PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 45, Fasc. 1 (2025), pp. 115–126
Published online 5.9.2025
doi:10.37190/0208-4147.00246

NOTES ON CONDITIONAL ENTROPY AND COMMITMENT CAPACITY*

BY

NEMANJA NEDIĆ (NOVI SAD), MLADEN KOVAČEVIĆ (NOVI SAD),
DARKO ČAPKO (NOVI SAD), AND SRĐAN VUKMIROVIĆ (NOVI SAD)

Abstract. The so-called commitment capacity of a discrete memoryless channel is given by the maximum of the conditional Shannon entropy $H(X|Y)$ over all input distributions. We examine in detail this optimization problem, motivated by its relevance in information-theoretic cryptography. In particular, we study the role of the channel's connected components in attaining the commitment capacity, and the questions of uniqueness and support of the optimal (capacity-achieving) input and output distributions. We also describe an iterative algorithm for computing the commitment capacity and the optimal input distribution.

2020 Mathematics Subject Classification: Primary 94A60; Secondary 94A15, 94A17.

Key words and phrases: conditional entropy, equivocation, commitment capacity, Blahut–Arimoto algorithm.

1. INTRODUCTION AND PRELIMINARIES

1.1. Definitions and notation. A discrete memoryless channel (DMC) is specified by an ordered triple $(\mathcal{X}, \mathcal{Y}, W)$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and the output alphabets, both assumed finite and of cardinality at least 2, and W is a family $\{W(\cdot|x)\colon x \in \mathcal{X}\}$ of conditional probability distributions, where $W(\cdot|x)$ stands for $(W(y|x))_{y\in\mathcal{Y}}$. To every DMC one can naturally assign a bipartite graph with $\mathcal{X}$ as the set of "left" vertices, $\mathcal{Y}$ as the set of "right" vertices, and an edge between any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ such that $W(y|x) > 0$. We say the channel $(\mathcal{X}, \mathcal{Y}, W)$ is *connected* if the associated graph is connected (in the usual sense that there exists a path between any two vertices). In general, any channel can be decomposed into $k \geqslant 1$ connected components. The ith connected component has $\mathcal{X}_i \subseteq \mathcal{X}$ as the set of left vertices and $\mathcal{Y}_i \subseteq \mathcal{Y}$ as the set of right vertices, and therefore describes

* This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (contract no. 451-03-137/2025-03/200156) and by the Faculty of Technical Sciences, University of Novi Sad, Serbia (project no. 01-50/295).

a channel $(\mathcal{X}_i, \mathcal{Y}_i, W_i)$, where W_i denotes the restriction of W to $\mathcal{X}_i \times \mathcal{Y}_i$. Note that $\{X_1, \ldots, \mathcal{X}_k\}$ is necessarily a partition of $\mathcal{X}$ and $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_k\}$ a partition of $\mathcal{Y}$.

Throughout the paper it is assumed that the channel is *nonredundant*, meaning that, for every $x \in \mathcal{X}$ and every probability distribution P on $\mathcal{X} \setminus \{x\}$, we have

$$W(\cdot|x) \neq \sum_{z \in \mathcal{X} \setminus \{x\}} P(z) W(\cdot|z). \tag{1.1}$$

(This is because the commitment capacity of redundant channels is trivially equal to zero [15].)

For a probability distribution S on $\mathcal{X} \times \mathcal{Y}$, its marginals on $\mathcal{X}$ and $\mathcal{Y}$ are denoted by S_{X} and S_{Y}, e.g., $S_{\mathrm{X}}(x) = \sum_{y \in \mathcal{Y}} S(x, y)$. The corresponding families of conditional distributions are denoted by $S_{\mathrm{X|Y}}, S_{\mathrm{Y|X}}$. E.g., $S_{\mathrm{X|Y}}(x|y) = \frac{S(x,y)}{S_{\mathrm{Y}}(y)}$ when $S_{\mathrm{Y}}(y) \neq 0$. For a given channel $(\mathcal{X}, \mathcal{Y}, W)$ and a given distribution P on $\mathcal{X}$ (also called the *input distribution*), we denote by $P \times W$ the induced joint distribution, i.e., $(P \times W)(x, y) = P(x)W(y|x)$.

