The "hallucination" bound for the BSC

Anant Sahai
Wireless Foundations, Dept. of EECS
University of California at Berkeley
email: sahai@eecs.berkeley.edu

Stark C. Draper
Dept. of Electrical and Computer Engr.
University of Wisconsin at Madison
email: sdraper@ece.wisc.edu

***Abstract*—Though the schemes are different, both Horstein's and Kudryashov's non-block strategies for communication with feedback over the binary-symmetric channel asymptotically achieve the identical reliability function (error exponent); a function that displays some curious features. For positive rates their reliability function is strictly larger than Burnashev's and transitions discontinuously at the channel capacity from a strictly positive value to zero. The purpose of this paper is to connect this reliability function to familiar coding contexts and to demonstrate that it provides an upper bound on the error exponents achievable in these contexts. We first show that this function gives an upper bound on the *minimum* probability of decoding error across codewords in a block-coding context with (or without) feedback.We then show that the same reliability function also gives an upper bound on the maximum probability of *bit* error in a non-block "streaming" context where noiseless feedback is available and the destination is (occasionally) allowed to declare erasures (per Forney). The basic insight underlying the bound leads to the moniker the "hallucination" bound.**

I. Introduction

Horstein [6] and Kudryashov [7] present coding strategies that attain seemingly impossibly high error exponents. While the authors present quite different coding strategies and define their error exponents slightly differently, they share three common features. First, they consider non-block communication in which the relevant probability of error is at the bit rather than block level. Second, the relevant "length" is viewed in an average sense rather than being a fixed number. Third, both strategies rely on the availability of noiseless feedback.

Recently, we have achieved the same error exponents in [2] by marrying the fixed-delay setting of [8] to the errors-and-erasure perspective of [4]. The source model we assume is central to understanding our results. We assume that data to be communicated arises as a stream generated in real time at the source (e.g. voice, video, or sensor measurements) and is useful to the destination in finely grained increments (e.g. a few milliseconds of voice, a single video frame, etc.) at some specified post-generation delay. The acceptable end-to-end delay is determined by the application and can often be much larger than the natural granularity of the information being communicated. This model differs from the traditional model – better matched to block-coding – in which information arises in large bursts and each burst needs to be received by the destination before the next burst is available at the source.

Rather than worrying what the appropriate granularity of information should be, we specify the problem at the level of individual bits. Message bits arrive at the encoder regularly at a spacing of $\frac{1}{R}$ (possibly non-integer) channel uses. There is a *fixed* latency requirement of Δ channel uses between when the bit arrives at the encoder and when it is required at the destination. However, rather than being forced to commit to a value for the bit as is done in [8], the decoder has the option of declaring an erasure as in [4]. When the bit-erasure probability is constrained to be kept below some (small, but positive) target ϵ, it is shown in [2] that the probability of undetected bit-error can be made to drop with Δ asymptotically as fast the error-exponents given in [6], [7], even if the feedback link connecting the decoder to the encoder is itself a noisy channel.

Even with noiseless feedback there are no known upper-bounds on the undetected-error exponent with delay in the non-block setting given an erasure option. Consequently, it is unclear whether even higher error exponents are attainable. The goal of this paper is to show that the Horstein/Kudryashov exponents are the best possible for the memoryless binary-symmetric channel with crossover probability p (the p-BSC).

II. The basic idea and outline of the paper

To illustrate the core tension in the errors-and-erasure context, consider the following coarse bound. First, conceive of the noise process in the p-BSC in the following manner. Each noise sample is drawn i.i.d. from a compound probability distribution. With probability $1-2p$ the noise sample equals zero and with probability $2p$ the noise sample is Bernoulli-(0.5). Thinking of the BSC noise process in this way makes it apparent that, over any Δ channel uses, with a probability $(2p)^{\Delta}$ the noise sequence is i.i.d. Bernoulli-(0.5) and independent of the channel input. With this probability the length-Δ noise sequence is a one-time-pad and the channel output is statistically independent of the input.

