

Reliable Robust Adaptive Steganographic Coding Based on Nested Polar Codes

Qiyi Yao, Kai Zeng, Weiming Zhang, Kejiang Chen

Abstract—Steganography is the art of covert communication that pursues the secrecy of concealment. In adaptive steganography, the most commonly used framework of steganography, the sender embeds a “secret message” signal within another “cover” signal with respect to a certain adaptive distortion function that measures the distortion incurred, contributing to the composite “stego” signal that resembles the cover, and the receiver extracts the “secret message” signal from the stego. When the communication channel between the sender and the receiver is noisy, robust steganography is needed, in which robust adaptive steganographic coding plays a central role. The existing robust adaptive steganographic coding methods can only provide very limited robustness, and they fail when the communication channel is bad. To ensure the success of covert communication, we propose a reliable robust adaptive steganographic coding scheme based on nested polar codes that possesses the highest robustness among the existing algorithms while the security performance is also maintained. Theoretically, we show that for the most important *binary embedding*, in the special case where the communication channel is a Binary Symmetric Channel (BSC), the proposed scheme is optimal under the constant distortion profile as the cover length N tends to infinity through powers of two when the *design embedding rate* is large enough. Experimentally, our method is capable of making sure the perfect extraction of the secret message in situations where the embedding rate is large or the communication channel is bad, while the existing algorithms are not applicable in these scenarios.

Index Terms—Robust adaptive steganographic coding, distortion function, binary embedding, noisy channel, nested polar codes, successive cancellation.

I. INTRODUCTION

INFORMATION embedding is about the reliable transmission of information embedded into a host signal [1], and it has a lot of applications including digital watermarking [2]–[6], [16] and steganography [7]–[17] where digital watermarking seeks high robustness and steganography pursues high embedding capacity. In steganography, the embedded signal is called the “secret message”, the host signal is called the “cover”, and the result composite signal is called the “stego” [20], [22]. The sender would send the stego through a communication channel to the receiver who extracts the secret message from the stego. In the most widely used framework, adaptive steganography, an adaptive distortion function is applied to measure the distortion incurred by the embedding

process, and the general aim is to minimize the average total distortion for a given *design embedding rate* [20], [25].

Conventionally, steganography does not strive for robustness like digital watermarking [18] as the communication channel between the sender and the receiver is assumed to be noise-free and lossless [19]–[22]. However, in real-world applications, the communication channel is usually noisy. Taking image steganography as an example, stego images uploaded onto social networks or image-sharing websites may suffer from lossy operations such as image recompression and visible watermarking [18], [23], [24], which would lead to errors in the extracted message at the receiver side, making the covert communication fail. Therefore, to make steganography practical, robust steganography is needed.

Taking real-world application scenarios into concern, many robust steganography methods have been developed to enhance the robustness of steganography [18], [23], [24], [26]–[30], in which robust adaptive steganographic coding plays a central role. Technically, there are two classes of robust adaptive steganographic coding methods that can be distinguished by the receiver behavior. In the first class, the secret message is encoded through error-correcting codes and then embedded into cover media. At the receiver side, the receiver first performs steganographic decoding (message extraction) on the received sequence. Then, the extracted sequence would go through the error correction process to obtain the final reconstructed secret message [23], [24], [27]–[30]. On the contrary, for methods in the second class, the receiver would perform error correction on the received sequence first, and steganographic decoding (message extraction) is performed after error correction [18], [26].

All the existing robust adaptive steganographic coding methods can only provide limited robustness, and they fail when the communication channel is bad. For instance, in *binary embedding* where the cover and stego are both binary sequences (e.g., the least significant bits of the pixel values of 8-bit grayscale images), all the existing methods fail when the communication channel is a Binary Symmetric Channel (BSC) whose transition probability is larger than 0.1. For methods in the first class, they mainly use Syndrome-Trellis coding (STC) [20] to perform the steganographic embedding and extraction. The bit error rate in the secret message after extraction is then susceptible to the constraint height h of the small submatrices used in STC [18] which determines the density of the STC encoding matrix [20]. Since an error that occurs in the stego would lead to several error bits in the secret message subject to the density of the STC encoding matrix, it is difficult to realize perfect extraction by performing error correction on the secret

The authors are with the CAS Key Laboratory of Electro-magnetic Space Information, Anhui Province Key Laboratory of Digital Security, University of Science and Technology of China, Hefei 230026, China (corresponding author: Weiming Zhang, e-mail: qyyao@mail.ustc.edu.cn; zk0128@mail.ustc.edu.cn; zhangwm@ustc.edu.cn; chenkj@ustc.edu.cn).

This work was supported in part by the Natural Science Foundation of China under Grant U2336206, 62102386, 62072421, 62372423, and 62121002.

message. Indeed, this is a common phenomenon which also happens when STC is replaced by other steganographic coding methods as the encoding matrices used in steganography are always not sparse. In comparison to methods in the first class, methods in the second class do not have a similar “error dispersal” problem. However, the research in the second class is still in the primary stage with the problem that the existing methods cannot provide reliable robustness.

To ensure the success of covert communication and make steganography practical, we propose a reliable robust adaptive steganographic coding scheme in the second class that realizes perfect extraction of the secret message even when the communication channel is fairly bad. It is a nested coding scheme [31] inspired by the success of polar codes [32] and Steganographic Polar Coding (SPC) [21] that show good performance in the long-length scenario where we note that the cover is usually very long in steganography. The scheme is flexible as it has a robustness parameter γ that controls the balance between security performance and robustness. We provide analysis on the robustness parameter γ and give a direction on how to determine the parameter based on the analysis, so that users are able to maintain security performance while ensuring that the secret message is transmitted without error. Furthermore, we prove that in the special case where the communication channel between the sender and the receiver is a BSC and the design embedding rate is large enough, the proposed method is optimal under the constant distortion profile as the cover length N tends to infinity through powers of two.

The rest of the paper is organized as follows. In the next section, we briefly introduce the background knowledge of adaptive steganographic coding, polar coding, and SPC. In Section III, we present the proposed robust adaptive steganographic coding scheme. In Section IV, we give an analysis of the proposed algorithm, including the guide on how to select the robustness parameter and the proof of optimality in the special case. Experimental results are presented in Section V. Finally, the paper is concluded in Section VI.

II. PRELIMINARIES

In this section, we first briefly review the fundamentals of adaptive steganographic coding and polar coding. Based on these, we introduce the polar codes-based adaptive steganographic coding method SPC.

A. Adaptive Steganographic Coding

Let $\mathcal{I}$ be the alphabet of elements in the cover and $\{\mathcal{I}_i\}$ be the alphabets of elements in the stego. When $\forall i, |\mathcal{I}_i| = 2$, the embedding process is called *binary* [20], [22]. Similarly, when $\forall i, |\mathcal{I}_i| = 3$, the embedding process is called *ternary* [20], [22], etc. Binary embedding plays the most important role in steganographic coding because the q -ary embedding can be implemented through multi-layered binary embedding [20], [33], [34]. Taking image steganography on 8-bit grayscale images as an example, binary embedding is conducted on the Least Significant Bits (LSBs) of the pixels of the cover image where the model can be simplified to $\forall i, \mathcal{I} = \mathcal{I}_i = \{0, 1\}$. We mainly consider binary embedding in this paper which naturally suits our goal of devising a polar codes-based scheme.

1) *Adaptive Distortion Model*: Suppose that $\forall i, \mathcal{I} = \mathcal{I}_i = \{0, 1\}$. Let $\mathbf{X} = (X_1, \dots, X_N)$ be the cover random vector of length N , $\mathbf{Y} = (Y_1, \dots, Y_N)$ be the stego random vector of length N , $\mathbf{M}$ be the secret message random variable, and $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{I}^N$, $\mathbf{y} = (y_1, \dots, y_N) \in \mathcal{I}^N$, $\mathbf{m} \in \{0, 1\}^m$ be their realizations, respectively. The real secret message is usually compressed or encrypted in steganography [11], [20], so it is commonly assumed that $\forall \mathbf{m} \in \{0, 1\}^m$, $\Pr(\mathbf{m}) = \frac{1}{2^m}$, where $\mathbf{m}$ is the secret message signal to be embedded. Adaptive steganography adopts an adaptive distortion function $D : \mathcal{I}^N \times \mathcal{I}^N \rightarrow [0, \infty)$ to measure the distortion incurred by the embedding process. It is defined by

$$D(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^N \rho_i(x_i, y_i), \quad (1)$$

where each $\rho_i : \mathcal{I} \times \mathcal{I} \rightarrow [0, \infty)$ represents the distortion incurred by modifying x_i to y_i . Usually when $x_i = y_i$, $\rho_i(x_i, y_i) = 0$, so we write $\boldsymbol{\rho} = \{\rho_i(x_i \neq y_i)\}$ to denote the distortion profile in the binary case. Note that the distortion function D is dependent on the cover characteristics as it measures the distortion incurred by modifying the cover.

Let $\{\pi(\mathbf{y}|\mathbf{x}) \triangleq \Pr(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})\}$ be the probability distribution of the modifications. Since the LSBs can be perceived as random noise [35], we usually have $\forall \mathbf{x} \in \mathcal{I}^N$, $\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{2^N}$. Then, the sender could send at most

$$\begin{aligned} H(\pi) &= - \sum_{\mathbf{y} \in \mathcal{I}^N} \pi(\mathbf{y}|\mathbf{x}) \log_2 \pi(\mathbf{y}|\mathbf{x}) \\ &= - \sum_{i=1}^N \sum_{y_i \in \mathcal{I}} \pi_i(y_i|x_i) \log_2 \pi_i(y_i|x_i) \end{aligned} \quad (2)$$

bits information on average where $\mathbf{x}$ can be any fixed cover vector [35]. The average distortion incurred by the embedding process is given by [35]

$$\begin{aligned} E_\pi(D) &= \sum_{\mathbf{y} \in \mathcal{I}^N} \pi(\mathbf{y}|\mathbf{x}) D(\mathbf{x}, \mathbf{y}) \\ &= \sum_{i=1}^N \sum_{y_i \in \mathcal{I}} \pi_i(y_i|x_i) \rho_i(x_i, y_i). \end{aligned} \quad (3)$$

2) *Performance Bounds*: Fig. 1 shows the adaptive steganographic coding model where the distortion between the cover and the stego is measured by an adaptive distortion function D . Adaptive steganographic coding aims to embed the secret message into the cover while minimizing the distortion incurred. This pursuit can be described in two ways: payload-limited sender (PLS) and distortion-limited sender (DLS) [20], where for the PLS problem, the sender aims to minimize $E_\pi(D)$ given a fixed design embedding rate $\frac{H(\pi)}{N} = \frac{m}{N}$; for the DLS problem, the sender aims to maximize $\frac{H(\pi)}{N}$ given a fixed design distortion $E_\pi(D) = d$.