$\log$ and $\exp$ denote the base-2 logarithm and exponential. We will use standard information-theoretic functionals [4], with $H(S)$, $H_{\mathrm{X|Y}}(S)$, $H_{\mathrm{Y|X}}(S)$, and $I_{\mathrm{X;Y}}(S)$ denoting the entropy, the conditional entropies, and the mutual information of two random variables with joint distribution S. For example,

$$H_{\mathrm{X|Y}}(S) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} S(x, y) \log \frac{S_{\mathrm{Y}}(y)}{S(x, y)}. \tag{1.2}$$

We will also use the familiar KL divergence, and the more general conditional KL divergence, which is defined as follows [4, (2.65)]: for two families of conditional distributions $P_{\mathrm{X|Y}}, Q_{\mathrm{X|Y}}$, and a distribution R on $\mathcal{Y}$,

$$D_{\mathrm{KL}}(P_{\mathrm{X|Y}} \parallel Q_{\mathrm{X|Y}} \,|\, R) := \sum_{y \in \mathcal{Y}} R(y) \sum_{x \in \mathcal{X}} P_{\mathrm{X|Y}}(x|y) \log \frac{P_{\mathrm{X|Y}}(x|y)}{Q_{\mathrm{X|Y}}(x|y)}. \tag{1.3}$$

This quantity is always nonnegative and equals zero if and only if the conditional distributions $P_{\mathrm{X|Y}}(\cdot|y)$ and $Q_{\mathrm{X|Y}}(\cdot|y)$ are identical for every y for which $R(y) > 0$.

The *equivocation* of the given channel is the conditional entropy of the input given the output. The *commitment capacity* of a nonredundant DMC $(\mathcal{X}, \mathcal{Y}, W)$ equals the maximum of the equivocation over all input distributions [15, 11]:

$$\mathsf{C}(W) = \max_{P} H_{\mathrm{X|Y}}(P \times W). \tag{1.4}$$

1.2. Results. Our aim in this article is to examine in detail the optimization problem (1.4), motivated by its relevance in information-theoretic cryptography. In Section 2 we investigate general properties such as the role of the channel's connected components in attaining the commitment capacity, properties of the

capacity-achieving input and output distributions, conditions under which they are unique, etc. In Section 3 we describe a simple iterative procedure for computing the commitment capacity and the optimal input distribution, analogous to the familiar Blahut–Arimoto algorithm [1, 2].

1.3. Motivation. A commitment scheme [3] is a cryptographic primitive enabling one party, usually called Alice, to choose a value m from a predefined set and send a "masked" version of it, $C(m)$, to another party, usually called Bob, so that the following conditions hold: (1) *binding property* – Alice cannot change her mind at a later point, i.e., after she decides to "open" $C(m)$, Bob will be confident (with probability $1-\delta$) that the value he now sees is indeed the one Alice originally committed to, and (2) *hiding property* – Bob cannot infer m from $C(m)$ alone (except with probability ϵ). These primitives are used as building blocks in various important cryptographic protocols such as zero-knowledge proofs [9].

It is well-known [5] that noisy channels can be used as resources to realize commitment schemes, in which case the problem may be analyzed by probabilistic/ information-theoretic methods. For example, one natural question in this context is that of characterizing the commitment capacity of a given channel, that is, the maximum asymptotic number of bits per symbol that Alice can commit to, with ϵ and δ being arbitrarily small. For the class of discrete memoryless channels this question was answered in [15] by showing that the commitment capacity of a DMC equals the maximum of the channel equivocation over all input distributions (see (1.4)). As noted above, our aim here is to study formal properties of this quantity and explicitly describe an iterative algorithm for its computation.

For more details on the subject of information-theoretic commitments we refer the reader to [15], as well as to some of the recent works [6, 10, 12, 14] and references therein.

2. PROPERTIES OF THE CHANNEL EQUIVOCATION AND THE COMMITMENT CAPACITY

That the problem (1.4) is tractable follows from the fact that $H_{\mathrm{X|Y}}(P\times W)$ is a concave function of the input distribution P for a fixed channel W. This fact is well-known and is stated (without proof) in, e.g., [12, Claim 13]. Note that this is not immediately obvious as the standard representations of the conditional entropy involve differences of two concave functions [4]:

$$H_{\mathrm{X|Y}}(S) = H(S) - H(S_{\mathrm{Y}}) = H(S_{\mathrm{X}}) - I_{\mathrm{X;Y}}(S). \tag{2.1}$$

A stronger claim that $H_{\mathrm{X|Y}}(S)$ is concave in the *joint* distribution S is also true and is surely known in the information theory community, but we nonetheless state and prove it for completeness. We emphasize that this statement does not assume a fixed channel, i.e., conditional distribution W; it holds for $H_{\mathrm{X|Y}}$ and $H_{\mathrm{Y|X}}$ regarded as functions on the space of all probability distributions on $\mathcal{X}\times\mathcal{Y}$.

