Greedy-Merge Degrading has Optimal Power-Law

Assaf Kartowsky, Ido Tal, *Senior Member, IEEE*

***Abstract*—Consider a channel with a given input alphabet size and a given input distribution. Our aim is to degrade or upgrade it to a channel with at most L output letters. A channel Q is degraded with respect to a channel W if Q can be obtained from W by processing the output of W. Upgrading is the inverse relation.**

The paper contains four main results. The first result, from which the paper title is derived, deals with the so called "greedy-merge" algorithm. We derive an upper bound on the reduction in mutual information between input and output, as a function of L. This upper bound is within a constant factor of an algorithm-independent lower bound. Thus, we establish that greedy-merge is optimal in the power-law sense (i.e. the power of L).

The other main results deal with upgrading. The second result shows that a certain sequence of channels that was previously shown to be "hard" for degrading, displays the same hardness in the context of upgrading. That is, suppose we are given such a channel and a corresponding input distribution. If we upgrade (degrade) to a new channel with L output letters, we incur an increase (decrease) in mutual information between input and output. We show that a previously derived bound on the decrease in mutual information for the degrading case is also a lower bound on the increase for the upgrading case.

The third result is an efficient algorithm for optimal upgrading, in the binary-input case. That is, we are given a channel and an input distribution. We must find an upgraded channel with L output letters, for which the increase in mutual information is minimal. We give a simple characterization of such a channel, which implies an efficient algorithm.

The fourth result is an analog of the first result for the upgrading case, when the input is binary. That is, we first present a sub-optimal algorithm for the setting considered in the third result. The main advantage of the sub-optimal algorithm is that it is amenable to analysis. We carry out the analysis, and show that the increase incurred in mutual information is within a constant factor of the lower bound derived in the second result.

***Index Terms*—Channel degrading, channel upgrading, greedy merge, greedy split, polar codes, quantization**

I. Introduction

In myriad digital processing contexts, quantization is used to map a large alphabet to a smaller one. For example, quantizers are an essential building block in receiver design, used to keep the complexity and resource consumption manageable. The quantizer used has a direct influence on the attainable code rate.

Another recent application is related to polar codes [1]. Polar code construction is equivalent to evaluating the misdecoding probability of each channel in a set of synthetic channels. This evaluation cannot be carried out naively, since the output alphabet size of a synthetic channel is intractably large. One approach to circumventing this difficulty is to either degrade or upgrade the evaluated synthetic channel to a channel with manageable output alphabet size [2]. We will fully define the relations of degrading and upgrading shortly, in Section II. In brief, a channel Q is degraded with respect to a channel W if Q can be obtained from W by processing the output of W. Upgrading is the inverse relation. When degrading (upgrading) a channel W to a channel Q, one obtains from the misdecoding probability of Q an upper (lower) bound on the misdecoding probability of W. In particular, for a wiretap channel, both upgrading and degrading are essential to ensure secure and reliable communications [3][4][5][6][7].

The general problem considered in this paper is the following. Given a design parameter L, we either degrade or upgrade an initial channel to a new one with output alphabet size at most L. We assume that the input distribution is specified, and note that degradation reduces the mutual information between the channel input and output, whereas upgradation increases it. This reduction (increase) in mutual information is roughly the loss (gain) in code rate due to quantization. We denote the smallest reduction (increase) possible by $\Delta I^*_{\downarrow}$ ($\Delta I^*_{\uparrow}$) or simply ΔI^*, when direction is clear from the context.

In this work we present four main results. For the sake of clarity, Table I lists previous related works along with our new results, marked by a "✓". In the table, BDMC and DMC are short for Binary Discrete Memoryless Channel and Discrete Memoryless Channel, respectively. We note that $\Delta I^* = \Omega(\cdot)$ denotes lower bounds on ΔI^* as a function of L for a specified channel and input distribution or a sequence of those two. On the other hand, $\Delta I^* = O(\cdot)$ are general upper bounds on ΔI^*, as a function of L, that are independent of channel and input distribution.

TABLE I
Previous related works and our new results

	Channel Type	Optimal Algorithm	$\Delta I^* = \Omega(\cdot)$	$\Delta I^* = O(\cdot)$
Degrading	BDMC	[8],[9]	[10]	[2],[11][1],✓
	DMC		[10]	[12],[13],[14],✓
Upgrading	BDMC	✓	✓	[2],[11][1],✓
	DMC		✓	[14]

Let $|\mathcal{X}|$ denote the channel input alphabet size, and treat it as a fixed quantity. In our first main result (Section III), we show that for any input distribution and any initial channel, $\Delta I^*_{\downarrow} = O(L^{-2/(|\mathcal{X}|-1)})$. Moreover, this bound is attained efficiently by the greedy-merge algorithm discussed already in [2] and [11] for binary-input memoryless symmetric channels (BMSCs), and

This work was supported by the Israel Science Foundation under Grant 1769/13 and by the United States–Israel Binational Science Foundation under grant number 2012016. An abbreviated version of this article has appeared in ISIT 2017.

A. Kartowsky is with Vayyar Imaging Ltd., I. Tal is with the Department of Electrical Engineering, Technion, Haifa 32000, Israel. Email: assaf.kartowsky@gmail.com, idotal@ee.technion.ac.il

[1] To be precise, the results in [2] and [11] were derived for binary-input memoryless symmetric channels (BMSCs).

in [13] for general discrete memoryless channels. This bound is tighter than the bounds derived in [2],[11],[12],[13] and [14]. In fact, up to a constant multiplier (dependent on $|\mathcal{X}|$), this bound is the tightest possible. Namely, [10] proves the existence of an input distribution and a sequence of channels for which $\Delta I_{\downarrow}^{*} = \Omega(L^{-2/(|\mathcal{X}|-1)})$. Both bounds have $-2/(|\mathcal{X}|-1)$ as the power of L, the same power-law. We mention a recent result [15] in which a different power-law is shown to be tight, for the special case in which the channel is very noisy. See also [16], which is especially relevant when L is small.

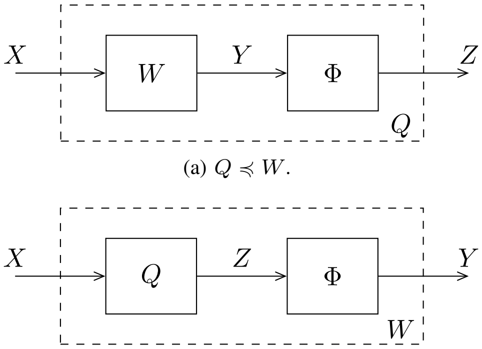
Our second main result (Section IV) is the analog of [10] to the upgrading problem. Namely, in [10], a sequence of channels is shown to have a degrading penalty of $\Delta I_{\downarrow}^{*} = \Omega(L^{-2/(|\mathcal{X}|-1)})$. We show that this same sequence of channels has an upgrading penalty $\Delta I_{\uparrow}^{*} = \Omega(L^{-2/(|\mathcal{X}|-1)})$, with the exact same constant. Similar to [10], we conclude that some channels with moderate $|\mathcal{X}|$ are "hard" to upgrade, in the sense that a very large L is required to keep $\Delta I_{\uparrow}^{*}$ small. Moreover, this result plays an important role in our fourth result.

An optimal degrading algorithm was presented in [8], for the binary-input case. Namely, a channel and input distribution are supplied, along with a target output alphabet size. The algorithm finds a degraded channel with the target output alphabet size, that has the largest possible mutual information between input and output. See also [9] for an improved implementation in terms of running time. In our third result (Section V), we present the upgrading analog: an optimal upgrading algorithm for binary-input discrete memoryless channels. Namely, we show that an optimal upgraded channel is a subset of the initial channel, when both are represented using posterior probability vectors. This characterization paves the way for an algorithm that efficiently finds the optimal subset.

In our fourth main result (Section VI), we use our previous results and techniques to obtain an upper bound on $\Delta I_{\uparrow}^{*}$, valid for any binary-input discrete memoryless channel and any input distribution. That is, a greedy version of the optimal upgrading algorithm, known as "greedy-split", is proved to obtain $\Delta I_{\uparrow}^{*} = O(L^{-2/(|\mathcal{X}|-1)}) = O(L^{-2})$. We note that the algorithm is a generalization of the one presented in [2] for the symmetric case. Our bound is tighter than the one previously derived in [11]. As in our first result, this new bound shares the same power-law as the lower bound from our second result, and is thus, up to a constant multiplier, the tightest possible.

II. Framework and Notation

We are given an input distribution and a DMC $W : \mathcal{X} \to \mathcal{Y}$. Both $|\mathcal{X}|$ and $|\mathcal{Y}|$ are assumed finite. Let X and Y denote the random variables that correspond to the channel input and output, respectively. Denote the corresponding distributions P_X and P_Y. The probability of receiving $y \in \mathcal{Y}$ as the channel output given that $x \in \mathcal{X}$ was transmitted, namely $\mathbb{P}\{Y = y|X = x\}$, is denoted by $W(y|x)$. The probability that $x \in \mathcal{X}$ is the channel input, namely $\mathbb{P}\{X = x\}$, is denoted by $\pi(x)$. We also assume that $\mathcal{X}$ and $\mathcal{Y}$ are disjoint, allowing ourselves to abuse notation and denote $\mathbb{P}\{X = x|Y = y\}$ and $\mathbb{P}\{Y = y\}$ as $W(x|y)$ and $\pi(y)$, respectively. We stress that unless stated otherwise, we will not assume that W is symmetric in any sense. Without loss of generality we assume that $\pi(x) > 0$ and $\pi(y) > 0$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.



(a) $Q \preccurlyeq W$.

(b) $Q \succcurlyeq W$.

Fig. 1. Degrading and upgrading W to Q.

The mutual information between the channel input and output is

$$I(W, P_X) \triangleq I(X;Y) = \sum_{x\in\mathcal{X}} \eta(\pi(x)) - \sum_{\substack{x\in\mathcal{X},\\ y\in\mathcal{Y}}} \pi(y)\eta(W(x|y)) ,$$

where

$$\eta(p) = \begin{cases} -p\log p & p > 0 , \\ 0 & p = 0 , \end{cases} \tag{1}$$

and the logarithm is taken in the natural basis. We note that the input distribution does not necessarily have to be the one that achieves the channel capacity.

We now define the relations of degradedness and upgradedness between channels. A channel $Q : \mathcal{X} \to \mathcal{Z}$ is said to be *degraded* with respect to a channel $W : \mathcal{X} \to \mathcal{Y}$, and we write $Q \preccurlyeq W$, if there exists a channel $\Phi : \mathcal{Y} \to \mathcal{Z}$ such that

$$Q(z|x) = \sum_{y\in\mathcal{Y}} W(y|x)\Phi(z|y) \tag{2}$$

for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ (see Figure 1a). The channel Φ is then defined as the *intermediate channel*. Conversely, the channel Q is said to be *upgraded* with respect to W, if W is degraded with respect to Q (see Figure 1b). We note that as a result of the data processing theorem, $Q \preccurlyeq W$ implies $\Delta I_{\downarrow} \triangleq I(W, P_X) - I(Q, P_X) \geq 0$, and similarly, $W \preccurlyeq Q$ implies $\Delta I_{\uparrow} \triangleq I(Q, P_X) - I(W, P_X) \geq 0$.

Since our focus is on approximating channels with a large output alphabet size using channels with a limited output alphabet size we define the *optimal degrading loss* for a given pair (W, P_X) and a target output alphabet size L as

$$\Delta I_{\downarrow}^{*}(W, P_X, L) = \Delta I_{\downarrow}^{*} \triangleq \min_{\substack{Q:Q\preccurlyeq W,\\ |Q|\leq L}} I(W, P_X) - I(Q, P_X) , \tag{3}$$

where $|Q|$ denotes the output alphabet size of the channel Q. The optimizer Q is the degraded channel that is "closest" to W in the sense of mutual information, yet has at most L output letters. In the same manner, we define the *optimal upgrading* *gain* for a given pair (W, P_X) and a target output alphabet size L as

$$\Delta I^*_{\uparrow}(W, P_X, L) = \Delta I^*_{\uparrow} \triangleq \min_{\substack{Q:Q \succcurlyeq W,\\ |Q| \le L}} I(Q, P_X) - I(W, P_X) \ . \tag{4}$$

As in [10], define the *degrading cost* in the following way:

$$\mathrm{DC}(|\mathcal{X}|, L) \triangleq \sup_{W, P_X} \Delta I^*_{\downarrow} \ .$$

The optimizers W and P_X are the channel and input distribution that yield the highest optimal degrading loss. In a way, they are the "worst" or "hardest" pair to degrade. We define the similar notion for the case of upgrading, namely the *upgrading cost* as

$$\mathrm{UC}(|\mathcal{X}|, L) \triangleq \sup_{W, P_X} \Delta I^*_{\uparrow} \ . \tag{5}$$

III. Upper bound on optimal degrading loss

A. Main result

Our first main result is an upper bound on $\Delta I^*_{\downarrow}$ and DC in terms of $|\mathcal{X}|$ and L that is tight in the power-law sense. This upper bound will follow from analyzing a sub-optimal degrading algorithm, called "greedy-merge". In each iteration of greedy-merge, we merge the two output letters $\alpha, \beta \in \mathcal{Y}$ that result in the smallest decrease of mutual information between input and output, denoted $\Delta I_{\downarrow}$. Namely, the intermediate channel Φ maps α and β to a new symbol, while all other symbols are unchanged by Φ. This is repeated $|\mathcal{Y}| - L$ times, to yield an output alphabet size of L. By upper bounding the $\Delta I_{\downarrow}$ of each iteration, we obtain an upper bound on $\Delta I^*_{\downarrow}$ and DC.

Theorem 1. *Let a DMC $W : \mathcal{X} \to \mathcal{Y}$ satisfy $|\mathcal{Y}| > 2|\mathcal{X}|$ and let $L \geq 2|\mathcal{X}|$. Then, for any fixed input distribution P_X,*

$$\Delta I^*_{\downarrow} = \min_{\substack{Q:Q \preccurlyeq W,\\ |Q| \le L}} I(W, P_X) - I(Q, P_X) = O\left(L^{-\frac{2}{|\mathcal{X}|-1}}\right) \ .$$

In particular,

$$\Delta I^*_{\downarrow} \leq \nu(|\mathcal{X}|) \cdot L^{-\frac{2}{|\mathcal{X}|-1}} \ ,$$

where

$$\nu(|\mathcal{X}|) \triangleq \frac{\pi \cdot |\mathcal{X}|(|\mathcal{X}|-1)}{2\left(\sqrt{1+\frac{1}{2(|\mathcal{X}|-1)}}-1\right)^2} \cdot \left(\frac{2|\mathcal{X}|}{\Gamma\left(1+\frac{|\mathcal{X}|-1}{2}\right)}\right)^{\frac{2}{|\mathcal{X}|-1}} \ ,$$

and $\Gamma(\cdot)$ is the Gamma function. That is, for an integer $n \geq 1$,

$$\Gamma(n) = (n-1)! \ , \quad \Gamma\left(n + \frac{1}{2}\right) = \frac{(2n)!}{4^n n!}\sqrt{\pi} \ . \tag{6}$$

This bound is attained by greedy-merge, and is tight in the power-law sense.

Note that for large values of $|\mathcal{X}|$ the Stirling approximation along with some other first order approximations can be applied to simplify $\nu(|\mathcal{X}|)$ to

$$\nu(|\mathcal{X}|) \approx 16\pi e|\mathcal{X}|^3 \ .$$

We stress that the greedy-merge algorithm is not optimal in the sense of $\Delta I^*_{\downarrow}$. To see this, consider the following example.

Example 1. Let $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{a, b, c, d\}$, and let the channel $W : \mathcal{X} \to \mathcal{Y}$ satisfy

$$\begin{aligned} W(a|0) &= 0, & W(b|0) &= \frac{1}{6}, & W(c|0) &= \frac{1}{3}, & W(d|0) &= \frac{1}{2}, \\ W(a|1) &= \frac{1}{2}, & W(b|1) &= \frac{1}{3}, & W(c|1) &= \frac{1}{6}, & W(d|1) &= 0. \end{aligned}$$

Assume also that $\pi(x) = \frac{1}{2}$ for all $x \in \mathcal{X}$. Suppose now that we wish to reduce the output alphabet from 4 to 2 (i.e., $L = 2$) by degrading. The greedy-merge algorithm would first merge the pair b, c to a new letter bc, and then merge a with bc, resulting in $\Delta I_{\downarrow} = 0.16$ in total. The optimal degrading algorithm, however, would merge a, b to ab and c, d to cd, resulting in a total reduction of $\Delta I^*_{\downarrow} = 0.13$. Clearly, the greedy-merge algorithm is not optimal.

To prove Theorem 1, we will use the following theorem which proves the existence of a pair of output letters whose merger yields a "small" $\Delta I_{\downarrow}$.

Theorem 2. *Let a DMC $W : \mathcal{X} \to \mathcal{Y}$ satisfy $|\mathcal{Y}| > 2|\mathcal{X}|$, and let the input distribution be fixed. There exists a pair $\alpha, \beta \in \mathcal{Y}$ whose merger results in a channel Q satisfying*

$$\Delta I_{\downarrow} = O\left(|\mathcal{Y}|^{-\frac{|\mathcal{X}|+1}{|\mathcal{X}|-1}}\right) \ .$$

In particular,

$$\Delta I_{\downarrow} \leq \mu(|\mathcal{X}|) \cdot |\mathcal{Y}|^{-\frac{|\mathcal{X}|+1}{|\mathcal{X}|-1}} \ , \tag{7}$$

where,

$$\mu(|\mathcal{X}|) \triangleq \frac{2}{|\mathcal{X}|-1}\nu(|\mathcal{X}|) \ ,$$

and $\nu(\cdot)$ was defined in Theorem 1.

