IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

On Optimum Parameter Modulation–Estimation From a Large Deviations Perspective

Neri Merhav

CCIT Report #806
March 2012

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT Report #806 March 2012

On Optimum Parameter Modulation–Estimation From a Large Deviations Perspective

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

Abstract

We consider the problem of jointly optimum modulation and estimation of a real–valued random parameter, conveyed over an additive white Gaussian noise (AWGN) channel, where the performance metric is the large deviations behavior of the estimator, namely, the exponential decay rate (as a function of the observation time) of the probability that the estimation error would exceed a certain threshold. Our basic result is in providing an exact characterization of the fastest achievable exponential decay rate, among all possible modulator–estimator (transmitter–receiver) pairs, where the modulator is limited only in the signal power, but not in bandwidth. This exponential rate turns out to be given by the reliability function of the AWGN channel. We also discuss several ways to achieve this optimum performance, and one of them is based on quantization of the parameter, followed by optimum channel coding and modulation, which gives rise to a separation–based transmitter, if one views this setting from the perspective of joint source–channel coding. This is in spite of the fact that, in general, when error exponents are considered, the source–channel separation theorem does not hold true. We also discuss several observations, modifications and extensions of this result in several directions, including other channels, and the case of multidimensional parameter vectors. One of our findings concerning the latter, is that there is an abrupt threshold effect in the dimensionality of the parameter vector: below a certain critical dimension, the probability of excess estimation error may still decay exponentially, but beyond this value, it must converge to unity.

Index Terms: Parameter estimation, modulation, AWGN, threshold effect, large deviations, reliability function, error exponents.

1 Introduction

The rich literature on parameter estimation includes a large variety of Bayesian and non–Bayesian lower bounds on the mean square error (MSE) in estimating parameters from signals corrupted by an additive white Gaussian noise (AWGN) channel, as well as other channels (see, e.g., the introductions of [1], [2], [20] for overviews on these bounds). Most of these bounds are amenable to calculation for a given form of dependence of the transmitted signal upon the parameter, i.e., a given modulator, and therefore they may give insights concerning optimum estimation for this specific modulator. They may not, however, lend themselves easily to the derivation of *universal* lower bounds, namely, lower bounds that depend neither on the modulator nor on the estimator, which are relevant when both optimum modulators and optimum estimators are sought. Two exceptions to this rule (although usually, not presented as such) are families of bounds that stem from generalized data processing theorems (DPT's) [14], [26], [28], and bounds based on hypothesis testing considerations [3], [27].

Consider, for a example, a random parameter U, uniformly distributed across the unit interval, which is to be conveyed across the AWGN channel with spectral density $N_0/2$, transmission power S, and no bandwidth limitation. Using the classical DPT, one views the random parameter U as a "source" and the MSE of an arbitrary estimator, $\boldsymbol{E}(\hat{U}-U)^2$, as the average distortion D, and then derives a lower bound on D from the inequality $R(D) \leq CT$, where $R(D)$ is the rate–distortion function of U, T is the transmission time, and C is the channel capacity, which for the AWGN with unlimited bandwidth, is given by $C = S/N_0$. Now, $R(D)$ is not known to have a closed–form expression in this case, but it can be further lower bounded by the Shannon lower bound (see, e.g., [9, Sect. 4.6, p. 101]):

$$R(D) \geq h(U) - \frac{1}{2}\ln(2\pi e D) = -\frac{1}{2}\ln(2\pi e D), \tag{1}$$

where $h(U) = 0$ is the differential entropy of U. This readily leads to the universal lower bound $\boldsymbol{E}(\hat{U}-U)^2 \geq \frac{1}{2\pi e}e^{-2CT} = \frac{1}{2\pi e}e^{-2\mathcal{E}/N_0}$, where $\mathcal{E} = ST$ is the signal energy. It turns out that this lower bound is not tight. In [26], it was shown that DPT's pertaining to generalized information measures, yield a tighter universal lower bound that decays (as $T \to \infty$) like e^{-CT}. In [14], this bound was further improved, by another generalized DPT, to behave like $e^{-2CT/3}$, and then yet further improved to $e^{-CT/2}$, using a universal lower bound based on signal detection considerations, in the spirit of the Ziv–Zakai bound [27]

and the Chazan–Zakai–Ziv bound [3].

Concerning upper bounds, it turns out that it is possible to achieve an MSE with an exponential decay rate of the order of $e^{-CT/3}$, which is quite close to the latter lower bound, but there is still some gap. As is shown in [21, Chap. 8], by using frequency position modulation (FPM) with central frequency and bandwidth that both grow like e^{RT}, where $R > 0$ is a fixed design parameter, the MSE of the maximum likelihood (ML) estimator turns out to be composed of two terms: a "small–error" term (or the "weak noise performance" in the terminology of [21]), that behaves essentially like the Cramér–Rao bound, and which is proportional to e^{-2RT}, and an anomalous error term (gross error due to the threshold effect) of the exponential order of $e^{-E(R)T}$, where $E(R)$ is the reliability function of the AWGN, given by

$$E(R) = \begin{cases} \frac{C}{2} - R & 0 \le R \le \frac{C}{4} \\ (\sqrt{C} - \sqrt{R})^2 & \frac{C}{4} \le R \le C \end{cases} \tag{2}$$

The optimum trade-off between these two terms is achieved for $R = C/6$, where they have the same exponential rate, $e^{-CT/3}$ (see also [13]). Similar things can be said about pulse position modulation (PPM) with exponentially growing bandwidth [13]. Yet another modulation scheme is based on simply quantizing the parameter U into one of $M = e^{RT}/2$ evenly spaced points in its interval (which are then far apart by $2e^{-RT}$) and then assigning, to each one of these points, one out of M orthogonal signals with energy $\mathcal{E}$ (see [15] for an analogue for the Poisson channel). Here, the MSE has the same two exponential terms as before, but now the first term, e^{-2RT}, is the contribution of the quantizer to the MSE and the second term, $e^{-E(R)T}$, is the contribution of channel decoding errors.

The quest for closing (or at least, further reducing) the gap between the best known lower bound, $e^{-CT/2}$, and the upper bound, $e^{-CT/3}$, remains unsatisfied at present. This challenge has, unfortunately, defied our best efforts thus far. We conjecture that it is the lower bound that is to be "blamed" for this gap, i.e., we believe that the above–mentioned modulation schemes are essentially optimal but there is room for further improvement of the lower bound that has not been exploited yet.

In this paper, instead of focusing on the MSE as our performance metric, we adopt a large deviations performance metric: We seek optimum modulation and estimation schemes in the sense of maximizing the exponential rate of decay of the probability that the estimation error $|\hat{U} - U|$ would exceed a given threshold. Motivated by the above discussion, we can afford to set this threshold to be exponentially decaying with T, i.e., e^{-RT}, where $R > 0$ is a

parameter whose value can be chosen freely in some range. More precisely, our asymptotic figure of merit for modulation–estimation is

$$E^*(R) = \limsup_{T\to\infty}\left[-\frac{1}{T}\log\inf\Pr\left\{|\hat{U}-U| > e^{-RT}\right\}\right], \tag{3}$$

where the infimum is over all modulator–estimator pairs with power S, and where we remind the reader that $\limsup_{T\to\infty} f(T)$, for a continuous–valued variable T (as opposed to a sequence $\{T_n\}$), is defined as $\lim_{T\to\infty}\sup_{T'\geq T} f(T')$.

Our basic result (asserted and proved in Section 2) is that the lim sup in eq. (3) is equal to the corresponding lim inf (and hence can be replaced by lim) and their common value has an exact characterization given by $E^*(R) = E(R)$, where $E(R)$ is as in (2). All three modulation schemes mentioned above, together with ML estimation, achieve this performance and hence are asymptotically optimum in the above sense.[1]

Beyond the fact that the large deviations performance metric has already been addressed in estimation theory (see, e.g., [10], [11, p. 4], [16], [18, p. 54], [24, eq. (32)], [27, Sect. IV]), a little thought suggests that it is actually natural in this particular setting of wide–band waveform communication, which exhibits threshold effects and anomalies. The reason is that it makes a clear distinction between 'small' errors, of the order of e^{-RT} ("allowed" under this metric), and gross errors, whose probabilistic weight is $e^{-E(R)T}$ at best.[2] A distinction in the same spirit (but not quite the same) was offered also in [21, Sect. 8.4], where it was shown that a non–anomalous MSE of about e^{-2CT} is the best that can be achieved (and again, by the same schemes) under the constraint that the probability of anomaly tends to zero. This has the flavor of our result for $R \approx C$, but here, we expand the spectrum of trade-offs to the entire range $0 \leq R \leq C$. For $R > C$, the error exponent vanishes in the strong sense, i.e., not only does the probability of the undesired error event cease to decay exponentially, it actually tends to unity. In that sense, the threshold effect is manifested in a clear way.