The PLS problem and the DLS problem are dual to each other and they share the same bound. Here we note that the PLS problem is the most commonly considered in practice as the secret message to be sent is usually predetermined in covert communication. On the contrary, the DLS problem plays an important role in theoretical analysis as it is a rate-distortion

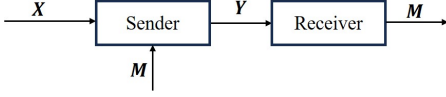


Fig. 1. The adaptive steganographic coding model. The signals $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{M}$ are respectively the cover, the stego, and the secret message.

problem that fits the conventional information-theoretic setting [36], [37]. The optimal probability distribution with respect to the two problems follows a form of Gibbs distribution [25]:

$$\pi(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^N \pi_i(y_i|x_i) \quad (4)$$

where

$$\pi_i(y_i|x_i) = \frac{\exp(-\lambda \rho_i(x_i, y_i))}{\sum_{t_i \in \mathcal{I}} \exp(-\lambda \rho_i(x_i, t_i))}. \quad (5)$$

By substituting (4), (5) into $H(\pi) = m$ for the PLS problem or into $E_\pi(D) = d$ for the DLS problem and solving the equation, the parameter λ is obtained. In practice, the parameter is usually obtained through a binary search due to the monotonicity of $H(\pi)$ and $E_\pi(D)$ with respect to λ [25].

B. Polar Coding

Suppose that $W : \{0, 1\} \rightarrow \mathcal{Y}$ is a Binary-input Discrete Memoryless Channel (B-DMC) with transition probabilities $\{W(y|x) \mid x \in \{0, 1\}, y \in \mathcal{Y}\}$. Define the Bhattacharyya parameter of W as [32]

$$Z(W) \triangleq \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}. \quad (6)$$

Define the symmetric capacity of W as [32]

$$I(W) \triangleq \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \frac{1}{2} W(y|x) \log_2 \frac{W(y|x)}{\frac{1}{2} W(y|0) + \frac{1}{2} W(y|1)}. \quad (7)$$

The symmetric capacity and the channel capacity are the same when W is symmetric. Both $I(W)$ and $Z(W)$ can be used to assess the channel quality of W since $I(W) = 0$ iff $Z(W) = 1$, and $I(W) = 1$ iff $Z(W) = 0$ [32].

For any $1 \leq i \leq j \leq N$, we use the notation $a_i^j \triangleq (a_i, a_{i+1}, \dots, a_j)$ to denote the partial vector. For a set $F \subseteq \{1, \dots, N\}$, $a_F \triangleq \{a_i \mid i \in F\}$. The core of polar coding is the *channel polarization* phenomenon which consists of two phases: the channel combining phase and the channel splitting phase [32]. In the channel combining phase, a synthesized channel $W_N : \{0, 1\}^N \rightarrow \mathcal{Y}^N$ defined by

$$W_N(y_1^N|u_1^N) \triangleq W^N(y_1^N|x_1^N) \triangleq \prod_{i=1}^N W(y_i|x_i) \quad (8)$$

is obtained where u_1^N is the information to be encoded and $x_1^N = u_1^N G_N$ with $G_N \in GL_N(GF(2))$ being the invertible generator matrix of polar codes [32]. In the channel splitting phase, the synthesized channel W_N is split into N coordinate channels $\{W_N^{(i)} : \{0, 1\} \rightarrow \mathcal{Y}^N \times \{0, 1\}^{i-1}, 1 \leq i \leq N\}$ with transition probabilities [32]

$$W_N^{(i)}(y_1^N, u_1^{i-1}|u_i) \triangleq \sum_{u_{i+1}^N \in \{0, 1\}^{N-i}} \frac{1}{2^{N-i}} W_N(y_1^N|u_1^N). \quad (9)$$

The blockwise channel combining and channel splitting can be broken recursively into Arikan's single-step polar transformations [32]. The basic single-step transformation $\{(W, W) \rightarrow (W^-, W^+)\}$ is defined as

$$W^-(y_1, y_2|u_1) \triangleq \frac{1}{2} \sum_{u_2} W(y_1|u_1 \oplus u_2) W(y_2|u_2) \quad (10)$$

$$W^+(y_1, y_2, u_1|u_2) \triangleq \frac{1}{2} W(y_1|u_1 \oplus u_2) W(y_2|u_2),$$

where

$$Z(W^+) = Z(W)^2 \quad (11)$$

$$Z(W^-) \leq 2Z(W) - Z(W)^2 \quad (12)$$

$$Z(W^-) \geq Z(W) \geq Z(W^+). \quad (13)$$

The equality in (12) holds iff W is a Binary Erasure Channel (BEC) and the equality in (13) holds iff $Z(W)$ equals 0 or 1 [32]. This transformation extends to $\{(W_N^{(i)}, W_N^{(i)}) \rightarrow (W_{2N}^{(2i-1)}, W_{2N}^{(2i)})\}$.

By (13), we can see that the single-step polar transformation turns two basic channel copies of W into a better channel W^+ and a worse channel W^- . The channel polarization phenomenon states that as the length N goes to infinity through powers of two, the coordinate channels polarize in the sense that the fraction of good coordinate channels with $\lim_{N \rightarrow \infty} I(W_N^{(i)}) = 1$, $\lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 0$ goes to $I(W)$ and the fraction of bad coordinate channels with $\lim_{N \rightarrow \infty} I(W_N^{(i)}) = 0$, $\lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 1$ goes to $1 - I(W)$ [32]. Moreover, it was shown in [38] that

$$\forall 0 < \beta < \frac{1}{2}, \lim_{N \rightarrow \infty} \frac{\left| \left\{ 1 \leq i \leq N : Z(W_N^{(i)}) < 2^{-N^\beta} \right\} \right|}{N} = I(W). \quad (14)$$

The set of indices corresponding to the bad coordinate channels F is called the *frozen set* and the construction of polar codes is to determine the frozen set F . Then, a polar code can be identified by a tuple (N, K, F^c, u_F) where $\frac{K}{N}$ is the coding rate and u_F is the fixed frozen bits assignment shared by both the encoder and the decoder. The encoder encodes the information bits to be sent by putting them on u_{F^c} , contributing to $u_1^N = u_{F^c} \cup u_F$, and obtaining the codeword $x_1^N = u_1^N G_N$. Polar codes achieve symmetric capacity of any B-DMC W with a successive cancellation (SC) decoder [32] that generates the decisions $\hat{u}_1^N$ sequentially by computing

$$\hat{u}_i \triangleq \begin{cases} u_i, & \text{if } i \in F \\ h_i(y_1^N, \hat{u}_1^{i-1}), & \text{if } i \in F^c, \end{cases} \quad (15)$$

where functions $h_i : \mathcal{Y}^N \times \{0, 1\}^{i-1} \rightarrow \{0, 1\}$ are defined as

$$h_i(y_1^N, \hat{u}_1^{i-1}) \triangleq \begin{cases} 0, & \text{if } \frac{W_N^{(i)}(y_1^N, \hat{u}_1^{i-1}|0)}{W_N^{(i)}(y_1^N, \hat{u}_1^{i-1}|1)} \geq 1 \\ 1, & \text{otherwise.} \end{cases} \quad (16)$$

C. Steganographic Polar Coding

SPC sees the steganographic modifications as the cover elements going through BSCs with transition probabilities $\{\pi_i\}_i$ which are the same as the optimal probabilities of

modifications (5). Then, the fixed channel W in channel coding is replaced by a sequence of independent BSCs $\{W_i\}_i$ with transition probabilities $\{\pi_i\}_i$. The synthesized channel $W_{Nst} : \{0, 1\}^N \rightarrow \{0, 1\}^N$ after channel combining of the non-stationary memoryless sequence of BSCs is defined by

$$W_{Nst}(y_1^N | u_1^N) \triangleq \prod_{i=1}^N W_i(y_i | x_i), \quad (17)$$

where $x_1^N = u_1^N G_N$. The coordinate channels after channel splitting are defined by

$$W_{Nst}^{(i)}(y_1^N, u_1^{i-1} | u_i) \triangleq \sum_{u_{i+1}^N \in \{0,1\}^{N-i}} \frac{1}{2^{N-i}} W_{Nst}(y_1^N | u_1^N). \quad (18)$$

Fortunately, the channel polarization phenomenon also happens for non-stationary memoryless channels [39]. So, by estimating the Bhattacharyya parameters of the coordinate channels after channel splitting and determining the frozen set, the steganographic code construction is accomplished. SPC employs a modified Arikan's heuristic method [21], [40] that perceives the sequence of BSCs as a single BEC with erasure probability $\epsilon = \frac{m}{N}$. Then, F is the frozen set containing m indices corresponding to coordinate channels with the largest Bhattacharyya parameters. Let $u_F = \mathbf{m}$ and use the SC encoder to "decode" $\mathbf{x}$ and then obtain u_{F^c} , thereby, the stego $\mathbf{y} = u_1^N G_N$ where $u_1^N = u_{F^c} \cup u_F$. At the receiver side, the secret message is extracted by calculating $u_1^N = \mathbf{y} G_N^{-1}$ and setting $\mathbf{m} = u_F$, where we note that $G_N^{-1} = G_N$ [32].

III. NESTED POLAR CODES-BASED CODING SCHEME

In this section, we first demonstrate the basic idea behind the proposed scheme and then look into the details.

A. Basic Idea

Suppose that the cover length $N = 2^n \in \mathbb{N}$. Let F_1, F_2 be the frozen sets of codes $\mathcal{C}_1, \mathcal{C}_2$ respectively such that $F_2^c \subset F_1^c, F_1 \subset F_2$. Then, we have a pair of polar codes $(\mathcal{C}_1, \mathcal{C}_2)$ identified by $(N, |F_1^c|, F_1^c, \bar{u}_{F_1})$ and $(N, |F_2^c|, F_2^c, u_{F_2})$, respectively. Our scheme is based on the following claim.

Claim 1: If $\bar{u}_{F_1} = u_{F_2 \cap F_1} = u_{F_1}$, then $\mathcal{C}_2 \subset \mathcal{C}_1$, i.e., the pair $(\mathcal{C}_1, \mathcal{C}_2)$ defines a nested polar code.

