Universal Polarization Codes

Mine Alsan

LTHI, I&C, EPFL

***Abstract*—A challenging problem related to the design of polar codes is "robustness against channel parameter variations", as stated in Arikan's original work. To build more robust systems, different models have been proposed in the literature. After we survey polar codes and the channel polarization principle, we present results on the capacity of a class of channels, and on the performance of a general class of additive d-decoders with a particular interest in mismatch decoders. We describe how the problem of robust polar code design can be viewed as a mismatch decoding problem over unknown channels or set of channels.**

***Index Terms*—channel polarization, polar codes, unknown channels, compound capacity, universal coding, d-decoders, mismatch decoders.**

I. Introduction

ONE major problem considered by the field of information theory is of reliable communication over a single point-to-point channel. To this objective, appropriate coding schemes are designed where channel encoders construct redundant messages, such that upon reception, the corresponding channel decoders can overcome the effects of noise introduced during transmission. An information rate is achievable over the channel if no matter how small the error probability is, we can find a code with a sufficiently large block-length of that rate. Shannon defined the channel capacity as the highest achievable rate that information can be sent through the channel. The error probability analysis over random code ensembles used with typicality decoders, tailored to jointly typical sequences, provides a simple way to show that the mutual information between the channel input and output, denoted by $I(P, W)$ where $P(x)$ is the input distribution and $W(y|x)$ the channel transition probability for any input $x \in \mathcal{X}$, and output $y \in \mathcal{Y}$, is a fundamental quantity [1].

Gallager refined the error analysis to provide a better characterization of the problem. The noisy channel coding theorem [2, Chapter 5] states that for a discrete memoryless channel (DMC), and a fixed rate $R > 0$, the average block decoding error probability P_e of an ensemble of block codes with codewords of length N using maximum likelihood (ML) decoding, can be upper bounded as $P_e \leq \exp\{-NE_r(R)\}$. The reliability exponent $E_r(R)$ establishes the compromise between P_e, N, and R. Thus, P_e vanishes exponentially with increasing N if $E_r(R) > 0$. Clearly, the approach here expresses different trade-offs involved in system design.

An important aspect of the above outcomes is that they are existence results, based on averaging over code ensembles. We know there are codes which are good when evaluated with respect to the achievability criteria. We can use these codes to communicate at rates up to the ultimate limit set by the probabilistic channel, namely the channel capacity. Even though a simple random draw from the "codebag" might give the good code the designer is looking for, no real system exists without cost. This last one is dictated by storage capabilities, and computational complexity limitations in case of communication systems. Consider a block code with exponentially many codewords in the block-length. Quickly we find ourselves out of memory for moderate values of the block-length, and of making unbounded pairwise comparisons to identify the correct message sent through the channel.

Coding theorists have spent decades aiming to find good codes with low encoding and decoding complexities. Linear codes give a partial answer to this problem as a class of codes with an efficient structure introduced for the codeword generation process. The encoder transforms the input message $u_1^N \triangleq (u_1, \ldots, u_N)$, to a codeword x_1^N using a generator matrix G_N, i.e. $x_1^N = u_1^N G_N$.

In 2007, Arıkan [3] proposed polar codes as an appealing error correction method. This class of codes are proved to achieve the symmetric capacity of any binary DMC (B-DMC) using low complexity encoders and decoders, and their block error probability is shown to decrease exponentially in the square root of the block-length [4]. A simple observation gives the idea of polarization. Assume we make N independent channel uses to transmit information after applying a one-to-

Proposal submitted to committee: September 12th, 2011; Candidacy exam date: September 20th, 2011; Candidacy exam committee: Michael Gastpar, Emre Telatar, Bixio Rimoldi.

This research plan has been approved:

Date: _____

Doctoral candidate: _____
(M. Alsan) (name and signature)

Thesis director: _____
(E. Telatar) (name and signature)

Thesis co-director: _____
(if applicable) (name and signature)

Doct. prog. director: _____
(R. Urbanke) (signature)

one transformation G_N to the data as above, we know that

$$I(U_1^N;Y_1^N) = I(X_1^N;Y_1^N) = \sum_{i=1}^{N} I(X_i;Y_i) = NI(W) \quad (1)$$

with the symmetric capacity of the channel $I(W)$ defined as

$$I(W) = \sum_{x,y} \frac{1}{2} W(y \mid x) \log \frac{W(y \mid x)}{\frac{1}{2}W(y \mid 0) + \frac{1}{2}W(y \mid 1)}.$$

When we apply the chain rule we obtain

$$I(U_1^N;Y_1^N) = \sum_{i=1}^{N} I(U_i;Y_1^N \mid U_1^{i-1}) = \sum_{i=1}^{N} I(U_i;Y_1^N U_1^{i-1}). \quad (2)$$

Therefore instead of the N independent channel uses of W, we can consider from the chain rule perspective N successive uses of the channels $W_N^{(1)}, \ldots, W_N^{(N)}$ given by the transition probabilities $W_N^{(1)}(y_1^N \mid u_1), \ldots, W_N^{(N)}(y_1^N u_1^{N-1} \mid u_N)$, respectively. Following this observation, channel polarization is defined as the case when each of these channels converges either to a perfect channel or completely noisy channel. Under this event, the empirical distribution of the mutual information terms in summation (2) satisfies

$$\frac{1}{N} \ \#\{i : I(U_i;Y_1^N U_1^{i-1}) \in (\gamma, 1-\gamma)\} \xrightarrow[n\to\infty]{} 0 \quad (3)$$

for any $\gamma \in (0,1)$. In fact, two basic channel transformations lie at the heart of channel polarization. The channels $W_N^{(i)}$ can be synthesized by the recursive applications of these transformations until sufficiently polarized.

This idea is exploited to propose polar codes, and the recursive process leads to efficient encoding, and decoding structures. On the encoder side, uncoded data bits are sent only through those perfect channels. Polar encoders' job seem to be rather simple given the knowledge of the good channels. On the decoder side, the chain rule channels lend themselves to a particular decoding procedure referred to as successive cancellation decoder (SCD). At the i-th stage, the SCD estimates the channel input u_i with law $W_N^{(i)}(y_1^N u_1^{i-1} \mid u_i)$ according to some suitable decision rule based on the current channel information, and supplies the estimates $\hat{u}_1, \ldots, \hat{u}_{i-1}$ of the previous channel inputs to the next stage. Though no genie exists to give the correct estimates, the analysis carried in [3] shows that the decoder still performs with vanishing error probability. More details on the paper are discussed in the first part of the survey section.