Having hopefully convinced the reader that the large deviations performance criterion is reasonable in the waveform communication setting considered here, there is considerable

[1] The fact that exponentially small error thresholds are exceeded with exponentially small probabilities is rather remarkable. It is thanks to the fact that the modulator is subjected to optimization. By contrast, for amplitude modulation (AM), where the estimation error of the ML estimator has variance $N_0/(2\mathcal{E}) = 1/(2CT)$, we have $\Pr\{|\hat{U}-U| > e^{-RT}\} = 2Q(e^{-RT}\sqrt{2CT}) \to 1$ for every $R > 0$, and only for $R = 0$ this probability decays exponentially.

[2] Typically, in the case of anomaly, the estimate $\hat{U}$ falls in a random point away from U, and so, it makes sense to assign to all gross error events the same cost, as is done by the proposed metric.

room for the speculation that this may not be the case with the MSE criterion, despite its popularity. The difficulty in capturing the threshold effect and in closing gaps between upper and lower bounds in this setting, as discussed above, may be attributed to the fact that the MSE does not distinguish between the small errors and the anomalous errors, which are so different in nature. Comments in the same spirit are made also in [21, p. 633, central paragraph].

We discuss several observations and implications of the above described basic result (Section 3) and several extensions (Section 4), including other channels, variable power, and the case of a multidimensional parameter vector $\boldsymbol{U} = (U_1, \ldots, U_d)$. In the vector case, our error exponent criterion becomes

$$E^*(R_1, \ldots, R_d) = \limsup_{T \to \infty} \left[-\frac{1}{T} \log \inf \Pr \left(\bigcup_{i=1}^{d} \left\{ |\hat{U}_i - U_i| > e^{-R_i T} \right\} \right) \right], \tag{4}$$

where our earlier characterization, in terms of the reliability function, extends to

$$E^*(R_1, \ldots, R_d) = E(R_1 + R_2 + \ldots + R_d). \tag{5}$$

One of the conclusions of this result is that there is an abrupt threshold effect in the dimensionality of the parameter vector: below a certain critical dimension, the probability of excess estimation error may still decay exponentially, but beyond this value, it must converge to unity. We also discuss several other implications of our results.

As a closing remark, we should point out that the criterion of excess estimation error probability was briefly discussed also in [27, Section IV], where a lower bound was given in terms of the error probability of an M–ary detection problem with optimum signaling. This is similar to the line of thought here, however, there are several differences: (i) We consider a Bayesian setting where U is a random variable, as opposed to the worst–case excess error probability, $\max_u \Pr\{\text{excess estimation error}|u\}$. (ii) We allow an arbitrary modulator, rather than focusing on PPM specifically. (iii) We allow an exponentially vanishing error threshold, e^{-RT} (as opposed to a fixed threshold in [27], corresponding to $R = 0$) and explore the entire spectrum of trade-offs between R and the excess estimation error exponent, which in turn is intimately related to the reliability function, $E(R)$. (iv) As described in the previous paragraph, we also expand the scope in several directions, like the multidimensional case and other channels. We also provide some insights from the perspectives of the threshold effect as well as joint source–channel coding and the separation theorem.

2 Problem Formulation and the Basic Result

Consider the signal model

$$y(t) = x(t,u) + n(t), \qquad t \in [0,T) \tag{6}$$

where $x(t,u)$ is a waveform with power S, which is parametrized by $u \in \mathcal{U} \subseteq \mathbb{R}$, and where $n(t)$ is AWGN with two–sided power spectral density $N_0/2$. Considering an arbitrary representation of $x(t,u)$ as a linear combination of orthonormal basis functions, then due to the power limitation, the length of the curve (locus) drawn by the vector of coefficients of this representation, $\{a_k(u),\ k = 1,2,\ldots\}$, as u exhausts $\mathcal{U}$, must be finite (and in fact, no larger than e^{CT} [21, Chap. 8]) in order to keep the anomalous error vanishingly small. It therefore makes sense to assume that $\mathcal{U}$ is a finite interval, which without loss of essential generality, will be taken to be the interval $[-1/2,+1/2)$, as any other interval can be obtained under re-parametrization using a simple affine transformation.

An estimator of u is any measurable mapping from $\{y(t),\ 0 \le t < T\}$ into $\mathcal{U}$. In order to avoid limitations on the class of estimators (e.g., unbiased estimators, etc.), we adopt the Bayesian setting, i.e., we assume that u is a realization of a random variable U, uniformly distributed over $[-1/2,+1/2)$. The uniform prior is assumed merely for convenience and it expresses the fact that no value of u has any preference *a-priori*. Any other prior, which is bounded away from zero and infinity, can be used as well.

A modulator with power S is a mapping from $\mathcal{U}$ into a family of waveforms $\{x(t,\cdot),\ 0 \le t < T\}$, whose power is exactly[3] S, i.e.,

$$\frac{1}{T}\int_0^T \mathrm{d}t \cdot x^2(t,u) = S \tag{7}$$

for all $u \in \mathcal{U}$. No bandwidth limitations are imposed on the waveforms in this family.

For a given $R > 0$, we are interested in characterizing the best achievable excess estimation error exponent

$$E^*(R) = \limsup_{T\to\infty}\left[-\frac{1}{T}\log\inf \Pr\left\{|\hat{U} - U| > e^{-RT}\right\}\right], \tag{8}$$

where the infimum is over all modulator–estimator pairs as defined as above.

We first provide a lower bound on the excess estimation error probability, that leads directly to a converse theorem concerning $E^*(R)$.

[3] In Subsection 4.4, we relax the restriction that the power would be exactly S for all u, and we allow instead the power $S(u)$ to vary with u, but we keep an average power constraint, $\boldsymbol{E}\{S(U)\} \le S$.

Theorem 1 *Consider the AWGN channel with noise power spectral density $N_0/2$. Let $R > 0$ be given and let $\epsilon > 0$ be arbitrarily small. For every modulator with power S and every estimator $\hat{U}$:*

$$Pr\left\{|\hat{U} - U| > e^{-RT}\right\} \geq (1 - e^{-\epsilon T}) \exp\{-T[E(R - \epsilon) + o(T)]\}, \tag{9}$$

where $E(R)$ is the reliability function of the AWGN, defined as in eq. (2) and where $o(T)$ designates a quantity that tends to zero as $T \to \infty$. Consequently,

$$E^*(R) \leq E(R). \tag{10}$$

While the lower bound in Theorem 1 applies, in principle, for every $\epsilon > 0$, quite obviously, for $T \to \infty$, the tightest lower bound is obtained as $\epsilon \to 0$, which yields an exponential decay rate of $E(R)$.