Proof: We write

$$\begin{aligned} \mathcal{C}_1 &= \{\bar{u}_1^N G_N\} = \{(\bar{u}_{F_1} \cup \bar{u}_{F_1^c}) G_N\}, \\ \mathcal{C}_2 &= \{u_1^N G_N\} = \{(u_{F_2} \cup u_{F_2^c}) G_N\}, \end{aligned} \quad (19)$$

where $\bar{u}_{F_1}$ and u_{F_2} are fixed sequences with $\bar{u}_{F_1} = u_{F_2 \cap F_1} = u_{F_1}$. So, we have that

$$\begin{aligned} \mathcal{C}_1 &= \{(\bar{u}_{F_1} \cup \bar{u}_{F_1^c}) G_N\} = \{(\bar{u}_{F_1} \cup \bar{u}_{F_2 \setminus F_1} \cup \bar{u}_{F_1^c \setminus F_2}) G_N\} \\ \mathcal{C}_2 &= \{(u_{F_2} \cup u_{F_2^c}) G_N\} = \{(u_{F_2 \setminus F_1} \cup u_{F_1} \cup u_{F_2^c}) G_N\}. \end{aligned} \quad (20)$$

Since $F_1^c \setminus F_2 = \{1, \dots, N\} \setminus F_2 = F_2^c$ as $F_1 \subset F_2$, we write

$$\begin{aligned} \mathcal{C}_1 &= \{(\bar{u}_{F_1} \cup \bar{u}_{F_2 \setminus F_1} \cup \bar{u}_{F_1^c \setminus F_2}) G_N\} \\ &= \{(u_{F_1} \cup \bar{u}_{F_2 \setminus F_1} \cup \bar{u}_{F_2^c}) G_N\} \\ &\supset \{(u_{F_1} \cup u_{F_2 \setminus F_1} \cup u_{F_2^c}) G_N\} = \mathcal{C}_2, \end{aligned} \quad (21)$$

where the third line is because G_N is invertible and $u_{F_2 \setminus F_1}$ is fixed but $\bar{u}_{F_2 \setminus F_1}$ is not. This concludes the proof. $\square$

By the above claim, we can split the set $\{1, \dots, N\}$ into three parts: $F_1, F_2 \setminus F_1$, and F_2^c , among which F_2^c is the set of information bits and F_1 is the frozen set shared by both $\mathcal{C}_1$ and $\mathcal{C}_2$. Since we want to devise a scheme in the second class where the receiver performs secret message extraction after error correction, we need the fine code (or outer code) $\mathcal{C}_1$ to work as an error-correction code and the coarse code (or inner code) $\mathcal{C}_2$ to work as an adaptive steganographic code. As a consequence, the frozen set F_1 of the outer code $\mathcal{C}_1$ should be the indices that are used for error correction where we can just set $\bar{u}_{F_1} = u_{F_1} = (0, \dots, 0)$ for simplicity. For the inner code $\mathcal{C}_2$, its frozen set F_2 should be the indices that are used for secret message embedding. As u_{F_1} is already assigned, we just set $u_{F_2 \setminus F_1} = \mathbf{m}$ to embed the secret message.

B. Code Construction

The main problem to be solved now is how to determine the sets $F_1, F_2 \setminus F_1$, and F_2^c . The basic idea here is to sort the coordinate channels after channel splitting by the channel quality and let F_1 contain the indices corresponding to the worst coordinate channels, let F_2^c contain the indices corresponding to the best coordinate channels, and let $F_2 \setminus F_1$ be in the middle. Fig. 2 shows the schematic of the basic idea of determining the three parts $F_1, F_2 \setminus F_1$, and F_2^c .

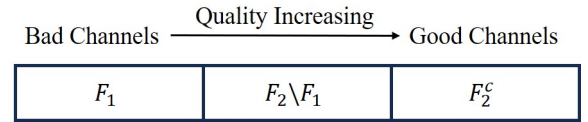


Fig. 2. The schematic of the three parts $F_1, F_2 \setminus F_1$, and F_2^c . By sorting the indices in ascending order of the quality of the coordinate channels, the indices sequentially belong to sets $F_1, F_2 \setminus F_1$, and F_2^c .

Here another problem arises: as we know that polar code construction is channel-dependent, since the steganographic embedding channels $\{W_i\}_i$ are non-stationary and memoryless which are unlikely to be similar to the communication channel W between the sender and the receiver, the order of the quality of the coordinate channels after channel splitting often differs for $\mathcal{C}_1$ and $\mathcal{C}_2$. This means that an index i may correspond to a bad channel for $\mathcal{C}_1$ but a good channel for $\mathcal{C}_2$, i.e., $F_1 \not\subset F_2$. So, we cannot simply apply the standard polar code construction by sorting the Bhattacharyya parameters or the symmetric capacities of the coordinate channels after channel splitting separately for $\mathcal{C}_1$ and $\mathcal{C}_2$ to attain our goal.

We do the code construction through a modified approach. Let $\gamma \triangleq \frac{|F_1|}{N}$ be a robust parameter. We can see that γ controls the fraction of the error-correction bits u_{F_1} , which means that the robustness of the method would increase as γ increases. Suppose that the secret message $\mathbf{m}$, the cover $\mathbf{x}$, the communication channel W , and the robust parameter γ are given. First, we estimate the channel quality of the coordinate channels $\{W_{Nst}^{(i)}\}$ based on (18) and choose $N - \gamma N - m$ indices corresponding to the best channels to form the set of information bits F_2^c . This step is because the inner code $\mathcal{C}_2$ should be a good adaptive steganographic

code, and we want lower distortion incurred. Consequently, the information bits $u_{F_2^c}$ should be determined based on the embedding channels $\{W_i\}_i$. Then, we estimate the channel quality of the coordinate channels $\{W_N^{(i)}\}$ based on (9) and choose γN indices corresponding to the worst channels from set F_2 to form F_1 where $\forall j \in F_1, j \notin F_2^c$. This step assures that the requirements of nested polar codes are met while keeping the outer code $\mathcal{C}_1$ a good error-correcting code. Finally, $F_2 \setminus F_1 = \{1, \dots, N\} \setminus F_2^c \setminus F_1$ will be the set of indices for the secret message $\mathbf{m}$.

In practice, we would see the non-stationary embedding channels $\{W_i\}_i$ as a single BSC whose transition probability $p = H^{-1}(\frac{m}{N} + \gamma) \leq \frac{1}{2}$ with $H(\cdot)$ being the binary entropy function, i.e., $\forall 1 \leq i \leq N, W_i$ is a BSC(p). This is because the probability distributions $\{W_i = \pi_i\}_i$ are dependent on the cover characteristics which are not synchronized to the receiver as the receiver can only see the stego. Except for that, the communication channel W is usually unknown to the sender and the receiver in real-world applications. In this situation, we would assume that the communication channel W is a BSC(q) where $q = H^{-1}(\gamma) \leq \frac{1}{2}$ for convenience. Our code construction method is optimal under the circumstances given above. To justify this, we first give the following definition of optimal channel pair for nested polar code construction.

Definition 1: Given two B-DMCs $\bar{W} : \{0, 1\} \rightarrow \bar{\mathcal{Y}}$ and $\tilde{W} : \{0, 1\} \rightarrow \tilde{\mathcal{Y}}$, if for any $1 \leq i \leq N$,

$$\lim_{N=2^n, n \rightarrow \infty} Z(\bar{W}_N^{(i)}) = 1 \Rightarrow \lim_{N=2^n, n \rightarrow \infty} Z(\tilde{W}_N^{(i)}) = 1, \quad (22)$$

we say that the pair $(\bar{W}, \tilde{W})$ is an *asymptotically optimal* channel pair for nested polar code construction.

Since the Bhattacharyya parameter of a bad coordinate channel would go to 1 as N tends to infinity through powers of two, the above definition essentially assures that $\bar{F} \subset \tilde{F}$ in an asymptotically optimal nested polar code construction $(\bar{\mathcal{C}}, \tilde{\mathcal{C}})$ even when $\bar{\mathcal{C}}$ and $\tilde{\mathcal{C}}$ are constructed separately. In the remainder of the paper, we would simply say a pair $(\bar{W}, \tilde{W})$ is *optimal* instead of asymptotically optimal for convenience.

The following proposition gives a sufficient condition for a channel pair to be optimal.

Proposition 1: A pair of B-DMCs $(\bar{W} : \{0, 1\} \rightarrow \bar{\mathcal{Y}}, \tilde{W} : \{0, 1\} \rightarrow \tilde{\mathcal{Y}})$ is asymptotically optimal for nested polar code construction if $\tilde{W}$ is *degraded* with respect to $\bar{W}$, i.e., if there exists a DMC $W' : \bar{\mathcal{Y}} \rightarrow \tilde{\mathcal{Y}}$ such that

$$\forall x \in \{0, 1\}, \tilde{y} \in \tilde{\mathcal{Y}}, \quad \tilde{W}(\tilde{y}|x) = \sum_{\bar{y} \in \bar{\mathcal{Y}}} \bar{W}(\bar{y}|x) W'(\tilde{y}|\bar{y}). \quad (23)$$

This proposition is proved in the Appendix. We make a remark here that a B-DMC $\tilde{W}$ is degraded with respect to B-DMC $\bar{W}$ basically means that the degraded channel $\tilde{W}$ is a concatenated channel whose first part is $\bar{W}$ [41]. Now, we have the following corollary about the optimality of our code construction method under the given circumstances.

Corollary 1: Suppose that $\frac{m}{N} > 0, \gamma > 0$ and $\frac{m}{N} + \gamma \leq 1$. Assume that $\forall 1 \leq i \leq N, W_i$ is a BSC(p) where $p = H^{-1}(\frac{m}{N} + \gamma) \leq \frac{1}{2}$ and the communication channel

W is a BSC(q) where $q = H^{-1}(\gamma) \leq \frac{1}{2}$. Then, the pair $(W, W_i) = (\text{BSC}(q), \text{BSC}(p))$ is optimal for nested polar code construction and the code $(\mathcal{C}_1, \mathcal{C}_2)$ obtained using our code construction method is optimal.

The corollary is proved in the Appendix. Indeed, the steganographic embedding channel and the communication channel are two BSCs with different transition probabilities, and the optimal construction comes from the fact that if $W_N^{(i)}$ is a bad coordinate channel, $W_{N_{st}}^{(i)}$ is worse, i.e., $F_1 \subset F_2$ always holds. In the experiments of this paper, we estimate the quality of the coordinate channels using the modified Arikan's heuristic method following SPC [21]. Therefore, the embedding channel is considered a BEC(ϵ_2) where $\epsilon_2 = \frac{m}{N} + \gamma$ and the communication channel is considered a BEC(ϵ_1) where $\epsilon_1 = \gamma$. Fortunately, the optimality also holds for BECs.

Proposition 2: Suppose that $\frac{m}{N} > 0, \gamma > 0$ and $\frac{m}{N} + \gamma \leq 1$. Assume that $\forall 1 \leq i \leq N, W_i$ is a BEC(ϵ_2) where $\epsilon_2 = \frac{m}{N} + \gamma$ and the communication channel W is a BEC(ϵ_1) where $\epsilon_1 = \gamma$. Then, for any fixed $N = 2^n, n \in \mathbb{N}$, the code $(\mathcal{C}_1, \mathcal{C}_2)$ obtained using our code construction method is optimal.

This proposition is proved in the Appendix. Here we remark that the optimality in Proposition 2 is not just asymptotic as it holds for any $N = 2^n, n \in \mathbb{N}$.

The main reason that the above problem of code construction for the nested polar codes-based scheme arise lies in the “nested” part by seeing the scheme as a robust extension of the original SPC. Recall that the conventional adaptive steganographic coding only consider the embedding channels as the communication channel is noise-free and lossless, and thereby, the code construction of SPC only splits the indices into two sets where the secret message is concealed in the frozen bits. By extending SPC into a nested polar codes-based scheme, the embedding channels and the communication channel are both taken into consideration, which would lead to the problem of how to divide the indices into three parts.