Beside their computational complexity, typicality decoders and ML decoders present another obstacle: they require the exact channel knowledge to function. To tackle the situations in which such a complete knowledge is missing, the study of reliable communication under channel uncertainty becomes relevant. The qualitative notion of channel uncertainty is made more concrete by the definition of more complex channel models. For instance, the unknown channel may be restricted to belong to some given class of channels. This model we will also be particularly interested in, is called the compound channel model. Here both the encoder and decoder are ignorant of which of the channels in the class occurs, but the channel remains fixed during the communication of a given codeword. Hence, the selected code should ensure good performance for all channels in the class. The paper by Blackwell et al. [5] that we survey in the next section, defined the notion of the capacity of a class of channels. Note that this last is allowed to contain infinitely many channels. A key tool will be described to attack coding within this context. Another more complex model is the arbitrarily varying channel (AVC) model. Here the channel law is allowed to vary over each use of the channel.

The importance of the decoders stands out even more in those more complex communication scenarios. To allow their study within a unified framework, decoders can be categorized based on generic definitions of the decision functions. Alpha (α) decoders are such a class where the decision rules are based on the joint type of the codewords and the received sequence. The derivation of the error exponent is based on a very useful tool called the method of types [6]. For the single point-to-point channel W, the random coding exponent of α decoders used with constant composition codes of type $\hat{P}$ and rate R, is given by Csiszár et al. [7] as

$$E_{r,\alpha}(R,\hat{P},W) = \inf_{\substack{\hat{P}_1 \in \mathcal{P}^n(\hat{P},W) \\ \hat{P}_2 \in \mathcal{P}^n(\hat{P},\hat{P}_1) \\ \alpha(\hat{P}\hat{P}_2) \le \alpha(\hat{P}\hat{P}_1)}} \left(D(\hat{P}_1||W|\hat{P}) + |I(\hat{P},\hat{P}_2) - R|^+ \right) \quad (4)$$

where $\hat{P}_1(y|x)$, $\hat{P}_2(y|x)$ are noise compositions, and we define

$$\mathcal{P}^n(\hat{P},W) \triangleq \{\hat{P}_1 : \hat{P}\hat{P}_1(x,y) \in \mathcal{P}^n(\mathcal{X},\mathcal{Y}), \hat{P}\hat{P}_1(y) = \hat{P}W(y)\}.$$

Each input distribution and channel transition probability induce a joint one denoted as $\hat{P}W(x,y) = \hat{P}(x)W(y|x)$, and the corresponding output marginals are found as usual from $\hat{P}W(y) = \sum_x \hat{P}W(x,y)$. $\mathcal{P}^n(\mathcal{X},\mathcal{Y})$ denotes the set of joint n-types, $D(\hat{P}_1||W|\hat{P})$ is the conditional divergence, and $|x|^+ = \max(x,0)$. Refer to [6], and [8] for more details on these notations.

The need to communicate reliably under the lack of channel knowledge led to the concept of universal coding. The maximum mutual information decoder (MMI) is such a decoder which declares an input message iff the corresponding codeword maximizes the empirical mutual information computed from the joint type of a given codeword and the received channel output sequence. Then, $E_{r,\alpha}(R,\hat{P},W)$ reduces to

$$\inf_{\hat{P}_1 \in \mathcal{P}^n(\hat{P},W)} \left(D(\hat{P}_1||W|\hat{P}) + |I(\hat{P},\hat{P}_1) - R|^+ \right) \quad (5)$$

The exponent is positive iff $R < I(\hat{P},W)$. Consequently, universally attainable transmission rates are obtained for any DMC with finite input, output alphabets.

Therefore, we encounter sub-optimal decoders in many cases whether this is for practical implementation purposes (complexity, feasibility requirements), or partial/missing channel information. Although this may result in capacity loss, this is in general unavoidable. The final selected paper of the survey section is concerned with the study of a more restricted class of decoders, and their performance in terms of capacity. The paper [7] focus on additive decoding rules called d-decoders. Even though this is a more restricted class, and the MMI decoder is no longer treated within this class, d-decoders still provide a broad enough framework to allow the study of

rules such as the optimal or mismatched (soon defined) ML decoding rules, or of difficult information-theoretical problems such as the Shannon capacity of a graph, and the zero-error capacity of a DMC [9].

In our research, we will focus on a particular type of d-decoder: mismatch decoders. The usual ML rule is kept, but instead of the true channel law a different one is employed in the decision procedure, whence the mismatch term is justified. These rules are particularly relevant for polar codes. The polar SCD meets the low complexity requirement, but the decision functions are still depending on the channel information. Moreover, we didn't consider how the good channel indices are found. As we will see later, it turns out that the polarization process is adjusted to the particular channel at hand. As a consequence, polar codes are channel specific designs. If the true channel is unknown, the code design should be based on a mismatched channel. We could further help the designer and give some partial information. Coming back to the compound model, the channel can be assumed to belong to a given class. Hence, it is important to assess the performance of polar codes over an unknown channel or given set of channels for mismatched designs.

Overall, these ideas can be viewed as an attempt to construct universal polarization codes. Up to our knowledge, only two papers appeared on the universality of polar codes: [10] studies the compound capacity of polar codes, and [11] is concerned with the universality in the source coding dual of polar codes.

In the next section, the article follows with a survey of the mentioned three papers. Finally, we conclude with the research proposal to introduce basic research ideas, and directions.

II. Survey of the Selected Papers:

A. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels

We have already introduced the channel polarization principle. In this section, we take a closer look at this phenomenon, and polar codes as described in [3]. We deal with B-DMCs unless specified otherwise.

A Closer Look at Polarization: To carry the discussion forward, we first introduce another channel parameter in companion to the symmetric capacity $I(W)$. The Bhattacharyya parameter $Z(W)$ of a channel is defined as

$$Z(W) = \sum_{y \in \mathcal{Y}} \sqrt{W(y \mid 0)W(y \mid 1)}.$$

This parameter gives an upper bound to ML decoding error probability of a single use of the channel W, and stands as a measure of reliability.

Consider the transformation matrix given as

$$G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{6}$$

The corresponding channel configuration is drawn in Fig. 1 combining two independent copies of W.