Proof. The proof is in the spirit of the derivation of the Ziv–Zakai bound [27] and the Chazan–Zakai–Ziv bound [3], but with M hypotheses (rather than 2), where M is exponentially large. Consider a given estimator $\hat{U}$ of U and a given modulator $\{x(t,\cdot),\ 0 \leq t < T\}$ with power S. For a given $u \in [-1/2, +1/2)$ and $\Delta > 0$, let $P_e(u, \Delta)$ denote the probability of error of the optimum (ML) detector for deciding among the M equiprobable hypotheses

$$\mathcal{H}_i: \quad y(t) = x(t, u + i\Delta) + n(t), \quad i = 0, 1, \ldots, M-1$$

where it is assumed that u and Δ are such that $u + i\Delta$, $i = 0, 1, \ldots, M-1$, are all in $[-1/2, +1/2)$. First, it is argued that

$$\begin{aligned} P_e(u,\Delta) \quad \leq \quad & \frac{1}{M}\left[\Pr\left\{\hat{U} - U > \frac{\Delta}{2}\middle| U = u\right\} + \right. \\ & \sum_{i=1}^{M-2} \Pr\left\{|\hat{U} - U| > \frac{\Delta}{2}\middle| U = u + i\Delta\right\} + \\ & \left. \Pr\left\{\hat{U} - U < -\frac{\Delta}{2}\middle| U = u + (M-1)\Delta\right\}\right]. \end{aligned} \tag{11}$$

To see why this is true, observe that the r.h.s. can be interpreted as the probability of error of a suboptimum M–ary detector that is based on first estimating U by $\hat{U}$ and then deciding on the hypothesis $\mathcal{H}_i$ whose corresponding grid point $u + i\Delta$ is nearest to $\hat{U}$. Next, we further upper bound the first and the last terms of the r.h.s. by $\Pr\{|\hat{U} - U| > \Delta/2 | U = u\}$ and $\Pr\{|\hat{U} - U| > \Delta/2 | U = u + (M-1)\Delta\}$, respectively, which yields

$$P_e(u,\Delta) \leq \frac{1}{M} \sum_{i=0}^{M-1} \Pr\left\{|\hat{U} - U| > \frac{\Delta}{2}\middle| U = u + i\Delta\right\}. \tag{12}$$

Integrating both sides over u, we get

$$\begin{aligned}
& \int_{-1/2}^{+1/2-(M-1)\Delta} \mathrm{d}u \cdot P_e(u,\Delta) \\
\leq\ & \int_{-1/2}^{+1/2-(M-1)\Delta} \mathrm{d}u \cdot \frac{1}{M} \sum_{i=0}^{M-1} \Pr\left\{|\hat{U}-U| > \frac{\Delta}{2}\middle| U = u + i\Delta\right\} \\
=\ & \frac{1}{M} \sum_{i=0}^{M-1} \int_{-1/2}^{+1/2-(M-1)\Delta} \mathrm{d}u \cdot \Pr\left\{|\hat{U}-U| > \frac{\Delta}{2}\middle| U = u + i\Delta\right\} \\
=\ & \frac{1}{M} \sum_{i=0}^{M-1} \int_{-1/2+i\Delta}^{+1/2-(M-1)\Delta+i\Delta} \mathrm{d}u \cdot \Pr\left\{|\hat{U}-U| > \frac{\Delta}{2}\middle| U = u\right\} \\
\leq\ & \frac{1}{M} \sum_{i=0}^{M-1} \int_{-1/2}^{+1/2} \mathrm{d}u \cdot \Pr\left\{|\hat{U}-U| > \frac{\Delta}{2}\middle| U = u\right\} \\
=\ & \int_{-1/2}^{+1/2} \mathrm{d}u \cdot \Pr\left\{|\hat{U}-U| \geq \frac{\Delta}{2}\middle| U = u\right\} \\
=\ & \Pr\left\{|\hat{U}-U| > \frac{\Delta}{2}\right\}. \qquad (13)
\end{aligned}$$

Now, let $\Delta = 2e^{-RT}$ and $M = e^{(R-\epsilon)T}/2 + 1$. Then, it is well known (see, e.g., [19, p. 168, eq. (3.6.26) and Section 3.8], [7, p. 383, eqs. (8.2.49), (8.2.50)], [21, pp. 345, eq. (5.106c)]) that

$$P_e(u,\Delta) \geq e^{-T[E(R-\epsilon)+o(T)]}, \qquad (14)$$

which, when substituted into the left–most side of (13), readily gives

$$\begin{aligned}
\Pr\left\{|\hat{U}-U| > e^{-RT}\right\} &\geq \int_{-1/2}^{+1/2-(M-1)\Delta} \mathrm{d}u \cdot e^{-T[E(R-\epsilon)+o(T)]} \\
&= [1-(M-1)\Delta]e^{-T[E(R-\epsilon)+o(T)]} \\
&= (1-e^{-\epsilon T})e^{-T[E(R-\epsilon)+o(T)]}, \qquad (15)
\end{aligned}$$

completing the proof of Theorem 1. □

Our next theorem, Theorem 2, provides a compatible achievability result.

Theorem 2 *Consider the AWGN channel with noise power spectral density $N_0/2$ and let $R > 0$ be given. Then there exists a modulator with power S and an estimator $\hat{U}$ for which*

$$Pr\left\{|\hat{U}-U| > e^{-RT}\right\} \leq e^{-E(R)T}. \qquad (16)$$

Consequently, the lim sup *in eq. (3) is equal to the* lim inf *(i.e., the limit exists) and*

$$E^*(R) = E(R). \qquad (17)$$

Proof. We first describe the modulator and estimator. Assume, without essential loss of generality, that $e^{RT}/2$ is integer (otherwise, alter the value of R slightly to make it such). The modulator first quantizes the parameter u to the nearest point in the grid $\{-1/2+e^{-RT}, -1/2+3e^{-RT}, -1/2+5e^{-RT}, \ldots, 1/2-e^{-RT}\}$. This grid, which consists of $M = e^{RT}/2$ points, is mapped into a set of M orthogonal signals, each with power S. Let $i(u)$ denote the index of the grid point nearest to u and let $x_i(t)$ be the signal corresponding to the i–th grid point, $i = 1, 2, \ldots, M$. Then the modulator is defined by

$$x(t, u) = x_{i(u)}(t). \tag{18}$$

Let $\hat{i}$ denote the output of the ML decoder for the signal set $\{x_i(t)\}_{i=1}^{M}$, namely,

$$\hat{i} = \text{argmax}_{1 \le i \le M} \int_0^T x_i(t) y(t) \mathrm{d}t. \tag{19}$$

Then, the estimator $\hat{u}$ is defined as the corresponding grid point, i.e.,

$$\hat{u} = -\frac{1}{2} + (2\hat{i} - 1)e^{-RT}. \tag{20}$$

Clearly, for this particular modulator–estimator pair, the event $\{|\hat{U} - U| > e^{-RT}\}$ implies $\hat{i} \neq i(U)$, namely, an error in decoding the index i of the transmitted signal $x_i(t)$. The probability of excess estimation error is therefore upper bounded by the probability of error for $M = e^{RT}/2$ orthogonal signals, each with energy $\mathcal{E} = ST$, which is well known (see, e.g., [19, p. 67, eq. (2.5.16)] or [7, p. 381, eqs. (8.2.43), (8.2.44)], [21, pp. 344–345, eqs. (5.104)–(5.106b)]) to be upper bounded in turn by $e^{-E[R-(\ln 2)/T]T} \le e^{-E(R)T}$. This completes the proof of Theorem 2. □

The Case $R = 0$

Theorems 1 and 2 refer to the case $R > 0$. The case $R = 0$ should be treated with caution as there is an inherent discontinuity of the *operational* reliability function at $R = 0$. As is well known, the operational reliability function for the infinite–bandwidth AWGN channel, which is defined as the asymptotic error exponent of the optimum rate–R code for this channel, agrees with $E(R)$, given in (2), only for $R > 0$. Concerning the point $R = 0$, there is a difference between the strong sense of this assignment, where the number of codewords M is fixed (independent of T), and the weak sense, where M grows (but in a subexponential rate). This is because for fixed M, the error exponent of the best signal

set (the simplex signal set) is determined by the minimum distance, which depends on M according to $d_{\min} = 2M\mathcal{E}/(M-1)$, where again, $\mathcal{E} = ST$ is the energy of all M signals. The error probability of the optimum code then decays according to $\exp[-T\frac{C}{2} \cdot \frac{M}{M-1}]$, which agrees with $E(0) = C/2$ only when M grows without bound.