A large part of this section is devoted to stating and proving claims cardinal to the proof of Theorem 2. Once Theorem 2 is proved, Theorem 1 will follow as a simple corollary. Thus, let us end this subsection by giving a short overview of the major steps and techniques involved in the proof of Theorem 2:

- For given output letters α and β that are to be merged, there is an explicit analytic expression for the reduction in mutual information, $\Delta I_{\downarrow}$. This is given in (9), below. Although the expression for $\Delta I_{\downarrow}$ is simple and compact, we were not able to easily manipulate it to suit our needs. Essentially, the difficulty stems from the fact that $\Delta I_{\downarrow}$ behaves very differently, depending on whether or not either one of the posterior probability vectors corresponding to α and β has an entry that is very close to 0. That is, the fact that the function η defined in (1) has an unbounded derivative close to 0 is an analytic complication.
- To overcome the above difficulty, we make use of a previously derived bound (10) on $\Delta I_{\downarrow}$, and also develop a new bound (11) on $\Delta I_{\downarrow}$. Essentially, the bound in (10) is useful when a vector of posterior probabilities has an

entry close to 0, while (11) is typically tighter than (10) when this is not the case.

- In (13)–(16), a new bound on $\Delta I_{\downarrow}$ is derived. In essence, we use the minimum (tightest) of (10) and (11), and replace a sum by a maximum in order to simplify the derivations later on. In (16), we define the function d, taking as arguments two posterior probability vectors. Although the notation is purposefully evocative, d is *not* a proper distance function: it does not satisfy the triangle inequality.
- Recall that our aim is to find a pair of output letters, α and β, for which $\Delta I_{\downarrow}$ is small, in the sense that it satisfies (7). Although searching over all the output alphabet $\mathcal{Y}$ is allowed, we limit ourselves to a subset $\mathcal{Y}_{\text{small}}$ of letters having a sufficiently small probability. Doing so does not reduce the search space by much (17), and affords us a simple bound on $\Delta I_{\downarrow}$ (18). Namely, the critical feature of (18) is that it is a linear function of the above mentioned d. That is, with respect to d, we must find a pair of output letters with posterior probabilities that are "close". Essentially, this is the same technique as [13], but with a tighter d.
- As in [13], our plan is to show the existence of such a pair by a sphere-packing argument. However, such arguments typically rely on an underlying distance function, and since this is not the case for our d, some care and effort are called for. As a first step, we calculate the "sphere" of radius r "centered" at some point $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the posterior probability vector corresponding to output letter α. A straightforward series of calculations yields the characterization of this sphere, $\mathcal{B}(\boldsymbol{\alpha}, r)$, given in (22) and (24).
- Most of the vectors contained in $\mathcal{B}(\boldsymbol{\alpha}, r)$ are irrelevant, since they do not correspond to a posterior probability, and thus lie outside of the probability simplex. To remedy this, we first consider the intersection of $\mathcal{B}(\boldsymbol{\alpha}, r)$ with the set of vectors with entries summing to 1. Doing so gives us a gain in the form of a reduced affine dimension, but results in a set that is unwieldy. Thus, we simplify by considering instead a simple set of vectors summing to 1, which is contained in $\mathcal{B}(\boldsymbol{\alpha}, r)$. This is the set $\mathcal{C}(\boldsymbol{\alpha}, r)$, defined in (26) and (27). Essentially, after a coordinate is removed, $\mathcal{C}(\boldsymbol{\alpha}, r)$ is simply a box.
- We now come to terms with d not being a proper distance function. Note that if d *were* a distance function, an intersection of $\mathcal{C}(\boldsymbol{\alpha}, r)$ and $\mathcal{C}(\boldsymbol{\beta}, r)$ would imply that $d(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is at most $2r$. This does not hold in our case. Our way around this difficulty is by carefully defining from $\mathcal{C}(\boldsymbol{\alpha}, r)$ the subset $\mathcal{Q}(\boldsymbol{\alpha}, r)$. This is done in (29). The redeeming property of $\mathcal{Q}(\boldsymbol{\alpha}, r)$ is this: if, up to technical conditions, $\mathcal{Q}(\boldsymbol{\alpha}, r)$ and $\mathcal{Q}(\boldsymbol{\beta}, r)$ intersect, then $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r$. Thus, we are well on our way to applying a sphere-packing argument: the intersection of sets $\mathcal{Q}$ is a suitable replacement for the "sphere intersection" in a typical sphere-packing argument.
- In many sphere-packing arguments, the volume of a sphere does not depend on the point it is centered at. However, this is not the case for $\mathcal{Q}(\boldsymbol{\alpha}, r)$, the volume of which is strongly dependent on $\boldsymbol{\alpha}$. This is a deficiency, which we counter by considering the "weight" of $\mathcal{Q}(\boldsymbol{\alpha}, r)$ as opposed to its volume. Namely, a "weight density" function φ is defined in (31) and (32), by which the weight of a set $\mathcal{Q}$ is defined in (33). The function φ is chosen so that all such $\mathcal{Q}$ have roughly the same weight. Lastly, a sphere-packing argument is applied.

B. An alternative "distance" function

We begin by addressing the merger of a pair of output letters $\alpha, \beta \in \mathcal{Y}$. Since this pair of letters is merged into a new letter γ (we assume $\gamma \notin \mathcal{X} \cup \mathcal{Y}$), the new output alphabet of Q is $\mathcal{Z} = \{\gamma\} \cup \mathcal{Y} \setminus \{\alpha, \beta\}$. The channel $Q : \mathcal{X} \to \mathcal{Z}$ then satisfies

$$Q(\gamma|x) = W(\alpha|x) + W(\beta|x) ,$$

whereas for all $y \in \mathcal{Z} \cap \mathcal{Y}$ we have $Q(y|x) = W(y|x)$. Using the shorthand

$$\pi_\gamma = \pi(\gamma) , \quad \pi_\alpha = \pi(\alpha) , \quad \pi_\beta = \pi(\beta) ,$$

one gets that

$$\pi_\gamma = \pi_\alpha + \pi_\beta .$$

Let us denote by $\boldsymbol{\alpha} = (\alpha_x)_{x \in \mathcal{X}}$, $\boldsymbol{\beta} = (\beta_x)_{x \in \mathcal{X}}$ and $\boldsymbol{\gamma} = (\gamma_x)_{x \in \mathcal{X}}$ the vectors corresponding to posterior probabilities associated with α, β and γ, respectively. Namely,

$$\alpha_x = W(x|\alpha) , \quad \beta_x = W(x|\beta) ,$$

and a short calculation shows that

$$\gamma_x = Q(x|\gamma) = \frac{\pi_\alpha \alpha_x + \pi_\beta \beta_x}{\pi_\gamma} = \frac{\pi_\alpha \alpha_x + \pi_\beta \beta_x}{\pi_\alpha + \pi_\beta} . \tag{8}$$

Thus, after canceling terms, one gets that

$$\Delta I_{\downarrow} = I(W, P_X) - I(Q, P_X) = \sum_{x \in \mathcal{X}} \Delta I_x , \tag{9}$$

where

$$\Delta I_x \triangleq \pi_\gamma \eta(\gamma_x) - \pi_\alpha \eta(\alpha_x) - \pi_\beta \eta(\beta_x) .$$

In order to bound $\Delta I_{\downarrow}$, we give two bounds on ΔI_x. The first bound was derived in [13],

$$\Delta I_x \leq (\pi_\alpha + \pi_\beta) \cdot d_1(\alpha_x, \beta_x) , \tag{10}$$

where for $\sigma \geq 0$ and $\zeta \in \mathbb{R}$, we define

$$d_1(\sigma, \zeta) \triangleq |\zeta - \sigma| .$$

The subscript "1" in d_1 is suggestive of the L_1 distance. Note that we will generally use α_x or σ to denote a probability associated with an input letter, while ζ will denote a "free" real variable, possibly negative. We will keep to this convention for the vector case as well. In addition, let us point out that the bound in (10) was derived assuming a uniform input distribution, however it remains valid for the general case.

We now derive the second bound on ΔI_x. For the case where $\alpha_x, \beta_x > 0$,

$$\begin{aligned}\Delta I_x &= \pi_\alpha(\eta(\gamma_x) - \eta(\alpha_x)) + \pi_\beta(\eta(\gamma_x) - \eta(\beta_x)) \\ &\overset{(a)}{\leq} \pi_\alpha \eta'(\alpha_x)(\gamma_x - \alpha_x) + \pi_\beta \eta'(\beta_x)(\gamma_x - \beta_x) \\ &\overset{(b)}{=} \frac{\pi_\alpha \pi_\beta}{\pi_\alpha + \pi_\beta}(\alpha_x - \beta_x)(\eta'(\beta_x) - \eta'(\alpha_x)) \\ &\overset{(c)}{\leq} \frac{1}{4}(\pi_\alpha + \pi_\beta)(\alpha_x - \beta_x)^2(-\eta''(\lambda)) \ ,\end{aligned}$$

where in (a) we used the concavity of $\eta(\cdot)$, in (b) the definition of γ_x (see (8)), and in (c) the AM-GM inequality and the mean value theorem where $\lambda = \theta\alpha_x + (1-\theta)\beta_x$ for some $\theta \in [0,1]$. Using the monotonicity of $-\eta''(p) = 1/p$ we get

$$-\eta''(\lambda) \leq \frac{1}{\min(\alpha_x, \beta_x)} \ .$$

Thus,

$$\Delta I_x \leq (\pi_\alpha + \pi_\beta) \cdot d_2(\alpha_x, \beta_x) \ , \tag{11}$$

where

$$d_2(\sigma, \zeta) \triangleq \begin{cases} \frac{(\zeta - \sigma)^2}{\min(\sigma, \zeta)} & \sigma, \zeta > 0 \ , \\ \infty & \text{otherwise} \ . \end{cases} \tag{12}$$

The subscript "2" in d_2 is suggestive of the squaring in the numerator. Combining (10) and (11) yields

$$\Delta I_x \leq (\pi_\alpha + \pi_\beta) \cdot d(\alpha_x, \beta_x) \ , \tag{13}$$

where

$$d(\sigma, \zeta) \triangleq \min(d_1(\sigma, \zeta), d_2(\sigma, \zeta)) \ . \tag{14}$$

Returning to (9) using (13) we get

$$\Delta I_\downarrow \leq (\pi_\alpha + \pi_\beta)|\mathcal{X}| \cdot d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \ , \tag{15}$$

where

$$d(\boldsymbol{\alpha}, \boldsymbol{\zeta}) \triangleq \max_{x \in \mathcal{X}} d(\alpha_x, \zeta_x) \ . \tag{16}$$

Before moving on, let us make a few remarks and give some motivation for (13), (15) and (16). First, the usage of "max" in (16) as opposed to summing over all x is to simplify upcoming derivations. Second, recall that we wish to prove the existence of a pair $\alpha, \beta \in \mathcal{Y}$ such that $\Delta I_\downarrow$ is "small". Then according to (15), it suffices to show the existence of a pair that is "close" in the sense of d, assuming that π_α, π_β are also small enough. Third, some intuition regarding the need for both d_1 and d_2 is the following. Using d_2 alone is not good enough since it diverges when either of its input arguments is in the vicinity of zero. Thus, the merger of a pair of close letters with a small entry yields a large value of d_2 instead of a small one. Using d_1 alone, on the other hand, would lead us to a looser bound than desired (see [13]).

Since we are interested in lowering the right hand side of (15), we limit our search to a subset of $\mathcal{Y}$, as was done in [13]. That is,

$$\mathcal{Y}_{\text{small}} \triangleq \left\{ y \in \mathcal{Y} : \pi(y) \leq \frac{2}{|\mathcal{Y}|} \right\} \ ,$$

which implies

$$|\mathcal{Y}_{\text{small}}| \geq \frac{|\mathcal{Y}|}{2} \ . \tag{17}$$

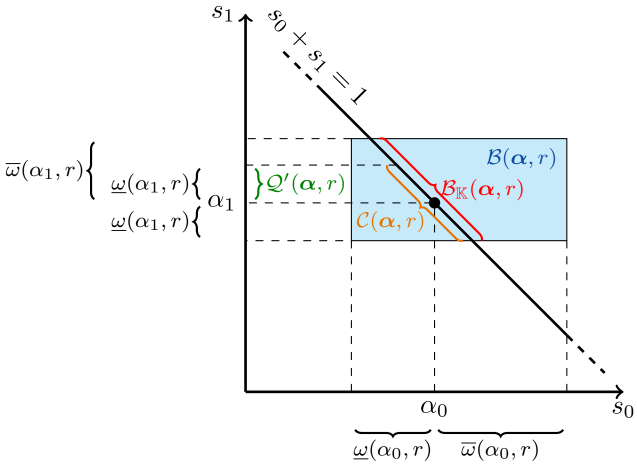

s_1
$s_0 + s_1 = 1$
$\overline{\omega}(\alpha_1, r)$
$\underline{\omega}(\alpha_1, r)$
$\underline{\omega}(\alpha_1, r)$
α_1
$\mathcal{Q}'(\boldsymbol{\alpha}, r)$
$\mathcal{B}(\boldsymbol{\alpha}, r)$
$\mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r)$
$\mathcal{C}(\boldsymbol{\alpha}, r)$
α_0
s_0
$\underline{\omega}(\alpha_0, r)$
$\overline{\omega}(\alpha_0, r)$

Fig. 2. The sets $\mathcal{B}(\boldsymbol{\alpha}, r)$, $\mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r)$, $\mathcal{C}(\boldsymbol{\alpha}, r)$ and $\mathcal{Q}'(\boldsymbol{\alpha}, r)$ for the binary case ($|\mathcal{X}| = 2$) assuming $\alpha_1 < \alpha_0$ in the (s_0, s_1) plane.

Namely, we have not reduced the search space by too much. The utility of confining ourselves to $\mathcal{Y}_{\text{small}}$ is that for all $\alpha, \beta \in \mathcal{Y}_{\text{small}}$, we have that $\pi_\alpha + \pi_\beta \leq 4/|\mathcal{Y}|$. Thus, by (15), we have the following lemma.

Lemma 3. *Let $\alpha, \beta \in \mathcal{Y}_{\text{small}}$. Then,*

$$\Delta I_\downarrow \leq \frac{4|\mathcal{X}|}{|\mathcal{Y}|} \cdot d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \ . \tag{18}$$

We still need to prove the existence of a pair $\alpha, \beta \in \mathcal{Y}_{\text{small}}$ that is "close" in the sense of d. To that end, as in [13], we would like to use a sphere-packing approach. A typical use of such an argument assumes a proper metric, yet d is not a metric. Specifically, the triangle-inequality does not hold:

$$\underbrace{d\left(\begin{bmatrix}0.1\\0.9\end{bmatrix}, \begin{bmatrix}0.2\\0.8\end{bmatrix}\right)}_{0.1} + \underbrace{d\left(\begin{bmatrix}0.2\\0.8\end{bmatrix}, \begin{bmatrix}0.3\\0.7\end{bmatrix}\right)}_{0.05} \not\geq \underbrace{d\left(\begin{bmatrix}0.1\\0.9\end{bmatrix}, \begin{bmatrix}0.3\\0.7\end{bmatrix}\right)}_{0.2} \ .$$

The absence of a triangle-inequality is a complication that we will overcome, but some care and effort are called for. Broadly speaking, as usually done in sphere-packing, we aim to show the existence of a critical "sphere" radius, $r_{\text{critical}} = r_{\text{critical}}(|\mathcal{X}|, |\mathcal{Y}|) > 0$. Such a critical radius will ensure the existence of $\alpha, \beta \in \mathcal{Y}_{\text{small}}$ for which $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r_{\text{critical}}$.

C. Non-intersecting "spheres"

In this section, we gradually derive our equivalent of a "sphere". For convenience, some of the sets defined in the process are illustrated in Figure 2 for the binary case. We start by giving explicit equations for the "spheres" corresponding to d_1 and d_2.