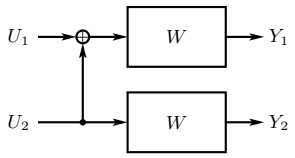

Fig. 1. Basic channel transformations

The two successive channels $W_N^{(1)}$ and $W_N^{(2)}$ are characterized by the transformations $W^- : \mathcal{X} \to \mathcal{Y}^2$ and $W^+ : \mathcal{X} \to \mathcal{Y}^2 \times \mathcal{X}$ respectively. Referred as the basic channel transformations, W^- and W^+ can be defined by the following transition probabilities

$$W^-(y_1 y_2 \mid u_1) = \sum_{u_2 \in \mathcal{X}} \frac{1}{2} W(y_1 \mid u_1 \oplus u_2) W(y_2 \mid u_2) \tag{7}$$

$$W^+(y_1 y_2 u_1 \mid u_2) = \frac{1}{2} W(y_1 \mid u_1 \oplus u_2) W(y_2 \mid u_2) \tag{8}$$

The following properties related to the above transformations are derived in [3]:

Properties 1 (W^-, W^+):

1) $I(W^-) + I(W^+) = 2I(W)$
2) $Z(W^-) + Z(W^+) \leq 2Z(W)$
3) $I(W^-) \leq I(W) \leq I(W^+)$
   $Z(W^-) \geq Z(W) \geq Z(W^+)$
4) $Z(W^+) = Z(W)^2$
   $Z(W^-) \leq 2Z(W) - Z(W)^2$ with equality iff W is a binary erasure channel (BEC).
5) $I(W)^2 + Z(W)^2 \leq 1$
   $I(W) + Z(W) \geq 1$ with equality iff W is a BEC.
6) Given a BEC W_{BEC}, the channels W_{BEC}^- and W_{BEC}^+ are also binary erasure channels.

Overall, we see that the mutual information is preserved, the overall reliability is improved, and $I(W) = 1$ iff $Z(W) = 0$ and vice-versa. The third property confirms that we have made progress towards polarization as while the channel W^+ is improved, the channel W^- is worsened. The idea now is to apply the same basic channel transformations to the channels W^- and W^+. As a result, four channels W^{--}, W^{-+}, W^{+-}, and W^{++} are obtained. However in general, we are no longer able to compare the parameters of these four channels in terms of rate and reliability, except the knowledge that the channel W^{++} is the best one and the channel W^{--} is the worst one. We keep applying the transformations to the obtained $\pm$ channels. To analyze the convergence properties of this recursion, the polarization process is defined.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. The random sequence $B_1, \ldots, B_\ell$ is drawn i.i.d. according to a Bernoulli distribution with probabilities equal to $\frac{1}{2}$. Let $\mathcal{F}_\ell$ be the σ-algebra generated by this Bernoulli sequence. Given a channel W, the random sequence of channels $\{W(\ell)\}$ is defined for $\ell \geq 0$ as

$$W(\ell) = \begin{cases} W & \text{if} \quad \ell = 0 \\ \{W(\ell-1)\}^- = & \text{if} \quad B_\ell = 0 \\ \{W(\ell-1)\}^+ = & \text{if} \quad B_\ell = 1 \end{cases}$$

In the sequel, the random processes $I_\ell = I(W(\ell))$, $Z_\ell = Z(W(\ell))$ are defined, and the next properties are proved in [3]:

Properties 2 (I_ℓ, Z_ℓ):

1) Properties 1 hold for I_ℓ and Z_ℓ at each $\ell \geq 0$.
2) Each $W(\ell)$ corresponds to one of the coordinate channels $W_{2^\ell}^{(i)}$. The binary expansion of $i-1 = \sum_{j=1}^{n} b_j 2^{n-j}$, where $b_1 \dots b_n$ is a sample of the Bernoulli process.
3) The process $\{I_\ell, \mathcal{F}_\ell\}$ is a bounded martingale on the interval $[0,1]$.
4) The process $\{Z_\ell, \mathcal{F}_\ell\}$ is a bounded supermartingale on the interval $[0,1]$.
5) The process $\{Z_\ell, \mathcal{F}_\ell\}$ converges a.s. to a $\{0,1\}$ valued random variable Z_∞.
6) The process $\{I_\ell, \mathcal{F}_\ell\}$ converges a.s. to a random variable I_∞ such that $\mathbb{E}[I_\infty] = I_0$ where I_∞ takes values a.s. in $\{0,1\}$.

The last property proves the next polarization theorem.

Theorem 2.1: For any B-DMC W, the sequence of channels $\{W(\ell)\}$ polarizes such that Eq. (3) holds for $N = 2^\ell$, for fixed $\gamma \in (0,1)$ and for each $i = 1, \dots, 2^\ell$.

A Closer Look at Polar Codes: Using the idea of channel polarization, polar codes are proposed as a new coding technique. Namely, a polar code (N, R) of block length N and rate $R < I(W)$ communicates information only on the $\lfloor NR \rfloor$ good channels whose mutual information values are close to 1 among all the $W_N^{(i)}$ channels with $i = 1, \dots, N$, and simply transmits randomly fixed bits known to both the encoder and decoder over the other bad channels.

Encoder: Given the data sequence u_1^N, the codeword x_1^N is computed from $x_1^N = u_1^N G_N$. The encoder structure is based on an operation called channel combining, consisting of combining the N independent channel realizations of W to distill the channels with transition probabilities given by $W_N(y_1^N|u_1^N) = W^N(y_1^N|u_1^N G_N)$. The recursion can be applied through the channel transformation matrix $G_N = B_N G^{\otimes N}$ where $\otimes$ denotes the Kronecker product, G is defined in Eq. (6), and B_N is the bit reversal permutation matrix. Depending on the direction of polarization of the channels, the data sequence $u_1 \dots u_N$ indexes are split into two sets before transmission. The first one includes the indexes of the data to be transmitted on the good channels, and is referred as the information set $\mathcal{A}_N$. The remaining one is the set $\mathcal{A}_N^c$ of the indexes corresponding to the frozen bits to be transmitted on the bad channels. The selection rule for $\mathcal{A}_N$ will be described after the error performance analysis.