C. The Detailed Method

At the sender side, given cover vector $\mathbf{x}$ of length $N = 2^n, n \in \mathbb{N}$, secret message $\mathbf{m}$ of length m , distortion profile ρ of length N , and robust parameter γ , the first step is to estimate the communication channel W . If W is known, we can do the code construction using the method we proposed in the last subsection. Otherwise, we would just assume that W is a BSC(q) where $q = H^{-1}(\gamma)$. So, we will see the communication channel W as a known input of the sender in the following content. After code construction, the sender will assign $u_{F_2 \setminus F_1} = \mathbf{m}$ and fix the frozen bits u_{F_1} . Usually, we can just let $u_{F_1} = (0, \dots, 0)$ since the choice of the frozen bits does not affect the error-correcting performance [32].

In the next step, the sender will calculate the optimal probability distributions of the modifications (5). Note that in the code construction step, we only need $\frac{H(\pi)}{N} + \gamma = \frac{m}{N} + \gamma$ where the modification probability distributions $\{\pi_i\}_i$ are not required. When doing the actual encoding, to minimize the

distortion incurred, we need the SC encoder to operate based on the following rule:

$$\hat{u}_i \triangleq \begin{cases} u_i, & \text{if } i \in F_2 \\ \hat{h}_i(y_1^N, \hat{u}_1^{i-1}), & \text{if } i \in F_2^c, \end{cases} \quad (24)$$

where each $\hat{h}_i : \{0, 1\}^N \times \{0, 1\}^{i-1} \rightarrow \{0, 1\}$ is defined as

$$\hat{h}_i(y_1^N, \hat{u}_1^{i-1}) \triangleq \begin{cases} 0, & \text{if } \frac{W_{N_{st}}^{(i)}(y_1^N, \hat{u}_1^{i-1}|0)}{W_{N_{st}}^{(i)}(y_1^N, \hat{u}_1^{i-1}|1)} \geq 1 \\ 1, & \text{otherwise,} \end{cases} \quad (25)$$

with $\forall 1 \leq i \leq N, W_i = \pi_i$ and $W_{N_{st}}^{(i)}$ following (17) and (18). Now, as $\hat{u}_{F_2^c}$ is determined through the SC encoder and $\hat{u}_{F_2} = u_{F_2}$, the whole sequence $\hat{u}_1^N$ is known. By performing polar encoding $\mathbf{y} = \hat{u}_1^N G_N$, the stego $\mathbf{y}$ is obtained. The procedures of the encoding method are presented in **Algorithm 1**.

Algorithm 1 Nested Polar Codes-Based Robust Adaptive Steganographic Encoding Algorithm

Input: cover vector $\mathbf{x}$ of length $N = 2^n, n \in \mathbb{N}$, secret message $\mathbf{m}$ of length m , distortion profile ρ of length N , communication channel W , robust parameter γ .

Output: stego vector $\mathbf{y}$ of length N .

- 1: Do code construction through the method proposed in Section III-B and obtain sets $F_1, F_2 \setminus F_1, F_2^c$;
 - 2: Set $u_{F_2 \setminus F_1} = \mathbf{m}, u_{F_1} = (0, \dots, 0)$;
 - 3: Calculate the optimal probability distributions of the modifications $\{\pi_i\}_i$ that follow (5);
 - 4: Use the SC encoder with rule (24), (25) to determine $\hat{u}_1^N$;
 - 5: Perform polar encoding $\mathbf{y} = \hat{u}_1^N G_N$;
 - 6: **return** $\mathbf{y}$.
-

At the receiver side, the input consists of $\hat{\mathbf{y}}$ which is the result of $\mathbf{y}$ going through the noisy channel W , embedding rate $\frac{m}{N}$, communication channel W , and robust parameter γ where we note that the rule of selecting γ is shared by the sender and the receiver. The first step is to identify sets $F_1, F_2 \setminus F_1, F_2^c$ through the same procedures employed by the sender. Then, the receiver assigns the frozen bits $u_{F_1} = (0, \dots, 0)$ and uses the SC decoder that works based on (15), (16) to decode $\hat{\mathbf{y}}$ and obtain $\hat{u}_1^N = \hat{u}_{F_1^c} \cup \hat{u}_{F_1}$ where $\hat{u}_{F_1} = u_{F_1} = (0, \dots, 0)$. Then, the secret message $\mathbf{m} = \hat{u}_{F_2 \setminus F_1}$ can be extracted. **Algorithm 2** presents the procedures of the decoding method.

Algorithm 2 Nested Polar Codes-Based Robust Adaptive Steganographic Decoding Algorithm

Input: received vector $\hat{\mathbf{y}}$ of length $N = 2^n, n \in \mathbb{N}$, embedding rate $\frac{m}{N}$, communication channel W , robust parameter γ .

Output: secret message $\mathbf{m}$ of length m .

- 1: Determine sets $F_1, F_2 \setminus F_1, F_2^c$ through the code construction method proposed in Section III-B;
 - 2: Set $u_{F_1} = (0, \dots, 0)$;
 - 3: Use the SC decoder with rule (15), (16) to determine $\hat{u}_1^N$;
 - 4: **return** $\mathbf{m} = \hat{u}_{F_2 \setminus F_1}$.
-

Finally, we consider the case where the cover length N is not a power of two. In this situation, we pad the cover

with a sequence of wet elements with modification probability 0 following [21] so that the padded cover is of length 2^n where $2^n > N$ and $2^{n-1} < N$. As this step is part of the algorithm, the padding sequence of wet elements is known by both the sender and the receiver. Then, by a shared permutation τ , the padding bits spread into the cover elements, which increases the embedding success rate because the bits that cannot be modified no longer stick together. At the sender side, the input cover vector of the proposed algorithm would become the padded and permuted vector, and the output stego vector would go through the inverse permutation τ^{-1} and padding bits removal. At the receiver side, the padding bits are appended to the received sequence $\hat{\mathbf{y}}$. Then, the permuted vector $\tau(\hat{\mathbf{y}})$ becomes the input of the decoding algorithm.

IV. THEORETICAL ANALYSIS

In this section, we provide theoretical analysis on the robust parameter γ and prove the optimality of the proposed scheme under the constant distortion profile as N tends to infinity through powers of two in the special case where the communication channel between the sender and the receiver is a BSC and the design embedding rate $\frac{m}{N}$ is large enough.

A. Analysis on Robust Parameter γ

We start by formalizing the communication problem model of robust adaptive steganographic coding. The sender embeds secret message signal $\mathbf{M}$ into cover signal $\mathbf{X}$ and contributes to the stego signal $\mathbf{Y}$ where the secret message and the cover are uniformly distributed in $\{0, 1\}^m$ and $\{0, 1\}^N$, respectively. The stego $\mathbf{Y}$ is then transmitted through a noisy B-DMC W and ends up in $\hat{\mathbf{Y}}$. The receiver extracts the secret message from $\hat{\mathbf{Y}}$ and obtains $\hat{\mathbf{M}}$. The schematic of the model is shown in Fig. 3 where the distortion between the cover and the stego is measured by an adaptive distortion function $D(\cdot, \cdot)$.

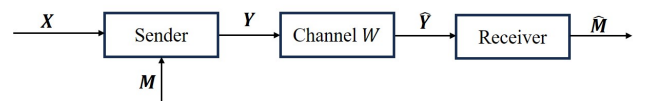


Fig. 3. The schematic of the communication problem model of robust adaptive steganographic coding where the distortion between the cover and the stego is measured by an adaptive distortion function $D(\cdot, \cdot)$.

This model is very similar to the information embedding model in [1] and the only difference lies in the way of measuring the distortion incurred by the embedding process: the classic model [1] considers a single-letter distortion measure while our model adopts an adaptive distortion function.

The aim of robust adaptive steganographic coding has two aspects: the first aspect is to make sure that $\mathbf{m} = \hat{\mathbf{m}}$ and the second aspect is to optimize the PLS/DLS problem. In our scheme, by employing a large robust parameter γ , one can guarantee the first aspect in most cases. However, the higher γ is, the more average distortion would be incurred by the embedding process for fixed embedding rate $\frac{m}{N}$ because as the number of information bits $|F_2^c|$ gets smaller along with the increasing γ , $|C_2|$ gets smaller as well. Therefore, by assuming a fixed robust adaptive steganographic coding

method, optimizing the PLS problem is equal to finding the smallest γ such that perfect extraction $\mathbf{m} = \hat{\mathbf{m}}$ is confirmed.

Intuitively, the choice of γ should depend on the channel W , and this is indeed true.

Proposition 3: Given B-DMC W , to attain perfect extraction of the secret message, $\gamma \geq 1 - I(W)$.

Proof: In the proposed robust adaptive steganographic decoding algorithm, the receiver performs channel decoding on the received sequence $\hat{\mathbf{y}}$ and then conducts adaptive steganographic extraction to obtain the secret message $\hat{\mathbf{m}}$. Consequently, perfect extraction is attained iff $\hat{\mathbf{m}} = \hat{u}_{F_2 \setminus F_1} = u_{F_2 \setminus F_1}$ after channel decoding. Let $\mathcal{E}_i$ denotes the event that $\hat{u}_i \neq u_i$. Then, following (15), (16), we have that

$$\forall i \in F_1^c, \quad \mathcal{E}_i = \left\{ (u_1^N, \hat{y}_1^N) \in \{0, 1\}^N \times \mathcal{Y}^N : \frac{W_N^{(i)}(\hat{y}_1^N, u_1^{i-1} | u_i)}{W_N^{(i)}(\hat{y}_1^N, u_1^{i-1} | u_i \oplus 1)} < 1 \right\}. \quad (26)$$

Let $\mathcal{E}$ denotes the event that $\hat{\mathbf{m}} \neq \mathbf{m}$. Then, we have that

$$\begin{aligned} \forall j \in F_2 \setminus F_1, \quad \mathcal{E}_j &\subset \mathcal{E} \subset \bigcup_{i \in F_2 \setminus F_1} \mathcal{E}_i, \\ P(\mathcal{E}_j) &\leq P(\mathcal{E}) \leq \sum_{i \in F_2 \setminus F_1} P(\mathcal{E}_i). \end{aligned} \quad (27)$$