Lemma 4. *For $\sigma \geq 0$ and $r > 0$, define the sets $\mathcal{B}_1(\sigma, r)$ and $\mathcal{B}_2(\sigma, r)$ as*

$$\begin{aligned}\mathcal{B}_1(\sigma, r) &\triangleq \{\zeta \in \mathbb{R} : d_1(\sigma, \zeta) \leq r\} \ , \\ \mathcal{B}_2(\sigma, r) &\triangleq \{\zeta \in \mathbb{R} : d_2(\sigma, \zeta) \leq r\} \ .\end{aligned}$$

Then,

$$\mathcal{B}_1(\sigma, r) = \{\zeta \in \mathbb{R} : -r \le \zeta - \sigma \le r\} \ ,$$

and

$$\mathcal{B}_2(\sigma, r) = \{\zeta \in \mathbb{R} : -\sqrt{\frac{r^2}{4} + \sigma \cdot r} + \frac{r}{2} \le \zeta - \sigma \le \sqrt{\sigma \cdot r}\} \ .$$

Proof: Assume $\zeta \in \mathcal{B}_1(\sigma, r)$. Then ζ satisfies $|\zeta - \sigma| \le r$, which is equivalent to $-r \le \zeta - \sigma \le r$, and we obtained the desired result for $\mathcal{B}_1(\sigma, r)$. Assume now $\zeta \in \mathcal{B}_2(\sigma, r)$. If $\zeta \ge \sigma$, then $\min(\sigma, \zeta) = \sigma$, and thus

$$\frac{(\zeta - \sigma)^2}{\sigma} \le r \ ,$$

which implies

$$0 \le \zeta - \sigma \le \sqrt{\sigma \cdot r} \ . \tag{19}$$

Next, consider the case $\zeta \le \sigma$. Note that since r is finite , we may further assume by the definition of d_2 in (12) that $\zeta > 0$. Now $\min(\sigma, \zeta) = \zeta$, which implies that

$$\frac{(\zeta - \sigma)^2}{\zeta} \le r \ .$$

Since $\zeta > 0$, we rearrange the above to yield

$$\zeta^2 - (2\sigma + r)\,\zeta + \sigma^2 \le 0 \ ,$$

and considering the above as a quadratic inequality in ζ yields

$$-\sqrt{\frac{r^2}{4} + \sigma \cdot r} + \frac{r}{2} \le \zeta - \sigma \le \sqrt{\frac{r^2}{4} + \sigma \cdot r} + \frac{r}{2} \ .$$

By assumption, $\zeta \le \sigma$. Thus, we can sharpen and simplify the rightmost inequality above, yielding

$$-\sqrt{\frac{r^2}{4} + \sigma \cdot r} + \frac{r}{2} \le \zeta - \sigma \le 0 \ . \tag{20}$$

The union of (19) and (20) yields the desired result for $\mathcal{B}_2(\sigma, r)$. ■

With $\mathcal{B}_1(\sigma, r)$ and $\mathcal{B}_2(\sigma, r)$ defined and calculated, we now move on to defining $\mathcal{B}(\sigma, r)$ as

$$\mathcal{B}(\sigma, r) \triangleq \{\zeta \in \mathbb{R} : d(\sigma, \zeta) \le r\} \ ,$$

and note that,

$$\mathcal{B}(\sigma, r) = \mathcal{B}_1(\sigma, r) \cup \mathcal{B}_2(\sigma, r) \ .$$

Namely, since d is the minimum of d_1 and d_2, a point ζ satisfies $\zeta \in \mathcal{B}(\sigma, r)$ if either $d_1(\zeta, \sigma) \le r$ or $d_2(\zeta, \sigma) \le r$. From our previous calculations in (19) and (20), we deduce that

$$\mathcal{B}(\sigma, r) = \{\zeta \in \mathbb{R} : -\underline{\omega}(\sigma, r) \le \zeta - \sigma \le \overline{\omega}(\sigma, r)\} \ , \tag{21}$$

where

$$\begin{aligned} \underline{\omega}(\sigma, r) &\triangleq \max\left(\sqrt{\frac{r^2}{4} + \sigma \cdot r} - \frac{r}{2}, r\right) \\ &= \begin{cases} \sqrt{\frac{r^2}{4} + \sigma \cdot r} - \frac{r}{2} & \sigma \ge 2r \ , \\ r & \sigma \le 2r \ , \end{cases} \\ \overline{\omega}(\sigma, r) &\triangleq \max\left(\sqrt{\sigma \cdot r}, r\right) \\ &= \begin{cases} \sqrt{\sigma \cdot r} & \sigma \ge r \ , \\ r & \sigma \le r \ . \end{cases} \end{aligned} \tag{22}$$

To extend $\mathcal{B}$ to vectors we define $\mathbb{R}^{|\mathcal{X}|}$ as the set of vectors with real entries that are indexed by $\mathcal{X}$,

$$\mathbb{R}^{|\mathcal{X}|} \triangleq \{\boldsymbol{\zeta} = (\zeta_x)_{x \in \mathcal{X}} : \zeta_x \in \mathbb{R}\} \ .$$

The set $\mathbb{K}^{|\mathcal{X}|}$ is defined as the set of vectors from $\mathbb{R}^{|\mathcal{X}|}$ with entries summing to 1,

$$\mathbb{K}^{|\mathcal{X}|} \triangleq \left\{\boldsymbol{\zeta} \in \mathbb{R}^{|\mathcal{X}|} : \sum_{x \in \mathcal{X}} \zeta_x = 1\right\} \ .$$

The set $\mathbb{K}_+^{|\mathcal{X}|}$ is the set of probability vectors. Namely, the set of vectors from $\mathbb{K}^{|\mathcal{X}|}$ with non-negative entries,

$$\mathbb{K}_+^{|\mathcal{X}|} \triangleq \left\{\boldsymbol{\zeta} \in \mathbb{K}^{|\mathcal{X}|} : \zeta_x \ge 0\right\} \ .$$

We can now define $\mathcal{B}(\boldsymbol{\alpha}, r)$. For $\boldsymbol{\alpha} \in \mathbb{K}_+^{|\mathcal{X}|}$ let

$$\mathcal{B}(\boldsymbol{\alpha}, r) \triangleq \left\{\boldsymbol{\zeta} \in \mathbb{R}^{|\mathcal{X}|} : d(\boldsymbol{\alpha}, \boldsymbol{\zeta}) \le r\right\} \ . \tag{23}$$

Recalling the maximum in (16), we deduce that $\boldsymbol{\zeta} \in \mathcal{B}(\boldsymbol{\alpha}, r)$ if and only if for all $x \in \mathcal{X}$ we have that $\zeta_x \in \mathcal{B}(\alpha_x, r)$. Thus, by (21) we have a simple characterization of $\mathcal{B}(\boldsymbol{\alpha}, r)$ as a box: a Cartesian product of segments. That is,

$$\begin{aligned} \mathcal{B}(\boldsymbol{\alpha}, r) = \Big\{\boldsymbol{\zeta} \in \mathbb{R}^{|\mathcal{X}|} :& \\ &- \underline{\omega}(\alpha_x, r) \le \zeta_x - \alpha_x \le \overline{\omega}(\alpha_x, r)\Big\} \ . \end{aligned} \tag{24}$$

We stress that the box $\mathcal{B}(\boldsymbol{\alpha}, r)$ contains $\boldsymbol{\alpha}$, but is not necessarily centered at it.

Recall that we have described r_{critical} at the end of Subsection III-B as a quantity ensuring the existence of unique $\alpha, \beta \in \mathcal{Y}_{\text{small}}$ for which $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \le r_{\text{critical}}$. Using our current notation, r_{critical} must imply the existence of distinct $\alpha, \beta \in \mathcal{Y}_{\text{small}}$ such that $\boldsymbol{\beta} \in \mathcal{B}(\boldsymbol{\alpha}, r_{\text{critical}})$. Naturally, in light of (15), we aim to pick r_{critical} as small as we can.

Note that $\mathcal{B}(\boldsymbol{\alpha}, r)$ is a set of vectors in $\mathbb{R}^{|\mathcal{X}|}$. However, the boxes $\mathcal{B}(\boldsymbol{\alpha}, r)$ are induced by points $\boldsymbol{\alpha}$ in $\mathbb{K}_+^{|\mathcal{X}|}$. Trivially, such points $\boldsymbol{\alpha}$ also belong to $\mathbb{K}^{|\mathcal{X}|}$, a translated subspace of $\mathbb{R}^{|\mathcal{X}|}$. Thus, the sphere-packing would yield a tighter result if performed in $\mathbb{K}^{|\mathcal{X}|}$ rather than in $\mathbb{R}^{|\mathcal{X}|}$. Then, for $\boldsymbol{\alpha} \in \mathbb{K}_+^{|\mathcal{X}|}$ and $r > 0$, let us define

$$\mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r) = \mathcal{B}(\boldsymbol{\alpha}, r) \cap \mathbb{K}^{|\mathcal{X}|} \ . \tag{25}$$

When considering $\mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r)$ in place of $\mathcal{B}(\boldsymbol{\alpha}, r)$, we have gained in that the affine dimension (see [17, Section 2.1.3]) of $\mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r)$ is $|\mathcal{X}| - 1$ while that of $\mathcal{B}(\boldsymbol{\alpha}, r)$ is $|\mathcal{X}|$. However, we have lost in simplicity: the set $\mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r)$ is not a box. Indeed, a moment's thought reveals that any subset of $\mathbb{K}^{|\mathcal{X}|}$ with more than one element cannot be a box.

We now show how to overcome the above loss. That is, we show a subset of $\mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r)$ which is — up to a simple transform — a box. Denote the index of the largest entry of a vector $\boldsymbol{\alpha} \in \mathbb{K}^{|\mathcal{X}|}$ as $x_{\max}(\boldsymbol{\alpha})$, namely,

$$x_{\max}(\boldsymbol{\alpha}) \triangleq \arg\max_{x \in \mathcal{X}} \alpha_x \ .$$

In case of ties, define $x_{\max}(\boldsymbol{\alpha})$ in an arbitrary yet consistent manner. For $x_{\max} = x_{\max}(\boldsymbol{\alpha})$ given, or clear from the context, define $\boldsymbol{\zeta}'$ as $\boldsymbol{\zeta}$, with index $x_{\max}$ deleted. That is, for a given $\boldsymbol{\zeta} \in \mathbb{K}^{|\mathcal{X}|}$,

$$\boldsymbol{\zeta}' \triangleq (\zeta_x)_{x \in \mathcal{X}'} \in \mathbb{R}^{|\mathcal{X}|-1} \ ,$$

where $\mathcal{X}' \triangleq \mathcal{X} \setminus \{x_{\max}\}$. Note that for $\boldsymbol{\zeta} \in \mathbb{K}^{|\mathcal{X}|}$, all the entries sum to one. Thus, given $\boldsymbol{\zeta}'$ and $x_{\max}$ we know $\boldsymbol{\zeta}$. Next, for $\boldsymbol{\alpha} \in \mathbb{K}_+^{|\mathcal{X}|}$ and $r > 0$, define the set

$$\begin{aligned} \mathcal{C}(\boldsymbol{\alpha}, r) \triangleq \{ & \boldsymbol{\zeta} \in \mathbb{K}^{|\mathcal{X}|} : \\ & \forall x \in \mathcal{X}' \ , \ -\omega'(\alpha_x, r) \leq \zeta_x - \alpha_x \leq \omega'(\alpha_x, r) \} \ , \end{aligned} \tag{26}$$

where $x_{\max} = x_{\max}(\boldsymbol{\alpha})$ and

$$\begin{aligned} \omega'(\sigma, r) &\triangleq \frac{\underline{\omega}(\sigma, r)}{|\mathcal{X}| - 1} \\ &= \frac{\max\left(\sqrt{r^2/4 + \sigma \cdot r} - r/2, r\right)}{|\mathcal{X}| - 1} \ . \end{aligned} \tag{27}$$

Note that unlike $\mathcal{B}(\boldsymbol{\alpha}, r)$, the center of $\mathcal{C}(\boldsymbol{\alpha}, r)$ is $\boldsymbol{\alpha}$.

Lemma 5. *Let $\boldsymbol{\alpha} \in \mathbb{K}_+^{|\mathcal{X}|}$ and $r > 0$ be given. Let $x_{\max} = x_{\max}(\boldsymbol{\alpha})$. Then,*

$$\mathcal{C}(\boldsymbol{\alpha}, r) \subset \mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r) \ .$$

Proof: By (22), we see that $0 \leq \underline{\omega}(\alpha_x, r) \leq \overline{\omega}(\alpha_x, r)$. Thus, according to (26), it suffices to show that

$$-\underline{\omega}(\alpha_{x_{\max}}, r) \leq \zeta_{x_{\max}} - \alpha_{x_{\max}} \leq \underline{\omega}(\alpha_{x_{\max}}, r) \ . \tag{28}$$

Indeed, summing the condition in (26) over all $x \in \mathcal{X}'$ gives

$$\sum_{x \in \mathcal{X}'} -\omega'(\alpha_x, r) \leq \sum_{x \in \mathcal{X}'} \zeta_x - \sum_{x \in \mathcal{X}'} \alpha_x \leq \sum_{x \in \mathcal{X}'} \omega'(\alpha_x, r) \ .$$

Since $\underline{\omega}(\alpha_x, r)$ is a monotonically non-decreasing function of α_x, we have that $\omega'(\alpha_x, r) \leq \omega'(\alpha_{x_{\max}}, r)$. Hence, the above can be simplified to

$$-\underline{\omega}(\alpha_{x_{\max}}, r) \leq \sum_{x \in \mathcal{X}'} \zeta_x - \sum_{x \in \mathcal{X}'} \alpha_x \leq \underline{\omega}(\alpha_{x_{\max}}, r) \ .$$

Since both $\boldsymbol{\zeta}$ and $\boldsymbol{\alpha}$ are in $\mathbb{K}^{|\mathcal{X}|}$, the middle term in the above is $\alpha_{x_{\max}} - \zeta_{x_{\max}}$. Thus, (28) follows. ■

We remind ourselves of our ultimate goal by stating the following corollary to Lemmas 3 and 5.

Corollary 6. *Let $\alpha, \beta \in \mathcal{Y}_{\text{small}}$ be such that $\boldsymbol{\beta} \in \mathcal{C}(\boldsymbol{\alpha}, r)$. Then, merging α and β induces a penalty $\Delta I_{\downarrow}$ of at most*

$$\Delta I_{\downarrow} \leq \frac{4|\mathcal{X}| r}{|\mathcal{Y}|} \ .$$

Proof: If $\boldsymbol{\beta} \in \mathcal{C}(\boldsymbol{\alpha}, r)$, then according to Lemma 5, we have that $\boldsymbol{\beta} \in \mathcal{B}_{\mathbb{K}}(\boldsymbol{\alpha}, r)$. Using (25) we get that $\boldsymbol{\beta} \in \mathcal{B}(\boldsymbol{\alpha}, r)$, which means that $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r$. Plugging it back in (18) yields the desired result. ■

As outlined before, our aim is to find an r_{critical} for which the conditions of the above corollary surely hold, for some α and β. A standard sphere-packing approach to finding such an r_{critical} is to consider the intersection of spheres of radius $r_{\text{critical}}/2$. Since the triangle inequality does not hold for d, we must use a somewhat different approach. Towards that end, define the positive quadrant associated with $\boldsymbol{\alpha}$ and r as

$$\begin{aligned} \mathcal{Q}'(\boldsymbol{\alpha}, r) \triangleq \{ & \boldsymbol{\zeta}' \in \mathbb{R}^{|\mathcal{X}|-1} : \\ & \forall x \in \mathcal{X}', \ 0 \leq \zeta_x - \alpha_x \leq \omega'(\alpha_x, r) \} \ , \end{aligned} \tag{29}$$

where $x_{\max} = x_{\max}(\boldsymbol{\alpha})$ and $\omega'(\alpha_x, r)$ is as defined in (27).

Lemma 7. *Let $\alpha, \beta \in \mathcal{Y}$ be such that $x_{\max}(\boldsymbol{\alpha}) = x_{\max}(\boldsymbol{\beta})$. If $\mathcal{Q}'(\boldsymbol{\alpha}, r)$ and $\mathcal{Q}'(\boldsymbol{\beta}, r)$ have a non-empty intersection, then $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r$.*

Proof: By (23), (25) and Lemma 5, it suffices to prove that $\boldsymbol{\beta} \in \mathcal{C}(\boldsymbol{\alpha}, r)$. Define $\mathcal{C}'(\boldsymbol{\alpha}, r)$ as the result of applying a prime operation on each member $\boldsymbol{\zeta}$ of $\mathcal{C}(\boldsymbol{\alpha}, r)$, where $x_{\max} = x_{\max}(\boldsymbol{\alpha})$. Namely, we remove the $x_{\max}$ entry from all $\boldsymbol{\zeta}$ in $\mathcal{C}(\boldsymbol{\alpha}, r)$. Hence, we must equivalently prove that $\boldsymbol{\beta}' \in \mathcal{C}'(\boldsymbol{\alpha}, r)$. By (26), we must show that for all $x \in \mathcal{X}'$,

$$-\omega'(\alpha_x, r) \leq \beta_x - \alpha_x \leq \omega'(\alpha_x, r) \ . \tag{30}$$

Since we know that the intersection of $\mathcal{Q}'(\boldsymbol{\alpha}, r)$ and $\mathcal{Q}'(\boldsymbol{\beta}, r)$ is non-empty, let $\boldsymbol{\zeta}'$ be a member of both sets. Thus, we know that for $x \in \mathcal{X}'$,

$$0 \leq \zeta_x - \alpha_x \leq \omega'(\alpha_x, r) \ ,$$

and

$$0 \leq \zeta_x - \beta_x \leq \omega'(\beta_x, r) \ .$$

For each $x \in \mathcal{X}'$ we must consider two cases: $\alpha_x \leq \beta_x$ and $\alpha_x > \beta_x$.

Consider first the case $\alpha_x \leq \beta_x$. Since $\zeta_x - \alpha_x \leq \omega'(\alpha_x, r)$ and $\beta_x - \zeta_x \leq 0$, we conclude that $\beta_x - \alpha_x \leq \omega'(\alpha_x, r)$. Conversely, since $\beta_x - \alpha_x \geq 0$ and, by (27), $\omega'(\alpha_x, r) \geq 0$, we have that $\beta_x - \alpha_x \geq -\omega'(\alpha_x, r)$. Thus we have shown that both inequalities in (30) hold.