LEMMA 2.1. *Conditional entropy is a concave function of the joint distribution.*

Proof. The log-sum inequality [4, Theorem 2.7.1] states that, for arbitrary nonnegative numbers a_1, a_2, b_1, b_2,

$$(a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2} \leqslant a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}. \tag{2.2}$$

Let S, T be arbitrary probability distributions on $\mathcal{X} \times \mathcal{Y}$. By (2.2), for any $\alpha \in (0, 1)$ and any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we have

$$\begin{aligned}(\alpha S(x,y) + (1-\alpha)T(x,y)) \log \frac{\alpha S(x,y) + (1-\alpha)T(x,y)}{\alpha S_{\mathrm{Y}}(y) + (1-\alpha)T_{\mathrm{Y}}(y)} \\ \leqslant \alpha S(x,y) \log \frac{S(x,y)}{S_{\mathrm{Y}}(y)} + (1-\alpha)T(x,y) \log \frac{T(x,y)}{T_{\mathrm{Y}}(y)}.\end{aligned} \tag{2.3}$$

Note that the second marginal of the bivariate distribution $\alpha S + (1-\alpha)T$ is precisely $\alpha S_{\mathrm{Y}} + (1-\alpha)T_{\mathrm{Y}}$ in the denominator. Multiplying the inequality (2.3) by -1 and summing both sides over all x, y we get

$$H_{\mathrm{X}|\mathrm{Y}}(\alpha S + (1-\alpha)T) \geqslant \alpha H_{\mathrm{X}|\mathrm{Y}}(S) + (1-\alpha) H_{\mathrm{X}|\mathrm{Y}}(T), \tag{2.4}$$

meaning that $H_{\mathrm{X}|\mathrm{Y}}$ is concave. ∎

The following theorem shows that, in order to determine the commitment capacity of a channel, one may consider each of its connected components separately.

THEOREM 2.1. *Let $(\mathcal{X}_i, \mathcal{Y}_i, W_i)$, $i = 1, \ldots, k$, be the connected components of the channel $(\mathcal{X}, \mathcal{Y}, W)$. Then*

$$\mathsf{C}(W) = \max\{\mathsf{C}(W_1), \ldots, \mathsf{C}(W_k)\}. \tag{2.5}$$

Proof. Any distribution on $\mathcal{X}$ can be thought of as a convex combination of distributions on $\mathcal{X}_i$. Namely, if we denote

$$\alpha_i = \sum_{x \in \mathcal{X}_i} P(x)$$

and define

$$P_i(x) = \begin{cases} \frac{P(x)}{\alpha_i}, & x \in \mathcal{X}_i, \\ 0, & x \in \mathcal{X} \setminus \mathcal{X}_i, \end{cases} \tag{2.6}$$

we will have

$$P(x) = \sum_{i=1}^{k} \alpha_i P_i(x), \quad \forall x \in \mathcal{X}. \tag{2.7}$$

We can then write

$$
\begin{aligned}
(2.8)\qquad & H_{\mathrm{X|Y}}(P \times W) \\
&= \sum_{i=1}^{k} \sum_{x \in \mathcal{X}_i} \sum_{y \in \mathcal{Y}_i} \alpha_i P_i(x) W(y|x) \log \frac{\sum_{z \in \mathcal{X}_i} \alpha_i P_i(z) W(y|z)}{\alpha_i P_i(x) W(y|x)} \\
&= \sum_{i=1}^{k} \alpha_i \sum_{x \in \mathcal{X}_i} \sum_{y \in \mathcal{Y}_i} P_i(x) W(y|x) \log \frac{\sum_{z \in \mathcal{X}_i} P_i(z) W(y|z)}{P_i(x) W(y|x)} \\
&= \sum_{i=1}^{k} \alpha_i H_{\mathrm{X|Y}}(P_i \times W_i).
\end{aligned}
$$

Denoting by $\boldsymbol{\alpha}$ the probability vector $(\alpha_1, \ldots, \alpha_k)$, and noticing that the maximization of $H_{\mathrm{X|Y}}(P \times W)$ over P is equivalent to the sequence of successive maximizations over $P_1, \ldots, P_k$, $\boldsymbol{\alpha}$, we conclude from (1.4) and (2.8) that

$$
\begin{aligned}
(2.9)\qquad \mathsf{C}(W) &= \max_{\boldsymbol{\alpha}} \max_{P_1, \ldots, P_k} H_{\mathrm{X|Y}}(P \times W) \\
&= \max_{\boldsymbol{\alpha}} \max_{P_1, \ldots, P_k} \sum_{i=1}^{k} \alpha_i H_{\mathrm{X|Y}}(P_i \times W_i) \\
&= \max_{\boldsymbol{\alpha}} \sum_{i=1}^{k} \alpha_i \mathsf{C}(W_i) = \max_{i} \mathsf{C}(W_i),
\end{aligned}
$$

which is what we wanted to show. ∎

Hence, when discussing commitment capacity, one can with no loss in generality restrict to connected channels.