An upper bound on $P(\mathcal{E}_i)$ is given by

$$\begin{aligned} P(\mathcal{E}_i) &= \sum_{u_1^N, \hat{y}_1^N} \frac{1}{2^N} W_N(\hat{y}_1^N | u_1^N) 1_{\mathcal{E}_i}(u_1^N, \hat{y}_1^N) \\ &= \sum_{u_i} \left(\frac{1}{2} \sum_{u_1^{i-1}, \hat{y}_1^N} \sum_{u_{i+1}^N} \frac{1}{2^{N-1}} W_N(\hat{y}_1^N | u_1^N) 1_{\mathcal{E}_i}(u_1^N, \hat{y}_1^N) \right) \\ &\leq \sum_{u_i} \frac{1}{2} \sum_{u_1^{i-1}, \hat{y}_1^N} \sqrt{\frac{W_N^{(i)}(\hat{y}_1^N, u_1^{i-1} | u_i \oplus 1)}{W_N^{(i)}(\hat{y}_1^N, u_1^{i-1} | u_i)}} \sum_{u_{i+1}^N} \frac{1}{2^{N-1}} W_N \\ &= \sum_{u_i} \frac{1}{2} Z(W_N^{(i)}) = Z(W_N^{(i)}), \end{aligned} \quad (28)$$

where the third line follows from (26) and the forth line follows from (9). Indeed, this is the same as the fact that the Bhattacharyya parameter upper bounds the probability of maximum-likelihood (ML) decision error [32]. By substituting (28) into (27), we have that

$$P(\mathcal{E}) \leq \sum_{i \in F_2 \setminus F_1} Z(W_N^{(i)}). \quad (29)$$

So, when $\gamma = \frac{|F_1|}{N} \geq 1 - I(W)$ and the optimal code construction is assumed, following (14), we have that

$$\lim_{N \rightarrow \infty} P(\mathcal{E}) < \lim_{N \rightarrow \infty} |F_2 \setminus F_1| \cdot 2^{-N^\beta} < \lim_{N \rightarrow \infty} N 2^{-N^\beta} = 0, \quad (30)$$

where we note that $\frac{|F_1^c|}{N} \leq I(W)$ and $\forall i \in F_1^c, Z(W_N^{(i)}) < 2^{-N^\beta}$ as N tends to infinity through powers of two. As a result, when $\gamma \geq 1 - I(W)$, $P(\mathcal{E}) \rightarrow 0$ in N .

The last thing to verify is that when $\gamma < 1 - I(W)$, $P(\mathcal{E}) > 0$. To prove this, we first note that $\frac{|F_2^c| + |F_2 \setminus F_1|}{N} > I(W)$, which means that there are indices corresponding to bad channels that are considered good. When optimal code construction is assumed, there exists $j \in \mathbb{N}$ such that $\lim_{N \rightarrow \infty} Z(W_N^{(j)}) = 1$, $\lim_{N \rightarrow \infty} I(W_N^{(j)}) = 0$, and $j \in F_2 \setminus F_1$. Then, by definition (7), we should have that

$$\forall \hat{y}_1^N \in \mathcal{Y}^N, u_1^{j-1} \in \{0, 1\}^{j-1}, W_N^{(j)}(\hat{y}_1^N, u_1^N | 0) = W_N^{(j)}(\hat{y}_1^N, u_1^N | 1). \quad (31)$$

By substituting (31) into (15) and (16), we know that $\hat{u}_j = 0$ after SC decoding no matter what the original value u_j takes. This suggests that

$$\lim_{N \rightarrow \infty} P(\mathcal{E}_j) = \frac{1}{2} \leq \lim_{N \rightarrow \infty} P(\mathcal{E}), \quad (32)$$

where we note that the secret message bits are random. Finally, when the code construction is not optimal, there may exist indices $\{j \in F_2 \setminus F_1 : \lim_{N \rightarrow \infty} Z(W_N^{(j)}) = 1\}$ even when $\gamma \geq 1 - I(W)$, which means that the error probability would only get larger than the situation when optimal code construction is assumed. This concludes the proof. $\square$

In practice, since a lot of approximations are made in our scheme, we may need larger γ to reach the goal of perfect extraction, but it should be at least larger than $1 - I(W)$ for an acceptable extraction precision.

B. Optimality in the Special Case

Our aim is to prove the following result in this subsection.

Theorem 1: Suppose that $\frac{m}{N} > 0$ is fixed and the communication channel W is a BSC(q) such that $H(q) + \frac{m}{N} \leq 1$. Then, under the constant distortion profile $\rho \triangleq (1, \dots, 1)$, if $\frac{m}{N} \geq H(1 - 2^{-H(q)}) - H(q)$, by letting $\gamma = H(q)$, the proposed method is optimal in the sense that it saturates the classic rate-distortion bound for information hiding as N tends to infinity through powers of two, i.e., the lowest possible average distortion is attained as $N = 2^n, n \rightarrow \infty$.

Proof: Since ρ_i remains the same for each $1 \leq i \leq N$, each $W_i = \pi_i$ following (5) is a BSC(p) where $p = H^{-1}(\frac{m}{N} + \gamma) = H^{-1}(\frac{m}{N} + H(q))$ in our scheme. Then, by Corollary 1, we have the optimal nested polar code construction $(\mathcal{C}_1, \mathcal{C}_2)$. As W is a BSC(q), $I(W) = C(W) = 1 - H(q)$ where $C(W)$ denotes the channel capacity of W . Similar to the proof of Proposition 3, when $\gamma = 1 - I(W) = H(q)$, by (30), we know that the error probability tends to zero as N tends to infinity through powers of two. So, perfect extraction is assured when the fixed embedding rate is $\frac{m}{N}$ and the robust parameter $\gamma = H(q)$ as N tends to infinity through powers of two.

The next step is to calculate the average distortion. As we know that steganographic polar codes are optimal under the constant distortion profile given the exact Bhattacharyya parameters of the coordinate channels after channel splitting [35, Cor. 1, 2], the average distortion incurred by the embedding process is indeed $E_\pi(D)$ following (3) where π with the form (4), (5) corresponds to the solution of the equation

$$\frac{H(\pi)}{N} = \frac{m}{N} + \gamma. \quad (33)$$

We write

$$\begin{aligned} \frac{m}{N} + \gamma &= \frac{H(\pi)}{N} \\ \frac{m}{N} + \gamma &= \frac{-\sum_{i=1}^N \sum_{y_i \in \mathcal{I}} \pi_i(y_i|x_i) \log_2 \pi_i(y_i|x_i)}{N} \\ \frac{m}{N} + \gamma &= -\sum_{y_i \in \mathcal{I}} \pi_i(y_i|x_i) \log_2 \pi_i(y_i|x_i) \\ \frac{m}{N} + \gamma &= H(\pi_i(\bar{x}_i|x_i)) = H(\pi_i(x_i|x_i)) = H(p), \end{aligned} \quad (34)$$

where the second line follows (2) and the third line comes from the fact that all $\{\pi_i\}$ are the same. As a result, $p = \pi_i(\bar{x}_i|x_i) = H^{-1}(\frac{m}{N} + \gamma)$ where we note that $p \leq \frac{1}{2}$. Then,

$$\begin{aligned} E_\pi(D) &= \sum_{i=1}^N \sum_{y_i \in \mathcal{I}} \pi_i(y_i|x_i) \rho_i(x_i, y_i) \\ &= N p \rho_i(x_i \neq y_i) = N p \end{aligned} \quad (35)$$

where the first line comes from (3) and the second line follows from the fact that $p = \pi_i(\bar{x}_i|x_i)$.

We have proved that pair $(\frac{m}{N}, \frac{E_\pi(D)}{N}) = (\frac{m}{N}, p)$ is achievable in our scheme under the condition of perfect extraction of the secret message. The classic rate-distortion bound of information hiding [1] states that

$$R_D(d) = H(d) - H(q), \quad \text{if } d_q \leq d \leq 0.5$$

in the binary symmetric case where $d_q = 1 - 2^{-H(q)}$. Let $R_D(d) = \frac{m}{N}$, we have that

$$\begin{aligned} \frac{m}{N} &= H(d) - H(q) \\ d &= H^{-1}\left(\frac{m}{N} + H(q)\right) = p. \end{aligned}$$

Since $\frac{m}{N} \geq H(1 - 2^{-H(q)}) - H(q)$, we write

$$\begin{aligned} d = p &= H^{-1}\left(\frac{m}{N} + H(q)\right) \\ &\geq H^{-1}\left(H(1 - 2^{-H(q)}) - H(q) + H(q)\right) \\ &= 1 - 2^{-H(q)}, \end{aligned}$$

where we note that $p \leq \frac{1}{2}$. As a consequence, the optimal pair $(\frac{m}{N}, d) = (\frac{m}{N}, p)$ is achieved by our scheme. $\square$

Theorem 1 shows the optimality of the proposed method in the special binary symmetric case under the constant distortion profile when the design embedding rate is large enough. Unfortunately, the optimality does not hold when the embedding rate is small. An approach to optimality in the low-rate scenario is through “time-sharing”. Specifically, suppose that we want to embed m bits message $\mathbf{m}$ into a length- N cover. Let $R = \frac{m}{N}$ and $R' = H(1 - 2^{-H(q)}) - H(q)$ where $R < R'$. Then, we have that $N' = \frac{m}{R'} < N$. Following Theorem 1, by embedding the secret message $\mathbf{m}$ into the partial cover of length N' , leaving the other elements of the cover unchanged, as N, N' tends to infinity, the pair $(R, H^{-1}(\frac{m}{N'} + H(q)) \frac{N'}{N})$ is achieved with perfect extraction assured. We write

$$H^{-1}\left(\frac{m}{N'} + H(q)\right) \frac{N'}{N} = \left(1 - 2^{-H(q)}\right) \frac{N'}{N} < 1 - 2^{-H(q)}. \quad (36)$$

Then, following the classic bound [1], we have that

$$\begin{aligned} R_D\left(H^{-1}\left(\frac{m}{N'} + H(q)\right) \frac{N'}{N}\right) &= \frac{H(1 - 2^{-H(q)}) - H(q)}{1 - 2^{-H(q)}} H^{-1}\left(\frac{m}{N'} + H(q)\right) \frac{N'}{N} \\ &= R' \frac{N'}{N} = \frac{m}{N} = R, \end{aligned} \quad (37)$$

which confirms the optimality.

V. EXPERIMENTS

In this section, we evaluate the proposed method through simulation experiments conducted on random binary sequences. In the code construction step, the quality of the coordinate channels is estimated using the modified Arikan's heuristic method following SPC [21] throughout this section.

A. Experimental Setup

The communication channel W is assumed to be a BSC(q) following the literature of robust adaptive steganographic coding [18]. Our method is denoted as RSPC(q') throughout this section where $\gamma = H(q')$. The covers and secret messages in the experiments are randomly generated binary sequences of lengths $N = 2^{12}$ and m , respectively, where the embedding rate is given by $\text{payload} = \frac{m}{N}$. Two distortion profiles are considered: the constant distortion profile $\rho = (1, \dots, 1)$ and the linear distortion profile $\rho = (1, 2, \dots, N)$.

All the experimental results presented in this subsection are averaged over 100 trials. To evaluate the security performance, we calculate the average embedding distortion $\bar{D}$ between cover $\mathbf{x}$ and stego $\mathbf{y}$ where the larger this distortion is, the worse the security performance is for fixed payload. To evaluate the robustness, we calculate the average bit-error rate (BER) $\bar{R}_{\text{error}}$ of the secret message. The extraction success rate $\bar{R}_{\text{success}}$ is also applied for evaluation which signifies the proportion of successful extractions.

