Channel Distribution Uncertainty", *IEEE Transactions on Information Theory*, vol. 58, no. 11, pp. 6825-6838, Nov. 2012.
[8] T. S. Han, Information-Spectrum Method in Information Theory, New York: Springer, 2003.
[9] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[10] K. Fan, "Minmax Theorems", *Proceedings of the National Academy of Science,* vol.39, pp.42-47, 1953.
[11] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecomm. ETT,* vol. 10, no. 6, pp. 585 - 596, Nov. 1999.
[12] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 2006.
[13] R. Horst et al, Introduction to Global Optimization, Kluwer, Dordrecht, 2000.
[14] M.A. Efgrafov, *Asymptotic Expansions and Entire Functions*, Moscow: GITTL, 1957.