Conditioned on this event, the channel outputs — being fair coin tosses — will likely turn out to be self-evident gibberish that the decoder will interpret as an erasure. However, there is a small probability that the decoder will "hallucinate" that decoding is progressing normally and will make a decoding error. At least one output sequence must correspond to "normal progress" else the message bit that arrived Δ channel uses before would never be decoded. Conditioned on the noise being a one-time-pad, all output strings are equally likely. This lower bounds the probability of undetected error (not an erasure) as $(2p)^{\Delta}2^{-\Delta} = p^{\Delta} = 2^{-\Delta \log \frac{1}{p}}$, giving the upper bound of $\log \frac{1}{p}$ on the error exponent. However, this bound does not depend on the data rate and therefore cannot be tight.

The above coarse discussion reveals the central tension that must be reflected in the hallucination bound. To minimize the probability of error, the decoder should be very strict in terms of categorizing received strings as being "normal progress." But, the need to keep the probability of erasure moderately small forces the decoder to accept at least a certain number of "typical" noise patterns as being compatible with normal progress for each possible message sequence. In the coarse bound just provided "typicality" is in the strictest sense – at least one sequence must decode. A positive-rate code mush have additional received strings that are compatible with normal progress. The higher the rate of the code, the more likely will be hallucinations.

To turn this idea into a bound, we proceed in stages. For clarity of exposition we restrict discussion to the BSC — the simplest possible channel that still allows for the possibility of hallucination. (Hallucinations cannot occur on erasure channels.) The block-coding version of the bound is derived in Section III. This shows quantitatively how the data rate impacts the error exponent. The case of streaming data with an end-to-end delay constraint is considered in Section IV. A pair of key lemmas bounding the probability of hallucination in terms of the number of candidate hallucinations and vice-versa are proved in the appendix. Conclusions are given in Section V.

III. Block Coding

In this section we connect the error-exponents approached asymptotically by Horstein [6] and Kudryashov [7] to a block-coding context. We show that their exponents provide lower bound on the *best-case* probability of (undetected) error across all messages. This result holds whether we consider the errors-and-erasure context or decoding without an erasure option.

In Appendix A we prove the following helper lemma that lower bounds the probability of a set in terms of a lower bound on the size of the set.

Lemma 3.1: Let d denote some number of channel uses of a p-BSC where $0 < p < \frac{1}{2}$. Consider any subset of output strings $\mathcal{D} \subseteq \{0,1\}^d$ where $|\mathcal{D}| \geq 2^{dE'}$ for some $0 < E' < 1$. Then, for any possible code function $\mathcal{F} = (f_1, f_2(y_1), \ldots, f_d(y^{d-1}))$, $\Pr(Y^d \in \mathcal{D}) \geq 2^{-dE^+_{hal}(E',d)}$ where

$$E^+_{hal}(E', d) = D(q'||p) + \frac{\log(d+1) + \log\frac{1}{p} + \ln 2}{d} \quad (1)$$

and $q' > \frac{1}{2}$ is the unique solution to

$$H_B(q') = E' - \frac{\log(d+1)}{d}. \quad (2)$$

$H_B(\cdot)$ is the binary entropy function and $D(\cdot||\cdot)$ is the binary divergence. If $E' < \frac{\log(d+1)}{d}$, then choose $q' = 1$.

Remark: As $d \to \infty$, the solution q' to (2) approaches the high-valued branch of the inverse binary entropy function $H_B^{-1}(E')$, and $E^+_{hal}(E')$ approaches $D(H_B^{-1}(E')||p)$. Lemma 3.1 is a central tool in proving the following result.

Theorem 3.1: Consider a sequence (in k) of increasing length-Δ_k block-codes with noiseless feedback that have $2^{\Delta_k R}$ messages each. Assume that for each code, each message is correctly decoded over a p-BSC ($0 < p < 0.5$) with an increasing probability of at least $1 - \epsilon_k$ where $\lim_{k\to\infty} \epsilon_k = 0$.

Then the best-case probability of undetected error has an exponent upper-bounded by:

$$\limsup_{k\to\infty} -\frac{1}{\Delta_k} \log\left[\min_m \Pr\left[\widehat{M} \neq m | M = m\right]\right] \leq E_{hal}(R) \quad (3)$$

where the "hallucination" exponent is defined as

$$E_{hal}(R) = D\Big(H_B^{-1}(R + H_B(p)) \,||p\Big), \quad (4)$$

and $H_B^{-1}(\cdot)$ is the branch of the inverse binary entropy function that returns values $\geq \frac{1}{2}$.

Remark: The condition $\lim_{k\to\infty} \epsilon_k = 0$ prevents all output strings being assigned to the same decoding region, which would not lead to a useful code.