Correspondingly, there is a parallel difference between the case where the error threshold, $\Delta/2$ (in the proof of Theorem 1) is *fixed*, as opposed to the weaker sense where Δ is allowed to vanish as T grows, but in a subexponential rate. Theorems 1 and 2 hold for $R > 0$, and the limit $R \to 0$ corresponds to the weaker meaning. What can be said about the stronger meaning? Repeating the proof of Theorem 1, but with a zero–rate lower bound on $P_e(u, \Delta)$ [19, p. 174, eqs. (3.7.2)–(3.7.5)], [21, pp. 345, eq. (5.106c)], we have (by choosing $M = \lfloor 1/\Delta \rfloor$)

$$\Pr\left\{|\hat{U} - U| > \frac{\Delta}{2}\right\} \geq \frac{1}{2}(1 + \Delta - \Delta\lfloor 1/\Delta \rfloor) \cdot Q\left(\sqrt{\frac{\mathcal{E}}{N_0} \cdot \frac{\lfloor 1/\Delta \rfloor}{\lfloor 1/\Delta \rfloor - 2}}\right). \tag{21}$$

In the limit of $T \to \infty$, this lower bound is of the exponential order of

$$\exp\left\{-\frac{CT}{2} \cdot \frac{\lfloor 1/\Delta \rfloor}{\lfloor 1/\Delta \rfloor - 2}\right\}.$$

As an upper bound we have, by a compatible upper bound on the probability of error (see proof of Theorem 2), the following:

$$\Pr\left\{|\hat{U} - U| > \frac{\Delta}{2}\right\} \leq (\lceil 1/\Delta \rceil - 1) \cdot Q\left(\sqrt{\frac{\mathcal{E}}{N_0} \cdot \frac{\lceil 1/\Delta \rceil}{\lceil 1/\Delta \rceil - 1}}\right), \tag{22}$$

which simply follows from the union bound on the probability of error in the detection of one out of $M = \lceil 1/\Delta \rceil$ simplex signals with energy $\mathcal{E} = ST$. Here, the exponential behavior is according to

$$\exp\left\{-\frac{CT}{2} \cdot \frac{\lceil 1/\Delta \rceil}{\lceil 1/\Delta \rceil - 1}\right\}.$$

While there is a gap in the error exponents for every finite Δ, this gap vanishes as $\Delta \to 0$, thus the best achievable asymptotic value of

$$\lim_{\Delta \to 0} \lim_{T \to \infty} \left[-\frac{\ln \Pr\{|\hat{U} - U| > \Delta/2\}}{T}\right]$$

is still $E(0) = C/2$.

In this context of large deviations for fixed Δ, it is appropriate to mention also the relation with the MSE criterion: The two criteria are easily related via the identity

$$\boldsymbol{E}(\hat{U} - U)^2 = 2\int_0^1 \mathrm{d}\Delta \cdot \Delta \cdot \Pr\{|\hat{U} - U| \geq \Delta\}, \tag{23}$$

and so, the MSE can be lower bounded via any lower bound on $\Pr\{|\hat{U} - U| \geq \Delta\}$ for all Δ in the appropriate range, which is exactly the line of thought that guides the Chazan–Zakai–Ziv bound [3] for two hypotheses. Here, as we consider M hypotheses rather than two,[4] and M is exponentially large, the integration range of Δ in the corresponding lower bound, where the integrand is $P_e(u, \Delta)$, must be limited to the interval $(0, 1/(M-1)]$, as otherwise, some grid points $\{u + i\Delta\}$ (in the proof of Theorem 1), would fall outside the interval $[-1/2, +1/2)$. This limitation on the range of Δ causes the resulting lower bound on the MSE to be relatively weak. One of the main points in this paper is that by considering the large deviations performance as our figure of merit in the first place, we actually avoid the need to integrate over Δ altogether. An interesting open question, in this context, is whether it is possible to devise a modulator–estimator pair, which would be independent of Δ, but yet achieve asymptotically optimum large deviations performance for all Δ in the interesting range. Such an estimator may also achieve asymptotically optimum MSE, in view of eq. (23).

3 Discussion

In this section, we pause to discuss a few observations, implications, and modifications of Theorems 1 and 2.

3.1 Strong Converse and the Threshold Effect

The case $R = 0$, discussed in Section 2, is one interesting extreme of the range of R. The other extreme is the point $R = C$, where $E(R)$ vanishes. Here, due to the strong converse to the channel coding theorem, $E(R)$ vanishes in the strong sense for $R > C$, namely, the probability of error tends to unity. Owing to the proof of Theorem 1, the large deviations estimation performance criterion, considered in this paper, 'inherits' this strong converse, and then the probability of excess estimation error tends to unity as $T \to \infty$, for $R > C$. This means an abrupt threshold effect in the limiting probability of excess error, from 0 to 1, as R crosses C.

[4] Here, two hypotheses correspond to antipodal signals, rather than orthogonal signals, and hence lead to non–tight exponential error bounds with a loss of 3dB.

3.2 Achievability by Other Schemes

As mentioned in the Introduction, alternative achievability proofs are possible by analyzing FPM and PPM systems. The FPM modulator (see, e.g., [21]) is defined as follows:

$$x(t,u) = \sqrt{2S}\cos[2\pi(f_0 + u\cdot\Delta f)t], \tag{24}$$

where in our case, both the central frequency f_0 and the frequency offset Δf ($\Delta f \ll f_0$) are taken to be proportional to e^{RT}. For the ML estimator $\hat{U}$, in this case, $\Pr\{|\hat{U}-U| > e^{-RT}\}$ is the probability of anomaly, which is essentially $e^{-E(R)T}$ (see [21, eqs. (8.175a)–(8.175c)]).

Another good modulator for our purpose is PPM, where

$$x(t,u) = s[t-(u+1/2)(T-\tau)], \tag{25}$$

$s[\cdot]$ being a pulse whose support is $[0,\tau]$, where τ is proportional to e^{-RT} (and hence the bandwidth is proportional to e^{RT}), and again, the large deviations event in question is the anomaly event (see, e.g., [13], [25] for more details).

3.3 Relation to Bounds on Moments of the Estimation Error

The combination of Theorem 1 with Chebyshev's inequality,

$$\Pr\left\{|\hat{U}-U| > e^{-RT}\right\} \le \frac{\boldsymbol{E}(\hat{U}-U)^2}{e^{-2RT}} \tag{26}$$

yields the following lower bound on the MSE

$$\boldsymbol{E}(\hat{U}-U)^2 \ge (1-e^{-\epsilon T})e^{-T[E(R-\epsilon)+2R+o(T)]}, \tag{27}$$

which is tightest for $R=\epsilon\to 0$, as $T\to\infty$. Thus, the MSE is lower bounded by an expression whose exponential order is $e^{-E(0)T} = e^{-CT/2}$, as discussed in the Introduction (see also [14]). The same comment applies, of course, to more general moments of the estimation error, $\boldsymbol{E}|\hat{U}-U|^\alpha$, in the range $\alpha \ge 1$ (see also [27, p. 388, Remark 1]). For $0<\alpha<1$, the best choice of R is near $R = C/(1+\alpha)^2$ and the resulting lower bound is of the exponential order of $\exp[-\alpha CT/(1+\alpha)]$.

3.4 Relation to the Joint Source–Channel Excess Distortion Exponent

Note that if we think of the random parameter U as a source variable, and then the modulation–estimation problem is considered as a joint source–channel coding problem,

then our conclusion from Theorem 2 is that *separate source- and channel coding is asymptotically optimum* in our setting: In the modulation scheme analyzed in the proof of Theorem 2, the transmitter first uses a source encoder that quantizes the parameter U, applying a simple uniform scalar quantizer – see also [15], and then maps the quantized version of U into a channel input waveform using a good channel code. The same comment applies to the case where the parameter is a vector $\boldsymbol{U} = (U_1, \ldots, U_d)$, as will be discussed in Subsection 4.1, where the source encoder will quantize each component U_i individually.

It is interesting to contrast this with the results of Csiszár [5] (see also [4]), where exponential rates of probabilities of excess end–to–end distortion between a source vector and its reconstruction vector were studied under a joint source–channel coding setting.[5] In that work, it was argued that, in general, separate source- and channel coding is suboptimum in the error exponent sense (see discussion at the second to the last paragraph in the Introduction of [5] as well as in [4, Introduction] and [7, Problem 5.16, pp. 534–535]). The natural question that arises is how do these two (seemingly contradicting) facts settle, if there is any contradiction. First, observe that there are some differences between our setting and the one in [5]:

1. In our setting, the source variable U is a scalar, namely, it remains of "block–length" 1, when T goes to infinity, whereas in [5] the analogous quantities grow together with a fixed ratio (which is known as the bandwidth expansion factor). Even in Subsection 4.1, where as mentioned earlier, we extend our setting to the case of a vector parameter, $\boldsymbol{U} = (U_1, \ldots, U_d)$, the dimension d will be assumed fixed while $T \to \infty$.

2. As another difference in the asymptotic regime, in our case, the allowed distortion threshold decays exponentially, whereas in [5] it is fixed.