To finish the proof, consider the case $\alpha_x > \beta_x$. We have already established that $\omega'(\alpha_x, r) \geq 0$. Thus, since by assumption $\beta_x - \alpha_x \leq 0$, we have that $\beta_x - \alpha_x \leq \omega'(\alpha_x, r)$. Conversely, since $\zeta_x - \beta_x \leq \omega'(\beta_x, r)$ and $\alpha_x - \zeta_x \leq 0$, we have that $\alpha_x - \beta_x \leq \omega'(\beta_x, r)$. We now recall that by (27), the fact that $\alpha_x \geq \beta_x$ implies that $\omega'(\beta_x, r) \leq \omega'(\alpha_x, r)$. Thus, $\alpha_x - \beta_x \leq \omega'(\alpha_x, r)$. Negating gives $\beta_x - \alpha_x \geq -\omega'(\alpha_x, r)$, and we have once again proved the two inequalities in (30). ■

D. Weighted "sphere"-packing

At first glance, thanks to Lemma 7, the quadrant $\mathcal{Q}'(\boldsymbol{\alpha}, r)$ could have been the equivalent of a "sphere" in the sense of d. However, recalling (27), we see that the dimensions of the quadrant $\mathcal{Q}'(\boldsymbol{\alpha}, r)$ are dependent on the point $\boldsymbol{\alpha}$. Specifically, for r fixed, a larger α_x is preferable, since the corresponding sphere will take up a larger volume. That is, the side length $\omega'(\alpha_x, r)$, corresponding to coordinate x, is increasing in α_x. We now show how to partially offset this dependence on $\boldsymbol{\alpha}$. Towards this end, we define a density over $\mathbb{R}^{|\mathcal{X}|-1}$ and derive a lower bound on the weight of a "sphere" that does not depend on $\boldsymbol{\alpha}$. That is, many sphere packing arguments typically involve an argument having to do with the volume of a sphere versus the volume of the space the spheres are placed in. In contrast, our argument will involve the *weight* of a sphere versus the weight of the space the spheres are placed in. We now define our density function φ. The utility of the following definition is that it partially offsets the above described dependence on α_x: the smaller the argument, the larger the density. Specifically, let $\varphi : \mathbb{R} \to \mathbb{R}$ be defined as

$$\varphi(\zeta) \triangleq \frac{1}{2\sqrt{\zeta}} \; . \tag{31}$$

Next, for $\boldsymbol{\zeta}' \in \mathbb{R}^{|\mathcal{X}|-1}$, abuse notation and define $\varphi : \mathbb{R}^{|\mathcal{X}|-1} \to \mathbb{R}$ as

$$\varphi(\boldsymbol{\zeta}') \triangleq \prod_{x \in \mathcal{X}'} \varphi(\zeta_x) \; . \tag{32}$$

The weight of $\mathcal{Q}'(\boldsymbol{\alpha}, r)$ is then defined as

$$M\left[\mathcal{Q}'(\boldsymbol{\alpha}, r)\right] \triangleq \int_{\mathcal{Q}'(\boldsymbol{\alpha},r)} \varphi(\boldsymbol{\zeta}')\, \mathrm{d}\boldsymbol{\zeta}' \; . \tag{33}$$

The following lemma proposes a lower bound on $M\left[\mathcal{Q}'(\boldsymbol{\alpha}, r)\right]$ that does not depend on $\boldsymbol{\alpha}$.

Lemma 8. *The weight* $M\left[\mathcal{Q}'(\boldsymbol{\alpha}, r)\right]$ *satisfies*

$$M\left[\mathcal{Q}'(\boldsymbol{\alpha}, r)\right] \geq r^{\frac{|\mathcal{X}|-1}{2}} \left(\sqrt{2 + \frac{1}{|\mathcal{X}|-1}} - \sqrt{2} \right)^{|\mathcal{X}|-1} \; . \tag{34}$$

Proof: Since $\varphi(\boldsymbol{\zeta}')$ is a product,

$$\begin{aligned} M\left[\mathcal{Q}'(\boldsymbol{\alpha}, r)\right] &= \prod_{x \in \mathcal{X}'} \int_{\alpha_x}^{\alpha_x + \omega'(\alpha_x, r)} \frac{\mathrm{d}\zeta_x}{2\sqrt{\zeta_x}} \\ &= \prod_{x \in \mathcal{X}'} \psi_r(\alpha_x) \; , \end{aligned}$$

where $\psi_r(\sigma) \triangleq \sqrt{\sigma + \omega'(\sigma, r)} - \sqrt{\sigma}$. It suffices to show that $\psi_r(\sigma)$ is decreasing for $\sigma < 2r$, and increasing for $\sigma > 2r$. Assume first that $\sigma < 2r$. Then,

$$\frac{\mathrm{d}\psi_r}{\mathrm{d}\sigma} = \frac{1}{2\sqrt{\sigma + r/(|\mathcal{X}|-1)}} - \frac{1}{2\sqrt{\sigma}} < 0 \; ,$$

for all $\sigma \geq 0$. Assume now that $\sigma > 2r$. Then,

$$\begin{aligned} \frac{\mathrm{d}\psi_r}{\mathrm{d}\sigma} = \frac{1}{2} & \left(\sqrt{\sigma + \frac{\sqrt{r^2/4 + \sigma \cdot r} - r/2}{|\mathcal{X}|-1}} \right)^{-1} \\ & \cdot \left(1 + \frac{r}{2(|\mathcal{X}|-1)\sqrt{r^2/4 + \sigma \cdot r}} \right) - \frac{1}{2\sqrt{\sigma}} \; . \end{aligned}$$

Comparing the derivative to zero and solving for σ yields

$$\begin{aligned} \sigma + & \frac{\sqrt{r^2/4 + \sigma \cdot r} - r/2}{|\mathcal{X}|-1} \\ & = \sigma \left(1 + \frac{r}{2(|\mathcal{X}|-1)\sqrt{r^2/4 + \sigma \cdot r}} \right)^2 , \end{aligned}$$

which is equivalent to

$$\begin{aligned} & \frac{\sqrt{r^2/4 + \sigma \cdot r} - r/2}{|\mathcal{X}|-1} = \\ & \frac{\sigma \cdot r}{(|\mathcal{X}|-1)\sqrt{r^2/4 + \sigma \cdot r}} + \frac{\sigma \cdot r^2}{4(|\mathcal{X}|-1)^2\,(r^2/4 + \sigma \cdot r)} \; . \end{aligned}$$

We define $\xi \triangleq \sqrt{\frac{r^2}{4} + \sigma \cdot r}$ and get

$$\xi - \frac{r}{2} = \frac{\xi^2 - \frac{r^2}{4}}{\xi} + \frac{r(\xi^2 - \frac{r^2}{4})}{4(|\mathcal{X}|-1)\xi^2} \; .$$

Now solving for ξ we have

$$(2|\mathcal{X}|-1)\xi^2 - (|\mathcal{X}|-1)r\xi - \frac{r^2}{4} = 0 \; ,$$

and since $\xi > 0$ the only solution is $\xi = r/2$ which implies $\sigma = 0$. We note that this is not a real solution since the derivative is not defined for $\sigma = 0$. Hence, the derivative is either non-negative or non-positive. By plugging $\sigma = 2r$ in the derivative we get

$$\begin{aligned} \left.\frac{\mathrm{d}\psi_r}{\mathrm{d}\sigma}\right|_{\sigma=2r} &= \frac{1}{2}\left(\sqrt{2r + \frac{r}{|\mathcal{X}|-1}} \right)^{-1} \cdot \left(1 + \frac{r}{(|\mathcal{X}|-1)3r} \right) \\ &\quad - \frac{1}{2\sqrt{2r}} \\ &= \frac{1}{2r\sqrt{2}} \left(\frac{1 + \frac{1}{3(|\mathcal{X}|-1)}}{\sqrt{1 + \frac{1}{2(|\mathcal{X}|-1)}}} - 1 \right) \\ &\geq 0 \; , \end{aligned}$$

for $|\mathcal{X}| \geq 2$ and $r > 0$. Thus, by continuity of ψ_r,

$$\psi_r(\alpha_x) \geq \psi_r(2r) \; ,$$

and since this lower bound does not depend on α_x we get (34). ■

We partition the letters in $\mathcal{Y}_{\text{small}}$ according to their $x_{\max}$ value to $|\mathcal{X}|$ subsets (at most). The largest subset is denoted by $\mathcal{Y}'$. We henceforth limit our search to $\mathcal{Y}'$. Thus, from this point onward, $x_{\max}$ is fixed.

Let $\mathcal{V}'$ be the union of all the quadrants corresponding to possible choices of $\boldsymbol{\alpha}$. Namely,

$$\mathcal{V}' \triangleq \bigcup_{\substack{\boldsymbol{\alpha} \in \mathbb{K}_+^{|\mathcal{X}|} \\ x_{\max}(\boldsymbol{\alpha}) = x_{\max}}} \mathcal{Q}'(\boldsymbol{\alpha}, r) \; .$$

In order to bound the weight of $\mathcal{V}'$, we introduce the simpler set

$$\mathcal{U}' \triangleq \left\{ \boldsymbol{\zeta}' \in \mathbb{R}^{|\mathcal{X}|-1} : \sum_{x \in \mathcal{X}'} \zeta_x \leq 2, \; \zeta_x \geq 0 \, \forall x \in \mathcal{X}' \right\} \; .$$

The constraint $r \leq 1$ in the following lemma will be motivated shortly.

Lemma 9. *Let* $r \leq 1$. *Then,* $\mathcal{V}' \subseteq \mathcal{U}'$.

Proof: Assume $\boldsymbol{\zeta}' \in \mathcal{V}'$. Then, there exists $\boldsymbol{\alpha} \in \mathbb{K}_+^{|\mathcal{X}|}$ such that $0 \leq \zeta_x - \alpha_x \leq \omega'(\alpha_x, r)$ for all $x \in \mathcal{X}'$. Hence, $\zeta_x \geq 0$ for all $x \in \mathcal{X}'$. Moreover,

$$\begin{aligned} \sum_{x \in \mathcal{X}'} \zeta_x &\leq \sum_{x \in \mathcal{X}'} \alpha_x + \sum_{x \in \mathcal{X}'} \omega'(\alpha_x, r) \\ &\leq 1 - \alpha_{x_{\max}} + \underline{\omega}(\alpha_{x_{\max}}, r) \; . \end{aligned} \tag{35}$$

There are two cases to consider. In the case where $\alpha_{x_{\max}} \geq 2r$ we have

$$\begin{aligned}\sum_{x \in \mathcal{X}'} \zeta_x &\leq 1 - \alpha_{x_{\max}} + \sqrt{\frac{r^2}{4} + \alpha_{x_{\max}} r} - \frac{r}{2} \\ &\leq 1 - \alpha_{x_{\max}} + \sqrt{\frac{\alpha_{x_{\max}}^2}{16} + \frac{\alpha_{x_{\max}}^2}{2}} - \frac{r}{2} \\ &= 1 - \frac{\alpha_{x_{\max}}}{4} - \frac{r}{2} \\ &\leq 2 \, ,\end{aligned}$$

where the second inequality is due to the assumption $\alpha_{x_{\max}} \geq 2r$. In the case where $\alpha_{x_{\max}} \leq 2r$, (35) becomes

$$\begin{aligned}\sum_{x \in \mathcal{X}'} \zeta_x &\leq 1 - \alpha_{x_{\max}} + r \\ &\leq 2 - \alpha_{x_{\max}} \\ &\leq 2 \, ,\end{aligned}$$

where we assumed $r \leq 1$. Therefore, $\boldsymbol{\zeta}' \in \mathcal{U}'$. ■

The lemma above and the non-negativity of φ, enable us to upper bound the weight of $\mathcal{V}'$, denoted by $M\left[\mathcal{V}'\right]$, using

$$M\left[\mathcal{V}'\right] \triangleq \int_{\mathcal{V}'} \varphi \, \mathrm{d}\boldsymbol{\zeta}' \leq \int_{\mathcal{U}'} \varphi \, \mathrm{d}\boldsymbol{\zeta}' \, . \tag{36}$$

We define the mapping $\rho_x = \sqrt{\zeta_x}$ for all $x \in \mathcal{X}'$ and perform a change of variables. As a result, $\mathcal{U}'$ is mapped to

$$\mathcal{S}' \triangleq \left\{ \boldsymbol{\rho}' \in \mathbb{R}^{|\mathcal{X}|-1} : \sum_{x \in \mathcal{X}'} \rho_x^2 \leq 2, \; \rho_x \geq 0 \right\} \, ,$$

which is a quadrant of a $|\mathcal{X}| - 1$ dimensional ball of a $\sqrt{2}$ radius. By (31), we have that

$$\frac{\mathrm{d}\zeta_x}{\mathrm{d}\rho_x} = 2\rho_x = 2\sqrt{\zeta_x} = \frac{1}{\varphi(\zeta_x)} \, .$$

Thus, we have by (32) that after the above change of variables, the Jacobian determinant exactly cancels φ,

$$\int_{\mathcal{U}'} \varphi \, \mathrm{d}\boldsymbol{\zeta}' = \int_{\mathcal{S}'} \mathrm{d}\boldsymbol{\rho}' \, . \tag{37}$$

Hence, by (36) and (37)

$$\begin{aligned}M\left[\mathcal{V}'\right] &\leq \int_{\mathcal{S}'} \mathrm{d}\boldsymbol{\rho}' \\ &= \frac{1}{2^{|\mathcal{X}|-1}} \frac{\pi^{\frac{|\mathcal{X}|-1}{2}}}{\Gamma\left(1 + \frac{|\mathcal{X}|-1}{2}\right)} 2^{\frac{|\mathcal{X}|-1}{2}} \\ &= \left(\frac{\pi}{2}\right)^{\frac{|\mathcal{X}|-1}{2}} \frac{1}{\Gamma\left(1 + \frac{|\mathcal{X}|-1}{2}\right)} \, ,\end{aligned} \tag{38}$$

where we have used the well known expression for the volume of a multidimensional ball. Thus, we are ready to prove Theorem 2.

Proof of Theorem 2: Recall that we are assuming $|\mathcal{Y}| > 2|\mathcal{X}|$. According to the definition of $\mathcal{Y}'$, we get

$$|\mathcal{Y}'| \geq \frac{|\mathcal{Y}_{\text{small}}|}{|\mathcal{X}|} \geq \frac{|\mathcal{Y}|}{2|\mathcal{X}|} > 1 \, , \tag{39}$$

where we used (17). As a result, we have at least two points in $\mathcal{Y}'$, and are therefore in a position to apply a sphere-packing argument. Towards this end, let r be such that the starred equality in the following derivation holds:

$$\begin{aligned}&\sum_{\boldsymbol{\alpha} \in \mathcal{Y}'} M\left[\mathcal{Q}'(\boldsymbol{\alpha}, r)\right] \\ &\quad \geq \frac{|\mathcal{Y}|}{2|\mathcal{X}|} \cdot r^{\frac{|\mathcal{X}|-1}{2}} \left(\sqrt{2 + \frac{1}{|\mathcal{X}|-1}} - \sqrt{2} \right)^{|\mathcal{X}|-1} \\ &\quad \overset{(*)}{=} \left(\frac{\pi}{2}\right)^{\frac{|\mathcal{X}|-1}{2}} \frac{1}{\Gamma\left(1 + \frac{|\mathcal{X}|-1}{2}\right)} \\ &\quad \geq M\left[\mathcal{V}'\right] \, .\end{aligned} \tag{40}$$

Namely,

$$\begin{aligned}r \triangleq \frac{\pi}{4} &\left(\sqrt{1 + \frac{1}{2(|\mathcal{X}|-1)}} - 1 \right)^{-2} \\ &\cdot \left(\frac{2|\mathcal{X}|}{\Gamma\left(1 + \frac{|\mathcal{X}|-1}{2}\right)} \right)^{\frac{2}{|\mathcal{X}|-1}} \cdot |\mathcal{Y}|^{-\frac{2}{|\mathcal{X}|-1}} \, .\end{aligned} \tag{41}$$

There are two cases to consider. If $r \leq 1$, then all of (40) holds, by (34), (38) and (39). We take $r_{\text{critical}} = r$, and deduce the existence of a pair $\alpha, \beta \in \mathcal{Y}'$ for which $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r$. Indeed, assuming otherwise would contradict (40), since each $\mathcal{Q}'$ in the sum is contained in $\mathcal{V}'$, and, by Lemma 7 and our assumption, all summed $\mathcal{Q}'$ are disjoint.

We next consider the case $r > 1$. Now, any pair of letters $\alpha, \beta \in \mathcal{Y}'$ satisfies $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r$. Indeed, by (14) and (16),

$$d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_\infty \leq 1 < r \, ,$$

where $\|\cdot\|_\infty$ is the maximum norm.

We have proved the existence of $\alpha, \beta \in \mathcal{Y}' \subset \mathcal{Y}_{\text{small}}$ for which $d(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq r$. By (18) and (41), the proof is finished. ■

Finally, we prove Theorem 1.

Proof of Theorem 1: If $L \geq |\mathcal{Y}|$, then obviously $\Delta I_{\downarrow}^* = 0$ which is not the interesting case. If $2|\mathcal{X}| \leq L < |\mathcal{Y}|$, then applying Theorem 2 repeatedly $|\mathcal{Y}| - L$ times yields

$$\begin{aligned}\Delta I_{\downarrow}^* &\leq \sum_{\ell = L+1}^{|\mathcal{Y}|} \mu(|\mathcal{X}|) \cdot \ell^{-\frac{|\mathcal{X}|+1}{|\mathcal{X}|-1}} \\ &\leq \mu(|\mathcal{X}|) \int_L^{|\mathcal{Y}|} \ell^{-\frac{|\mathcal{X}|+1}{|\mathcal{X}|-1}} \, \mathrm{d}\ell \\ &= \nu(|\mathcal{X}|) \left(L^{-\frac{2}{|\mathcal{X}|-1}} - |\mathcal{Y}|^{-\frac{2}{|\mathcal{X}|-1}} \right) \\ &\leq \nu(|\mathcal{X}|) \cdot L^{-\frac{2}{|\mathcal{X}|-1}} \, ,\end{aligned}$$

by the monotonicity of $\ell^{-(|\mathcal{X}|+1)/(|\mathcal{X}|-1)}$. The bound is tight in the power-law sense by [10, Theorem 2], in which a sequence of channels is proved to obtain

$$\Delta I_{\downarrow}^* \geq \delta(|\mathcal{X}|) \cdot L^{-\frac{2}{|\mathcal{X}|-1}} \, ,$$

for a specified $\delta(\cdot)$. ■

We note that Theorem 1 can be generalized to channels with a continuous output alphabet. This is done using arbitrarily close approximating degraded channels [12][14], with corresponding large finite output alphabets.

E. Symmetric channels

We note that degrading a symmetric channel optimally does not necessarily yield a symmetric channel [18, Theorem 2][19]. However, often, the fact that the resultant channel is symmetric is important. For example, [1, Theorem 4] proves that if the underlying binary-input channel is symmetric, then randomness is not required in some parts of the process (namely, to borrow the nomenclature of [1], any choice of frozen bits is as good as any other). Thus, we would like an analog of Theorem 1, for the case in which all channels involved are symmetric. We have indeed found such an analog, for a restricted set of symmetric channels, defined as *cyclo-symmetric* channels.