Decoder: The SCD described in the introduction is closely tied to a channel splitting operation. After channel combining, the splitting synthesizes the channels whose transition probabilities are given by:

$$W_N^{(i)}(y_1^N u_1^{i-1}|u_i) = \sum_{u_{i+1}^N} \frac{1}{2^{N-1}} W_N(y_1^N|u_1^N).$$

Decision functions similar to ML decoding rule can then be defined by

$$d_i(y_1^N, \hat{u}_1^{i-1}) = \begin{cases} 0, & \text{if } \frac{W_N^{(i)}(y_1^N, \hat{u}_1^{i-1}|0)}{W_N^{(i)}(y_1^N, \hat{u}_1^{i-1}|1)} \geq 1 \\ 1, & \text{otherwise} \end{cases}$$

The polar SCD will decode the received output in N stages using a chain of estimators from $i = 1, \dots, N$, each depending on the previous ones. The estimators are defined as

$$\hat{\mathcal{U}}_i = \begin{cases} u_i, & \text{if } \ i \in \mathcal{A}_N^c \\ d_i(y_1^N, \hat{u}_1^{i-1}), & \text{if } \ i \in \mathcal{A}_N \end{cases}$$

Code Performance: Let $P_{e,\mathrm{SCD}}(W)$ denotes the best achievable block error probability under successive cancellation decoding over the ensemble of all possible choices of the set $\mathcal{A}_N^c$ when $|\mathcal{A}_N| = \lfloor NR \rfloor$. Then,

$$\begin{aligned} P_{e,\mathrm{SCD}}(W) &= \mathbb{P}\left[\bigcup_{i \in \mathcal{A}_N} \left\{\hat{\mathcal{U}}_1^{i-1} = u_1^{i-1} \cap \hat{\mathcal{U}}_i \neq u_i\right\}\right] \\ &= \mathbb{P}\left[\bigcup_{i \in \mathcal{A}_N} \left\{\ \hat{\mathcal{U}}_1^{i-1} = u_1^{i-1} \cap d_i(y_1^N, \hat{\mathcal{U}}_1^{i-1}) \neq u_i\right\}\right] \\ &\leq \mathbb{P}\left[\bigcup_{i \in \mathcal{A}_N} \left\{d_i(y_1^N, u_1^{i-1}) \neq u_i\right\}\right] \\ &\leq \sum_{i \in \mathcal{A}_N} \sum_{y_1^N, u_1^N} \frac{1}{2} W_N(y_1^N|u_1^N) \sqrt{\frac{W_N^{(i)}(y_1^N, u_1^{i-1} \mid u_i \oplus 1)}{W_N^{(i)}(y_1^N, u_1^{i-1} \mid u_i)}} \\ &\leq \sum_{i \in \mathcal{A}_N} Z(W_N^{(i)}) \end{aligned}$$

where $Z(W_N^{(i)}) \triangleq \sum_{y_1^N, u_1^{i-1}} \sqrt{W_N^{(i)}(y_1^i u_1^{i-1} \mid 0) W_N^{(i)}(y_1^i u_1^{i-1} \mid 1)}$.

We remark that the standalone estimators $\hat{\mathcal{U}}_i$ will perform well as long as the system depending on a genie to provide the correct data values does so. Moreover, the following selection rule for the information set is defined: pick $\mathcal{A}_N \subset \{1, \dots, N\}$ with $\lfloor NR \rfloor$ elements such that $Z(W_N^{(i)}) \leq Z(W_N^{(j)})$ for all $i \in \mathcal{A}_N$ and $j \in \mathcal{A}_N^c$.

Complexities: The described encoder and SCD can be implemented in $O(N \log N)$ complexity. However, no general low complexity algorithm has been proposed to identify the information set for a given B-DMC, except the BEC.

Another important result concerns symmetric channels. It can be shown that the frozen bits can be selected in an arbitrary way because they don't affect the code performance. This suggests to set the frozen bits to 0 for symmetric channels, and results in a deterministic code where the given error bound still holds.

The rate of polarization is crucial to make polar codes part of any real application. We need to ensure that polarization takes place fast enough as, in practice, the performance of polar codes with finite block-lengths are important. The next theorem characterizes the speed of convergence of $P_{e,\mathrm{SCD}}(W)$.

Theorem 2.2: [4] Given any rate $R \geq 0$ such that $R \leq I(W)$ and a constant $\beta \leq \frac{1}{2}$, consider the polar code (N, R) of block length N. Then,

$$P_{e,\mathrm{SCD}}(W) = o(2^{-N^\beta}). \tag{9}$$

Finally, we close this section with some results stated in [12]. The minimum distance of polar codes is given to be $O(\sqrt{N})$ [Lemma 6.3]. Comparing with the error exponent in Eq. (9), one sees that with the proposed SCD, decoding beyond

the minimum distance of the code is possible. Moreover, it is shown that the SCD has a performance comparable to ML decoding of polar codes, up to the first order in the exponent, as $P_{e,\mathrm{ML}}(W) > 2^{-N^\beta}$ for any $\beta > \frac{1}{2}$ [Lemma 6.4].

B. The Capacity of a Class of Channels

Blackwell et al. [5] defined the capacity of a class of DMCs $\mathcal{W}$ by

$$C(\mathcal{W}) = C = \max_{Q(x)} \inf_{W \in \mathcal{W}} I(Q, W).$$

They showed that this quantity corresponds to the supremum of all achievable transmission rates $T(\mathcal{W})$, when the code is required to perform well in all the channels in the class. To be more specific, a maximum error probability criteria is used to asses performance: A sequence of (e^{Rn}, ϵ_n, n) codes of block-length n for the class with e^{nR} codewords and corresponding sequence of disjoint decoding regions $B_i \subset \mathcal{Y}^n$ is required to satisfy for all the codewords $u_i \in \mathcal{X}^n$, and for every channel $W \in \mathcal{W}$, an upper bound $W(B_i^c|u_i) \leq \epsilon_n$ where $\epsilon_n \to 0$. The part (a), and (b) of the next theorem proves, respectively, the direct part, i.e. $T(\mathcal{W}) \geq C$, and the converse part, i.e. $T(\mathcal{W}) \leq C$ of the result.

Theorem 2.3: Let $\mathcal{W}$ be any class of channels. For any integer n, and any $R > 0$

(a) If $0 \leq C - R \leq \frac{1}{2}$, there is an (e^{Rn}, ϵ_n, n) code for $\mathcal{W}$ with

$$\epsilon_n = A\exp\left\{-n\frac{(C-R)^2}{B}\right\} \tag{10}$$

where A, and B are constants depending solely on $|\mathcal{X}|$, $|\mathcal{Y}|$, C, and R.

(b) If $R > C$, and $e^{nR} \geq 2$, then any (e^{Rn}, ϵ_n, n) code for $\mathcal{W}$ cannot have $\epsilon_n \to 0$ whenever $n \to \infty$.

The key point is that the exponential error bound does not depend on the actual channel that occurs during the transmission of a codeword, but depends only on C, R, and the input-output alphabet sizes. Note that the condition $0 \leq C - R \leq \frac{1}{2}$ is for convenience. A code with a higher rate can always be reduced to another one with a lower rate.