3. For the AWGN with infinite bandwidth, the reliability function is fully known, as opposed to that of a general DMC.

Nonetheless, in spite of these differences, our results can be understood in the framework of [5]. It turns out that while in general, there is no separation theorem for error exponents,

[5] In other words, instead of analyzing the performance of the communication system under the criterion of average distortion, it was analyzed in [5] under the probability that the block distortion would exceed a certain threshold in the large deviations regime.

the parameter modulation–estimation problem considered here is analogous to a special case, where a separation theorem holds true for error exponents nevertheless.

To be more specific, Csiszár's main result in [5] can be presented essentially as follows: The best excess distortion exponent of joint source–channel coding is upper bound by

$$e(D) = \min_R [F(D,R) + E(R)], \tag{28}$$

where

$$F(D,R) = \min_{\{Q':\ R(D,Q') \geq R\}} D(Q'\|Q) \tag{29}$$

is Marton's source coding (excess distortion) exponent of the source Q [12], $R(D,Q')$ being the rate–distortion function of a source Q', and $E(R)$ is the reliability function of the channel. Now, consider the source Q^* that maximizes $R(D,Q)$ (which is the uniform source in many cases, in analogy to our continuous–valued uniform source U). For this source,

$$F(D,R) = \begin{cases} 0 & R \leq R(D,Q^*) \\ \infty & R > R(D,Q^*) \end{cases} \tag{30}$$

This is the case where the entire source space can be fully covered by spheres of noramlized radius D. In this case, the minimization range in the expression of $e(D)$ obviously reduces to the range $R \leq R(D,Q^*)$, where the contribution of the source coding exponent vanishes and hence we are left with $e(D) = E[R(D,Q^*)]$. This can be seen as follows:

$$\begin{aligned} e(D) &= \min_R [F(D,R) + E(R)] \\ &= \min_{R \leq R(D,Q^*)} [0 + E(R)] \\ &= E[R(D,Q^*)]. \end{aligned} \tag{31}$$

We now argue that this is a case where separate source– and channel coding happens to be optimal: If the source sequence space is fully covered by spheres of radius D, the source encoder contributes nothing to the excess distortion event and so, excess distortion may happen only in the event of a channel error whose exponent is $E(R)$, computed at $R = R(D,Q^*)$, which is exactly the above mentioned expression of $e(D)$. Indeed, from the mathematical point of view, the source–channel excess distortion exponent pertaining to separate source- and channel coding, denoted by $e_{sep}(D)$, and given by $\sup_R \min\{F(D,R), E(R)\}$, is also equal to $E[R(D,Q^*)]$ in this case. This is easily shown as follows:

$$e_{sep}(D) = \sup_R \min\{F(D,R), E(R)\}$$

$$= \sup_R \begin{cases} 0 & R \le R(D,Q^*) \\ E(R) & R > R(D,Q^*) \end{cases}$$
$$= E[R(D,Q^*)]. \tag{32}$$

This is clearly analogous to our case: We fully cover the unit interval with small intervals of size $2e^{-RT}$ using a rate–R source code. Similarly, in the d–dimensional case to be described in Subsection 4.1, we perfectly cover the unit cube by boxes of sizes $2e^{-R_1T} \times \ldots \times 2e^{-R_dT}$ using a code of rate $R_1 + \ldots + R_d$.

4 Extensions

In this section, we extend Theorems 1 and 2 in several directions (one at a time). These include the multidimensional case, more general channels, and allowing a variable power that depends on the parameter.

4.1 The Multidimensional Case

The extension to a multidimensional parameter vector is conceptually quite straightforward. Suppose now that the parameter is a vector $\boldsymbol{u} = (u_1, \ldots, u_d) \in [-1/2, +1/2)^d$, which is a realization of a random vector $\boldsymbol{U} = (U_1, \ldots, U_d)$, uniformly distributed over the d–dimensional unit hypercube $[-1/2, +1/2)^d$. Consider now the probability

$$\Pr\left[\bigcup_{i=1}^{d} \left\{|\hat{U}_i - U_i| > e^{-R_iT}\right\}\right].$$

Then, here both in the upper bound and the lower bound, the d–dimensional unit cube is divided by a Cartesian grid with about e^{R_iT} points in each dimension, $i = 1, 2, \ldots, d$, thus a total of $e^{(R_1+R_2+\ldots+R_d)T}$ points, which means an effective rate of $R_1 + R_2 + \ldots + R_d$. More precisely, the lower bound is now given by

$$\Pr\left[\bigcup_{i=1}^{d} \left\{|\hat{U}_i - U_i| > e^{-R_iT}\right\}\right] \ge (1-e^{-\epsilon T})^d \exp\{-T[E(R_1+R_2+\ldots+R_d-\epsilon d)+o(T)]\} \tag{33}$$

since the integration in eq. (13) now becomes d–dimensional. In the upper bound, we are again quantizing and transmitting one of $e^{(R_1+R_2+\ldots+R_d)T}$ orthogonal codewords, the one which represents the corresponding quantization cell. Thus, the probability of the undesired event in question is of the exponential order of $e^{-E(R_1+R_2+\ldots+R_d)T}$. Considering the case $R_i = R$ for all $i \in \{1, 2, \ldots, d\}$ (hence $\sum_i R_i = R \cdot d$), there is then an interesting threshold effect in the dimensionality of the problem: For $R = 0$ (in the weak sense), the exponential

rate of decay of the probability of the large deviations event $\cup_{i=1}^{d}\{|\hat{U}_i - U_i| \geq e^{-0\cdot T}\}$ is essentially $E(0) = C/2$, independently of d. For $R > 0$, the behavior is as follows: As long as

$$d < d_c \stackrel{\triangle}{=} \lfloor C/R \rfloor, \tag{34}$$

the probability of the event $\cup_{i=1}^{d}\{|\hat{U}_i - U_i| \geq e^{-RT}\}$ tends to zero as $T \to \infty$. But when d exceeds d_c and hence the effective rate $R \cdot d$ exceeds C, the probability tends to unity. Thus, d_c is a critical dimension in this sense. This abrupt transition from 0 to 1 in the limiting probability of excess error is another aspect of the threshold effect. In most estimation problems we normally encounter, the estimation performance degrades with the dimensionality (an effect known as the "curse of dimensionality"), but usually the degradation is graceful and not abrupt as here.

All this discussion can be extended, in principle, from Cartesian lattices in the parameter space to general lattices, where the undesired excess error event is defined as the event where the estimated parameter vector falls outside the respective Voronoi cell centered at the true parameter vector. Here, the effective rate to be used as the argument of the reliability function is determined by the normalized logarithm of the ratio between the volume of the source vector space and the volume of a basic cell.

4.2 Other Channels

The assumption of an AWGN channel with unlimited bandwidth was not used very strongly beyond the fact that for this particular channel, the reliability function is fully known for the entire range of rates, $0 \leq R \leq C$. But the reliability function is also known for the Poisson channel with unlimited bandwidth [22], [23]. Here too, the idea would be to first quantize the parameter and then to use a good code for the Poisson channel, that asymptotically achieves the reliability function, e.g., the Wyner code (see also [15]). Similar comments apply also to more general channels in the limit of the infinite bandwidth regime [8].

In the discrete–time case, the reliability function may not be known for the entire range of rates, but it is known for all rates above the critical rate, where it is also achievable by random coding. Moreover, even if the channel is not fully known, we can derive a universal estimator that relies on a universal decoder for memoryless channels (see, e.g., [6] and references therein), on the basis of the proof of the achievability in Theorem 2. But even at rates below the critical rate, where the reliability function is not known, the basic

principle of optimum modulation–estimation using a separation–based scheme continues to hold: First quantize U uniformly and then apply an optimum channel code.

The modification of our results to discrete memoryless channels also enables to handle, at least partially, the case of the AWGN channel with limited bandwidth. This is because the case of limitation to finite bandwidth W is asymptotically equivalent to the discrete memoryless Gaussian channel with $N = 2WT$ channel uses (pertaining to $N = 2WT$ orthonormal basis functions that span the subspace of allowable signals). In this case, $E(R)$ for high rates agrees with the sphere–packing bound, which in the Gaussian band–limited case is given by

$$E_{sp}(R) = \max_{\rho \geq 0} \left\{ \rho W \ln \left[1 + \frac{S}{N_0 W (1+\rho)} \right] - \rho R \right\}. \tag{35}$$

The critical rate beyond which $E_{sp}(R) = E(R)$ is given by

$$\begin{aligned} R_c(W) &= \frac{\partial}{\partial \rho} \left\{ \rho W \ln \left[1 + \frac{S}{N_0 W (1+\rho)} \right] \right\} \Big|_{\rho=1} \\ &= W \left[\ln \left(1 + \frac{S}{2N_0 W} \right) - \frac{1}{2} \cdot \frac{S}{S + 2N_0 W} \right], \end{aligned} \tag{36}$$

where the maximum over ρ in achieved within the interval $[0, 1]$.