Proof: Since the maximal probability of error is tending to 0, $R \leq C = 1 - H_B(p)$ and $H_B^{-1}(R + H_B(p))$ is well defined.

For any $\epsilon > 0$, consider the set of ϵ-weakly-typical noise:

$$\mathcal{T}_{\Delta_k} = \left\{ z^{\Delta_k} | 2^{-d\{H_B(p)+\epsilon\}} \leq p(z^{\Delta_k}) \leq 2^{-d\{H_B(p)-\epsilon\}} \right\}.$$

Let $\delta_k = 1 - \Pr(\mathcal{T}_{\Delta_k})$. Clearly $\lim_{k\to\infty} \delta_k = 0$ and $\Pr[\mathcal{T}_{\Delta_k}] \leq \sum_{z^{\Delta_k} \in \mathcal{T}_{\Delta_k}} 2^{-d\{H_B(p)-\epsilon\}}$ means $|\mathcal{T}_{\Delta_k}| \geq (1 - \delta_k) 2^{d\{H_B(p)-\epsilon)\}}$.

For any given message m, there is a one-to-one and onto mapping between a channel noise sequence z^{Δ_k} and channel outputs y^{Δ_k}. Consider the decoding region $\mathcal{D}_m$ for message m. Because decoding must be correct most of the time, $\mathcal{D}_m$ must include a large part of the channel outputs corresponding to the typical noise set. For each message m, we bound the decoding region size $|\mathcal{D}_m|$ using

$$\begin{aligned} \Pr[\mathcal{D}_m \cap \mathcal{T}_{\Delta_k}] &= 1 - \Pr[\mathcal{D}_m^c \cup \mathcal{T}_{\Delta_k}^c] \\ &\geq 1 - \Pr[\mathcal{D}_m] - \Pr[\mathcal{T}_{\Delta_k}^c] \\ &\geq 1 - \epsilon_k - \delta_k, \quad (5) \\ \Pr[\mathcal{D}_m \cap \mathcal{T}_{\Delta_k}] &\leq \sum_{z^{\Delta_k} \in \mathcal{D}_m} \Pr\left[z^{\Delta_k} | z^{\Delta_k} \in \mathcal{T}_{\Delta_k}\right] \\ &\leq |\mathcal{D}_m| 2^{-d(H_B(p)-\epsilon)}. \quad (6) \end{aligned}$$

Together the relations give $|\mathcal{D}_m| \geq (1 - \epsilon_k - \delta_k) 2^{\Delta_k (H_B(p)-\epsilon)}$.

Since decoding regions are disjoint, the cardinality of the union of the decoding regions corresponding to the other $2^{\Delta_k R} - 1$ messages

$$\begin{aligned} \Big| \bigcup_{i \neq m} \mathcal{D}_i \Big| &\geq (2^{\Delta_k R} - 1)(1 - \epsilon_k - \delta_k) 2^{\Delta_k \{H_B(p)-\epsilon\}} \\ &> 2^{\Delta_k R - 1} 2^{\Delta_k \left\{ H_B(p) - \epsilon + \frac{\log(1-\epsilon_k-\delta_k)}{\Delta_k} \right\}} \\ &= 2^{\Delta_k \left\{ R + H_B(p) - \epsilon - \frac{1 - \log(1-\epsilon_k-\delta_k)}{\Delta_k} \right\}} \end{aligned}$$

There exists a $\underline{k}$ large enough so that for all $k > \underline{k}$, the term $\frac{1-\log(1-\epsilon_k-\delta_k)}{\Delta_k} \leq \epsilon$ and so

$$\Big| \bigcup_{i \neq m} \mathcal{D}_i \Big| > 2^{\Delta_k \{R + H_B(p) - 2\epsilon\}}$$

Applying Lemma 3.1 tells us that the probability of decoding to one of these other messages is at least:

$$\Pr\left[\widehat{M} \neq m | M = m\right] \geq 2^{-\Delta_k E^+_{hal}(R+H_B(p)-2\epsilon,\, \Delta_k)}$$
$$= 2^{-\Delta_k \left\{ D\left(H_B^{-1}(H_B(p)+R-2\epsilon)||p\right) + \frac{\log(\Delta_k+1)+\log\frac{1}{p}+\ln 2}{\Delta_k} \right\}}$$