Two main robust adaptive steganographic coding schemes are used for comparison: the widely used STC combined with Reed-Solomon (RS) codes in the first class and Dual-STC (DSTC) [18] in the second class. For STC+RS, we set the STC submatrix height to 10 and the RS code is denoted as RS(n^*, k^*). Dual-STC [18] is the state-of-the-art binary robust steganographic coding scheme, and in this paper, we use an improved version [26], denoted as DSTC(H_m, H_s), where H_m and H_s are heights of submatrices used for message embedding and error correction, respectively.

B. Comparison of Robustness

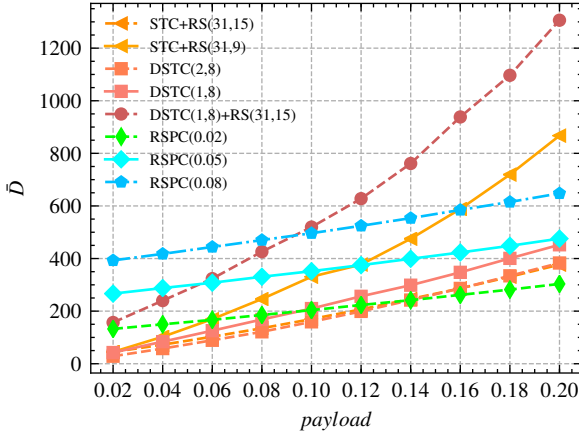
The channel we use for the robustness experiments is BSC(0.02). We calculated $\bar{R}_{\text{error}}$ and $\bar{R}_{\text{success}}$ of different algorithms in this scenario and the results are shown in Table I and Table II for the constant distortion profile and the linear distortion profile, respectively. It can be seen that the traditional STC+RS and DSTC are not qualified in this scenario, while the proposed method basically achieves successful extraction under various payloads by adjusting the

TABLE I
COMPARISON OF ROBUSTNESS UNDER CONSTANT DISTORTION PROFILE AND VARIOUS PAYLOADS (EMBEDDING RATES)

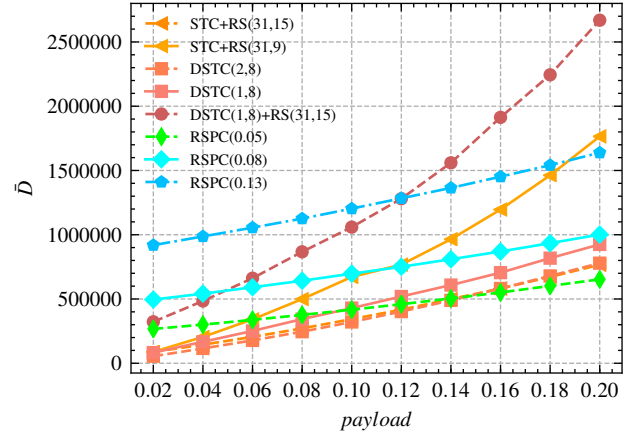
Algorithms	$\bar{R}_{error} (\times 10^{-3} \downarrow) \setminus \bar{R}_{success} (\% \uparrow)$									
	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2
STC+RS(31,15)	198.4\0	229.4\0	244.0\0	269.4\0	300.0\0	334.0\0	373.2\0	411.1\0	432.1\0	464.0\0
STC+RS(31,9)	456.6\0	404.6\0	327.6\0	263.7\0	204.4\0	174.6\0	131.3\0	104.6\0	69.6\0	55.0\1
DSTC(2,8)	466.3\0	415.6\0	351.0\0	297.5\0	230.1\0	172.0\0	115.9\0	81.8\0	48.6\0	31.2\2
DSTC(1,8)	394.5\0	292.2\0	241.2\0	186.8\0	148.0\0	114.4\0	74.6\0	48.1\0	33.5\0	22.2\0
DSTC(1,8)+RS(31,15)	188.9\0	97.6\2	24.5\42	5.1\76	0.4\97	0.2\99	0\100	0\100	0.1\99	0\100
RSPC(0.02)	249.5\0	262.1\0	269.8\0	302.7\0	296.6\0	307.2\0	322.8\0	324.3\0	341.5\0	335.2\0
RSPC(0.05)	4.3\81	3.9\80	3.3\76	3.2\80	3.8\76	4.6\74	7.8\73	10.2\74	5.6\75	6.2\77
RSPC(0.08)	0\100	0\99	0\100	0.1\98	0\100	0\100	0\100	0\100	0\100	0\100

TABLE II
COMPARISON OF ROBUSTNESS UNDER LINEAR DISTORTION PROFILE AND VARIOUS PAYLOADS (EMBEDDING RATES)

Algorithms	$\bar{R}_{error} (\times 10^{-3} \downarrow) \setminus \bar{R}_{success} (\% \uparrow)$									
	0.02	0.04	0.06	0.08	0.1	0.12	0.14	0.16	0.18	0.2
STC+RS(31,15)	477.7\0	431.3\0	410.1\0	373.8\0	328.4\0	295.5\0	262.0\0	238.6\0	216.6\0	196.7\0
STC+RS(31,9)	465.2\0	406.4\0	324.2\0	260.9\0	203.6\0	179.8\0	134.5\0	97.6\0	76.6\0	50.6\0
DSTC(2,8)	473.3\0	420.8\0	348.5\0	307.0\0	221.3\0	171.6\0	113.6\0	79.4\0	50.7\0	30.2\0
DSTC(1,8)	392.0\0	292.4\0	239.4\0	182.8\0	142.7\0	107.0\0	75.4\0	48.7\0	31.9\0	19.1\0
DSTC(1,8)+RS(31,15)	187.7\0	104.5\4	32.2\29	5.1\82	0.8\94	0.3\99	0\100	0\100	0\99	0\100
RSPC(0.05)	190.7\0	184.9\0	201.8\0	208.4\0	217.5\0	216.4\0	236.0\0	247.5\0	242.0\0	251.6\0
RSPC(0.08)	18.3\49	17.6\46	15.9\52	22.2\35	25.6\42	32.9\40	27.0\47	30.5\41	30.4\46	34.5\45
RSPC(0.13)	0\100	0\100	0\100	0\100	0.1\99	0.2\99	0\100	0\100	0\100	0\100



(a) Constant Distortion Profile



(b) Linear Distortion Profile

Fig. 4. Comparison of the average embedding distortion for different algorithms under various payloads (embedding rates) where $W = \text{BSC}(0.02)$.

parameter p' . Furthermore, we combine DSTC and RS codes to achieve the strongest robustness achievable by conventional methods and it can be seen that successful extraction can only be realized when the payload ranges from 0.14 to 0.2.

C. Comparison of Security Performance

To evaluate the security performance, we calculate the average embedding distortion of the above algorithms in the same setting of Section V-B and the results are shown in Fig. 4. It can be seen that the average distortion incurred by RSPC is much less than that of DSTC(1,8)+RS(31,15) when successful extraction is achieved. Here we note that methods with average distortion lower than RSPC are not feasible in practical scenarios as they cannot provide reliable robustness. Overall, the proposed algorithm provides reliable robustness

and also has distinct advantages when it comes to security performance compared to the existing methods.

D. Evaluation under High Payloads and Bad Channels

In real-world applications, we may need high embedding rates when the secret message is long. Table III and Table IV show the experimental results of robustness comparison in this scenario where DSTC++ denotes DSTC(1,8)+RS(31,15). It can be observed that combining DSTC and RS codes is not suitable for the scenario when the embedding rate is high due to the fact that error-correcting codes significantly increase the message length which would exceed the bound of the maximum information that can be embedded using DSTC. On the contrary, the proposed method can provide reliable robustness by adjusting the robust parameter, and thus, it is

TABLE III
EVALUATION OF ROBUSTNESS WITH HIGH PAYLOADS (EMBEDDING RATES) UNDER CONSTANT DISTORTION PROFILE

Channels	Algorithms	$\bar{R}_{error} (\times 10^{-3} \downarrow) \setminus \bar{R}_{success} (\% \uparrow)$					
		0.1	0.2	0.3	0.4	0.5	0.6
BSC(0.02)	DSTC++	0.4\97	0\100	-	-	-	-
	RSPC(0.08)	0\100	0\100	0\100	0\100	0.1\99	0.1\99
	RSPC(0.1)	0\100	0\100	0\100	0\100	0\100	0.5\99
BSC(0.05)	DSTC++	147.3\0	0.7\94	-	-	-	-
	RSPC(0.1)	2.6\88	2.6\91	5.8\86	4.2\88	6.9\84	82.9\33
	RSPC(0.15)	0\100	0\100	0\100	0\100	1.9\96	171.1\31

TABLE IV
EVALUATION OF ROBUSTNESS WITH HIGH PAYLOADS (EMBEDDING RATES) UNDER LINEAR DISTORTION PROFILE

Channels	Algorithms	$\bar{R}_{error} (\times 10^{-3} \downarrow) \setminus \bar{R}_{success} (\% \uparrow)$				
		0.1	0.2	0.3	0.4	0.5
BSC(0.02)	DSTC++	1.1\94	0\100	-	-	-
	RSPC(0.1)	0\90	0\91	0\84	0\93	0.1\86
	RSPC(0.13)	0.1\99	0\100	0.3\98	0\100	3.0\94
BSC(0.05)	DSTC++	144.5\0	0.5\92	-	-	-
	RSPC(0.15)	3.4\88	2.4\93	4.1\89	6.4\91	18.3\8
	RSPC(0.2)	0\100	0\100	0.1\99	4.6\94	44.6\0

TABLE V
EVALUATION OF ROBUSTNESS WITH BAD CHANNELS UNDER
CONSTANT DISTORTION PROFILE

Channels	Algorithms	$\bar{R}_{error} (\times 10^{-3} \downarrow) \setminus \bar{R}_{success} (\% \uparrow)$		
		0.05	0.1	0.15
BSC(0.1)	DSTC++	399.2\0	303.8\0	222.7\0
	RSPC(0.15)	7.5\74	6.8\74	8.8\76
	RSPC(0.2)	0\100	0\100	0\100
BSC(0.15)	RSPC(0.2)	11.1\72	8.8\85	16.5\78
	RSPC(0.25)	0\100	0\100	0\100
BSC(0.2)	RSPC(0.25)	16.3\74	20.6\73	14.7\80
	RSPC(0.3)	0\100	0\100	1.1\97

TABLE VI
EVALUATION OF ROBUSTNESS WITH BAD CHANNELS UNDER
LINEAR DISTORTION PROFILE

Channels	Algorithms	$\bar{R}_{error} (\times 10^{-3} \downarrow) \setminus \bar{R}_{success} (\% \uparrow)$		
		0.05	0.1	0.15
BSC(0.1)	DSTC++	392.2\0	303.1\0	223.9\0
	RSPC(0.2)	2.6\88	4.4\88	6.7\85
	RSPC(0.25)	0\100	0\100	0\100
BSC(0.15)	RSPC(0.25)	4.1\91	3.3\93	9.1\86
	RSPC(0.3)	0\100	0\100	0.4\98
BSC(0.2)	RSPC(0.3)	2.9\55	1.5\97	89.6\94

superior to the existing algorithms. Here we remark that our method does not show the expected performance in the lower-right corner of the tables because $H(q') + \frac{m}{N} > 1$ which exceeds the theoretical parameter range.