A channel $W : \mathcal{X} \to \mathcal{Y}$ is cyclo-symmetric if the following holds.

1) The input alphabet is labeled $\mathcal{X} \triangleq \{0, 1, \ldots, |\mathcal{X}| - 1\}$.
2) The output alphabet size is a multiple of $|\mathcal{X}|$, and partitioned into $|\mathcal{Y}|/|\mathcal{X}|$ disjoint sets $\{\mathcal{Y}_i\}_{i=1}^{|\mathcal{Y}|/|\mathcal{X}|}$. Each such $\mathcal{Y}_i$ contains $|\mathcal{X}|$ members,
$$\mathcal{Y}_i \triangleq \left\{ y_i^{(0)}, y_i^{(1)}, \ldots, y_i^{(|\mathcal{X}|-1)} \right\} .$$
3) For $0 \le \theta \le |\mathcal{X}| - 1$,
$$W(y_i^{(0)}|x) = W(y_i^{(\theta)}|x + \theta) , \tag{42}$$
where $x + \theta$ is short for $x + \theta \mod |\mathcal{X}|$.

Note that a cyclo-symmetric channel is symmetric, according to the definition in [20, page 94]. Hence, the capacity-achieving input distribution is the uniform input distribution $\pi(x) = 1/|\mathcal{X}|$ [20, Theorem 4.5.2]. We remark in passing that in the binary-input case, $|\mathcal{X}| = 2$, a symmetric channel is essentially cyclo-symmetric as well: the only problematic symbol is the erasure symbol, which can be split, as discussed in [2, Lemma 4].

Theorem 10. *Let a DMC $W : \mathcal{X} \to \mathcal{Y}$ be cyclo-symmetric and satisfy $|\mathcal{Y}| > 2|\mathcal{X}|$ and let $L \ge 2|\mathcal{X}|$ be a multiple of $|\mathcal{X}|$. Fix the input distribution P_X to be uniform. Then,*
$$\Delta I^*_\downarrow = \min_{\substack{Q:Q\preccurlyeq W,\\ |Q|\le L}} I(W, P_X) - I(Q, P_X) = O\left(L^{-\frac{2}{|\mathcal{X}|-1}}\right) ,$$
where the optimization is over Q that are cyclo-symmetric. In particular,
$$\Delta I^*_\downarrow \le \nu(|\mathcal{X}|) \cdot L^{-\frac{2}{|\mathcal{X}|-1}} ,$$
where $\nu(|\mathcal{X}|)$ is as defined in Theorem 1. This bound is attained by a simple modification of greedy-merge, and is tight in the power-law sense.

Before getting into the proof, let us explain the modification of greedy-merge mentioned in the theorem. Using the above notation, in greedy-merge we are to choose the $y_i^{(t)}$ and $y_j^{(t')}$ whose merger results in the smallest drop in mutual information between input and output. In our modified algorithm, we limit our search to the case in which $i \ne j$. Namely, the symbols are taken from $\mathcal{Y}_i$ and $\mathcal{Y}_j$, and these sets are distinct. After we have completed our search and found the above $y_i^{(t)}$ and $y_j^{(t')}$, we merge $\mathcal{Y}_i$ and $\mathcal{Y}_j$ into a new set $\mathcal{Y}_{ij}$, in the following sense: for $0 \le \theta \le |\mathcal{X}| - 1$, we merge $y_i^{(t+\theta)}$ and $y_j^{(t'+\theta)}$ into $y_{ij}^{(\theta)}$, where $t + \theta$ and $t' + \theta$ are calculated modulo $|\mathcal{X}|$. As in greedy-merge, the operation of merging $\mathcal{Y}_i$ and $\mathcal{Y}_j$ is repeated until the output alphabet size is small enough.

Proof of Theorem 10: We start by proving that the resulting channel Q is cyclo-symmetric. To do so, we prove that each merging iteration — merging of sets $\mathcal{Y}_i$ and $\mathcal{Y}_j$ — preserves cyclo-symmetry. Suppose for notational convenience that only one such merge iteration is needed, taking us from channel W to channel Q. Let the merging be carried out using the notation above: $\mathcal{Y}_i$ and $\mathcal{Y}_j$ are merged to form $\mathcal{Y}_{ij}$, with $y_i^{(t)}$ and $y_j^{(t')}$ as the pair initiating the merger. To prove that cyclo-symmetry is preserved, we must show that (42) holds. Namely, for all $x \in \mathcal{X}$,
$$Q(y_{ij}^{(0)}|x) = Q(y_{ij}^{(\theta)}|x + \theta) .$$
The above is equivalent to showing that
$$W(y_i^{(t)}|x) + W(y_j^{(t')}|x) = W(y_i^{(t+\theta)}|x+\theta) + W(y_j^{(t'+\theta)}|x+\theta) ,$$
which follows immediately from (42). Obviously, if the output alphabet of W is a multiple of $|\mathcal{X}|$, then the output alphabet of Q is smaller by $|\mathcal{X}|$, and is thus still a multiple of $|\mathcal{X}|$.

We now move to proving the upper-bound on the difference in mutual information. Since Theorem 1 is a direct consequence of Theorem 2, it suffices to prove that each sub-merger of $y_i^{(t+\theta)}$ and $y_j^{(t'+\theta)}$ attains the bound in Theorem 2. Namely, the bound corresponding to $\theta = 0, 1, \ldots, |\mathcal{X}| - 1$ must hold with respect to $|\mathcal{Y}| - \theta$ output letters.

Let us first consider the case $\theta = 0$. Recall that for $\theta = 0$, the only constraint imposed by our version of greedy-merge is that the two symbols merged, $y_i^{(t)}$ and $y_j^{(t')}$, must have $i \ne j$. Apart from this, as in regular greedy-merge, we pick the pair $y_i^{(t)}$ and $y_j^{(t')}$ for which the reduction in mutual information is minimal. Thus, we must show that this added constraint is still compatible with the proof of Theorem 2. Recall that in the proof of Theorem 2, only symbols with the same $x_{\max}$ are considered. Thus, the proof will indeed be compatible with our version of greedy-merge if we manage to show that all the symbols in a generic subset $\mathcal{Y}_i$ have distinct $x_{\max}$. Indeed, by (42), the $x_{\max}$ corresponding to $y_i^{(\theta)}$ is simply the $x_{\max}$ corresponding to $y_i^{(0)}$, shifted by θ places (where, if needed, ties are broken accordingly).

Recall that we are considering (7), with $|\mathcal{Y}|$ replaced by $|\mathcal{Y}| - \theta$. Let us term this (7)'. We have just proved that (7)' holds for $\theta = 0$, and our aim now is to prove it for $1 \le \theta < |\mathcal{X}|$. Since the input distribution is uniform, we have by (42) that the difference in mutual information between input and output resulting from merging $y_i^{(t+\theta)}$ and $y_j^{(t'+\theta)}$ equals that from merging $y_i^{(t)}$ and $y_j^{(t')}$. That is, the LHS of (7)' is independent of θ. Since the RHS of (7)' is increasing in θ, and the claim holds for $\theta = 0$, we are done.

Lastly, we must prove the claim of tightness, in the power-law sense. This is so, since the channels in [10, Theorem 2] are essentially cyclo-symmetric. That is, consider the output symbols in such a channel. All symbols having a corresponding posterior probability vector with period $|\mathcal{X}|$ can be grouped into subsets satisfying (42). The remaining symbols (a vanishing

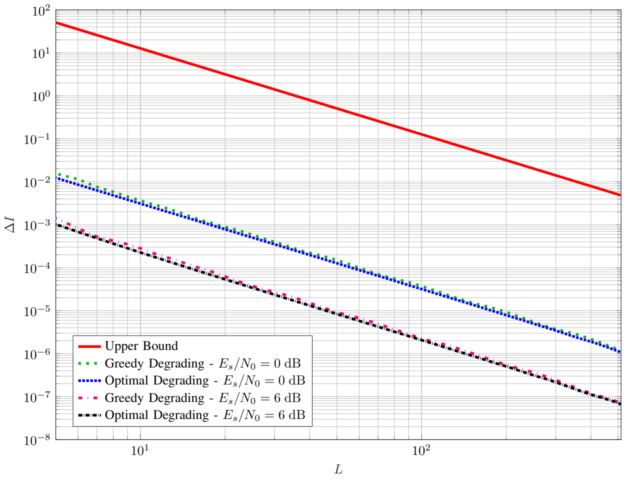


Fig. 3. The performance of optimal and greedy degrading algorithms for a binary-input AWGN channel with $E_s/N_0 = 0$ dB and $E_s/N_0 = 6$ dB, and the upper bound from Theorem 1.

fraction) will have posterior probability vectors with period dividing $|\mathcal{X}|$. These can be "split", similarly to how erasure symbols are split in [2, Lemma 4], to yield an equivalent channel which is cyclo-symmetric. ■

F. Numerical example

In this subsection we compare the performance of the optimal and greedy algorithms applied to specific channels, along with the universal upper bound we proved in Theorem 1. The results shown in Figure 3 are for a binary-input AWGN[2] channel, with $E_s/N_0 = 0$ dB and $E_s/N_0 = 6$ dB, for every $L \in \{5, 6, \ldots, 512\}$. Specifically, we have chosen some rather typical "real world" channel, in stark contrast to the somewhat contrived channels used in [10]. Nevertheless, although the power law is only currently known to be tight for the channels in [10], the graph gives a very strong numerical indication that it also holds for the considered AWGN channels as well. That is, in our log-log plot, all curves have the same slope. It is also remarkable how close each pair of curves are. That is, how close to optimal the greedy approach was in this case.

IV. Lower bound on upgrading cost

Recall the definition of upgrading-cost given in (4) and (5). In this section, we derive a lower bound on the upgrading cost.

Theorem 11. *Given $|\mathcal{X}|$ and L, the upgrading cost defined in* (5) *satisfies*

$$\mathrm{UC} \geq \kappa(|\mathcal{X}|) \cdot L^{-\frac{2}{|\mathcal{X}|-1}} \ , \tag{43}$$

where

$$\kappa(|\mathcal{X}|) \triangleq \frac{|\mathcal{X}|-1}{2\pi \cdot (|\mathcal{X}|+1)} \cdot \left(\frac{\Gamma\left(1 + \frac{|\mathcal{X}|-1}{2}\right)}{(|\mathcal{X}|-1)!} \right)^{\frac{2}{|\mathcal{X}|-1}} ,$$

and $\Gamma(\cdot)$ is the Gamma function, defined in (6).

Note that for large values of $|\mathcal{X}|$ the Stirling approximation can be applied to simplify $\kappa(|\mathcal{X}|)$ to

$$\kappa(|\mathcal{X}|) \approx \frac{e}{4\pi \cdot (|\mathcal{X}|-1)} \ .$$

The proof of the above theorem will rely on a sequence of channels that are "hard" to upgrade. It turns out that these channels are exactly the channels that [10] proved were hard to *degrade*. In fact, more is true: the lower bound of Theorem 11 is exactly equal to the lower bound proved in [10, Theorem 2]. As a result, this section will be rather short: we will first prove two lemmas which are specific to the upgrading case, and then use them to show that a key part of the proof of [10, Theorem 2] is applicable to our setting.

We now fix some notation. Let $W : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{X} \to \mathcal{Z}$ be two DMCs such that Q is upgraded with respect to W, that is $W \preccurlyeq Q$. We assume, again, without loss of generality, that $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ are disjoint and that an input distribution is fixed. We can thus abuse notation and define

$$\begin{aligned} y_x &\triangleq W(x|y) \ , \\ z_x &\triangleq Q(x|z) \ , \end{aligned} \tag{44}$$

and the corresponding vectors $\boldsymbol{y} \triangleq (y_x)_{x \in \mathcal{X}}$, $\boldsymbol{z} \triangleq (z_x)_{x \in \mathcal{X}}$. Since we will be using this notation heavily, we stress that both $\boldsymbol{y}$ and $\boldsymbol{z}$ (boldfaced) are vectors whose x entries, $x \in \mathcal{X}$, are the probabilities y_x and z_x, respectively, corresponding to the (non-boldfaced) output letters $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$.

Let us think of $\mathcal{Y}$ as a "large" alphabet, that is reduced to a "small" alphabet $\mathcal{Z}$. For each $z \in \mathcal{Z}$, we define $\mathcal{A}_z$ as the set of letters in $\mathcal{Y}$ that are closest to z, with respect to Euclidean distance between posterior vectors. That is, for $z \in \mathcal{Z}$

$$\mathcal{A}_z \triangleq \left\{ y \in \mathcal{Y} : z = \operatorname*{arg\,min}_{z' \in \mathcal{Z}} \|\boldsymbol{z}' - \boldsymbol{y}\|_2^2 \right\} \ , \tag{45}$$

where $\|\cdot\|_2$ is the Euclidean norm. We stress that the sets $\{\mathcal{A}_z\}_{z \in \mathcal{Z}}$ are disjoint. That is, "arg min" ties are broken in an arbitrary yet consistent manner.

We now show how the sets $\mathcal{A}_z$ can be used to derive a lower bound on the cost of upgrading W to Q. As before, we use the shorthand π_y to denote $\pi(y)$.

Lemma 12. *Let the DMCs $W : \mathcal{X} \to \mathcal{Y}$ and $Q : \mathcal{X} \to \mathcal{Z}$ satisfy $W \preccurlyeq Q$. Assume a fixed input distribution. Then,*

$$\Delta I_{\uparrow} \triangleq I(Q, P_X) - I(W, P_X) \geq \sum_{z \in \mathcal{Z}} \Delta(\mathcal{A}_z) \ ,$$

where

$$\Delta(\mathcal{A}_z) \triangleq \frac{1}{2} \sum_{y \in \mathcal{A}_z} \pi_y \, \|\boldsymbol{z} - \boldsymbol{y}\|_2^2 \ . \tag{46}$$

Proof: Using our notation,

$$\Delta I_{\uparrow} = \sum_{y \in \mathcal{Y}} \pi_y h(\boldsymbol{y}) - \sum_{z \in \mathcal{Z}} \pi_z h(\boldsymbol{z}) \ , \tag{47}$$

where

$$h(\boldsymbol{y}) \triangleq \sum_{x \in \mathcal{X}} \eta(y_x) \ ,$$

[2] Recall that the optimal degrading and greedy merge degrading algorithms operate on channels with a finite output alphabet size. Thus, as a pre-processing step, the AWGN output alphabet was first very finely sampled (degraded) to have a discrete output alphabet size containing 4000 points.

and $\eta(\cdot)$ was defined in (1). Since $W \preccurlyeq Q$, there exists an intermediate channel $\Phi : \mathcal{Z} \to \mathcal{Y}$ such that concatenating Φ to Q results in W. We now claim that this concatenation applies also to posterior probabilities,

$$\boldsymbol{y} = \sum_{z\in\mathcal{Z}} \Phi_{z|y}\boldsymbol{z} \;, \tag{48}$$

where for $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$

$$\Phi_{z|y} = \Phi(z|y) \triangleq \frac{\Phi(y|z)\cdot\pi_z}{\pi_y} \tag{49}$$

is the "reverse" or "posterior" channel, often also called the "test" channel. Note that (48) follows from (2) and an application of Bayes' rule. Moreover, by (49),

$$\pi_z = \sum_{y\in\mathcal{Y}} \Phi_{z|y}\pi_y \;. \tag{50}$$

Plugging (48) and (50) in (47) yields

$$\Delta I_{\uparrow} = \sum_{y\in\mathcal{Y}} \pi_y \left(h\left(\sum_{z\in\mathcal{Z}} \Phi_{z|y}\boldsymbol{z} \right) - \sum_{z\in\mathcal{Z}} \Phi_{z|y} h\left(\boldsymbol{z}\right) \right) \;. \tag{51}$$

It easily follows from (49) that $\sum_{z\in\mathcal{Z}} \Phi_{z|y} = 1$. Hence, since $h(\cdot)$ is concave, we can apply Jensen's inequality to the expression contained by the outer parentheses of (51) and conclude that it is non-negative. However, as in [10, Corollary 5], we invoke a stronger inequality, known as Hölder's defect formula [21, Page 94]. This yields

$$\begin{aligned} h\left(\sum_{z\in\mathcal{Z}} \Phi_{z|y}\boldsymbol{z} \right) - \sum_{z\in\mathcal{Z}} \Phi_{z|y} h\left(\boldsymbol{z}\right) \geq \\ \frac{1}{2}\lambda_{\min}(-\nabla^2 h) \sum_{z\in\mathcal{Z}} \Phi_{z|y} \left\| \boldsymbol{z} - \sum_{z'\in\mathcal{Z}} \Phi_{z'|y}\boldsymbol{z}' \right\|_2^2 \;, \end{aligned}$$

where $-\nabla^2 h$ is the negated Hessian matrix of h, and $\lambda_{\min}(-\nabla^2 h)$ is its smallest eigenvalue. Using (48) for the term inside the norm and $\lambda_{\min}(-\nabla^2 h) \geq 1$ (proved in [10, Corollary 5]), we get

$$h\left(\sum_{z\in\mathcal{Z}} \Phi_{z|y}\boldsymbol{z} \right) - \sum_{z\in\mathcal{Z}} \Phi_{z|y} h\left(\boldsymbol{z}\right) \geq \frac{1}{2} \sum_{z\in\mathcal{Z}} \Phi_{z|y} \left\| \boldsymbol{z} - \boldsymbol{y} \right\|_2^2 \;.$$

Thus, by (51) and the above,

$$\begin{aligned} \Delta I_{\uparrow} &\geq \frac{1}{2} \sum_{y\in\mathcal{Y}} \pi_y \sum_{z\in\mathcal{Z}} \Phi_{z|y} \left\| \boldsymbol{z} - \boldsymbol{y} \right\|_2^2 \\ &\geq \frac{1}{2} \sum_{y\in\mathcal{Y}} \pi_y \sum_{z\in\mathcal{Z}} \Phi_{z|y} \left(\min_{z'\in\mathcal{Z}} \left\| \boldsymbol{z}' - \boldsymbol{y} \right\|_2^2 \right) \\ &= \frac{1}{2} \sum_{y\in\mathcal{Y}} \pi_y \min_{z'\in\mathcal{Z}} \left\| \boldsymbol{z}' - \boldsymbol{y} \right\|_2^2 \;. \end{aligned}$$

Recall that the sets $\{\mathcal{A}_z\}_{z\in\mathcal{Z}}$ partition $\mathcal{Y}$. Thus, continuing the above,

$$\begin{aligned} \Delta I_{\uparrow} &\geq \frac{1}{2} \sum_{z\in\mathcal{Z}} \sum_{y\in\mathcal{A}_z} \pi_y \min_{z'\in\mathcal{Z}} \left\| \boldsymbol{z}' - \boldsymbol{y} \right\|_2^2 \\ &= \frac{1}{2} \sum_{z\in\mathcal{Z}} \sum_{y\in\mathcal{A}_z} \pi_y \left\| \boldsymbol{z} - \boldsymbol{y} \right\|_2^2 \\ &= \sum_{z\in\mathcal{Z}} \Delta(\mathcal{A}_z) \;, \end{aligned}$$

where the first and second equalities follow from (45) and (46), respectively. ■

To coincide with the proof in [10, Theorem 2] we will further lower bound $\Delta I_{\uparrow}$ by making use of the following lemma.