REMARK 2.1. Theorem 2.1 implies that the commitment capacity is achieved by a probability distribution supported on the "best" connected component, i.e., the component having the largest maximum equivocation (or any one of them if there exist multiple connected components of capacity $\mathsf{C}(W)$ each). The remaining connected components are not used in capacity-achieving commitment schemes. This implies that, in general, a capacity-achieving input distribution need not assign positive probabilities to all symbols from $\mathcal{X}$, which may at first appear counter-intuitive because we are trying to maximize the uncertainty of the input given the output. The corresponding output distribution will then also assign zero probability to all symbols from unused connected components. For comparison we note that, in the case of *information capacity* ($\max_P I_{\mathrm{X;Y}}(P \times W)$), all output symbols have positive probability under an optimal distribution [8, p. 95].

In Theorem 2.2 below we show that, if the channel is connected, then the commitment-capacity-achieving input (resp. output) distribution is supported on all of $\mathcal{X}$ (resp. $\mathcal{Y}$).

REMARK 2.2. In general, the capacity-achieving input distribution is not unique, and neither is the corresponding output distribution. For example, consider the channel $(\mathcal{X}, \mathcal{Y}, W)$ consisting of two connected components, $(\mathcal{X}_i, \mathcal{Y}_i, W_i)$, $i = 1, 2$, having the same capacity, $\mathsf{C}(W_1) = \mathsf{C}(W_2)$. Suppose that P_i^* is a distribution on $\mathcal{X}_i$ that achieves the capacity of W_i, $i = 1, 2$. Then the convex combination $\alpha P_1^* + (1-\alpha)P_2^*$ is a distribution on $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ that achieves the capacity of W for any $\alpha \in [0, 1]$ (see (2.9)). The output distributions on $\mathcal{Y}$ induced by these input distributions are also clearly different. For comparison we note that, in the case of information capacity of a DMC, there may in general be multiple capacity-achieving input distributions, but they all induce the same output distribution [8, p. 96].

Our Theorems 2.1 and 2.2 demonstrate that the above scenario – a DMC containing distinct connected components with the same commitment capacity – is the only case in which the optimal input distribution is not unique.

The following claim is inspired by [13, Lemma 1(b)], which states an analogous result for mutual information.

LEMMA 2.2. *Let $(\mathcal{X}, \mathcal{Y}, W)$ be a connected channel. Let P, P^* be two probability distributions on $\mathcal{X}$, and assume that P^* achieves the maximum in* (1.4). *Denote $S = P \times W$, $S^* = P^* \times W$, and let $S_{\mathrm{X|Y}}, S^*_{\mathrm{X|Y}}$ be the families of conditional distributions induced by S, S^*. Then*

$$H_{\mathrm{X|Y}}(S^*) - H_{\mathrm{X|Y}}(S) \geqslant D_{\mathrm{KL}}\big(S_{\mathrm{X|Y}} \parallel S^*_{\mathrm{X|Y}} \,|\, S_{\mathrm{Y}}\big). \tag{2.10}$$

Proof. Let S, S^* be as in the statement and suppose, for the sake of contradiction, that

$$\begin{aligned} H_{\mathrm{X|Y}}(S^*) - H_{\mathrm{X|Y}}(S) - D_{\mathrm{KL}}\big(S_{\mathrm{X|Y}} \parallel S^*_{\mathrm{X|Y}} \,|\, S_{\mathrm{Y}}\big) \\ = H_{\mathrm{X|Y}}(S^*) + \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} S(x,y) \log S^*_{\mathrm{X|Y}}(x|y) < 0. \end{aligned} \tag{2.11}$$