Aside from the above scenario, we may also have to face bad channels in real-world applications. Table V and Table VI show the robustness comparison results in this scenario. It can be noticed that combining DSTC and RS codes cannot provide reliable robustness when the communication channel W is bad as the extraction success rate is always 0, while the proposed method guarantees the perfect extraction of the secret message by lifting the robust parameter γ (equivalently, q'). These experimental results are as expected as our scheme has a theoretical basis and a structure consistent with the communication problem model shown in fig. 3 while the design of the existing methods are quite empirical and heuristic.

Experimental results on the comparison of the average embedding distortion are shown in Fig. 5. We can see that the proposed method provides more robustness than DSTC++ at a similar distortion level. More importantly, our method is capable when the embedding rate is high while DSTC++ is not applicable in this scenario. By lifting the robust parameter, the proposed method provides much more robustness than DSTC++ which can even assure successful extraction when

the communication channel is bad.

In the last of this subsection, we would like to explain why the performance of the proposed method seems worse when it comes to the linear distortion profile compared to the constant distortion profile. The key reason is that when doing code construction, the adaptive steganographic embedding channels $\{W_i\}_i$ are considered all the same BSC(p) while they are actually not the same at all. Since the receiver does not have access to the cover, the distortion profile ρ based on the cover characteristics and the optimal transition probabilities $\{W_i(\cdot|\cdot)\}_i$ that are dependent on ρ are not synchronized to the receiver. As the embedding rate and the robust parameter are known by both sides, to make sure that the receiver can conduct the same code construction procedures, $\{W_i\}_i$ are considered all the same subject to the embedding rate and the robust parameter. As $\{W_i\}_i$ are indeed all the same under the constant distortion profile, the performance of the proposed method would appear to be better in this situation compared to other general distortion profiles.

E. Complexity Analysis

In this subsection, we compare the computational complexity of various robust adaptive steganographic coding methods. For RSPC, the modified SC encoding and SC decoding both

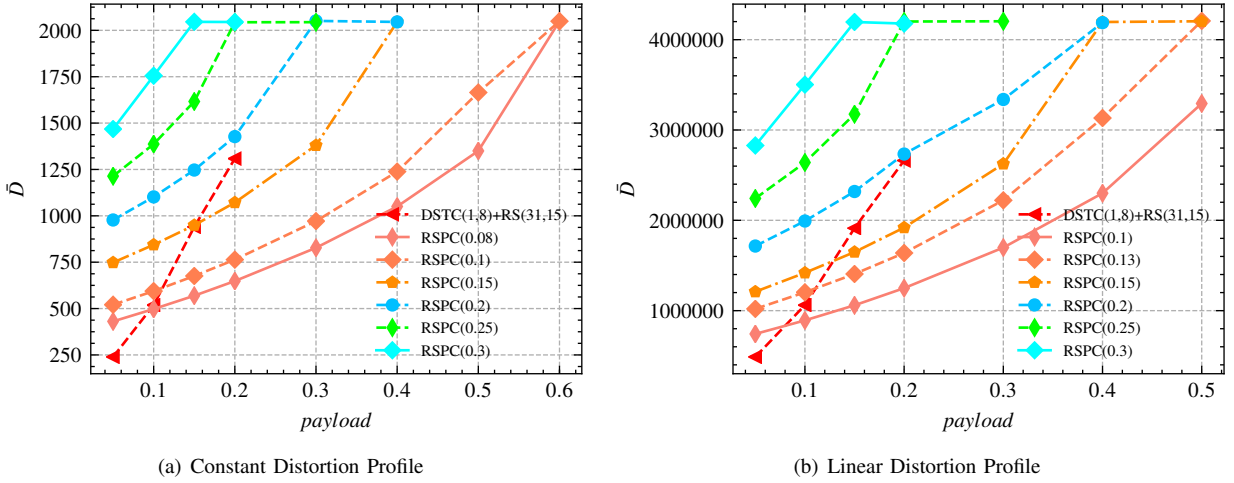


Fig. 5. Comparison of the average embedding distortion of the proposed algorithm and DSTC++ under high payloads (embedding rates) and bad channels.

have complexity $O(N \log N)$. Compared to the original SPC, the main difference is that the Bhattacharyya parameters have to be estimated two times for C_1 and C_2 , and the decoder has to conduct an extra SC decoding. Since all these operations only lead to constant multiples of the complexity, the computational complexity of RSPC remains $O(N \log N)$.

TABLE VII
COMPARISON OF COMPUTATIONAL COMPLEXITY

RSPC	DSTC	STC+RS / DSTC+RS
$O(N \log N)$	$O(2^h N)$	$O(2^h N + N \log^2 N)$

The complexity comparison of various methods are shown in Table VII where h denotes the constraint height of the STC encoding matrix. Note that DSTC shares the same complexity with STC while STC+RS shares the same complexity with DSTC+RS. Here we make a remark that although the RS coding is conducted on the secret message, by fixing the embedding rate R , $O(NR \log^2(NR)) = O(N \log^2 N)$. So, RSPC also holds lower complexity compared with DSTC+RS.

VI. CONCLUSION

We have proposed a reliable robust adaptive steganographic coding scheme based on nested polar codes in this paper. The proposed method is very flexible with a robust parameter γ that controls the balance between robustness and the average distortion incurred by the embedding process. Compared to the existing robust adaptive steganographic coding algorithms which can only provide limited robustness, the proposed method can guarantee the perfect extraction of the secret message by lifting parameter γ even when the communication channel between the sender and the receiver is fairly bad. Aside from the practical concerns, we have theoretically analyzed the robust parameter γ and provided a direction on how to select the parameter for an acceptable extraction precision. Moreover, we proved the optimality of the proposed scheme under the constant distortion profile as cover length N tends to infinity through powers of two in the special case

where the communication channel is a BSC and the design embedding rate $\frac{m}{N}$ is sufficiently large. The overall performance of the proposed scheme would increase by applying better Bhattacharyya parameter estimation methods like [41], [45] and the Successive Cancellation List (SCL) [43] algorithm.

There are several directions of improvement that we are going to work on in the future. The first possible improvement is to consider real-world applications for specific covers and lossy processing conducted on the stego. By utilizing the properties of the cover, the lossy processing, and the proposed coding method, we can devise supporting methods to support the steganographic coding scheme for specific real-world scenarios, e.g., JPEG compression. By combining the specific supporting methods with the proposed coding method, the algorithm would be more practicable in real-world applications. Another way of improvement is to devise “time-sharing” strategies for the low-rate scenario. Although the “time-sharing” strategy for the constant distortion profile is straightforward, for general adaptive distortion functions, the design of such strategies still need more investigations from both the theoretical and practical aspects.

APPENDIX PROOF OF PROPOSITION 1

To prove the result, we first introduce a lemma.

Lemma 1: If B-DMC $\tilde{W} : \{0, 1\} \rightarrow \tilde{\mathcal{Y}}$ is degraded with respect to B-DMC $\bar{W} : \{0, 1\} \rightarrow \bar{\mathcal{Y}}$, then, $\forall 1 \leq i \leq N$, $\bar{W}_N^{(i)}$ is degraded with respect to $\tilde{W}_N^{(i)}$ and $Z(\bar{W}_N^{(i)}) \leq Z(\tilde{W}_N^{(i)})$.

This lemma is a restatement of [42, Lemma 4.7] and the proof is omitted. Now, using Lemma 1, we have that

$$\forall 1 \leq i \leq N, \quad Z(\bar{W}_N^{(i)}) \leq Z(\tilde{W}_N^{(i)}). \quad (38)$$

Taking limits on both sides of (38), we have that $\forall 1 \leq i \leq N$

$$\lim_{N=2^n, n \rightarrow \infty} Z(\bar{W}_N^{(i)}) \leq \lim_{N=2^n, n \rightarrow \infty} Z(\tilde{W}_N^{(i)}). \quad (39)$$

Then, when $\lim_{N=2^n, n \rightarrow \infty} Z(\bar{W}_N^{(i)}) = 1$,

$$\begin{aligned} 1 &= \lim_{N=2^n, n \rightarrow \infty} Z(\bar{W}_N^{(i)}) \leq \lim_{N=2^n, n \rightarrow \infty} Z(\tilde{W}_N^{(i)}) \leq 1 \\ &\Rightarrow \lim_{N=2^n, n \rightarrow \infty} Z(\tilde{W}_N^{(i)}) = 1. \end{aligned} \quad (40)$$

By Definition 1, we conclude that the pair $(\bar{W}, \tilde{W})$ is an optimal channel pair for nested polar code construction.

PROOF OF COROLLARY 1

Since $\frac{m}{N} > 0, \gamma > 0$, apparently, we have that $\frac{m}{N} + \gamma > \gamma$ and $q < p \leq \frac{1}{2}$. Let $W' : \{0, 1\} \rightarrow \{0, 1\}$ be a BSC($\frac{p-q}{1-2q}$). Then, following (23), we write

$$\begin{aligned} \sum_{y \in \{0,1\}} W(y|0)W'(0|y) &= \sum_{y \in \{0,1\}} W(y|1)W'(1|y) \\ &= (1-q) \left(1 - \frac{p-q}{1-2q}\right) + q \frac{p-q}{1-2q} = 1-p \end{aligned} \quad (41)$$

and

$$\begin{aligned} \sum_{y \in \{0,1\}} W(y|0)W'(1|y) &= \sum_{y \in \{0,1\}} W(y|1)W'(0|y) \\ &= (1-q) \frac{p-q}{1-2q} + q \left(1 - \frac{p-q}{1-2q}\right) = p. \end{aligned} \quad (42)$$

As a consequence, the channel $W_i = \text{BSC}(p)$ is degraded with respect to $W = \text{BSC}(q)$. Therefore by Proposition 1, we have that the pair $(W, W_i) = (\text{BSC}(q), \text{BSC}(p))$ is optimal for nested polar code construction.