Lemma 13. *Let $\Delta(\mathcal{A}_z)$ be as defined in Lemma 12. Then,*

$$\Delta(\mathcal{A}_z) \geq \tilde{\Delta}(\mathcal{A}_z) \;,$$

where

$$\tilde{\Delta}(\mathcal{A}_z) \triangleq \frac{1}{2} \sum_{y\in\mathcal{A}_z} \pi_y \left\| \boldsymbol{y} - \bar{\boldsymbol{y}}_z \right\|_2^2 \;, \tag{52}$$

and $\bar{\boldsymbol{y}}_z$ is the weighted center of $\mathcal{A}_z$,

$$\bar{\boldsymbol{y}}_z \triangleq \frac{\sum_{y\in\mathcal{A}_z} \pi_y \boldsymbol{y}}{\sum_{y\in\mathcal{A}_z} \pi_y} \;.$$

Proof: Define the function

$$V(\boldsymbol{u}) \triangleq \frac{1}{2} \sum_{y\in\mathcal{A}_z} \pi_y \left\| \boldsymbol{y} - \boldsymbol{u} \right\|_2^2 \;.$$

Since the π_y are non-negative, $V(\boldsymbol{u})$ is easily seen to be convex in $\boldsymbol{u}$. Thus, the minimum is calculated by differentiating with respect to $\boldsymbol{u}$ and equating to 0. Since $V(\boldsymbol{u})$ is quadratic in $\boldsymbol{u}$, we have a simple closed-form solution,

$$\boldsymbol{u} = \frac{\sum_{y\in\mathcal{A}_z} \pi_y \boldsymbol{y}}{\sum_{y\in\mathcal{A}_z} \pi_y} = \bar{\boldsymbol{y}}_z \;,$$

and the proof is finished. ■

We return to proving the main theorem of this section.

Proof of Theorem 11: According to [10, Claim 1], a DMC $W : \mathcal{X} \to \mathcal{Y}$ is optimally degraded to a channel $Q : \mathcal{X} \to \mathcal{Z}$ by partitioning $\mathcal{Y}$ to $|\mathcal{Z}|$ disjoint subsets, denoted by $\{A_z\}_{z\in\mathcal{Z}}$, and merging all the letters in each subset. It is then shown in [10, Corollary 5] that the loss in mutual information, as a result of this operation, can be lower bounded by $\sum_{z\in\mathcal{Z}} \tilde{\Delta}(A_z)$, where $\tilde{\Delta}(A_z)$ is defined as in (52). As a final step, in [10, Section V] a specific sequence of channels and input distributions is introduced and analyzed. For this sequence, $\sum_{z\in\mathcal{Z}} \tilde{\Delta}(A_z)$ is lower bounded by the same bound as in (43).

In our case, as a result of Lemma 12 and Lemma 13,

$$\Delta I_{\uparrow} \geq \sum_{z\in\mathcal{Z}} \tilde{\Delta}(\mathcal{A}_z) \;. \tag{53}$$

and we get the same expression as in [10, Corollary 5]. From this point on, the proof is identical to [10, Section V]. ■

V. Optimal binary upgrading

A. Main result

In this section we show an efficient algorithmic implementation of optimal upgrading for the BDMC case, $|\mathcal{X}| = 2$. As in the degrading case [8], the algorithm will be an instance of dynamic programming. The following theorem is cardinal, and is the main result of the section.

Theorem 14. *Let a BDMC $W : \mathcal{X} \to \mathcal{Y}$ and an input distribution be given. Denote by $Q : \mathcal{X} \to \mathcal{Z}$ and $\Phi : \mathcal{Z} \to \mathcal{Y}$ the optimizers of (4), for a given L. Denote by $\{\boldsymbol{y}\}_{y\in\mathcal{Y}}$ and $\{\boldsymbol{z}\}_{z\in\mathcal{Z}}$ the posterior probabilities associated with the output letters of W and Q, respectively. Assume without loss of generality that all the $\{\boldsymbol{z}\}_{z\in\mathcal{Z}}$ are distinct. Then,*

$$\{\boldsymbol{z}\}_{z\in\mathcal{Z}} \subseteq \{\boldsymbol{y}\}_{y\in\mathcal{Y}} \ . \tag{54}$$

Moreover, recalling the notation in (44), $\Phi(y|z) > 0$ implies that z has either the largest z_0 such that $z_0 \le y_0$ or the smallest z_0 such that $z_0 \ge y_0$.

The theorem essentially states that the optimal upgraded channel contains a subset of the output letters of the original channel, each such letter retaining its posterior probability vector. Moreover, any output letter $y \in \mathcal{Y}$ is generated by the two output letters $z \in \mathcal{Z}$ neighboring it, when ordered on the line which is the posterior probability simplex. Thus, if we are told the optimal subset $\{\boldsymbol{z}\}_{z\in\mathcal{Z}}$ we can efficiently deduce from it the optimal Q and Φ. That is, to find Q, note that we can calculate the the probabilities $\{\pi_z\}$ using the RHS of (50). After the above calculations are carried out, we have full knowledge of the reverse Q, and can hence deduce Q. From here, finding Φ is immediate: we can find the reverse Φ as in equation (60), and then apply (49) to get the forward channel.

Note that if $\boldsymbol{z}^{(a)} = \boldsymbol{z}^{(b)}$ for some two distinct letters $z^{(a)}, z^{(b)} \in \mathcal{Z}$, then we can merge these letters and obtain an equivalent channel [2, Section III] with $L-1$ letters. Repeating this until all the $\{\boldsymbol{z}\}_{z\in\mathcal{Z}}$ are distinct allows us to assume distinction while retaining generality.

We stress that Theorem 14 is only valid for BDMCs, while for larger input alphabet sizes it can be disproved. For example, let $|\mathcal{X}| = 3$, and assume that the points $\{\boldsymbol{y}\}_{y\in\mathcal{Y}}$ are arranged on a circle in the simplex plane. Note now that no point can be expressed as a convex combination of the other points. Hence, (48) cannot be satisfied if $\{\boldsymbol{z}\}_{z\in\mathcal{Z}}$ is a strict subset of $\{\boldsymbol{y}\}_{y\in\mathcal{Y}}$. This example can be easily extended to larger input alphabet sizes using higher dimensional spheres.

B. Optimal intermediate channel

By definition, if $W \preccurlyeq Q$ then there exists a corresponding intermediate channel Φ. Our aim now is to characterize the optimal Φ by which $\Delta I_{\uparrow}^{*}$ in (4) is attained. Recall from (48) and (49) our definition of the "reverse" channel $\Phi(z|y) = \Phi_{z|y}$. From (51),

$$\Delta I_{\uparrow} = \sum_{y\in\mathcal{Y}} \pi_y h(\boldsymbol{y}) - \sum_{y\in\mathcal{Y}} \pi_y i_y \ , \tag{55}$$

where

$$i_y \triangleq \sum_{z\in\mathcal{Z}} \Phi_{z|y} h(\boldsymbol{z}) \ . \tag{56}$$

To recap, we have gained a simplification by considering reversed channels: each output letter $y \in \mathcal{Y}$ decreases $\Delta I_{\uparrow}$ by $\pi_y i_y$.

In the following lemma we consider a simple yet important case: an output letter y of the original channel W is obtained by combining exactly two output letters of the upgraded channel Q, denoted z_1 and z_2. Informally, the lemma states that the closer the posterior probabilities of z_1 and z_2 are to y, the better we are in terms of i_y.

Lemma 15. *Let $\boldsymbol{y} = [\sigma, 1-\sigma]^T$ be fixed. For*

$$0 < \zeta_1 < \sigma < \zeta_2 < 1 \ ,$$

define $\boldsymbol{z}_1 = [\zeta_1, 1-\zeta_1]^T$ and $\boldsymbol{z}_2 = [\zeta_2, 1-\zeta_2]^T$. Next, let $\phi(\zeta_1, \zeta_2)$ be such that

$$\boldsymbol{y} = \phi(\zeta_1,\zeta_2)\cdot \boldsymbol{z}_1 + (1-\phi(\zeta_1,\zeta_2))\cdot \boldsymbol{z}_2 \ . \tag{57}$$

Define

$$i_y(\zeta_1,\zeta_2) \triangleq \phi(\zeta_1,\zeta_2)\cdot h(\boldsymbol{z}_1) + (1-\phi(\zeta_1,\zeta_2))\cdot h(\boldsymbol{z}_2) \ .$$

Then, $i_y(\zeta_1,\zeta_2)$ is increasing with respect to ζ_1 and decreasing with respect to ζ_2.

Proof: To satisfy (57) we have

$$\phi(\zeta_1,\zeta_2) = \frac{\zeta_2-\sigma}{\zeta_2-\zeta_1} \ .$$

Now using the derivative of i_y we get,

$$\begin{aligned}
\frac{\partial i_y}{\partial \zeta_1} &= \frac{\partial \phi}{\partial \zeta_1}\cdot h(\boldsymbol{z}_1) + \phi\cdot(\eta'(\zeta_1) - \eta'(1-\zeta_1)) - \frac{\partial \phi}{\partial \zeta_1}\cdot h(\boldsymbol{z}_2) \\
&= \frac{\partial \phi}{\partial \zeta_1}\cdot [h(\boldsymbol{z}_1) + (\zeta_2-\zeta_1)(\eta'(\zeta_1)-\eta'(1-\zeta_1)) \\
&\quad - h(\boldsymbol{z}_2)] \\
&= \frac{\partial \phi}{\partial \zeta_1}\cdot [-h(\boldsymbol{z}_2) - \zeta_2\log(\zeta_1) - (1-\zeta_2)\log(1-\zeta_1)] \\
&= \frac{\zeta_2-\sigma}{(\zeta_2-\zeta_1)^2} d_{\mathrm{KL}}(\zeta_2||\zeta_1) \\
&> 0 \ ,
\end{aligned}$$

where we defined

$$d_{\mathrm{KL}}(p||q) \triangleq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \ , \tag{58}$$

which is the binary Kullback-Leibler divergence. In the same manner,

$$\begin{aligned}
\frac{\partial i_y}{\partial \zeta_2} &= \frac{\partial \phi}{\partial \zeta_2}\cdot h(\boldsymbol{z}_1) - \frac{\partial \phi}{\partial \zeta_2}\cdot h(\boldsymbol{z}_2) \\
&\quad + (1-\phi)\cdot(\eta'(\zeta_2)-\eta'(1-\zeta_2)) \\
&= \frac{\partial \phi}{\partial \zeta_2}\cdot [h(\boldsymbol{z}_1) - h(\boldsymbol{z}_2) \\
&\quad + (\zeta_2-\zeta_1)\,(\eta'(\zeta_2)-\eta'(1-\zeta_2))] \\
&= \frac{\partial \phi}{\partial \zeta_2}\cdot [h(\boldsymbol{z}_1) + \zeta_1\log(\zeta_2) + (1-\zeta_1)\log(1-\zeta_2)] \\
&= -\frac{\sigma-\zeta_1}{(\zeta_2-\zeta_1)^2} d(\zeta_1||\zeta_2) \\
&< 0 \ ,
\end{aligned}$$

and the proof is finished. ■

The following lemma states that the second assertion of Theorem 14 holds.

Lemma 16. *Let a BDMC $W : \mathcal{X} \to \mathcal{Y}$ and an input distribution be given. Denote by $Q : \mathcal{X} \to \mathcal{Z}$ and $\Phi : \mathcal{Z} \to \mathcal{Y}$ the optimizers of (4), for a given L. Assume without loss of generality that all the $\{\boldsymbol{z}\}_{z \in \mathcal{Z}}$ are distinct. Then, $\Phi(y|z) > 0$ implies that z has either the largest z_0 such that $z_0 \leq y_0$ or the smallest z_0 such that $z_0 \geq y_0$.*

Proof: Note that the input distribution is given. Thus, the reverse channels corresponding to W, Q, and Φ are well defined, and our proof will revolve around them. Since $W \preccurlyeq Q$, the reverse channel $\Phi_{z|y}$ satisfies (48), (50) and (55). Let us assume to the contrary that $y \in \mathcal{Y}$ does not satisfy the assertion in the lemma. Our plan is to find a reverse channel $\Phi'_{z|y}$ that attains a greater i_y than the one attained by Φ, and thus arrive at a contradiction.

As a first case, assume that there exists $z^* \in \mathcal{Z}$ for which $\boldsymbol{z}^* = \boldsymbol{y}$. In this case, $z_0^* = y_0$, and thus the lemma states that the only non-zero term in $\{\Phi(y|z)\}_{z \in \mathcal{Z}}$ is $\Phi(y|z^*)$. Assume the contrary. Thus, switching to the reverse channel, we note that $\left\{\Phi_{z|y}\right\}_{z \in \mathcal{Z}}$ has at least two non-zero terms. Then, using (56) and the strict concavity of h we get

$$i_y < h\left(\sum_{z \in \mathcal{Z}} \Phi_{z|y} \boldsymbol{z}\right) = h(\boldsymbol{y}) \; .$$

Note that this upper bound can be attained by simply choosing $\Phi'_{z|y} = 1$ when $z = z^*$ and $\Phi'_{z|y} = 0$ otherwise. For all other $y' \neq y$, we define $\Phi'_{z|y'}$ as equal to $\Phi_{z|y'}$. Thus, using $\Phi'_{z|y}$ and (50), we obtain a new set of probabilities $\{\pi'_z\}_{z \in \mathcal{Z}}$ that together with the reverse channel Q generate a new "forward" channel $Q' : \mathcal{X} \to \mathcal{Z}$ (applying Bayes' rule). Since the reverse channels satisfy $Q' \succcurlyeq W$, we have by (48) and the explanation following it that the "forward" channels satisfy the same upgrading relation. Yet Q' attains a strictly lower $\Delta I_{\uparrow}$, a contradiction.