Taking $\log$ of both sides and dividing by $\Delta_k \to \infty$, gives an exponent of $D(H_B^{-1}(H_B(p)+R-2\epsilon)||p)$. Since $\epsilon > 0$ was arbitrary, we get the desired bound. □

This bound reveals the combined impact of the rate and the crossover probability on the best-case error exponent. It is interesting to evaluate the hallucination bound at $R = C = 1 - H_B(p)$. There it gives an exponent of $D(\frac{1}{2}||p) > 0$. This is interesting because it is the same as the sphere-packing bound at rate $R = 0$. Evaluating the hallucination bound at $R = 0$ gives an exponent of $D(H_B^{-1}(H_B(p))||p) = D(1-p||p)$. This is the same as the Burnashev exponent [1] at $R = 0$.

IV. Streaming case with delay

The hallucination bound in the block-coding context corresponds to the non-standard problem of bounding the *best-case* probability of (undetected) block error across messages. In an errors-and-erasures context this bound holds when also subject to a constraint on erasure probability. This contrasts with the more familiar *worst-case* probability of error. Subject to a constraint on erasure probability Burnashev's result [1] bounds the worst-case probability of error (and is achievable when feedback is available). It turns out that the hallucination bound governs the worst-case probability of *bit error* with end-to-end delay subject to a constraint on the erasure probability. While space constraints prevent us from providing the full proof, we state the main theorem and provide one of the main technical lemmas.

Prior to stating the main result we introduce a technical definition that captures the non-block nature of the codes we consider. The important aspect to capture is that the code typically has the channel inputs depend on the data-bits soon after they arrive rather than usually waiting in a buffer for some time proportional to the end-to-end delay — as they would in a block-code. As mentioned above, block-coding with feedback in an errors-and-erasure context is governed by Burnashev's bound [1]. Therefore, to really understand what Horstein and Kudryashov's bounds are telling us we need to restrict attention to non-block codes. We conjecture, however, that codes not meeting the specific technical condition stated can only have worse performance than those covered by this definition.

Definition 4.1: A sequence (in k) of rate-R delay-Δ_k codes with noiseless feedback is called *guess-friendly* if there exists a positive constant K_1 (the same for all k) such that for all n and all l such that $n - \Delta_k < l < n$

$$\lim_{k\to\infty} \Pr\left[\widehat{B}^l_{l-1}(n) \neq B^l_{l-1}\right] \leq \frac{K_1}{n-l}. \tag{7}$$

where

$$b_i^j \in \left\{0,1\right\}^{\lfloor jR \rfloor}_{\lfloor iR \rfloor} \text{ and } \hat{b}_i^j \in \left\{*,0,1\right\}^{\lfloor jR \rfloor}_{\lfloor iR \rfloor}$$

are, respectively, realizations of the rate-R streaming data sequence B_i^j and the rate-R data sequence estimate $\widehat{B}_i^j$, and where $*$ indicates the erasure option. For conciseness we define $b_i^j = \hat{b}_i^j = \emptyset$ if $j \leq i$. We sometimes need to indicate explicitly at what time the estimate is made, e.g., $\hat{b}_i^j(t)$ is the estimate made at time t.

In [3] Definition 4.1 is relaxed somewhat to the concept of *coarse-guess-possible* codes using the ideas of list decoding. Space limitations herein prevent us from stating that relaxation. Intuitively, the concept of *guess-friendly* codes is designed to capture the idea that the conditional entropy of message bits given the channel outputs must drop reasonably fast even before their individual deadlines expire. The schemes of Horstein [6] and Kudryashov [7] are both guess-friendly.

Theorem 4.1: Consider a sequence of rate-R increasing delay-Δ_k coarse-guess-possible codes with noiseless feedback. Assume that for each code and all possible data-bits each bit is correctly decoded over a p-BSC ($0 < p < 0.5$) with an increasing probability of at least $1 - \frac{1}{\Delta_k^2}$, so the probability of declared bit-erasure drops as $o(\frac{1}{\Delta_k})$. Then,

$$\limsup_{k\to\infty} -\frac{1}{\Delta_k} \log\Big(\max_{t, b^{t+\Delta_k}} \Pr\left[\widehat{B}^t_{t-1}(t+\Delta_k) \neq B^t_{t-1} \,\middle|\, B^{t+\Delta_k} = b^{t+\Delta_k}\right]\Big) \leq E_{hal}(R) \tag{8}$$

where $E_{hal}(R)$ is as defined in (4).