Next, we give a brief overview of the proof of the direct part with an emphasis on some approaches which might be useful in our research.

Recipe for a class of channels: Two Approximations:

Step 1: If $\mathcal{W}$ is an infinite class of channels, we first approximate it with a finite class denoted by $\mathcal{W}_Q$ using a quantization procedure. Given a channel $W \in \mathcal{W}$, for each input $x \in \mathcal{X}$, we carry out the next steps:

- Select a number $M > 2|\mathcal{Y}|^2$,
- $\forall y \in \mathcal{Y}$ sort $W(y|x)$ in ascending order $p_1 \leq \ldots \leq p_{|\mathcal{Y}|}$,
- For $i = 1, \ldots, |\mathcal{Y}| - 1$, assign a number $p_i^\cdot$ uniquely such that $p_i \leq p_i^\cdot \leq p_i + \frac{1}{M}$, and $Mp_i^\cdot =$integer,
- Set $p_{|\mathcal{Y}|} = 1 - \sum_{i=1}^{|\mathcal{Y}|-1} p_i^\cdot$, and add W_Q with law $p_i^\cdot$ to $\mathcal{W}_Q$.

Step 2: From this point on, we have a finite class to work with. We index the channels in $\mathcal{W}_Q$ from $i = 1, \ldots, L \triangleq |\mathcal{W}_Q|$. We will soon see that L is not very large. We construct an "average" channel $W_{Q,A}$ with transition probability given by $W_{Q,A}(y|x) = \frac{1}{L}\sum_{i=1}^{L} W_i(y|x)$.

Step 3: For the channel $W_{Q,A}$, and any $P(x)$ on $\mathcal{X}$, compute the joint density $PW_{Q,A}(x, y)$. For each $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{y} \in \mathcal{Y}^n$ calculate the realizations of the random variable J defined as

$$J(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \log \frac{PW_{Q,A}(x_i, y_i)}{P(x_i)PW_{Q,A}(y_i)} \triangleq \sum_{i=1}^{n} J(x_i, y_i)$$

Note that $\mathbb{E}_{PW_{Q,A}(x,y)}[J(X, Y)] = I(P, W_{Q,A})$. Then for a given block-length n, fix a number $\alpha > 0$, and select G codewords using the described codeword selection process:

- Compute the number $\epsilon = Ge^{-\alpha} + \mathbb{P}(J \leq \alpha)$,
- Find a codeword $\mathbf{x}(1)$ such that $\mathbb{P}\{J > \alpha|\mathbf{x}(1)\} \geq 1 - \epsilon$. Set the first decoding region as $B_1 = A_{\mathbf{x}(1)} \triangleq \{(\mathbf{x}, \mathbf{y})|J(\mathbf{x}(1), \mathbf{y}) > \alpha\}$,
- At the k-th step, keep selecting a codeword $\mathbf{x}(k)$ if there exists a subset $B_k = \cup_{i=1}^{k} A_{\mathbf{x}(i)} - \cup_{i=1}^{k-1} A_{\mathbf{x}(i)}$ such that $P(B_k|\mathbf{x}(k)) \geq 1 - \epsilon$ until G codewords are found.

Analysis of the error probability: Now, we proceed backward to analyze the error probability of the resulting code. Let $C = \inf_{W \in \mathcal{W}} I(P^*, W)$, $C_Q = \inf_{W_Q \in \mathcal{W}_Q} I(P^{**}, W_Q)$, and R, R_Q be positive numbers.

Step 4: Note that $W_i(y|x) \leq LW_{Q,A}(y|x)$ for any $W_i \in \mathcal{W}_Q$. So, for any codeword $\mathbf{x}(k)$, and its corresponding decoding region B_k, we have $W_i(B_k^c|\mathbf{x}(k)) \leq LW_{Q,A}(B_k^c|\mathbf{x}(k))$.

Step 5: By the construction in Step 3, the elements of the codewords are drawn from an i.i.d distribution. As we assume a DMC, the random variables $\{J(X_i, Y_i)\}_{i=1}^n$ also form an i.i.d ensemble. Using a Chernoff bound, a proper choice of α and $G = e^{nR}$, the above code can be shown to be a code for the class $\mathcal{W}_Q$ with the maximum probability of error upper bounded by

$$\epsilon_n = 2L^2\exp\left\{-n\frac{(C_Q - R_Q)^2}{16|\mathcal{X}||\mathcal{Y}|}\right\}$$

as long as $0 \leq C_Q - R_Q \leq 1/2$.

Step 6: The channels in the finite class $\mathcal{W}_Q$ can be shown to satisfy the next properties.

Properties 3 (Quantization): For any $W \in \mathcal{W}$, there exists a $W_Q \in \mathcal{W}_Q$ such that for $\forall x \in \mathcal{X}, y \in \mathcal{Y}$:

1) $|W(y|x) - W_Q(y|x)| \leq \frac{|\mathcal{Y}|}{M}$,
2) $W(y|x) \leq \exp\left\{\frac{2|\mathcal{Y}|^2}{M}\right\} W_Q(y|x)$,
3) $\forall P(x)$ on $\mathcal{X}$, $|I(P, W) - I(P, W_Q)| \leq 2|\mathcal{Y}|\left(\frac{|\mathcal{Y}|}{M}\right)^{1/2}$,
4) Any n-length code for $W_Q$with e^{nR_Q}codewords and error probability $\leq \epsilon_n$, can be used as code for $W$with the error probability $\leq \exp\left\{\frac{2|\mathcal{Y}|^2}{M}\right\}\epsilon_n$,
5) $|\mathcal{W}_Q| \leq (M+1)^{|\mathcal{X}||\mathcal{Y}|}$.

It only remains to relate the quantities C, C_Q, R, and R_Q. Define R_Q by $C_Q - R_Q = C - R - \frac{2|\mathcal{Y}|^2}{M}$. Then, a few manipulations give Eq. (10) Furthermore, we observe that $C \leq C_Q + 2|\mathcal{Y}|\left(\frac{|\mathcal{Y}|}{M}\right)^{1/2}$, by Properties 3-3). Hence $R_Q \geq R$ and

the codeword set with e^{nR_Q} codewords can be reduced to obtain a code for the class $\mathcal{W}$ with e^{nR} codewords.

Converse: The converse part can be proved using types of the codeword set, and two well-known inequalities in information theory: Fano's, and data processing. We do not present the proof due to space limitations.