In the second case, we assume that $\boldsymbol{z} \neq \boldsymbol{y}$ for all $z \in \mathcal{Z}$ and define the sets

$$\mathcal{Z}_R = \{z \in \mathcal{Z} : z_0 > y_0\} \; ,$$
$$\mathcal{Z}_L = \{z \in \mathcal{Z} : z_0 < y_0\} \; .$$

Geometrically, if we draw $\boldsymbol{y}$ and $\{\boldsymbol{z}\}_{z \in \mathcal{Z}}$ as points on the line $s_0 + s_1 = 1$ in the first quadrant of the (s_0, s_1) plane (the two dimensional simplex), then the letters in $\mathcal{Z}_R$ would be on the right side of $\boldsymbol{y}$ and the letters in $\mathcal{Z}_L$ on its left (see Figure 4). Define also

$$z_L^* \triangleq \operatorname*{arg\,min}_{z \in \mathcal{Z}_L} \|\boldsymbol{y} - \boldsymbol{z}\|_2 \; , \quad z_R^* \triangleq \operatorname*{arg\,min}_{z \in \mathcal{Z}_R} \|\boldsymbol{y} - \boldsymbol{z}\|_2 \; .$$

Namely, the closest neighboring letters from $\mathcal{Z}$, one from each side. The lemma states that the only non-zero terms in $\{\Phi(y|z)\}_{z \in \mathcal{Z}}$ are $\Phi(y|z_L^*)$ and $\Phi(y|z_R^*)$. Assume again the contrary. Thus, switching again to the reverse channel, both $\left\{\Phi_{z|y}\right\}_{z \in \mathcal{Z}_L}$ and $\left\{\Phi_{z|y}\right\}_{z \in \mathcal{Z}_R}$ have non-zero terms. By assumption, one such term corresponds to neither z_L^* nor z_R^*. Note that we can write

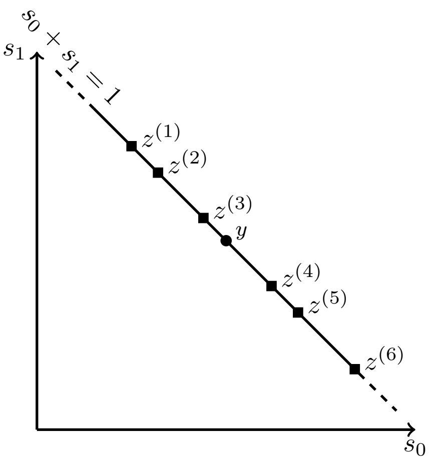


Fig. 4. The letters $\left\{z^{(i)}\right\}_{i=1}^{6}$ of an upgraded BDMC Q and a letter y from an initial BDMC W. In this example $\mathcal{Z}_L = \left\{z^{(1)}, z^{(2)}, z^{(3)}\right\}$, $\mathcal{Z}_R = \left\{z^{(4)}, z^{(5)}, z^{(6)}\right\}$, $z_L^* = z^{(3)}$, $z_R^* = z^{(4)}$.

$$\begin{aligned} i_y &= \sum_{z \in \mathcal{Z}_L} \Phi_{z|y} h(\boldsymbol{z}) + \sum_{z \in \mathcal{Z}_R} \Phi_{z|y} h(\boldsymbol{z}) \\ &= \phi \sum_{z \in \mathcal{Z}_L} \frac{\Phi_{z|y}}{\phi} h(\boldsymbol{z}) + (1 - \phi) \sum_{z \in \mathcal{Z}_R} \frac{\Phi_{z|y}}{1 - \phi} h(\boldsymbol{z}) \; , \end{aligned}$$

where $\phi \triangleq \sum_{z \in \mathcal{Z}_L} \Phi_{z|y}$, and $0 < \phi < 1$. Using the strict concavity of h we get

$$\begin{aligned} i_y &\leq \phi \cdot h\left(\sum_{z \in \mathcal{Z}_L} \frac{\Phi_{z|y}}{\phi} \boldsymbol{z}\right) + (1 - \phi) \cdot h\left(\sum_{z \in \mathcal{Z}_R} \frac{\Phi_{z|y}}{1 - \phi} \boldsymbol{z}\right) \\ &= \phi \cdot h(\boldsymbol{z}_L) + (1 - \phi) \cdot h(\boldsymbol{z}_R) \; , \end{aligned} \tag{59}$$

where

$$\boldsymbol{z}_L \triangleq \sum_{z \in \mathcal{Z}_L} \frac{\Phi_{z|y}}{\phi} \boldsymbol{z} \; , \quad \boldsymbol{z}_R \triangleq \sum_{z \in \mathcal{Z}_R} \frac{\Phi_{z|y}}{1 - \phi} \boldsymbol{z} \; .$$

Note that the locations of $\boldsymbol{z}_L$ and $\boldsymbol{z}_R$ depend on the probabilities $\left\{\Phi_{z|y}\right\}_{z \in \mathcal{Z}}$. Since both $\boldsymbol{z}_L$ and $\boldsymbol{z}_R$ are convex combinations of the letters in $\mathcal{Z}_L, \mathcal{Z}_R$, respectively, they have to reside in the convex hull of those sets. Then, by assumption, either $\boldsymbol{z}_L \neq \boldsymbol{z}_L^*$ or $\boldsymbol{z}_R \neq \boldsymbol{z}_R^*$. Recall now that according to Lemma 15, any choice of $\left\{\Phi_{z|y}\right\}_{z \in \mathcal{Z}}$ could be improved as long as $\boldsymbol{z}_L$ and $\boldsymbol{z}_R$ are not the closest letters to $\boldsymbol{y}$. Hence,

$$i_y < \phi' \cdot h(\boldsymbol{z}_L^*) + (1 - \phi') \cdot h(\boldsymbol{z}_R^*) \; ,$$

for the corresponding ϕ'. Once again, this upper bound can be attained by choosing

$$\Phi'_{z|y} = \begin{cases} \frac{\|\boldsymbol{z}_R^* - \boldsymbol{y}\|_2}{\|\boldsymbol{z}_R^* - \boldsymbol{z}_L^*\|_2} & z = z_L^* \; , \\ \frac{\|\boldsymbol{y} - \boldsymbol{z}_L^*\|_2}{\|\boldsymbol{z}_R^* - \boldsymbol{z}_L^*\|_2} & z = z_R^* \; , \\ 0 & \text{Otherwise} \; . \end{cases} \tag{60}$$

Thus, as before, we have found a channel $Q' \succcurlyeq W$ that attains a strictly lower $\Delta I_{\uparrow}$, a contradiction. ■

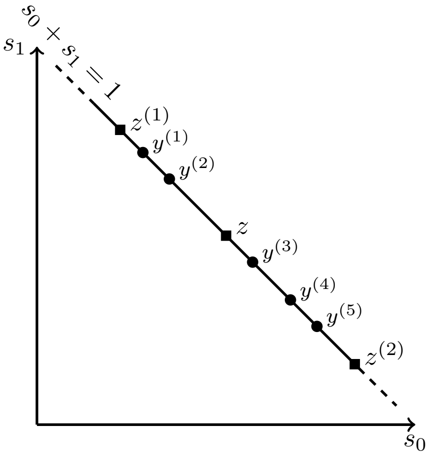


Fig. 5. The letters $z^{(1)}$, z and $z^{(2)}$ of an upgraded BDMC Q, and $\{y^{(i)}\}_{i=1}^{5}$ of an initial BDMC W. In this example, when z is picked as illustrated: $\tilde{\mathcal{Y}} = \{y^{(i)}\}_{i=1}^{5}$, $\tilde{\mathcal{Y}}_L = \{y^{(1)}, y^{(2)}\}$, $\tilde{\mathcal{Y}}_R = \{y^{(3)}, y^{(4)}, y^{(5)}\}$, $y^L = y^{(2)}$, $y^R = y^{(3)}$.

C. Optimal upgraded channel

Lemma 16 is meaningful for two reasons. First, now that we know the optimal Φ for any given Q, we can minimize over Q alone. Equivalently, as per the discussion after Theorem 14, we can minimize over the subset $\{\boldsymbol{z}\}_{z\in\mathcal{Z}}$. Second, any two adjacent letters in $\{\boldsymbol{z}\}_{z\in\mathcal{Z}}$ (on the simplex line) exclusively generate all the letters $\{\boldsymbol{y}\}$ that reside in the segment between the two of them. Hence, our proof of Theorem 14 can be "localized" in the sense that we can focus on two adjacent segments and move the separating letter $\boldsymbol{z}$ between them, thus only affecting the letters $\{\boldsymbol{y}\}$ in those two segments. We return to the proof of our theorem.

Proof of Theorem 14: Since Lemma 16 has been proved, all that is left is to prove the inclusion (54). Note that using (48) and $\sum_{z\in\mathcal{Z}} \Phi_{z|y} = 1$, if $W \preccurlyeq Q$ then

$$\{\boldsymbol{y}\}_{y\in\mathcal{Y}} \subseteq \operatorname{conv}\left\{\{\boldsymbol{z}\}_{z\in\mathcal{Z}}\right\} ,$$

where conv denotes convex hull. Namely, each $\boldsymbol{y}$ can be expressed as a convex combination of vectors $\{\boldsymbol{z}\}$ that correspond to letters in $\mathcal{Z}$.

Let us assume first that $L = 2$. That is, we need to find two letters $\boldsymbol{z}_{\min}$ and $\boldsymbol{z}_{\max}$ whose convex hull contains all the letters $\{\boldsymbol{y}\}_{y\in\mathcal{Y}}$ and $\Delta I_{\uparrow}$ is minimized. Then, according to Lemma 15, it is optimal to choose $\boldsymbol{z}_{\min}$ and $\boldsymbol{z}_{\max}$ with the smallest possible convex hull containing $\{\boldsymbol{y}\}_{y\in\mathcal{Y}}$. Hence, the letters of the optimal Q are

$$\boldsymbol{z}_{\min} = \boldsymbol{y}_{\min} , \quad \boldsymbol{z}_{\max} = \boldsymbol{y}_{\max} , \tag{61}$$

where

$$y_{\min} \triangleq \operatorname*{arg\,min}_{y\in\mathcal{Y}} y_0 , \quad y_{\max} \triangleq \operatorname*{arg\,max}_{y\in\mathcal{Y}} y_0 .$$

Namely, the leftmost and rightmost letters in $\{\boldsymbol{y}\}_{y\in\mathcal{Y}}$, respectively.

Assume now that $L > 2$ and let

$$\boldsymbol{z}^{(1)} = [\zeta_1, 1-\zeta_1]^T , \quad \boldsymbol{z} = [\zeta, 1-\zeta]^T , \quad \boldsymbol{z}^{(2)} = [\zeta_2, 1-\zeta_2]^T ,$$

be three adjacent points on the simplex line satisfying

$$0 \le \zeta_1 < \zeta < \zeta_2 \le 1 .$$

Assume there is a subset $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$ of M letters in the *interior* of $\operatorname{conv}\left\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}\right\}$. Thus, we stress that $z^{(1)}$ and $z^{(2)}$ are not contained in $\tilde{\mathcal{Y}}$. Our aim is to show that an optimal choice for $\boldsymbol{z}$ satisfies $\boldsymbol{z} \in \tilde{\mathcal{Y}}$. That will ensure that there cannot be a letter $z \in \mathcal{Z}$ such that $\boldsymbol{z} \notin \{\boldsymbol{y}\}_{y\in\mathcal{Y}}$, except maybe for the two extreme ones (since $\boldsymbol{z}$ is internal). However, as in the $L = 2$ case discussed previously, these two extreme letters must be $\boldsymbol{y}_{\min}$ and $\boldsymbol{y}_{\max}$, as defined above.

Note that if $M = 0$ then $\pi_z = 0$, by Lemma 16. In this case, without loss of generality, $\boldsymbol{z}$ can be removed from $\mathcal{Z}$. Thus, we henceforth assume that $M > 0$. Figure 5 illustrates some of the sets and letters we will shortly define.

If $\boldsymbol{z} \in \{\boldsymbol{y}\}_{y\in\mathcal{Y}}$, then we are done. Henceforth, let us assume this is not the case. We express each $\boldsymbol{y}$ as $\boldsymbol{y} = [y_0, 1-y_0]^T$, and partition $\tilde{\mathcal{Y}}$ into the sets

$$\tilde{\mathcal{Y}}_L \triangleq \left\{y \in \tilde{\mathcal{Y}} : y_0 < \zeta\right\} , \quad \tilde{\mathcal{Y}}_R \triangleq \left\{y \in \tilde{\mathcal{Y}} : y_0 > \zeta\right\} ,$$

for a given ζ. Since $M > 0$, at least one subset is non-empty. In fact, by similar reasoning to the $L = 2$ case, we deduce that both $\tilde{\mathcal{Y}}_L$ and $\tilde{\mathcal{Y}}_R$ are non-empty. By (56) and Lemma 16, the contribution of these subsets to $\Delta I_{\uparrow}$ is

$$\begin{aligned}
\Delta I_{\uparrow}(\tilde{\mathcal{Y}}) &\triangleq \sum_{y\in\tilde{\mathcal{Y}}} \pi_y h(\boldsymbol{y}) - \sum_{y\in\tilde{\mathcal{Y}}_L} \pi_y i_y - \sum_{y\in\tilde{\mathcal{Y}}_R} \pi_y i_y \\
&= \sum_{y\in\tilde{\mathcal{Y}}} \pi_y h(\boldsymbol{y}) \\
&\quad - \sum_{y\in\tilde{\mathcal{Y}}_L} \pi_y \left(\frac{\zeta - y_0}{\zeta - \zeta_1} h(\boldsymbol{z}^{(1)}) + \frac{y_0 - \zeta_1}{\zeta - \zeta_1} h(\boldsymbol{z}) \right) \\
&\quad - \sum_{y\in\tilde{\mathcal{Y}}_R} \pi_y \left(\frac{\zeta_2 - y_0}{\zeta_2 - \zeta} h(\boldsymbol{z}) + \frac{y_0 - \zeta}{\zeta_2 - \zeta} h(\boldsymbol{z}^{(2)}) \right) .
\end{aligned}$$

Thus,

$$\begin{aligned}
\Delta I_{\uparrow}(\tilde{\mathcal{Y}}) &= \sum_{y\in\tilde{\mathcal{Y}}} \pi_y \left(h(\boldsymbol{y}) - h(\boldsymbol{z}^{(1)}) - h(\boldsymbol{z}^{(2)}) \right) \\
&\quad - C_1 \frac{h(\boldsymbol{z}) - h(\boldsymbol{z}^{(1)})}{\zeta - \zeta_1} + C_2 \frac{h(\boldsymbol{z}^{(2)}) - h(\boldsymbol{z})}{\zeta_2 - \zeta} ,
\end{aligned} \tag{62}$$

where

$$C_1 \triangleq \sum_{y\in\tilde{\mathcal{Y}}_L} \pi_y (y_0 - \zeta_1) , \quad C_2 \triangleq \sum_{y\in\tilde{\mathcal{Y}}_R} \pi_y (\zeta_2 - y_0) .$$

Recall that $\tilde{\mathcal{Y}}_L, \tilde{\mathcal{Y}}_R$ were defined as a function of $\boldsymbol{z}$. Hence, they remain the same as long as $\boldsymbol{z}$ is strictly between the rightmost letter in $\tilde{\mathcal{Y}}_L$ and the leftmost letter in $\tilde{\mathcal{Y}}_R$, denoted by $\boldsymbol{y}^L \triangleq [y_0^L, 1-y_0^L]^T$ and $\boldsymbol{y}^R \triangleq [y_0^R, 1-y_0^R]^T$, respectively.

By definition, $\tilde{\mathcal{Y}}$ does not contain $z^{(1)}$ and $z^{(2)}$. This implies the following two points. First, we deduce that C_1 and C_2 are both positive. Next, by (62), $\Delta I_{\uparrow}(\tilde{\mathcal{Y}})$ is continuous and bounded for $y_0^L \le \zeta \le y_0^R$. The theorem will be proved if we show that the minimum is attained by setting ζ to either y_0^L or y_0^R. Thus, we show that the minimum is not attained when

$\zeta \in (y_0^L, y_0^R)$. To that end, we take the derivative of $\Delta I_{\uparrow}(\tilde{\mathcal{Y}})$ with respect to ζ and get

$$\begin{aligned}\frac{\partial \Delta I_{\uparrow}(\tilde{\mathcal{Y}})}{\partial \zeta} &= -C_1 \frac{(\eta'(\zeta) - \eta'(1-\zeta))(\zeta - \zeta_1)}{(\zeta - \zeta_1)^2} \\ &\quad - C_2 \frac{(\eta'(\zeta) - \eta'(1-\zeta))(\zeta_2 - \zeta)}{(\zeta_2 - \zeta)^2} \\ &\quad + C_1 \frac{h(\boldsymbol{z}) - h(\boldsymbol{z}^{(1)})}{(\zeta - \zeta_1)^2} + C_2 \frac{h(\boldsymbol{z}^{(2)}) - h(\boldsymbol{z})}{(\zeta_2 - \zeta)^2} \\ &= C_1 \frac{d_{\mathrm{KL}}(\zeta_1 || \zeta)}{(\zeta - \zeta_1)^2} - C_2 \frac{d_{\mathrm{KL}}(\zeta_2 || \zeta)}{(\zeta_2 - \zeta)^2} \ .\end{aligned}$$

We now recall that both C_1 and C_2 are positive. Thus,

$$\frac{\partial \Delta I_{\uparrow}(\tilde{\mathcal{Y}})}{\partial \zeta} = C_2 \frac{d_{\mathrm{KL}}(\zeta_1 || \zeta)}{(\zeta - \zeta_1)^2} \left(\frac{C_1}{C_2} - q(\zeta) \right) ,$$

where we defined

$$q(\zeta) \triangleq \frac{d_{\mathrm{KL}}(\zeta_2 || \zeta)(\zeta - \zeta_1)^2}{d_{\mathrm{KL}}(\zeta_1 || \zeta)(\zeta_2 - \zeta)^2} \ .$$

To achieve our goal, it suffices to show that $q(\zeta)$ is non-decreasing in (ζ_1, ζ_2), thus ensuring that $\Delta I_{\uparrow}(\tilde{\mathcal{Y}})$ is either monotonic or has a single maximum. The proof is given in Appendix A. ■

D. Dynamic programming implementation

Theorem 14 simplifies the task of channel upgrading to finding the optimal L-sized subset of $\mathcal{Y}$. Note that such a subset must contain the leftmost and rightmost letters of $\mathcal{Y}$.

We can now use dynamic programming to efficiently find the optimal subset. The key idea is to use the structure of (55) and the ordering of the letters on the simplex. Each possible L-sized subset partitions $\mathcal{Y}$ to $L-1$ contiguous sets on the simplex, with overlapping borders (the internal $\{\boldsymbol{z}\}$). Since the cost function (55) is additive in the letters $\mathcal{Y}$, we can employ a dynamic programming approach, similar to [8, Section IV] and [9, Section III].

VI. Upper bound on binary optimal upgrading gain

Our final result is the fruit of combining Section III, Section IV and Section V. Namely, an upper bound on $\Delta I_{\uparrow}^*$ and UC for $|\mathcal{X}| = 2$. Once again, we use an iterative sub-optimal upgrading algorithm called "greedy-split", similar to the one proposed in [11]. Implicitly, we apply the optimal upgrading algorithm, to get from an alphabet of $|\mathcal{Y}|$ output letters to one with $|\mathcal{Y}| - 1$ output letters. This is done iteratively, until the required number of letters, L, is reached. This simplifies to the following. In each iteration we find the letter $y \in \mathcal{Y}$ that minimizes

$$\Delta I_{\uparrow} = \pi_y h(\boldsymbol{y}) - \pi_y \phi h(\boldsymbol{y}_L) - \pi_y (1-\phi) h(\boldsymbol{y}_R) \ , \qquad (63)$$

where y_L and y_R are the left and right adjacent letters to y, respectively, and

$$\phi = \frac{\|\boldsymbol{y}_R - \boldsymbol{y}\|_2}{\|\boldsymbol{y}_R - \boldsymbol{y}_L\|_2} \ ,$$

as in Lemma 16. The minimizing letter y is then split between y_L and y_R by updating

$$\pi_{y_L} \leftarrow \pi_{y_L} + \phi \cdot \pi_y \ , \quad \pi_{y_R} \leftarrow \pi_{y_R} + (1-\phi) \cdot \pi_y \ ,$$

and then eliminating y. The following theorem is the main result of this section.