The full proof is given in [3]. However, the intuition can be summarized as follows and uses techniques from [8]. First, the code is treated like an appropriately long block-code and the total volume of channel outputs that correspond to "normal decoding" is bounded below using typicality arguments and the assumption that erasures happen infrequently. Next, the natural tree-structure of these decoding regions is exploited to upper-bound the total volume of channel outputs assuming that bit errors are very rare. These two ways of bounding the total volume of channel outputs are then compared to establish the bound on the error exponent.

A main lemma needed to prove Thm. 4.1 is stated next, and its proof is given in Appendix A. This lemma upper bounds the size of a set given only an upper bound on that set's probability. For a single code-function, this can be obtained by considering the contra positive of Lemma 3.1. Because our interest is in codes with feedback, the past channel outputs act like a randomization of the code function for any given message.

Lemma 4.1: Let d denote some number of channel uses of a p-BSC where $0 < p < \frac{1}{2}$. Consider a list (indexed by i) of K sets $\mathcal{D}_i \subseteq \{0,1\}^d$ of possible channel outputs. Let p_i be a corresponding list of positive real numbers in a tight band (within a tolerance factor of $2^{-d\epsilon}$) so that for all i, j:

$$2^{-d\epsilon} \leq \frac{p_i}{p_j} \leq 2^{+d\epsilon} \tag{9}$$

and let the corresponding code functions be denoted $\mathcal{F}_i$. Let

$q^* \geq \frac{1}{2} + \frac{1}{d}$ be the unique solution to:

$$D(q^*||p) = E - \epsilon + \frac{\log(d+1)}{d}. \tag{10}$$

Then, as long as $d \geq \max\left\{\frac{2}{q^*}, \frac{1}{1-q^*}\right\}$, if the weighted average probability

$$\frac{1}{\sum_{i=1}^K p_i} \sum_{i=1}^K p_i \Pr\left[Y^d \in \mathcal{D}_i \text{ using } \mathcal{F}_i\right] \leq 2^{-dE} \tag{11}$$

then the average size of the sets is also small:

$$\frac{\sum_{i=1}^K |\mathcal{D}_i|}{K} \leq 2^{d\left\{H_B(q^*) + \frac{1+\log(q^*/(1-q^*))+\log(d+1)}{d}\right\}}. \tag{12}$$

When $D(1||p) + \epsilon - \frac{\log(d+1)}{d} > E > D(\frac{1}{2}||p) + \epsilon - \frac{\log(d+1)}{d}$, $q^* = \frac{1}{2}$ if E is too small, and $q^* = 1$ if E is too big.

As $d \to \infty$ and $\epsilon \to 0$, the exponent in (12) approaches $H_B(q^*)$ where the error exponent E approaches $D(q^*||p)$. Thus, the bound here asymptotically matches the bound in Lemma 3.1.

V. Conclusion

The very high error exponents of Horstein [6] and Kudryashov [7] were originally stated in the context of bit-error probability for non-block coding with an average delay constraint. However, they can be more easily understood if we adopt Forney's errors-and-erasures perspective and focus on the probability of undetected error. For that problem, the hallucination bound stated herein provides the natural converse which shows that no larger error exponents are possible.

In the block-coding context, the hallucination bound complements the traditional sphere-packing bound. Whereas the traditional sphere-packing bound tells how bad the maximum or average probability of error must be, the hallucination bound considers the minimum probability of error. This allows the code to implicitly have a "favorite message" that is extra-reliable. The reason this error exponent can be bounded away from zero even at capacity is that the problem at the decoder becomes one of multi-stage hypothesis testing. The decoder first asks "is this my favorite message" before trying to decode to any other message. Since the favorite message is unique, this first stage can have a positive error exponent at any rate for the overall code.

When feedback is allowed in the block-coding context, the encoder can optimize its channel input distribution based on how the channel is behaving. This logic leads to the Haroutunian bound [5]. When feedback is allowed in the streaming context the encoder has the option of changing its message mid-stream to its favorite message. As is developed in [2] when erasure are permitted the favorite message can be assigned the role of warning the decoder that the channel is behaving badly and instructing it to declare an erasure. The hallucination bound shows that the performance thereby achieved is the best possible.