Discussions: A more modern analysis could be carried using the method of types. Actually, the code constructed could have been substituted with a constant composition code from a random type ensemble based on the error exponent given in Eq. (5). Though both techniques would produce the same capacity expression, some differences are worth citing. First the MMI decoder is more robust, in the sense that it does not even need the knowledge of the compound set as required otherwise. On the other hand, the proof in the paper is constructive and provides an explicit method on code construction though not practically implementable.

In general, $C(\mathcal{W})$ is smaller than the infimum of the capacities of any channel in $\mathcal{W}$. This explains partially why instead of designing the code sequence with respect to the worst channel in the class, the second approximation is introduced. However, if we restrict the analysis to symmetric class of channels $\mathcal{W}_s$, then $C(\mathcal{W}_s) = \inf_{W\in\mathcal{W}_s} I(Q^*, W)$, where $Q^*(x)$ is the uniform input distribution. To show this, note that $\max_{Q(x)} \inf_{W\in\mathcal{W}_s} I(Q,W) \leq \inf_{W\in\mathcal{W}_s} \max_{Q(x)} I(Q,W)$ holds as a general fact. As the uniform input distribution is the capacity achieving input distribution for all symmetric channels, for any $W \in \mathcal{W}_s$, we have $I(Q^*, W) = \max_{Q(x)} I(Q,W)$. Thus, the inequality in the opposite direction also holds,

$$C(\mathcal{W}_s) = \max_{Q(x)} \inf_{W\in\mathcal{W}_s} I(Q,W) \geq \inf_{W\in\mathcal{W}_s} I(Q^*,W).$$

Even in this case, there is no clear-cut evidence that a code designed for the worst channel will satisfy the error probability criteria. The decoding regions given for the worst channel may not be adequate for the other channels in the class.

C. Channel Capacity for a Given Decoding Metric

Csiszár, and Narayan [7] studied the performance of an alternative class of decoders. Given a codeword set $\{\mathbf{x}(1), \ldots, \mathbf{x}(M)\} \in \mathcal{X}^n$, a d-decoder assigns a received sequence $\mathbf{y}$, the message $i = 1, \ldots, M$ such that $d(\mathbf{x}(i), \mathbf{y}) < d(\mathbf{x}(j), \mathbf{y})$ for $\forall j \neq i$, otherwise declares an error. The decision function is computed using the additive extension of a single letter metric $d(x, y)$. We limit the survey on finite valued ones.

Definition 1: The d-capacity of a DMC W, denoted by $C_d(W)$, is defined as the supremum of rates R for which, for every $\epsilon > 0$ and sufficiently large block length n, there exist codes with d-decoding such that the code rate $\frac{\log M}{n} > R$, and the average/maximum probability of error is less than ϵ.

The analysis of d-decoders can be carried using constant composition codes based on the usual expurgation argument. Indeed, for a given $\mathbf{x} \in \mathcal{X}^n$, and $\mathbf{y} \in \mathcal{Y}^n$ with joint types $P_{\mathbf{x},\mathbf{y}}$, the metric computes

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} d(x_i, y_i) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} nP_{\mathbf{x},\mathbf{y}}(x,y)d(x,y).$$

Based on this fact, the authors note that for any DMC W, $C_d(W) = C_{\tilde{d}}(W)$ holds if d, and $\tilde{d}$ are such that

$$d(x,y) = c(a(x) + b(y) + \tilde{d}(x,y)) \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

for some finite numbers $a(x)$, $b(y)$, and $c > 0$. This in turn leads to an "equivalence" concept among decoders, useful in proving capacity results for a given d-decoding rule. For instance, for binary channels with $W(0|0) + W(1|1) > 0$, a d-decoder employing minimum distance decoding rule is equivalent to maximum likelihood decoding used in conjunction with constant composition codes. (see Lemma 6.5 of [8]) As a result, $C_d(W) = C(W)$. Another result of the same flavor is established in the paper; the next theorem provides a sufficient condition for a DMC W to have $C_d(W) = C(W)$ when d is an erasure only (e.o.) metric, i.e. $d(x, y) = 0$ iff $W(y|x) > 0$.

Theorem 2.4: Suppose that there exist positive numbers $A(x)$, $x \in \mathcal{X}$, $B(y)$, $y \in \mathcal{Y}$, such that $W(y|x) = A(x)B(y)$ whenever $W(y|x) > 0$. Then $C_{eo}(W) = C(W)$.

As no general expression on $C_d(W)$ is available, the major concern of the paper is the study of lower bounds on $C_d(W)$, and their tightness.

Lower Bounds on $C_d(W)$: The random coding bound on constant composition codes given for α decoders in Eq. (4) can be specialized to d-decoders. We define

$$I_d(\hat{P}, W) \triangleq \min_{\substack{\hat{P}_2:\ \hat{P}\hat{P}_2(y)=\hat{P}W(y) \\ \mathbb{E}_{\hat{P}\hat{P}_2(x,y)}[d(X,Y)]>\Delta}} I(\hat{P}, \hat{P}_2) \tag{11}$$

where $\Delta \triangleq \mathbb{E}_{\hat{P}W(x,y)}[d(X,Y)]$. Then, $E_{r,d}(R, \hat{P}, W) > 0$ iff $R < I_d(\hat{P}, W)$.

Next, we discuss some properties of $I_d(P, W)$.

Properties 4 ($I_d(P,W)$):

1) $I_d(P,W)$is a continuous function of the pair (P,W).
2) $I_d(P,W)$is positive iff
$$\mathbb{E}_{P(x)PW(y)}[d(X,Y)] > \Delta \tag{12}$$
In this case, the inequality constraint in Eq. (11) is replaced by equality.
3) For binary channels, given any d-decoding rule $I_d(P,W) = I(P,W)\mathbf{1}_{\{A\}}$where $A = \{\text{sign}(1 - W(0|0) + W(1|1)) = \text{sign}(d(0,0) + d(1,1) - d(0,1) - d(1,0))\}$. So, $C_d(W) = C(W)$ or 0. This property is due to [13].

Based on Properties 4-1), the following lower bound on the d-capacity of a DMC is obtained.

Proposition 1: For any decoding metric d,

$$C_d(W) \geq \max_P I_d(P,W).$$

Moreover, for any distribution P on $\mathcal{X}$ and $R > 0$, there exist constant composition codes with d-decoder, with codeword type approaching P, rate approaching R, and probability of error going to zero exponentially for every DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$ such that $R < I_d(P,W)$.