Consider the distribution

$$S_\alpha = \alpha S + (1-\alpha) S^*$$

and its conditional entropy $H_{\mathrm{X|Y}}(S_\alpha)$. For $\alpha = 0$, we have

$$H_{\mathrm{X|Y}}(S_0) = H_{\mathrm{X|Y}}(S^*) = \mathsf{C}(W), \tag{2.12}$$

and

$$\left.\frac{\mathrm{d}H_{\mathrm{X|Y}}(S_\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=0} = -H_{\mathrm{X|Y}}(S^*) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} S(x,y) \log S^*_{\mathrm{X|Y}}(x|y), \tag{2.13}$$

which is positive by our assumption (2.11). This implies that, when α is in a sufficiently small neighborhood of zero, $H_{\mathrm{X|Y}}(S_\alpha) > H_{\mathrm{X|Y}}(S^*)$, contradicting the assumption that S^* maximizes $H_{\mathrm{X|Y}}$. ∎

THEOREM 2.2. *Let $(\mathcal{X}, \mathcal{Y}, W)$ be a connected channel. Then there is exactly one distribution P^* on $\mathcal{X}$ that achieves the maximum in* (1.4)*, and this distribution is supported on all of $\mathcal{X}$, i.e., $P^*(x) > 0$ for every $x \in \mathcal{X}$.*

Proof. We first show the second part of the statement. Consider $H_{\mathrm{X|Y}}(P \times W)$ as a function of the probability vector P over $\mathcal{X}$. This function is concave (see Lemma 2.1) and its derivatives

$$\frac{\partial H_{\mathrm{X|Y}}(P \times W)}{\partial P(x)} = \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{\sum_{z \in \mathcal{X}} P(z) W(y|z)}{P(x) W(y|x)} \tag{2.14}$$

are continuous. It therefore satisfies the assumptions of [8, Theorem 4.4.1], which states that necessary and sufficient conditions for P to maximize $H_{\mathrm{X|Y}}(P \times W)$ are

$$\frac{\partial H_{\mathrm{X|Y}}(P \times W)}{\partial P(x)} = \lambda \quad \text{for } x \text{ such that } P(x) > 0, \tag{2.15a}$$

$$\frac{\partial H_{\mathrm{X|Y}}(P \times W)}{\partial P(x)} \leqslant \lambda \quad \text{for } x \text{ such that } P(x) = 0, \tag{2.15b}$$

for some $\lambda \in \mathbb{R}$ (these are the well-known KKT conditions in optimization theory). Let P^* be such a maximizer and let $x \in \mathcal{X}$ be a symbol such that $P^*(x) > 0$. Then every $y \in \mathcal{Y}$ that can be produced by x, i.e., such that $W(y|x) > 0$, also has positive probability, $\sum_{z \in \mathcal{X}} P^*(z) W(y|z) \geqslant P^*(x) W(y|x) > 0$. Since the channel has at least two input symbols, and is connected by assumption, for at least one such y there exists another $x' \in \mathcal{X} \setminus \{x\}$ such that $W(y|x') > 0$. Now, if $P^*(x')$ were equal to zero, we would have $\frac{\partial H_{\mathrm{X|Y}}(P^* \times W)}{\partial P^*(x')} = \infty$ (see (2.14)), contradicting (2.15b). Therefore, it must be the case that $P^*(x') > 0$. Continuing in this way, using the assumption that the channel is connected, we conclude that $P^*(x) > 0$ for every $x \in \mathcal{X}$.

It remains to prove uniqueness. Suppose that $P^\star, P^*$ both achieve the maximum in (1.4), and let $S^\star = P^\star \times W$, $S^* = P^* \times W$. We know from the previous paragraph that $S^\star_{\mathrm{X}} \equiv P^\star, S^*_{\mathrm{X}} \equiv P^*$ and $S^\star_{\mathrm{Y}}, S^*_{\mathrm{Y}}$ are strictly positive over $\mathcal{X}$ and $\mathcal{Y}$, respectively. We now invoke Lemma 2.2. Since $H_{\mathrm{X|Y}}(S^\star) = H_{\mathrm{X|Y}}(S^*) = \mathsf{C}(W)$ by assumption, and since the KL divergence is always nonnegative, we conclude from (2.10) that $D(S^\star_{\mathrm{X|Y}} \parallel S^*_{\mathrm{X|Y}} \,|\, S^\star_{\mathrm{Y}}) = 0$, implying that $S^\star_{\mathrm{X|Y}}(\cdot|y) = S^*_{\mathrm{X|Y}}(\cdot|y)$ for every $y \in \mathcal{Y}$, i.e.,

$$\frac{S^*(x,y)}{S^*_{\mathrm{Y}}(y)} = \frac{S^\star(x,y)}{S^\star_{\mathrm{Y}}(y)}, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{2.16}$$

We also know that $S^{\star}_{\mathrm{Y|X}} = S^{*}_{\mathrm{Y|X}} = W$, i.e.,

$$\frac{S^{*}(x,y)}{P^{*}(x)} = \frac{S^{\star}(x,y)}{P^{\star}(x)}, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{2.17}$$