4.3 The AWGN Channel With Rayleigh Fading

Another important channel model is the AWGN channel with Rayleigh fading. Here, the signal model is

$$y(t) = a \cdot x(t, u) + n(t), \qquad t \in [0, T) \tag{37}$$

where a is a realization of a Rayleigh random variable A, whose pdf is given by

$$f_A(a) = \frac{a}{\sigma^2} e^{-\frac{a^2}{2\sigma^2}}, \qquad a \geq 0. \tag{38}$$

It is assumed that A is independent of U, as well as of the noise $\{n(t),\ 0 \leq t < T\}$. It is instructive to examine the best achievable behavior of the probability of excess estimation error under this fading model.

For a given $A = a$, the received signal has power $a^2 S$, which implies that the channel capacity is $a^2 S/N_0 = a^2 C$. Correspondingly, the reliability function is given by

$$E_a(R) = \begin{cases} a^2 \frac{C}{2} - R, & 0 \leq R \leq a^2 \frac{C}{4} \\ (a\sqrt{C} - \sqrt{R})^2, & a^2 \frac{C}{4} \leq R \leq a^2 C \\ 0, & R \geq a^2 C \end{cases} \tag{39}$$

Equivalently, if we think of $E_a(R)$ as a function of a parametrized by R, then

$$E_a(R) = \begin{cases} a^2\frac{C}{2} - R, & a \geq 2\sqrt{\frac{R}{C}} \\ (a\sqrt{C} - \sqrt{R})^2, & \sqrt{\frac{R}{C}} \leq a \leq 2\sqrt{\frac{R}{C}} \\ 0, & a \leq \sqrt{\frac{R}{C}} \end{cases} \tag{40}$$

In view of Theorems 1 and 2, averaging the upper and lower bounds on the probability of decoding error given a, would yield respective bounds for the fading channel. For the lower bound, this averaging is legitimate as it corresponds to a receiver that is informed of the realization a of the random variable A. For the upper bound, this is legitimate too since the ML decoder does not depend on (the possibly unknown value of) a in the regime of equal–energy signals considered here.

As before, one should distinguish between the cases $R > 0$ and $R = 0$ (in the strong sense). The following two results are shown in Appendix A. For the case $R > 0$, the probability of excess estimation error is essentially equal (for large T) to the probability of channel outage, which is

$$\Pr\{A \leq \sqrt{R/C}\} = 1 - e^{-R/2\bar{C}}, \tag{41}$$

where $\bar{C} = \sigma^2 C$ designates the average capacity of the channel. In other words, there is no decay as $T \to \infty$. For $R = 0$, the best achievable probability of excess estimation error decays at the rate of $1/T$ rather than exponentially with T.

4.4 Variable Transmission Power

In Section 2, we have restricted the class of modulators in a manner that the power of the transmitted signal, $\{x(t,u),\ 0 \leq t < T\}$, is always S, independently of u. Consider the somewhat broader setting, where the power of $\{x(t,u),\ 0 \leq t < T\}$, denoted $S(u)$, is allowed to depend on u, and we only limit the average power according to

$$\boldsymbol{E}\{S(U)\} = \int_{-1/2}^{+1/2} \mathrm{d}u \cdot S(u) \leq S. \tag{42}$$

We argue that our results apply to this wider class of modulators as well.

Concerning the achievability, we continue to use the same modulator and estimator as in the proof of Theorem 2, where the power is $S(u) = S$ for every u. The proof of Theorem 1, on the other hand, has to be extended to allow variable power. The point is that the proof of Theorem 1 in Section 2 relies heavily on the lower bound on the probability of error in M–ary signal detection, which in [19, Section 3.6.1], is derived under the assumption of

equal–energy signals, and we are not aware of an existing extension of this result to allow sets of signals with different energies, where the limitation is on the average energy only. In Appendix B, we extend the proof of Theorem 1 to accommodate a given average energy constraint, or equivalently, an average power constraint (42). In a nutshell, the intuition is that when some of the signals have higher power and some have lower power, the probability of error is basically dominated by the those with the lower power, which is, of course, smaller than the average S. Thus, variable power signal sets offer no improvement relative to fixed power signal sets in terms of achievable error exponents.

Acknowledgments

I would like to thank Yariv Ephraim for many useful discussions and comments in the course of this work. Interesting discussions with Tsachy Weissman and Yonina Eldar are also acknowledged with thanks.

Appendix A

In this appendix, we derive the results for the fading channel for the case $R > 0$ and the case $R = 0$.

Consider the case $R > 0$ first. Here, there is a positive probability that a would be small enough that the corresponding capacity a^2C would fall below the given R, which is exactly the event of channel outage. This happens with probability

$$\Pr\{A^2C < R\} = \int_0^{\sqrt{R/C}} \mathrm{d}a \cdot \frac{a}{\sigma^2} e^{-a^2/2\sigma^2} = 1 - e^{-R/2\sigma^2 C} = 1 - e^{-R/2\bar{C}}. \tag{A.1}$$

Owing to the discussion in Subsection 3.1, in the event of outage, the probability of excess estimation error is very close to unity, and so, the overall probability of excess estimation error is essentially lower bounded by the outage probability, i.e.,

$$\Pr\{|\hat{U} - U| > e^{-RT}\} \geq [1 - o(T)] \cdot (1 - e^{-R/2\bar{C}}), \tag{A.2}$$

that is, the probability of excess estimation error no longer decays as T grows without bound. Concerning the upper bound, we have from eq. (40)

$$\begin{aligned} \Pr\{|\hat{U} - U| > e^{-RT}\} &\leq \int_0^{\infty} \mathrm{d}a \frac{a}{\sigma^2} e^{-a^2/2\sigma^2} e^{-TE_a(R)} \\ &= \int_0^{\sqrt{R/C}} \mathrm{d}a \cdot \frac{a}{\sigma^2} e^{-a^2/2\sigma^2} + \end{aligned}$$

$$\begin{aligned} & \int_{\sqrt{R/C}}^{2\sqrt{R/C}} \mathrm{d}a \cdot \frac{a}{\sigma^2} e^{-a^2/2\sigma^2} e^{-T(a\sqrt{C}-\sqrt{R})^2} + \\ & \int_{2\sqrt{R/C}}^{\infty} \mathrm{d}a \cdot \frac{a}{\sigma^2} e^{-a^2/2\sigma^2} e^{-T(a^2C/2-R)} \\ = \;& 1 - e^{-R/2\bar{C}} + o(T), \end{aligned} \tag{A.3}$$

where the last line follows from the fact that the above two last integrals, over the ranges $[\sqrt{R/C}, 2\sqrt{R/C})$ and $[2\sqrt{R/C}, \infty)$, both vanish as $T \to \infty$, as can easily be shown. Thus, the lower bound and the upper bound asymptotically coincide.