Theorem 17. *Let a BDMC $W : \mathcal{X} \to \mathcal{Y}$ satisfy $|\mathcal{Y}| > 8$, and let $L \geq 8$. Then, for any fixed input distribution P_X,*

$$\Delta I_{\uparrow}^* = \min_{\substack{Q : Q \succcurlyeq W, \\ |Q| \leq L}} I(Q, P_X) - I(W, P_X) = O\left(L^{-2}\right) \ .$$

In particular,

$$\Delta I_{\uparrow}^* \leq 2\nu(2) \cdot L^{-2} \ ,$$

where $\nu(\cdot)$ was defined in Theorem 1. This bound is attained by greedy-split and is tight in the power-law sense.

As in Section III, we first prove the following theorem.

Theorem 18. *Let a BDMC $W : \mathcal{X} \to \mathcal{Y}$ satisfy $|\mathcal{Y}| > 8$, and let a BDMC $Q : \mathcal{X} \to \mathcal{Z}$ be the result of upgrading W by splitting a letter $y \in \mathcal{Y}$. Then, for any fixed input distribution, there exists a letter $y \in \mathcal{Y}$ for which the resulting Q satisfies*

$$\Delta I_{\uparrow} = O\left(|\mathcal{Y}|^{-3}\right) \ .$$

In particular,

$$\Delta I_{\uparrow} \leq 2\mu(2) \cdot |\mathcal{Y}|^{-3} \ , \qquad (64)$$

where $\mu(\cdot)$ was defined in Theorem 2.

Proof: Note that (63) has the form of (9). That is, $\boldsymbol{y}_L$ plays the role of $\boldsymbol{\alpha}$; $\boldsymbol{y}_R$ plays the role of $\boldsymbol{\beta}$; $\pi_y \phi$ plays the role of π_α; $\pi_y(1-\phi)$ plays the role of π_β; $\boldsymbol{y}$ plays the role of $\boldsymbol{\gamma}$; the first π_y in (63) plays the role of π_γ.

Thus, (15) applies to our case as well, yielding

$$\Delta I_{\uparrow} \leq 2\pi_y \cdot d(\boldsymbol{y}_L, \boldsymbol{y}_R) \ , \qquad (65)$$

where d was defined in (16). As we did before, we narrow the search to letters in $\mathcal{Y}_{\text{small}}$. Note that y_L and y_R cannot be adjacent, hence we define $\mathcal{Y}_{\text{punctured}}$ as the subset of $\mathcal{Y}_{\text{small}}$ one gets by eliminating every other letter, starting from the second letter, when drawing $\{\boldsymbol{y}\}_{y \in \mathcal{Y}_{\text{small}}}$ on the two dimensional simplex. Thus, each pair of adjacent letters in $\mathcal{Y}_{\text{punctured}}$ has a letter from $\mathcal{Y}_{\text{small}}$ in its convex hull. This operation of puncturing results in

$$|\mathcal{Y}_{\text{punctured}}| \geq \frac{|\mathcal{Y}_{\text{small}}|}{2} \geq \frac{|\mathcal{Y}|}{4} \ . \qquad (66)$$

Using the same method as in Theorem 2, there exists a pair $y_L, y_R \in \mathcal{Y}_{\text{punctured}}$ for which

$$d(\boldsymbol{y}_L, \boldsymbol{y}_R) \leq 4 \cdot r \ , \qquad (67)$$

where r was defined in (41). The factor of 4 is due to the following. Since we are using $\mathcal{Y}_{\text{punctured}}$, the proof of Theorem 2 must be slightly changed towards the end. Specifically, (39) changes to

$$|\mathcal{Y}'| \geq \frac{|\mathcal{Y}_{\text{punctured}}|}{|\mathcal{X}|} \geq \frac{|\mathcal{Y}_{\text{small}}|}{2|\mathcal{X}|} \geq \frac{|\mathcal{Y}|}{4|\mathcal{X}|} > 1 \ .$$

Then, the term $|\mathcal{Y}|/(2|\mathcal{X}|)$ in (40) becomes $|\mathcal{Y}|/(4|\mathcal{X}|)$. Thus, the r needed for the starred equality in (40) to hold is 4 times the original r. Now, plugging (67) in (65) and recalling that $\pi_y \leq 2/|\mathcal{Y}|$ for $y \in \mathcal{Y}_{\text{small}}$, we get (64). ■

Proof of Theorem 17: The proof uses Theorem 18 iteratively, and is similar to Theorem 1, hence omitted. The tightness in power-law is due to Theorem 11. ■

Note that the greedy-split algorithm can be implemented in a similar manner to [2, Algorithm C]. Hence, it has the same complexity, that is, $O(|\mathcal{Y}| \log |\mathcal{Y}|)$.

We now discuss the application of Theorem 17 to symmetric channels, as we did in Section III-E. Assume that W is a symmetric channel according to our definition in Section III-E. To upgrade W to a symmetric channel Q, we slightly modify the greedy-split algorithm as follows. Instead of searching over all $y \in \mathcal{Y}$, we limit the search to all $y \in \mathcal{Y}$ for which $W(0|y) < \frac{1}{2}$ and $W(0|y_R) \leq \frac{1}{2}$. Then, after splitting y to y_L and y_R, we split $\bar{y}$ to $\bar{y}_L$ and $\bar{y}_R$ where $\bar{y}$, $\bar{y}_L$ and $\bar{y}_R$ are the corresponding "conjugate" letters in the partition of $\mathcal{Y}$, as defined in [2, Section II]. Using the same arguments as in the proof of Theorem 10, we deduce that Theorem 18 holds for the symmetric case as well.

Appendix A
Last claim in the proof of Theorem 14

Recall that the proof of Theorem 14 will be complete, once the last step in it is justified. Thus, we state and prove the following lemma.

Lemma 19. *Fix $0 \leq \zeta_1 < \zeta_2 \leq 1$. For $\zeta_1 < \zeta < \zeta_2$, define*

$$q(\zeta) \triangleq \frac{d_{\mathrm{KL}}(\zeta_2||\zeta)(\zeta-\zeta_1)^2}{d_{\mathrm{KL}}(\zeta_1||\zeta)(\zeta_2-\zeta)^2} \, .$$

Then, $q(\zeta)$ is non-decreasing.

Proof: To prove that $q(\zeta)$ is non-decreasing, we consider its derivative. Straightforward algebraic manipulations yield

$$\begin{aligned}\frac{\mathrm{d}q}{\mathrm{d}\zeta} = &\frac{2(\zeta-\zeta_1)(\zeta_2-\zeta)d_{\mathrm{KL}}(\zeta_1||\zeta)d_{\mathrm{KL}}(\zeta_2||\zeta)(\zeta_2-\zeta_1)}{[d_{\mathrm{KL}}(\zeta_1||\zeta)(\zeta-\zeta_2)^2]^2} \\ &\cdot \left[1 - \frac{(\zeta_2-\zeta)^2}{2\zeta(1-\zeta)d_{\mathrm{KL}}(\zeta_2||\zeta)} \cdot \frac{\zeta-\zeta_1}{\zeta_2-\zeta_1} - \right. \\ &\left. \frac{(\zeta_1-\zeta)^2}{2\zeta(1-\zeta)d_{\mathrm{KL}}(\zeta_1||\zeta)} \cdot \frac{\zeta_2-\zeta}{\zeta_2-\zeta_1}\right] . \end{aligned} \tag{68}$$

Note that the term outside the brackets is non-negative. For the inner term, let us define the two-variable function

$$q_2(\tau,\zeta) \triangleq \begin{cases} \frac{(\tau-\zeta)^2}{2\zeta(1-\zeta)d_{\mathrm{KL}}(\tau||\zeta)} & \zeta \neq \tau \, , \\ 1 & \zeta = \tau \, . \end{cases}$$

For $0 < \zeta < 1$ fixed, $q_2(\tau,\zeta)$ is a continuous function of τ, where $0 < \tau < 1$. Indeed, this can be deduced by a double application of L'Hôpital's rule. As we will show, more is true: for ζ and τ as above, $q_2(\tau,\zeta)$ is concave as a function of τ. Namely, for every $\tau_1, \tau_2, \theta \in [0,1]$,

$$q_2(\theta\tau_1 + (1-\theta)\tau_2, \zeta) - \theta \cdot q_2(\tau_1,\zeta) - (1-\theta) \cdot q_2(\tau_2,\zeta) \geq 0 \, .$$

By choosing $\tau_1 = \zeta_2$, $\tau_2 = \zeta_1$, and $\theta = \frac{\zeta-\zeta_1}{\zeta_2-\zeta_1}$, we get the non-negativity of the inner term in (68).

To prove the concavity of $q_2(\tau,\zeta)$ with respect to τ, we will show that the second derivative $\frac{\partial^2 q_2}{\partial \tau^2}$ is non-positive. The following identity is easily proved, and will be used to simplify many otherwise unwieldy expressions:

$$\boxed{(p-q)\frac{\partial d_{\mathrm{KL}}(p||q)}{\partial p} = d_{\mathrm{KL}}(p||q) + d_{\mathrm{KL}}(q||p) \, .} \tag{69}$$

Using the above, we deduce that

$$\frac{\partial q_2}{\partial \tau} = \begin{cases} \frac{(\tau-\zeta)[d_{\mathrm{KL}}(\tau||\zeta) - d_{\mathrm{KL}}(\zeta||\tau)]}{2\zeta(1-\zeta)(d_{\mathrm{KL}}(\tau||\zeta))^2} & \zeta \neq \tau \, , \\ \frac{1-2\zeta}{3\zeta(1-\zeta)} & \zeta = \tau \, , \end{cases}$$

where the derivative for the case $\zeta = \tau$ is obtained by considering the limit

$$\lim_{\tau \to \zeta} \frac{q_2(\tau,\zeta) - q_2(\zeta,\zeta)}{\tau - \zeta} \, .$$

By an application of L'Hôpital's rule, we deduce that $\frac{\partial q_2}{\partial \tau}$ is continuous in τ.

We differentiate once again, using (69), simple algebra, and similar reasoning to what was employed before. The result is

$$\frac{\partial^2 q_2}{\partial \tau^2} = \begin{cases} \frac{(\tau-\zeta)^2}{2\zeta(1-\zeta)d_{\mathrm{KL}}(\tau||\zeta)^3\tau(1-\tau)} \\ \quad \cdot \left(\frac{2\tau(1-\tau)d_{\mathrm{KL}}(\zeta||\tau)^2}{(\tau-\zeta)^2} - d_{\mathrm{KL}}(\tau||\zeta)\right) & \zeta \neq \tau \, , \\ \frac{\zeta-\zeta^2-1}{9(1-\zeta)^2\zeta^2} & \zeta = \tau \, , \end{cases}$$

a continuous function.

Clearly, $\frac{\partial^2 q_2}{\partial \tau^2}$ is negative when $\zeta = \tau$, since $0 < \zeta < 1$. Considering the case $\zeta \neq \tau$, we see that $\frac{\partial^2 q_2}{\partial \tau^2}$ is non-positive when the bracketed term is non-positive,

$$\frac{2\tau(1-\tau)d_{\mathrm{KL}}(\zeta||\tau)^2}{(\tau-\zeta)^2} \leq d_{\mathrm{KL}}(\tau||\zeta) \, .$$

Note now that both sides go to zero as $\zeta \to \tau$. The derivatives of the LHS and RHS with respect to ζ are

$$\frac{4\tau(1-\tau)d_{\mathrm{KL}}(\zeta||\tau)d_{\mathrm{KL}}(\tau||\zeta)}{(\zeta-\tau)^3}$$

and

$$\frac{\zeta-\tau}{\zeta(1-\zeta)} \, ,$$

respectively. Both derivatives are positive when $\zeta > \tau$, negative when $\zeta < \tau$ and zero when $\zeta = \tau$. Thus, it suffices to show that the ratio between the derivatives (left over right) is less than 1. That is,

$$\frac{2\zeta(1-\zeta)d_{\mathrm{KL}}(\tau||\zeta)}{(\zeta-\tau)^2} \cdot \frac{2\tau(1-\tau)d_{\mathrm{KL}}(\zeta||\tau)}{(\tau-\zeta)^2} \leq 1 \, ,$$

which is equivalent when taking the square root of both sides. Using the AM-GM inequality it is enough to show that

$$\frac{1}{2}\frac{2\zeta(1-\zeta)d_{\mathrm{KL}}(\tau||\zeta)}{(\zeta-\tau)^2} + \frac{1}{2}\frac{2\tau(1-\tau)d_{\mathrm{KL}}(\zeta||\tau)}{(\tau-\zeta)^2} \leq 1 \, ,$$

which is equivalent to showing

$$\begin{aligned} q_3(\tau,\zeta) \triangleq \zeta(1-\zeta)d_{\mathrm{KL}}(\tau||\zeta) + \tau(1-\tau)d_{\mathrm{KL}}(\zeta||\tau) - (\tau-\zeta)^2 \\ \leq 0 \, . \end{aligned}$$

The function q_3 can be shown to be twice continuously differentiable with respect to τ for a fixed $\zeta \in (0,1)$. Its second derivative satisfies

$$\frac{\partial^2 q_3}{\partial \tau^2}(\tau,\zeta) = -\frac{(\zeta-\tau)^2 + 2\tau(1-\tau)d_{\mathrm{KL}}(\zeta||\tau)}{\tau(1-\tau)} \leq 0 \ ,$$

and thus q_3 is concave with respect to τ. Moreover,

$$\frac{\partial q_3}{\partial \tau}(\zeta,\zeta) = 0 \ ,$$

which means that q_3 is maximal when $\tau = \zeta$, namely,

$$q_3(\tau,\zeta) \leq q_3(\zeta,\zeta) = 0 \ ,$$

and the proof is finished. ■

Acknowledgements

We thank Boaz Shuval and the anonymous reviewers for helpful comments and suggestions.

References

[1] E. Arıkan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.

[2] I. Tal and A. Vardy, “How to construct polar codes,” *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6562–6582, October 2013.

[3] H. Mahdavifar and A. Vardy, “Achieving the secrecy capacity of wiretap channels using polar codes,” *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6428–6443, October 2011.

[4] E. Hof and S. Shamai, “Secrecy-achieving polar-coding for binary-input memoryless symmetric wire-tap channels,” `arXiv:1005.2759v2`, 2010.

[5] M. Andersson, V. Rathi, R. Thobaben, J. Kliewer, and M. Skoglund, “Nested polar codes for wiretap and relay channels,” *IEEE Communications Letters*, vol. 14, no. 8, pp. 752–754, August 2010.

[6] O. O. Koyluoglu and H. El Gamal, “Polar coding for secure transmission and key agreement,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1472–1483, October 2012.

[7] I. Tal and A. Vardy, “Channel upgrading for semantically-secure encryption on wiretap channels,” in *2013 IEEE International Symposium on Information Theory (ISIT)*, July 2013, pp. 1561–1565.

[8] B. M. Kurkoski and H. Yagi, “Quantization of binary-input discrete memoryless channels,” *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4544–4552, August 2014.

[9] K. I. Iwata and S. Y. Ozawa, “Quantizer design for outputs of binary-input discrete memoryless channels using SMAWK algorithm,” in *2014 IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 191–195.

[10] I. Tal, “On the construction of polar codes for channels with moderate input alphabet sizes,” *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1501–1509, March 2017.

[11] R. Pedarsani, S. H. Hassani, I. Tal, and E. Telatar, “On the construction of polar codes,” in *2011 IEEE International Symposium on Information Theory (ISIT)*, July 2011, pp. 11–15.

[12] I. Tal, A. Sharov, and A. Vardy, “Constructing polar codes for non-binary alphabets and MACs,” in *2012 IEEE International Symposium on Information Theory (ISIT)*, July 2012, pp. 2132–2136.

[13] T. C. Gulcu, M. Ye, and A. Barg, “Construction of polar codes for arbitrary discrete memoryless channels,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 51–55.

[14] U. Pereg and I. Tal, “Channel upgradation for non-binary input alphabets and MACs,” *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1410–1424, March 2017.

[15] B. Nazer, O. Ordentlich, and Y. Polyanskiy, “Information-distilling quantizers,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 96–100.

[16] J. A. Zhang and B. M. Kurkoski, “Low-complexity quantization of discrete memoryless channels,” in *2016 International Symposium on Information Theory and Its Applications (ISITA)*, October 2016, pp. 448–452.

[17] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[18] T. Koch and A. Lapidoth, “At low SNR, asymmetric quantizers are better,” *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5421–5445, September 2013.

[19] G. Alirezaei and R. Mathar, “Optimal one-bit quantizers are asymmetric for additive uniform noise,” in *2017 Information Theory and Applications Workshop (ITA)*, 2017.

[20] R. G. Gallager, *Information theory and reliable communication*. Springer, 1968.

[21] J. M. Steele, *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004.

Assaf Kartowsky was born in Afula, Israel, in 1985. He received the B.Sc., and M.Sc. degrees in electrical engineering, and the B.Sc. degree in Physics from Technion — Israel Institute of Technology, Haifa, Israel, in 2007, 2017 and 2007, respectively.

Ido Tal (S'05–M'08–SM'18) was born in Haifa, Israel, in 1975. He received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Technion — Israel Institute of Technology, Haifa, Israel, in 1998, 2003 and 2009, respectively. During 2010–2012 he was a postdoctoral scholar at the University of California at San Diego. In 2012 he joined the Electrical Engineering Department at Technion. His research interests include constrained coding and error-control coding. He received the IEEE Joint Communications Society/Information Theory Society Paper Award (jointly with Alexander Vardy) for the year 2017.