This bound can be improved by substituting instead of W the n-extension channel W^n, with super-letter alphabets $\mathcal{X}^n$, $\mathcal{Y}^n$, obtained from the n-independent uses of W. This product space characterization yields the following sequence of lower bounds

$$C_d(W) \geq C_d^{(n)}(W) \triangleq \frac{1}{n} \max_{\tilde{P}} I_d(\tilde{P}, W^n) \tag{13}$$

where $\tilde{P}$ ranges over all distributions on $\mathcal{X}^n$.

Tightness of the Lower Bounds: No converse for the general case on any of the lower bounds is given in the paper. The existence of codes with d-decoding with rates approaching to $R < C_d(W)$, and exponentially vanishing error probability is left as an open problem. Note that in general, we do not expect $I_d(P, W)$ to be a concave function of the input distribution.

Binary input memoryless output channels are mentioned as a special case where $C_d^{(1)}(W)$ is tight. We have already stated in Properties 4-3) the binary channel case, where $C_d(W) = C(W)$ or 0 depending on whether or not the mismatch metric is in "harmony" with the channel behavior. Indeed, Balakirsky et al. [13] extended this result to general output alphabets. The paper provides another special case where $C_d^{(\infty)}(W)$ is tight.

Theorem 2.5: For a DMC W, $C_d(W) = C_d^{(\infty)}(W)$ holds for every e.o. metric.

Subsequent research revealed more aspects on the subject. As opposed to the case of binary channels, Lapidoth et al. [14] showed that $C_d(W) = C(W)$ does not always imply $C_d(W) = C_d^{(1)}(W)$. The author provided a counter-example based on d-decoders for multiple-access channels.

Discussions: We close this section with two remarks in the paper which drew our attention. The first one points out a quantity which might be useful to assess the performance of polar codes in the presence of channel mismatch.

Remark 1: Consider the mismatched decoding metric $d(x, y) = -\log(V(y|x))$, then $I_d(P, W) \geq I(P, W, V)$ where

$$I(P, W, V) \triangleq \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x)W(y|x) \log\left(\frac{V(y|x)}{PV(y)}\right) \tag{14}$$

A closer look at the constraints in Eq. (11) shows that they are equivalent to $I(P, \hat{P}_2, V) \geq I(P, W, V)$. After some simple manipulations, one can derive the inequality $I(P, \hat{P}_2) \geq I(P, \hat{P}_2, V)$, from which the lower bound follows. Note that $I(P, W, V)$ can also be derived using Gallager's error exponent derivation technique. As opposed to the matched case, we see that the method of types yields a stronger bound here.

The second remark is concerned with the achievability of the compound capacity with a d-decoding metric.

Remark 2: When the class of channels $\mathcal{W}$ discussed in the previous section is a convex compact set, the compound capacity $C(\mathcal{W}) = I(P^*, W^*)$ can be achieved by d decoding with respect to $d(x, y) = -\log W^*(y|x)$.

To show this, we first need a result from [8] stated in the next proposition.

Proposition 2: Given a set of channels $\mathcal{W}$, and $\bar{\mathcal{W}}$ its convex closure, if $V = \arg\min_{W \in \bar{\mathcal{W}}} I(W)$, then

$$I(P, \alpha W + (1-\alpha)V) \geq I(V) \quad \forall \alpha \in [0, 1], \forall W \in \mathcal{W}$$

which in turns implies that $I(P, W, V) \geq I(P, V)$.

Using the first remark, we have $I_d(P^*, W) \geq I(P^*, W, W^*)$. Then, by assumption the channel W^* corresponds to the channel which minimizes $I(P^*, W)$ for any $W \in \mathcal{W} = \bar{\mathcal{W}}$. In addition, we know

$$C(\mathcal{W}) = \max_P \min_{W \in \mathcal{W}} I(P, W) = \min_{W \in \mathcal{W}} I(P^*, W)$$

Hence, $I_d(P^*, W) \geq I(P^*, W, W^*) \geq I(P^*, W^*) = C(\mathcal{W})$ by Proposition 2, and the achievability is proved.

III. Research Proposal:

In a previous work [15], we showed that linear codes can indeed achieve the compound capacity of a class of symmetric channels. The first question we pose is: Does the class of polar codes, as a subset of the linear class, achieve the compound capacity of a class of symmetric channels? Upper and lower bounds derived on the compound capacity of polar codes with SCD rule $I_{P,SC}(\mathcal{W})$ in [10] yield a negative result to this question. The underlying assumption throughout the analysis is the availability of the actual communication channel among the compound set to the decoder. Although the practicality of such an assumption is to be defended, the negative result saves the day as the lack of knowledge cannot possibly reverse the end result.

Then, another question is raised: Do polar codes achieve the d-capacity of any discrete memoryless symmetric channel W when the SCD mechanism as described in the previous section employs mismatch decision rules $d_i(y_1^N, \hat{u}_1^{i-1}) = -\log(V_N^{(i)}(y_1^N, \hat{u}_1^{i-1}|u_i))$ based on the channels obtained after applying the polarization transformations to another symmetric B-DMC V? No general answer is contended, but we give an example where the answer turns out to be no as a direct consequence of the fact that polar codes do not achieve the compound capacity of symmetric channels. Let the compound set be defined as $\mathcal{W} = \{\alpha W + (1-\alpha)V, \forall \alpha \in [0, 1]\}$. Since $\mathcal{W}$ is convex (or since the uniform input distribution is the capacity achieving input distribution for any symmetric channel) the min-max (or inf-max) equality holds

$$C(\mathcal{W}) = \max_P \min_{W \in \mathcal{W}} I(P, W) = \min_{W \in \mathcal{W}} \max_P I(P, W)$$

and by Remark 2, if $C(\mathcal{W}) = I(P^*, W^*)$, for every $W \in \mathcal{W}$ we have

$$C_d(W) \geq I_d(P^*, W) \geq I(P^*, W, W^*) \geq I(P^*, W^*) = C(\mathcal{W}).$$

Now, if we assume $W^* = \arg\min_{W \in \mathcal{W}} I(W) = V$, we conclude that since polar codes cannot achieve $C(\mathcal{W})$, $C_d(W)$ is not either achievable with mismatched SCD with d_i.

On the other hand, nothing precludes the option of achieving $I_{P,SC}(\mathcal{W})$ via mismatched SCD with respect to W^*, i.e. the channel with minimum mutual information parameter in the convex set $\mathcal{W}$, as $I(P^*, W, W^*) \geq C(\mathcal{W}) \geq I_{P,SC}(\mathcal{W})$ holds. However, we would also require that the mutual information minimizing property of the channel W^* is preserved under the basic polarization transformations.