The relations (2.16) and (2.17) together imply that

$$\frac{P^{\star}(x)}{P^{*}(x)} = \frac{S^{\star}_{\mathrm{Y}}(y)}{S^{*}_{\mathrm{Y}}(y)} \tag{2.18}$$

for all x, y with $W(y|x) > 0$ (whence $S^{\star}(x,y) > 0$ and $S^{*}(x,y) > 0$). Given that the channel is connected, however, it is easy to see that (2.18) must hold for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Finally, since $P^{\star}, P^{*}$ are probability distributions, the ratio $P^{\star}(x)/P^{*}(x)$ being a constant independent of x implies that this constant is in fact 1, i.e., $P^{\star} = P^{*}$. ∎

We note for later reference that the constant λ in (2.15) is precisely the capacity $\mathsf{C}(W)$, namely

$$\lambda = \sum_{x \in \mathcal{X}} P^{*}(x) \frac{\partial H_{\mathrm{X|Y}}(P^{*} \times W)}{\partial P^{*}(x)} = H_{\mathrm{X|Y}}(P^{*} \times W) = \mathsf{C}(W), \tag{2.19}$$

which follows from (2.15a), (2.14), (1.2) and (1.4).

The following theorem provides an alternative way of writing the commitment capacity, which appears more complicated than (1.4) but will prove useful in what follows. It is analogous to the corresponding double-maximization form of the information capacity of a DMC [2].

THEOREM 2.3. *For a channel $(\mathcal{X}, \mathcal{Y}, W)$, a distribution P on $\mathcal{X}$, and a distribution Q on $\mathcal{Y}$, denote*

$$f(P,W,Q) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x)W(y|x) \log \frac{Q(y)}{P(x)W(y|x)}. \tag{2.20}$$

Then:

(a) $\mathsf{C}(W) = \max_P \max_Q f(P,W,Q)$.

(b) *For fixed P and W, $f(P,W,Q)$ is uniquely maximized when $Q = (P \times W)_{\mathrm{Y}}$, i.e., $Q(y) = \sum_{x \in \mathcal{X}} P(x)W(y|x)$.*

(c) *For fixed W and Q, $f(P,W,Q)$ is uniquely maximized when*

$$P(x) = \frac{\exp\bigl(\sum_{y \in \mathcal{Y}} W(y|x) \log \frac{Q(y)}{W(y|x)}\bigr)}{\sum_{z \in \mathcal{X}} \exp\bigl(\sum_{y \in \mathcal{Y}} W(y|z) \log \frac{Q(y)}{W(y|z)}\bigr)}. \tag{2.21}$$

Proof. Since

$$H_{\mathrm{X|Y}}(S) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} S(x,y)\log\frac{Q(y)}{S(x,y)} = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} S(x,y)\log\frac{S_{\mathrm{Y}}(y)}{Q(y)} = D_{\mathrm{KL}}(S_{\mathrm{Y}} \parallel Q) \geqslant 0, \tag{2.22}$$

with equality if and only if $Q = S_{\mathrm{Y}}$, it follows that

$$H_{\mathrm{X|Y}}(S) = \max_{Q} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} S(x,y)\log\frac{Q(y)}{S(x,y)}. \tag{2.23}$$

This proves both (a) (see (1.4)) and (b). Part (c) is obtained by applying the standard Lagrange multiplier method to the function $f(\cdot, W, Q)$. ∎

3. COMPUTING THE COMMITMENT CAPACITY AND THE OPTIMAL INPUT DISTRIBUTION

In this section we describe an iterative procedure for computing the commitment capacity and the optimal input distribution, analogous to the Blahut–Arimoto algorithm [1, 2] for computing the information capacity of a DMC. We should note that the general conditions for such an alternating optimization procedure to be possible are well-known [7], and thus Theorem 3.1 below may be seen as a special case thereof. Nonetheless, having in mind the operational interpretation of the problem (1.4) mentioned in Section 1.3, and a large number of works on the subject, we believe it is useful to describe this procedure explicitly.