As for the case $R = 0$ (i.e., Δ fixed, but small), then in view of the derivations in Section 2 (see eqs. (21) and (22)), for a given a, both the upper bound and the lower bound on the probability of excess estimation error probability admit the form $\alpha \cdot Q(a\sqrt{CT\beta})$, where α and β are constants. The respective constants, α and β, pertaining to the upper bound and the lower bound, are different. However, α is just a multiplicative constant, which is of secondary importance here, because we are primarily interested in the rate of decay of both bounds as $T \to \infty$. On the other hand, β is very close to unity in both bounds when Δ is small. Thus, the quantity of interest is basically the expectation of $Q(A\sqrt{CT})$ w.r.t. the randomness of A. We next show that for large T, this quantity is well–approximated by

$$\boldsymbol{E}\{Q(A\sqrt{CT})\} = \int_0^{\infty} \mathrm{d}a \frac{a}{\sigma^2} e^{-a^2/2\sigma^2} Q(a\sqrt{CT}) \approx \frac{1}{4\bar{C}T}, \tag{A.4}$$

that is, the minimum achievable excess estimation error probability decays algebraically rather than exponentially. Using Craig's formula (see, e.g., [17]),

$$Q(x) = \frac{1}{\pi} \int_0^{\pi/2} \mathrm{d}\theta \exp\left(-\frac{x^2}{2\sin^2\theta}\right), \tag{A.5}$$

we have the following:

$$\begin{aligned} \boldsymbol{E}\{Q(A \cdot \sqrt{CT})\} &= \int_0^{\infty} \frac{\mathrm{d}a}{\sigma^2} \cdot a e^{-a^2/2\sigma^2} Q(a \cdot \sqrt{CT}) \\ &= \int_0^{\infty} \frac{\mathrm{d}a}{\sigma^2} \cdot a e^{-a^2/2\sigma^2} \cdot \frac{1}{\pi} \int_0^{\pi/2} \mathrm{d}\theta \exp\left(-\frac{a^2 CT}{2\sin^2\theta}\right) \\ &= \frac{1}{\pi} \int_0^{\pi/2} \mathrm{d}\theta \int_0^{\infty} \mathrm{d}\left(\frac{a^2}{2\sigma^2}\right) \exp\left(-\frac{a^2}{2\sigma^2}\left[1 + \frac{\bar{C}T}{\sin^2\theta}\right]\right) \\ &= \frac{1}{\pi} \int_0^{\pi/2} \frac{\mathrm{d}\theta}{1 + \bar{C}T/\sin^2\theta} \\ &= \frac{1}{\pi} \int_0^{\pi/2} \frac{\mathrm{d}\theta \cdot \sin^2\theta}{\bar{C}T + \sin^2\theta}. \end{aligned} \tag{A.6}$$

The last expression can be upper bounded and lower bounded by bounding the $\sin^2\theta$ term of the denominator by 0 and 1, respectively. Both bounds are well approximated by $1/(4\bar{C}T)$ for large T.

Appendix B

In this appendix, we provide an outline of the extension of Theorem 1 to the variable power case.

For a given u and Δ, consider again the grid $\{u + i\Delta\}_{i=0}^{M-1}$, which is assumed to lie entirely in $[-1/2, +1/2)$, and let $\Delta = 2e^{-RT}$ and $M = e^{(R-\epsilon)T}/2 + 1$, as before. Let $S_{\min} \stackrel{\triangle}{=} \min_u S(u)$ and $S_{\max} \stackrel{\triangle}{=} \max_u S(u)$. We first argue that for modulators whose power function $S(u)$ is continuous (or at least, left- or right–continuous) in the vicinity of its minimum, the assertion of Theorem 1 is rather straightforward in the range $C_{\min} \leq R < C$, where $C_{\min} = S_{\min}/N_0$. The reason is that the grid points, $\{u+i\Delta\}_{i=0}^{M-1}$, where u is near the minimum of the power function (and so are all other grid points, with the above assignment of M and Δ), constitute a signal set whose rate, $R-\epsilon$, is very close to (or even exceeds) its capacity, which is about $C_{\min}$, since all signals in this grid have power near $S_{\min}$. Thus, this grid dominates the probability of error (and hence also the probability of excess estimation error) and it dictates a sub-exponential decay at best, which is trivially lower bounded by the exponent $\exp[-TE(R)]$. In view of this, we shall confine attention throughout to the range of rates $0 < R < C_{\min}$.

Now, consider the partition of the range of powers $[S_{\min}, S_{\max}]$ into small bins of width δ, where δ is assumed to divide $S_{\max} - S_{\min}$. For a given u, let $\mathcal{C}_i$ denote the subset of integers $\{j\}$ for which $S(u + j\Delta)$ falls in the i–th bin, that is,

$$S_{\min} + i\delta \leq S(u + j\Delta) < S_{\min} + (i+1)\delta, \quad i = 0, 1, \ldots, r-1 \tag{B.1}$$

where $r = (S_{\max} - S_{\min})/\delta$. First, observe that

$$P_e(u, \Delta) \geq \sum_{i=0}^{r-1} \frac{|\mathcal{C}_i|}{M} P_e(\mathcal{C}_i), \tag{B.2}$$

where $P_e(\mathcal{C}_i)$ is the probability of error pertaining to the subset of signals $\{x(t, u+j\Delta)\}_{j\in\mathcal{C}_i}$ alone. The reason for the inequality is that error events associated with confusion between pairs of signals that belong to different bins are not counted in the r.h.s. Next, for a given

$\epsilon > 0$, let $\mathcal{I}_\epsilon$ denote the index set $\{i:\ |\mathcal{C}_i| \geq e^{-\epsilon T} M\}$. Then, obviously, $P_e(u,\Delta)$ is further lower bounded by

$$P_e(u,\Delta) \geq \sum_{i\in\mathcal{I}_\epsilon} \frac{|\mathcal{C}_i|}{M} P_e(\mathcal{C}_i). \tag{B.3}$$

Now, let us slightly alter the powers of all signals in $\mathcal{C}_i$ to be $S_i \stackrel{\triangle}{=} S_{\min}+(i+1/2)\delta$, neglecting the effect that this may have on the exponent of $P_e(\mathcal{C}_i)$.[6] Let us denote here the reliability function of a rate–R code with power S by $E(R,S)$, to emphasize the dependence on the power (via the dependence on the capacity). Then, for every $i \in \mathcal{I}_\epsilon$, we have

$$P_e(\mathcal{C}_i) \geq e^{-T[E(R-2\epsilon,S_i)+o(T)]}, \tag{B.4}$$

since the size of $\mathcal{C}_i$ is of the exponential order of at least $e^{(R-2\epsilon)T}$. Also, let us denote

$$\pi_i = \frac{|\mathcal{C}_i|}{\sum_{j\in\mathcal{I}_\epsilon}|\mathcal{C}_j|}. \tag{B.5}$$

Then,

$$P_e(u,\Delta) \geq \sum_{i\in\mathcal{I}_\epsilon} \frac{|\mathcal{C}_i|}{M} \cdot \sum_{i\in\mathcal{I}_\epsilon} \pi_i e^{-T[E(R-2\epsilon,S_i)+o(T)]}. \tag{B.6}$$

As for the first factor on the r.h.s. of (B.6), we have

$$1 = \sum_{i\in\mathcal{I}_\epsilon} \frac{|\mathcal{C}_i|}{M} + \sum_{i\in\mathcal{I}_\epsilon^c} \frac{|\mathcal{C}_i|}{M} \leq \sum_{i\in\mathcal{I}_\epsilon} \frac{|\mathcal{C}_i|}{M} + re^{-\epsilon T}, \tag{B.7}$$

and so, this factor is lower bounded by $(1-re^{-\epsilon T})$. Now, observe that the function $e^{-TE(R-2\epsilon,S)}$ is convex[7] in S for all $T > \sqrt{R}/[2\sqrt{C_{\min}}(\sqrt{C_{\min}}-\sqrt{R})^2]$. It follows then from (B.6) that

$$P_e(u,\Delta) \geq (1-re^{-\epsilon T}) \cdot \exp\left[-TE\left(R-2\epsilon, \sum_{i\in\mathcal{I}_\epsilon}\pi_i S_i\right) + o(T)\right]. \tag{B.8}$$

Next, we need an upper bound on $\sum_{i\in\mathcal{I}_\epsilon}\pi_i S_i$. This is accomplished as follows:

$$\begin{aligned}
\bar{S}(u) &\stackrel{\triangle}{=} \frac{1}{M}\sum_{i=0}^{M-1} S(u+i\Delta) \\
&\geq \sum_{i\in\mathcal{I}_\epsilon} \frac{|\mathcal{C}_i|}{M}\cdot S_i + \sum_{i\in\mathcal{I}_\epsilon^c} \frac{|\mathcal{C}_i|}{M}\cdot S_i - \frac{\delta}{2} \\
&\geq \sum_{i\in\mathcal{I}_\epsilon} \frac{|\mathcal{C}_i|}{M}\cdot S_i - \frac{\delta}{2}
\end{aligned} \tag{B.9}$$

[6] The lower bounds on the probability of error of M equal–energy signals are straightforwardly extended to allow almost equal powers (within $\pm\delta/2$), with only a small degradation in the exponential rate, which depends on δ.