Moreover, the mismatch mutual information parameter can be related to the error probability of the SCD scheme of polar codes. Indeed, the error probability that results from a single

use of the channel W to transmit a 0 or 1 when ML decoding with respect to the channel V is used as a decoding metric can be upper bounded as:

$$P_{e,\mathrm{ML}}(W,V) \leq 1 - I(W,V).$$

This can be extended, as done in the previous section, to get an upper bound on the error probability of polar codes of block-length N under mismatched SCD

$$P_{e,\mathrm{ML}}(W,V) \leq \sum_{i \in \mathcal{A}_N} 1 - I(W_N^{(i)}, V_N^{(i)}).$$

All these considerations point out to analyze $I(W,V)$ from channel polarization perspective, as a priori we don't know if the quantity $I(W,V)$ also polarizes or not as its matched counterpart. In analogy to the matched case $Z(W)$, a mismatch Bhattacharyya parameter $Z(W,V)$ can also be defined. Moreover, as $I(W,V)$ is the offspring of a random coding error exponent analysis based on ML decoding, that the SCD has comparable performance to the ML decoder might be useful.

Note that this problem is of much more practical flavor. A solution would give an operational scheme for using polar codes over an unknown channel W or a given compound set of channels $\mathcal{W}$ with both the encoder and decoder ignorant of the true channel information, except its membership in the compound set. There is no trivial answer to this problem. On the encoder side, the designer needs to ensure the good channel indices picked based on the channel V is good for W or $\mathcal{W}$. Hence an ordering on the channel Bhattacharyya parameters is required. On the decoder side, the designer needs to guarantee that the error probability of the mismatched communication scenario for W or $\mathcal{W}$ will be upper bounded by the error probability of the matched scenario the true channel had been V. For an arbitrary channel, we don't expect to be able to use the mismatched polar code designed with respect to this channel to achieve good rates with vanishing error probability over the unknown channel W or the given set of channels $\mathcal{W}$. However, some simulation results shown in Fig. 2 suggest that a positive answer might be expected in certain cases.

To clarify the problem, we give the example of the binary erasure channel. When we have a set of channels specified with a fixed Bhattacharyya parameter, by Properties 1-5), we know that the BEC minimizes the mutual information among the channels in the set. Hence, the BEC plays the role of the channel W^*. In this situation, by Properties 1-6), we also know that the mutual information minimizing property is preserved under the basic polarization transformations applied to W^*. Hence, we can use a polar code (both encoder and decoder) designed for a BEC in all other DMCs with the same Bhattacharyya parameter.

One property that could be considered is channel degradation. It is known from [12] that in the case W is unknown, but V is expected to be a degraded version of W, the information sets satisfy $\mathcal{A}_\mathcal{V} \subset \mathcal{A}_\mathcal{W}$. That result helps the designer to use safely the polar encoder designed for the channel V for communication over W. However, whether the same result holds for the decoder is a question yet to be investigated.

Another idea would be to allow the decoder to run the SCD for each of the channels in the compound set. Then, some procedure such as the generalized likelihood ratio test (GRLT) might be applied to decide which of the estimates to declare. We could first consider the performance when the class of channel is finite, and then try to extend the analysis to the infinite case via some suitable quantization procedure.

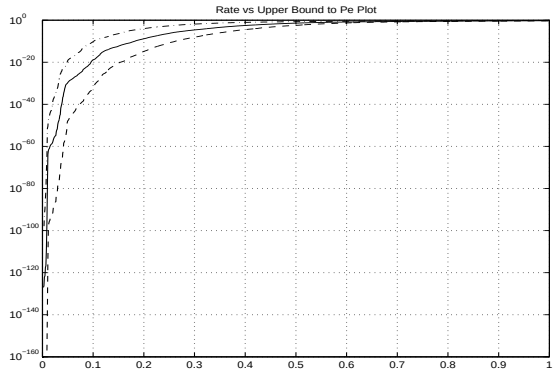

Fig. 2. Simulation results with blocklenght 2^{10} over binary symmetric channels (BSC): Dashed line = true channel and true metric BSC(0.1), Solid line = true channel BSC(0.1) and mismatched metric BSC(0.2), Dashed and dotted line = mismatched channel and mismatched metric BSC(0.2).

In conclusion, we propose to investigate in our research the performance of polar codes for mismatched designs, and identify channel relationships under which the mismatch design goal is met. We believe this study will provide insight in and contribution to the search for universal polarization codes.

REFERENCES

[1] T.M. Cover, and J.A. Thomas, "Elements of information theory," New York: Wiley, 1991.
[2] R. Gallager, "Information Theory and Reliable Communication," New York: Wiley, 1968.
[3] E. Arıkan, "Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theor.*, vol. 55, no. 7, pp. 3051-3073, 2009.
[4] E. Arıkan, and E. Telatar, "On the Rate of Channel Polarization," 2008, http://arxiv.org/abs/0807.3806v2 [cs.IT].
[5] D. Blackwell, and L. Breiman, and A.J. Thomasian, "The capacity of a class of channels," *The Annals of Mathematical Statistics*, vol. 3, no. 4, pp. 1229-1241, 1959.
[6] I. Csiszár, "The method of types [information theory]," *IEEE Trans. Inf. Theor.*, vol. 44, no. 6, pp. 2505 -2523, 1998.
[7] I. Csiszár, and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theor.*, vol. 41, no. 1, pp. 35 -43, 1995.
[8] I. Csiszár, and J. Körner, "Information theory: coding theorems for discrete memoryless systems," New York: Academic, 1981.
[9] C. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theor.*, vol. IT-2, pp. 8-19, 1956.
[10] S.H. Hassani, and S.B. Korada, and R. Urbanke, "The compound capacity of polar codes," *Proc. of the Allerton Conf. on Commun., Control, and Computing*, 2009.
[11] E. Abbe, "Universal source polarization and sparse recovery," *In. Theor. Workshop*, pp. 1-5, 2010.
[12] S.B. Korada, "Polar codes for channel and source coding," PhD Thesis, EPFL, Switzerland, 2009.
[13] V.B. Balakirsky, "Coding theorem for discrete memoryless channels with given decision rule," *Lecture Notes in Comp. Sci. Proc. 1st French-Soviet Workshop on Algebraic Coding*, pp.142-150, 1991.
[14] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inf. Theor.*, vol. 42, no. 5, pp. 1439 -1452, 1996.
[15] M. Alsan, "Channel Polarization and Polar codes," Master Thesis, EPFL, Switzerland, 2010.