The basis of the algorithm is the double maximization form of $\mathsf{C}(W)$ stated in Theorem 2.3 and the properties of the function f being maximized. To be more concrete, the algorithm is suggested by Theorem 2.3(b–c): start with some input distribution $P^{(0)}$, maximize $f(P^{(0)}, W, \cdot)$ over its third argument by setting its value to $(P^{(0)}\times W)_{\mathrm{Y}}$, then maximize $f(\cdot, W, (P^{(0)}\times W)_{\mathrm{Y}})$ over its first argument by setting its value to $P^{(1)}$ according to (2.21), and repeat. Theorem 2.3 implies that the resulting sequence of values $f(P^{(\ell)}, W, (P^{(\ell)}\times W)_{\mathrm{Y}}) \equiv H_{\mathrm{X|Y}}(P^{(\ell)}\times W)$, $\ell = 0, 1, \ldots$, is nondecreasing, and since it is bounded from above by $\mathsf{C}(W)$, it converges. The following theorem states that it converges to the desired value.

Denote

$$G(P, W; x) = \exp\biggl(\sum_{y\in\mathcal{Y}} W(y|x)\log\frac{\sum_{z\in\mathcal{X}} P(z)W(y|z)}{W(y|x)}\biggr), \tag{3.1}$$

and note that

$$H_{\mathrm{X|Y}}(P\times W) = \sum_{x\in\mathcal{X}} P(x)\log\frac{G(P, W; x)}{P(x)}. \tag{3.2}$$

THEOREM 3.1. *Let $(\mathcal{X}, \mathcal{Y}, W)$ be a connected channel and P^* the unique input distribution that attains $\mathsf{C}(W)$. Let $P^{(0)}$ be any distribution on $\mathcal{X}$, and define the sequence of distributions $(P^{(\ell)})_{\ell \geqslant 0}$ by*

$$P^{(\ell+1)}(x) = \frac{G(P^{(\ell)}, W; x)}{\sum_{z \in \mathcal{X}} G(P^{(\ell)}, W; z)}, \quad x \in \mathcal{X}. \tag{3.3}$$

Then, as $\ell \to \infty$, $P^{(\ell)}$ converges to P^*, and $H_{\mathrm{X|Y}}(P^{(\ell)} \times W)$ monotonically converges to $H_{\mathrm{X|Y}}(P^* \times W) = \mathsf{C}(W)$.

Proof. The update rule (3.3) expresses the two maximization steps described in the first paragraph of this section, performed simultaneously. As noted above, we know from Theorem 2.3 that the sequence $(H_{\mathrm{X|Y}}(P^{(\ell)} \times W))_{\ell \geqslant 0}$ monotonically converges to some value, and we need to show that this value is $H_{\mathrm{X|Y}}(P^* \times W) = \mathsf{C}(W)$. Take any convergent subsequence of the sequence $(P^{(\ell)})_{\ell \geqslant 0}$, and denote its limit by $\overline{P}$. (The sequence $(P^{(\ell)})_{\ell \geqslant 0}$ is defined in the $|\mathcal{X}|$-dimensional probability simplex, so by the Bolzano–Weierstrass theorem it must have a convergent subsequence.) Then it follows from (3.3) that $\overline{P}$ satisfies the equations

$$\overline{P}(x) = \frac{G(\overline{P}, W; x)}{\sum_{z \in \mathcal{X}} G(\overline{P}, W; z)}, \quad x \in \mathcal{X}, \tag{3.4}$$

which can be rewritten as

$$\sum_{y \in \mathcal{Y}} W(y|x) \log \frac{\sum_{z \in \mathcal{X}} \overline{P}(z) W(y|z)}{\overline{P}(x) W(y|x)} = \lambda, \quad x \in \mathcal{X}, \tag{3.5}$$

where $\lambda = \log \sum_{z \in \mathcal{X}} G(\overline{P}, W; z)$. Similarly to the proof of Theorem 2.2, one can now argue that (3.5) implies that $\overline{P}(x) > 0$ for all $x \in \mathcal{X}$. Now notice that (3.5) is equivalent to the necessary and sufficient conditions (2.15) for a distribution to be the maximizer of $H_{\mathrm{X|Y}}$, and therefore $\overline{P} = P^*$. ∎

For the iterative procedure (3.3) to be useful, it is left to specify the termination criterion. To this end note that

$$H_{\mathrm{X|Y}}(P^{(\ell)} \times W) \leqslant \mathsf{C}(W) \tag{3.6a}$$

$$= \sum_{x \in \mathcal{X}} P^*(x) \log \frac{G(P^*, W; x)}{P^*(x)} \tag{3.6b}$$

$$\leqslant \sum_{x \in \mathcal{X}} P^*(x) \log \frac{G(P^{(\ell)}, W; x)}{P^{(\ell)}(x)} \tag{3.6c}$$

$$\leqslant \max_{x \in \mathcal{X}} \log \frac{G(P^{(\ell)}, W; x)}{P^{(\ell)}(x)} \tag{3.6d}$$