[7] The function $e^{-Tf(x)}$ is convex in $x \in \mathcal{X}$ whenever f is twice differentiable and $T \geq \sup_{x\in\mathcal{X}} f''(x)/|f'(x)|^2$, as can easily be seen from the second derivative of $e^{-Tf(x)}$. An alternative consideration is that for large T, the average of $e^{-Tf(x)}$ is dominated by $e^{-T\inf_{x\in\mathcal{X}} f(x)}$, and that $\inf_{x\in\mathcal{X}} f(x)$ is smaller than the average of $f(x)$ over $\mathcal{X}$.

and so,

$$\sum_{i\in\mathcal{I}_\epsilon}\pi_i S_i \le \frac{\bar{S}(u)+\delta/2}{\sum_{i\in\mathcal{I}_\epsilon}|\mathcal{C}_i|/M} \le \frac{\bar{S}(u)+\delta/2}{1-re^{-\epsilon T}}. \tag{B.10}$$

Thus, from (B.8), we have

$$P_e(u,\Delta) \ge (1-re^{-\epsilon T})\cdot\exp\left[-TE\left(R-2\epsilon,\frac{\bar{S}(u)+\delta/2}{1-re^{-\epsilon T}}\right)+o(T)\right]. \tag{B.11}$$

Finally, we integrate both sides of the last inequality w.r.t. u, in order to relate it to the probability of excess estimation error, as in the proof of Theorem 1. To this end, we first observe the following:

$$\begin{aligned}
\int_{-1/2}^{1/2-(M-1)\Delta}\mathrm{d}u\cdot\bar{S}(u) &= \int_{-1/2}^{1/2-(M-1)\Delta}\mathrm{d}u\cdot\frac{1}{M}\sum_{i=0}^{M-1}S(u+i\Delta)\\
&= \frac{1}{M}\sum_{i=0}^{M-1}\int_{-1/2}^{1/2-(M-1)\Delta}\mathrm{d}u\cdot S(u+i\Delta)\\
&= \frac{1}{M}\sum_{i=0}^{M-1}\int_{-1/2+i\Delta}^{1/2-(M-1)\Delta+i\Delta}\mathrm{d}u\cdot S(u)\\
&\le \frac{1}{M}\sum_{i=0}^{M-1}\int_{-1/2}^{1/2}\mathrm{d}u\cdot S(u)\\
&= \int_{-1/2}^{1/2}\mathrm{d}u\cdot S(u)\\
&\le S,
\end{aligned} \tag{B.12}$$

and therefore, for the above defined assignments of Δ and M, we have:

$$\frac{1}{1-e^{-\epsilon T}}\int_{-1/2}^{1/2-e^{-\epsilon T}}\mathrm{d}u\cdot\bar{S}(u) \le \frac{S}{1-e^{-\epsilon T}}. \tag{B.13}$$

Thus,

$$\begin{aligned}
&\Pr\{|\hat{U}-U|>e^{-RT}\}\\
\ge\ &\int_{-1/2}^{1/2-e^{-\epsilon T}}\mathrm{d}u\cdot P_e(u,2e^{-RT})\\
\ge\ &(1-re^{-\epsilon T})(1-e^{-\epsilon T})\int_{-1/2}^{1/2-e^{-\epsilon T}}\frac{\mathrm{d}u}{1-e^{-\epsilon T}}\cdot\exp\left[-TE\left(R-2\epsilon,\frac{\bar{S}(u)+\delta/2}{1-re^{-\epsilon T}}\right)+o(T)\right]\\
\ge\ &(1-re^{-\epsilon T})^2\exp\left[-TE\left(R-2\epsilon,\frac{S+\delta/2}{(1-re^{-\epsilon T})(1-e^{-\epsilon T})}\right)+o(T)\right]\\
=\ &(1-re^{-\epsilon T})^2e^{-T[E(R,S)+o'(T)]},
\end{aligned} \tag{B.14}$$

where in the last line, $o'(T)$ means another function (other than $o(T)$ of the previous lines) that tends to 0 as $T\to\infty$, which is obtained by letting ϵ and δ tend to zero as $T\to\infty$ at the appropriate rates (e.g., $\epsilon=\delta=1/\sqrt{T}$).

References

[1] K. L. Bell, Y. Steinberg, Y. Ephraim, and H. L. van Trees, "Extended Ziv—Zakai lower bounds for vector parameter estimation," *IEEE Trans. Inform. Theory*, vol. 43, no. 2, pp. 626–637, March 1997.

[2] Z. Ben–Haim and Y. C. Eldar, "A lower bound on the Bayesian MSE based on the optimal bias function," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5179–5196, November 2009.

[3] D. Chazan, M. Zakai, and J. Ziv, "Improved lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT–21, no. 1, pp. 90–93, January 1975.

[4] I. Csiszár, "Joint source–channel error exponent," *Problems of Control and Information Theory*, vol. 9, no. 5, pp. 315–328, 1980.

[5] I. Csiszár, "On the error exponent of source-channel transmission with a distortion threshold," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 6, pp. 823–828, November 1982.

[6] M. Feder and N. Merhav, "Universal composite hypothesis testing: a competitive minimax approach," *IEEE Trans. Inform. Theory*, special issue in memory of Aaron D. Wyner, vol. 48, no. 6, pp. 1504–1517, June 2002.

[7] R. G. Gallager, *Information Theory and Reliable Communication*, New York, Wiley 1968.

[8] R. G. Gallager, "Energy limited channels: coding, multiaccess, and spread spectrum," LIDS Report, LIDS–P–1714, M.I.T., November 1988.

[9] R. M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, 1990.

[10] A. D. M. Kester and W. C. M. Kallenberg, "Large deviations of estimators," *Ann. Statist.*, vol. 14, no. 2, pp. 848–664, 1986.

[11] E. L. Lehmann, *Theory of Point Estimation*, John Wiley & Sons, 1983.

[12] K. Marton, "Error exponent for source coding with a fidelity criterion, " *IEEE Trans. Inform. Theory*, vol. IT–20, no. 2, pp. 197–199, March 1974.

[13] N. Merhav, "Threshold effects in parameter estimation as phase transitions in statistical mechanics," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 7000–7010, October 2011.

[14] N. Merhav, "Data processing inequalities based on a certain structured class of information measures with application to estimation theory," submitted to *IEEE Trans. Inform. Theory*, September 2011. [http://arxiv.org/pdf/1109.5351.pdf]

[15] A. No, K. Venkat, and T. Weissman, "Joint source–channel coding of one random variable over the Poisson channel," submitted to *ISIT 2012*, February 2012.

[16] S. Sherman, "Non-mean-square error criteria," *IRE Trans. Inform. Theory*, pp. 125–126, September 1958.

[17] C. Tellambura and A. Annamalai, "Derivation of Craig's formula for Gaussian probability function," *Electronic Letters*, vol. 35, no. 17, pp. 1424–1425, August 19, 1999.

[18] H. van Trees, *Detection, Estimation, and Modulation Theory*, Part I, New York, John Wiley & Sons, 1968.

[19] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw–Hill, 1979.

[20] A. J. Weiss, *Fundamental Bounds in Parameter Estimation*, Ph.D. dissertation, Tel Aviv University, Tel Aviv, Israel, June 1985.

[21] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons, 1965. Reissued by Waveland Press, 1990.

[22] A. D. Wyner, "Capacity and error exponent for the direct detection photon channel – part I," *IEEE Trans. Inform. Theory*, vol. 34, no. 6, pp. 1449–1461, November 1988.

[23] A. D. Wyner, "Capacity and error exponent for the direct detection photon channel – part II," *IEEE Trans. Inform. Theory*, vol. 34, no. 6, pp. 1462–1471, November 1988.

[24] A. D. Wyner and J. Ziv, "On communication of analog data from a bounded source space," *Bell System Technical Journal*, vol. 48, no. 10, pp. 3139–3172, December 1969.

[25] M. Zakai and J. Ziv, "On the threshold effect in radar range estimation," *IEEE Trans. Inform. Theory*, vol. IT–15, pp. 167–170, January 1969.

[26] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," in: *Information Theory New Trends and Open Problems*, edited by G. Longo, Springer-Verlag, 1975, pp. 87–123.

[27] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Transactions on Information Theory*, vol. IT–15, no. 3, pp. 386–391, May 1969.

[28] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Trans. Inform. Theory*, vol. IT–19, no. 3, pp. 275–283, May 1973.