Appendix

Proof of Lemma 3.1: The binary entropy function $H_B(\cdot)$ is continuous in its argument and so the second part of the lemma is clearly true since $\lim_{d\to\infty} \frac{\log(d+1)}{d} = 0$ as is $\lim_{d\to\infty} \frac{\log(d+1)+\log\frac{1}{p}+\ln 2}{d} = 0$.

For the first part, it is useful to consider the BSC as a modulo 2 additive noise channel with noise Z_i drawn i.i.d. Bernoulli-p. For any code function the mapping from channel noise z^d to channel outputs y^d is one-to-one and onto. Thus, the probability $\Pr\left[Y^d \in \mathcal{D}\right]$ must be larger than the minimum probability of a set $\mathcal{D}_Z$ of channel noise satisfying $|\mathcal{D}_Z| \geq 2^{dE'}$ giving

$$\Pr\left[Y^d \in \mathcal{D}\right] \geq \min_{\mathcal{D}_Z \text{ s.t. } |\mathcal{D}_Z| \geq 2^{\mathrm{dE}'}} \Pr\left[Z^d \in \mathcal{D}_Z\right]$$

Define $\mathcal{D}_Z^{\bar{q}}$ to be the set of all length-d binary strings with average weight greater than or equal to $\bar{q} \geq \frac{1}{2}$. Then

$$\min_{\mathcal{D}_Z \text{ s.t. } |\mathcal{D}_Z| \geq 2^{\mathrm{dE}'}} \Pr\left[Z^d \in \mathcal{D}_Z\right] \quad \geq \quad \Pr\left[Z^d \in \mathcal{D}_Z^{\bar{q}}\right]$$

for any value of $\bar{q}$ chosen so that

$$\left|\mathcal{D}_Z^{\bar{q}}\right| \leq 2^{dE'}. \tag{13}$$

This holds because the least likely strings for a Bernoulli process with $p < \frac{1}{2}$ are the ones with the largest weights.

Set $\bar{q} = \frac{\lceil dq' \rceil}{d}$ where q' is defined in (2). To see that this choice of $\bar{q}$ satisfies (13) observe that

$$\begin{aligned} |\mathcal{D}_Z^{\bar{q}}| &< (d+1)|\mathcal{T}_{\bar{q}}| \leq (d+1)2^{dH_B(\bar{q})} \\ &\leq (d+1)2^{dH_B(q')} = 2^{dE'} \end{aligned}$$

where $\mathcal{T}_{\bar{q}}$ is the set of sequences of type $\bar{q}$. The first inequality holds since type-classes closer to 0.5 have more members than those further away and there are $d+1$ binary types in total. The second inequality holds because since $q' > 0.5$, $\bar{q} > q' > 0.5$ and so $H_B(\bar{q}) < H_B(q')$. The third inequality holds by the definition of q' in (2).

Finally, to lower bound the probability itself, observe that $q' \leq \bar{q} \leq \min\{1, q' + \frac{1}{d}\}$ and so:

$$\begin{aligned} &\Pr\left[Z^d \in \mathcal{D}_Z^{\bar{q}}\right] > \Pr\left[Z^d \in \mathcal{T}_{\bar{q}}\right] \\ &\quad \geq \frac{1}{d+1} 2^{-dD(\bar{q}||p)} \\ &\quad = \frac{1}{d+1} 2^{-d\left\{\bar{q}\log\frac{\bar{q}}{p} + (1-\bar{q})\log\frac{1-\bar{q}}{1-p}\right\}} \\ &\quad \geq \frac{1}{d+1} 2^{-d\left\{(q'+\frac{1}{d})\log\frac{\min(1,q'+\frac{1}{d})}{p} + (1-q')\log\frac{1-q'}{1-p}\right\}} \\ &\quad \geq \frac{1}{d+1} 2^{-d\left\{\frac{1}{d}(\log\frac{1}{p}) + q'(\log\frac{q'}{p} + \log\frac{q'+\frac{1}{d}}{q'}) + (1-q')\log\frac{1-q'}{1-p}\right\}} \\ &\quad = \frac{1}{d+1} 2^{-d\left\{D(q'||p) + \frac{1}{d}(\log\frac{1}{p}) + q'\log(1+\frac{1}{q'd})\right\}} \\ &\quad \geq \frac{1}{d+1} 2^{-d\left\{D(q'||p) + \frac{1}{d}(\log\frac{1}{p} + \ln 2)\right\}} \\ &\quad = 2^{-d\left\{D(q'||p) + \frac{1}{d}(\log(d+1) + \log\frac{1}{p} + \ln 2)\right\}} \end{aligned}$$

which proves the desired result. □

Proof of Lemma 4.1: Since $D(\cdot||p)$ is continuous in its first argument the second part of the lemma holds for reasons identical to Lemma 3.1.