where the inequality in (3.6c) holds because the difference of the expressions in (3.6c) and (3.6b) equals $D_{\mathrm{KL}}\big((P^* \times W)_{\mathrm{X|Y}} \parallel (P^{(\ell)} \times W)_{\mathrm{X|Y}} | (P^* \times W)_{\mathrm{Y}}\big)$ and is therefore nonnegative. This shows that the capacity $\mathsf{C}(W)$ is sandwiched between $H_{\mathrm{X|Y}}(P^{(\ell)} \times W)$ and $\log \max_{x \in \mathcal{X}} \frac{G(P^{(\ell)}, W; x)}{P^{(\ell)}(x)}$ for every $\ell \geqslant 0$. Moreover, both of these quantities tend to $\mathsf{C}(W)$ because $\log \frac{G(P^*, W; x)}{P^*(x)} = \mathsf{C}(W)$ for every $x \in \mathcal{X}$ (see (3.1), (2.14), (2.15), and (2.19)). This implies that, given any $\epsilon > 0$, for sufficiently large ℓ we are guaranteed to have

$$\log \max_{x \in \mathcal{X}} \frac{G(P^{(\ell)}, W; x)}{P^{(\ell)}(x)} - H_{\mathrm{X|Y}}(P^{(\ell)} \times W) \leqslant \epsilon, \tag{3.7}$$

in which case

$$\mathsf{C}(W) - \epsilon \leqslant H_{\mathrm{X|Y}}(P^{(\ell)} \times W) \leqslant \mathsf{C}(W). \tag{3.8}$$

Hence, if ϵ is the desired accuracy, the iteration (3.3) may be terminated at the first value of ℓ for which (3.7) holds.

REFERENCES

[1] S. Arimoto, *An algorithm for computing the capacity of arbitrary discrete memoryless channels*, IEEE Trans. Information Theory 18 (1972), 14–20.

[2] R. Blahut, *Computation of channel capacity and rate-distortion functions*, IEEE Trans. Information Theory 18 (1972), 460–473.

[3] M. Blum, *Coin flipping by telephone a protocol for solving impossible problems*, ACM SIGACT News 15 (1983), 23–27.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley, 2006.

[5] C. Crépeau, *Efficient cryptographic protocols based on noisy channels*, in: Proc. 1997 Int. Conf. Theory Appl. Cryptogr. Tech. (EUROCRYPT), Konstanz, 306–317.

[6] C. Crépeau, R. Dowsley, and A. C. A. Nascimento, *On the commitment capacity of unfair noisy channels*, IEEE Trans. Information Theory 66 (2020), 3745–3752.

[7] I. Csiszár and G. Túsnady, *Information geometry and alternating minimization procedures*, Statistics and Decisions 1984, Suppl. Issue, 205–237.

[8] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

[9] O. Goldreich, S. Micali, and A. Wigderson, *Proofs that yield nothing but their validity, or all languages in NP have zero-knowledge proof systems*, J. ACM 38 (1991), 690–728.

[10] M. Hayashi and N. A. Warsi, *Commitment capacity of classical-quantum channels*, IEEE Trans. Information Theory 69 (2023), 5083–5099.

[11] H. Imai, J. Müller-Quade, A. Nascimento, and A. Winter, *Rates for bit commitment and coin tossing from noisy correlation*, in: Proc. 2004 IEEE Int. Symp. Information Theory (ISIT), Chicago, 1 p.

[12] M. Mamindlapally, A. K. Yadav, M. Mishra, and A. J. Budkuley, *Commitment capacity under cost constraints*, in: Proc. 2021 IEEE Int. Symp. Information Theory (ISIT), Melbourne, 3208–3213.

[13] S. Shamai and S. Verdú, *The empirical distribution of good codes*, IEEE Trans. Information Theory 43 (1997), 836–846.

[14] H. Tyagi and S. Watanabe, *Information-Theoretic Cryptography*, Cambridge Univ. Press, 2023.

[15] A. Winter, A. Nascimento, and H. Imai, *Commitment capacity of discrete memoryless channels*, in: Proc. 2003 IMA Int. Conf. Cryptogr. and Coding, Cirencester, 35–51.

Nemanja Nedić, Mladen Kovačević, Darko Čapko, Srđan Vukmirović
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
and
Ethernal
Novi Sad, Serbia
E-mail: nemanja.nedic@uns.ac.rs
kmladen@uns.ac.rs
dcapko@uns.ac.rs
srdjanvu@uns.ac.rs

Received 24.4.2024;
accepted 13.8.2025