Because of the channel polarization phenomenon, the fraction of coordinate channels such that $\lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 1$ goes to $1 - I(W) = H(q) = \gamma$ and the fraction of coordinate channels such that $\lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 0$ goes to $I(W) = 1 - \gamma$. Similarly, the fraction of channels such that $\lim_{N \rightarrow \infty} Z(W_{N_{st}}^{(i)}) = 1$ goes to $1 - I(W_i) = H(p) = \frac{m}{N} + \gamma$ and the fraction of channels such that $\lim_{N \rightarrow \infty} Z(W_{N_{st}}^{(i)}) = 0$ goes to $I(W_i) = 1 - \frac{m}{N} - \gamma$. Using our code construction method, F_2 contains $\gamma N + m$ indices corresponding to coordinate channels with the largest Bhattacharyya parameters $Z(W_{N_{st}}^{(i)})$. Then, as N tends to infinity through powers of two, $\forall i \in F_2$, $\lim_{N \rightarrow \infty} Z(W_{N_{st}}^{(i)}) = 1$; $\forall i \in F_2^c$, $\lim_{N \rightarrow \infty} Z(W_{N_{st}}^{(i)}) = 0$. Similarly, F_1 consists of γN indices corresponding to coordinate channels with the largest Bhattacharyya parameters $Z(W_N^{(i)})$ that are contained in F_2 . By optimality of pair (W, W_i) ,

$$\begin{aligned} \lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 1 &\Rightarrow \lim_{N \rightarrow \infty} Z(W_{N_{st}}^{(i)}) = 1, \\ \lim_{N \rightarrow \infty} Z(W_{N_{st}}^{(i)}) = 0 &\Rightarrow \lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 0, \end{aligned}$$

which means that $\forall i \in F_2^c$, $\lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 0$ and $\forall i \in F_1$, $\lim_{N \rightarrow \infty} Z(W_N^{(i)}) = 1$. This concludes the proof.

PROOF OF PROPOSITION 2

Apparently, we have that $0 < \epsilon_1 < \epsilon_2 \leq 1$ following the proposition statement. By (6), we write

$$\begin{aligned} Z(W) &= \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)} \\ &= 0 + 0 + \epsilon_1 = \epsilon_1. \end{aligned} \quad (43)$$

Similarly, we also have

$$Z(W_i) = \epsilon_2.$$

Consequently, $Z(W) < Z(W_i)$.

Let us consider Arikan's single-step polar transformation $(\bar{W}, \tilde{W}) \rightarrow (\bar{W}^-, \bar{W}^+)$ where $\bar{W}$ is a BEC. Following (11), we write

$$\begin{aligned} Z(\bar{W}^+) &= Z(\bar{W})^2 \\ \frac{dZ(\bar{W}^+)}{dZ(\bar{W})} &= 2Z(\bar{W}) \geq 0. \end{aligned} \quad (44)$$

This means that as $Z(\bar{W})$ increases, $Z(\bar{W}^+)$ increases as well. Similarly, following (12), we write

$$\begin{aligned} Z(\bar{W}^-) &= 2Z(\bar{W}) - Z(\bar{W})^2 \\ \frac{dZ(\bar{W}^-)}{dZ(\bar{W})} &= 2 - 2Z(\bar{W}) \geq 0. \end{aligned} \quad (45)$$

So, $Z(W^-)$ also increases along with $Z(W)$. Since $Z(W)$ increases continuously along with the erasure probability of the BEC, $Z(W^+)$ and $Z(W^-)$ increases continuously along with the erasure probability as well. Therefore,

$$\epsilon_1 < \epsilon_2 \Rightarrow Z(W^+) < Z(W_i^+), Z(W^-) < Z(W_i^-). \quad (46)$$

As we know that if $\bar{W}$ is a BEC, $\bar{W}^+$ and $\bar{W}^-$ are also BECs [32, Proposition 6], all the coordinate channels after channel splitting are BECs by induction. Therefore, for any fixed $N = 2^n, n \in \mathbb{N}, \forall 1 \leq i \leq N, Z(W_N^{(i)}) < Z(W_{N_{st}}^{(i)})$. This means that if $W_N^{(i)}$ is a bad coordinate channel, $W_{N_{st}}^{(i)}$ is worse. Consequently, by similar means of the proof of Corollary 1, we can always make sure that $i \in F_1 \Rightarrow i \in F_2$, $F_1 \subset F_2$. This concludes the proof.

REFERENCES

- [1] R. J. Barron, Brian Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1159-1180, May 2003.
- [2] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423-1443, May 2001.
- [3] W. Trappe, M. Wu, Z. J. Wang, and K. J. R. Liu, "Anti-collusion fingerprinting for multimedia," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 1069-1087, Apr. 2003.
- [4] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inf. Theory*, vol. 49, pp. 563-593, Mar. 2003.
- [5] F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: Theory and practice," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3976-3987, Oct. 2005.
- [6] P. Moulin and R. Koetter, "Data-Hiding Codes," *Proc. IEEE*, vol. 93, no. 12, pp. 2083-2126, Dec. 2005.
- [7] J. Fridrich, M. Goljan, P. Lisonek, and D. Soukal, "Writing on wet paper," *IEEE Trans. Signal Process.*, vol. 53, pp. 3923-3935, Oct. 2005.
- [8] C. Munuera, "Steganography and error-correcting codes," *Signal Process.*, vol. 87, pp. 1528-1533, 2007.
- [9] W. Zhang, X. Zhang, and S. Wang, "A Double Layered 'Plus-Minus One' Data Embedding Scheme," *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 848-851, Nov. 2007.
- [10] I. Cox et al., *Digital Watermarking and Steganography*. San Mateo, CA, USA: Morgan Kaufmann, 2008.
- [11] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

- [12] W. Zhang, X. Zhang, and S. Wang, "Near-Optimal Codes for Information Embedding in Gray-Scale Signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1262-1270, March 2010.
- [13] A. D. Ker, "The square root law of steganography," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2017, pp. 33-44.
- [14] Kai Zeng, Kejiang Chen, Weiming Zhang, Yaofei Wang, and Nenghai Yu, "Improving robust adaptive steganography via minimizing channel errors," *Signal Process.*, vol. 195, pp. 108498, 2022.
- [15] P. Wei, Z. Zhu, G. Luo, Z. Qian, and X. Zhang, "Breaking Robust Data Hiding in Online Social Networks," *IEEE Signal Process. Lett.*, vol. 29, pp. 2682-2686, 2022.
- [16] M. Barni et al., "Information Forensics and Security: A quarter-century-long journey," *IEEE Signal Process. Mag.*, vol. 40, no. 5, pp. 67-79, July 2023.
- [17] Guoliang Xie, Jinchang Ren, Stephen Marshall, Huimin Zhao, and Rui Li, "A Novel Gradient-guided Post-processing Method for Adaptive Image Steganography," *Signal Process.*, vol. 203, pp. 108813, 2023.
- [18] C. Kin-Cleaves and A. D. Ker, "Adaptive steganography in the noisy channel with dual-syndrome trellis codes," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1-7.
- [19] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 2, pp. 221-234, Feb. 2016.
- [20] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Trans. Inf. Forensics Secur.*, vol. 6, no. 3, pp. 920-935, Sep. 2011.
- [21] W. Li, W. Zhang, L. Li, H. Zhou, and N. Yu, "Designing near-optimal steganographic codes in practice based on polar codes," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 3948-3962, Jul. 2020.
- [22] Q. Yao, W. Zhang, K. Chen, and N. Yu, "LDGM Codes-Based Near-Optimal Coding for Adaptive Steganography," *IEEE Trans. Commun.*, vol. 72, no. 4, pp. 2138-2151, April 2024.
- [23] F. Li, K. Wu, C. Qin, and J. Lei, "Anti-compression JPEG steganography over repetitive compression networks," *Signal Process.*, vol. 170, May 2020, Art. no. 107454.
- [24] L. Zhu, X. Luo, C. Yang, Y. Zhang, and F. Liu, "Invariances of JPEG-quantized DCT coefficients and their application in robust image steganography," *Signal Process.*, vol. 183, Jun. 2021, Art. no. 108015.
- [25] T. Filler and J. Fridrich, "Gibbs construction in steganography," *IEEE Trans. Inf. Forensics Secur.*, vol. 5, no. 4, pp. 705-720, Dec. 2010.
- [26] Q. Guan, P. Liu, W. Zhang, W. Lu, and X. Zhang, "Double-Layered Dual-Syndrome Trellis Codes Utilizing Channel Knowledge for Robust Steganography," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 501-516, 2023.
- [27] Z. Zhao, Q. Guan, H. Zhang, and X. Zhao, "Improving the robustness of adaptive steganographic algorithms based on transport channel matching," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 7, pp. 1843-1856, Jul. 2019.
- [28] Y. Zhang, X. Luo, Y. Guo, C. Qin, and F. Liu, "Multiple robustness enhancements for image adaptive steganography in lossy channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2750-2764, Aug. 2020.
- [29] W. Lu, J. Zhang, X. Zhao, W. Zhang, and J. Huang, "Secure robust JPEG steganography based on AutoEncoder with adaptive BCH encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2909-2922, Jul. 2021.
- [30] X. Yu, K. Chen, Y. Wang, W. Li, W. Zhang, and N. Yu, "Robust adaptive steganography based on generalized dither modulation and expanded embedding domain," *Signal Process.*, vol. 168, Mar. 2020, Art. no. 107343.
- [31] R. Zamir, S. Shamai and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1250-1276, June 2002.
- [32] E. Arıkan, "Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051-3073, Jul. 2009.
- [33] T. Filler and J. Fridrich, "Using non-binary embedding operation to minimize additive distortion functions in steganography," in *Proc. 2nd IEEE Int. Workshop Inf. Forensics Secur.*, Seattle, WA, USA, Dec. 2010.
- [34] X. Zhang, W. Zhang, and S. Wang, "Efficient double-layered steganographic embedding," *Electron. Lett.*, vol. 43, no. 8, pp. 482-483, 2007.
- [35] Q. Yao, W. Zhang, and N. Yu, "Optimality of polar codes in additive steganography under constant distortion profile," in *Proc. 14th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nov. 2022, pp. 404-408.
- [36] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, vol. 27, pt. 4, pp. 142-163, 1959.
- [37] T. Berger, *Rate Distortion Theory*. London, U.K.: Prentice-Hall, 1971.
- [38] E. Arıkan and E. Telatar, "On the rate of channel polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, South Korea, Jul. 2009, pp. 1493-1495.
- [39] M. Alsan and E. Telatar, "A Simple Proof of Polarization and Polarization for Non-Stationary Memoryless Channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 4873-4878, Sept. 2016.
- [40] E. Arıkan, "A performance comparison of polar codes and Reed-Muller codes," *IEEE Commun. Lett.*, vol. 12, no. 6, pp. 447-449, Jun. 2008.
- [41] I. Tal and A. Vardy, "How to Construct Polar Codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6562-6582, Oct. 2013.
- [42] S. B. Korada, "Polar Codes for Channel and Source Coding," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2009.
- [43] A. Balatsoukas-Stimming, M. B. Parizi, and A. Burg, "LLR-based successive cancellation list decoding of polar codes," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5165-5179, Oct. 2015.
- [44] Y. Niu, X. Li, Y. Zhao, and R. Ni, "Detection of Double JPEG Compression With the Same Quantization Matrix via Convergence Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3279-3290, May 2022.
- [45] R. Mori and T. Tanaka, "Performance and construction of polar codes on symmetric binary-input memoryless channels," in *Proc. IEEE Symp. Inf. Theory*, Seoul, Korea, 2009, pp. 1496-1500.