For the first part, notice that the bound is trivial whenever $q^* = \frac{1}{2}$. Similarly $E > D(1||p)$ cannot happen unless the set $\mathcal{D}$ is empty so just concentrate on the case $q^* > \frac{1}{2}$.

As in Lemma 3.1, for each code function, the one-to-one mapping from binary channel noise z^d to channel outputs y^d justifies solely considering sets $\mathcal{D}_{Z,i}$ of channel noise sequences instead of the original sets of channel outputs. We need to bound the total size $\sum |\mathcal{D}_{Z,i}|$, subject to

$$\frac{1}{\sum_{i=1}^K p_i} \sum_{i=1}^K p_i \Pr\left[Z^d \in \mathcal{D}_{Z,i}\right] \leq 2^{-dE}. \tag{14}$$

By calculating the average set size subject to a looser constraint we upper bound the total size of the sets. The constraint (14) is relaxed by noticing $p_{\min} \leq p_i \leq p_{\max}$ where $p_{\min} \geq 2^{-d\epsilon} p_{\max}$. Since $2^{-d\epsilon} p_{\max} \leq p_i \leq p_{\max}$,

$$\begin{aligned}
&\frac{1}{\sum_{i=1}^K p_i} \sum_{i=1}^K p_i \Pr\left[Z^d \in \mathcal{D}_{Z,i}\right] \\
&\quad \geq \frac{1}{\sum_{i=1}^K p_{\max}} \sum_{i=1}^K p_i \Pr\left[Z^d \in \mathcal{D}_{Z,i}\right] \\
&\quad \geq \frac{1}{\sum_{i=1}^K p_{\max}} \sum_{i=1}^K 2^{-d\epsilon} p_{\max} \Pr\left[Z^d \in \mathcal{D}_{Z,i}\right] \\
&\quad = 2^{-d\epsilon} \frac{1}{K} \sum_{i=1}^K \Pr\left[Z^d \in \mathcal{D}_{Z,i}\right].
\end{aligned}$$

Therefore the constraint (14) implies

$$\frac{1}{K} \sum_{i=1}^K \Pr\left[Z^d \in \mathcal{D}_{Z,i}\right] \leq 2^{-d(E-\epsilon)}. \tag{15}$$

Since the probability of a noisy sequence is strictly decreasing in its weight, it suffices to consider $\mathcal{D}_{Z,i}$ satisfying the property that if $z^d \in \mathcal{D}_{Z,i}$, then any $\tilde{z}^d$ with weight strictly larger than z^d is also in $\mathcal{D}_{Z,i}$. If a set does not satisfy this property, sequences can be swapped in/out so that the set remains the same size but does satisfy this property. Meanwhile, the total probability after the swaps is strictly smaller and hence still satisfies the constraint (15).

Thus, for each set $\mathcal{D}_{Z,i}$ there must be a minimal weight and corresponding type q_i. Each string z^d of type q_i has probability $2^{-dH_B(q_i)}$ and so its contribution to the un-normalized sum in (15) is $2^{-dH_B(q_i)}$. These $2^{-dH_B(q_i)}$ can be assumed to be not far apart across i since otherwise, the average probability could be reduced by swapping a string from one $\mathcal{D}_{Z,i}$ to another $\mathcal{D}_{Z,j}$. The only way they could not be identical is if $\underline{q} = \min_i q_i < \max_i q_i = \bar{q}$ were adjacent types and the type $\bar{q}$, and all less likely types, are completely full in all the $\mathcal{D}_{Z_i}$. So,

$$\mathcal{D}_Z^{\bar{q}} \subseteq \mathcal{D}_{Z,i} \subseteq \mathcal{D}_Z^{\underline{q}} \tag{16}$$

Thus, (15) can be further relaxed to

$$\Pr\left[Z^d \in \mathcal{D}_Z^{\bar{q}}\right] \leq 2^{-d(E-\epsilon)} \tag{17}$$

and then

$$\begin{aligned}
\sum_{i=1}^K |\mathcal{D}_i| &\leq K|\mathcal{D}_Z^{\underline{q}}| \leq K(d+1)2^{dH_B(\underline{q})} \\
&\leq K 2^{d\left\{H_B(\underline{q}) + \frac{\log(d+1)}{d}\right\}}.
\end{aligned}$$

Let $\bar{q}^*$ be the smallest $\bar{q}$ that satisfies (17). Then for any $q \leq \bar{q}^*$ such that $q - \frac{1}{d} \geq \frac{1}{2}$ we have

$$K|\mathcal{D}_Z^{\underline{q}}| \leq K 2^{d\left\{H_B\left(q-\frac{1}{d}\right) + \frac{\log(d+1)}{d}\right\}}. \tag{18}$$

In particular, any $q > \frac{1}{2} + \frac{1}{d}$ that satisfies $\Pr\left[Z^d \in \mathcal{D}_Z^q\right] \geq 2^{-d(E-\epsilon)}$ will do. Such a q can be found by using a lower-bound on the probability of a single type class.

$$\Pr\left[Z^d \in \mathcal{D}_Z^q\right] \geq \frac{1}{d+1} 2^{-dD(q||p)}$$

and so picking q^* from (10) is good enough.

All that remains is to upper bound $H_B\left(q^* - \frac{1}{d}\right)$ from (18):

$$\begin{aligned}
H_B\left(q^* - \frac{1}{d}\right) =& \frac{1}{d} \log\left[\frac{q^* - \frac{1}{d}}{1 - q^* + \frac{1}{d}}\right] - q^* \left(\log[q^*] + \log\left[1 - \frac{1}{dq^*}\right]\right) \\
&- (1-q^*)\left(\log[1-q^*] + \log\left[1 + \frac{1}{d(1-q^*)}\right]\right) \\
\leq& H_B(q^*) + \frac{1}{d} \log\left[\frac{q^*}{1-q^*}\right] - q^* \log\left[1 - \frac{1}{dq^*}\right] \\
&- (1-q^*) \log\left[1 + \frac{1}{d(1-q^*)}\right]
\end{aligned} \tag{19}$$

Assuming $d \geq \max\{\frac{2}{q^*}, \frac{1}{1-q^*}\}$, we know $\frac{1}{dq^*} < \frac{1}{2}$ and $\frac{1}{d(1-q^*)} < 1$. Since $\log$ is concave–$\cap$ with $\log 1 = 0$, $\log \frac{1}{2} = -1$ and $\log 2 = 1$, we know $\log(1 - \frac{1}{dq^*}) \geq -\frac{2}{dq*}$ and $\log(1 + \frac{1}{d(1-q^*)})) \geq \frac{1}{d(1-q^*)}$. Continuing from (19) we get

$$\begin{aligned}
H_B\left(q^* - \frac{1}{d}\right) &\leq H_B(q^*) + \frac{1}{d} \log\left[\frac{q^*}{1-q^*}\right] + \frac{2}{d} - \frac{1}{d} \\
&= H_B(q^*) + \frac{1}{d} \log\left[\frac{q^*}{1-q^*}\right] + \frac{1}{d}
\end{aligned}$$

Combining this with (18), gives (12). □

REFERENCES

[1] M. V. Burnashev. Data transmission over a discrete channel with feedback. Random transmission time. *Prob. Peredachi Informatsii*, 12(4):10–30, 1976.

[2] S. C. Draper and A. Sahai. Beating the Burnashev bound using noisy feedback. In *Proc. 44th Allerton Conf. on Communication, Control and Computing*, September 2006.

[3] S. C. Draper and A. Sahai. *IEEE Trans. Inform. Theory*, To be submitted.

[4] G. D. Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Trans. Inform. Theory*, 14:206–220, March 1968.

[5] E. A. Haroutunian. Lower bound for error probability in channels with feedback. *Problemy Peredachi Imformatsii*, 13(2):36–44, 1977.

[6] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Trans. Inform. Theory*, 1963.

[7] B. D. Kudryashov. Message transmission over a discrete channel with noiseless feedback. *Problemy Peredachi Imformatsii*, 21(1):3–13, 1979.

[8] A. Sahai. Why block length and delay behave differently if feedback is present. *IEEE Trans. Inform. Theory*, Submitted.