One-Shot Information Hiding

Yanxiao Liu and Cheuk Ting Li
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
Email: yanxiaoliu@link.cuhk.edu.hk, ctli@ie.cuhk.edu.hk

***Abstract*—We present a one-shot information-theoretic analysis of the information hiding problem, which has a wide range of applications including watermarking, fingerprinting, steganography and copyright protection. The problem can be viewed as a game: one party includes an information hider and a decoder, where the former embeds a message into a host data source and introduces some tolerable distortion, and the latter wishes to reconstruct the message; another party is an attacker that is modeled as a noisy channel which aims at removing the hidden information. We derive a one-shot achievability result using the Poisson matching lemma. Unlike previous asymptotic results, our result applies to any distribution of the host data, and any class of attack channels (not necessarily memoryless or ergodic).**

***Index Terms*—Information hiding, one-shot achievability, finite blocklength, network information theory, watermarking.**

I. Introduction

Information hiding has been a widely studied topic in the past decades, due to its wide range of applications including watermarking, fingerprinting, data embedding, steganography and copyright protection. It borrows techniques from various areas, e.g., wireless communication, signal processing, cryptography and game theory [1]–[5]. The information hiding problem can be formulated as a communication system [1], where the goal is to characterize the maximum rate of reliable transmissions under attacks. More specifically, a message M is expected to be reliably transmitted to a decoder. To protect M from attacks during the transmission, it is embedded into a host source S, and X is the encoded signal. Upon receiving Y from the attack channel $A(Y|X)$, the decoder decodes M.

Two main classes of applications of information hiding are watermarking and fingerprinting [2]. In watermarking, the message usually contains personal identification and the goal is usually to protect copyright. The message is expected to be embedded in the host data, but the secrecy is not always required [3], [4] and sometimes the host data is also fully available at the decoder [2], [6], [7]. In fingerprinting, the message is a fingerprint inside the host data that can identify a unique user, and collusion between users is usually considered [2], [8]. In [1], a comprehensive study on the fundamental limits of asymptotic information hiding systems has been addressed, and the hiding capacity is proved by borrowing techniques from the celebrated Gelfand-Pinsker coding [9], [10].

However, the hiding capacity provided by [1] is the *asymptotic* capacity, i.e., the law of large number is utilized while assuming that the signal has a blocklength approaching infinity. In the past decade, since packets have bounded lengths in practice, *finite blocklength information theory* has been widely studied [11]–[13]. More generally, we are interested in the *one-shot* achievability results, i.e., the channel is arbitrary and used only *once*. Various one-shot coding techniques have been proposed [13]–[20], yielding one-shot bounds that imply existing (first-order and second-order) asymptotic results when applied to memoryless channels.

In this paper, we provide one-shot achievability results of the information hiding problem. We utilize the Poisson matching lemma [19], which is rooted in the Poisson functional representation [21]. Compared with the asymptotic results, [1] has assumptions that the attack channels are memoryless or blockwise-memoryless and the decoder has complete knowledge about the attacker, which are dropped in this paper; [4] assumes the side information K is a shared key of unlimited size that is independent of S, M and can be chosen as a part of the coding scheme, while we assume the given K is correlated with S and cannot be changed (as in [1]).

II. Related Literature

We briefly review three lines of research: the information hiding problem, one-shot information theory, and applications on copyright protection in modern scenarios.

A. Information Hiding, Watermarking and Fingerprinting

The information hiding problem has been discussed for a long time, due to its wide range of applications on watermarking, fingerprinting, steganography, data embedding, audio/image/video processing and copyright protection [1]–[5]. In [1], a guiding theory for the fundamental information-theoretic limits of information hiding has been proposed. The information hiding system is modeled as a communication problem, where a message is to be embedded and hidden in a host data (by introducing some tolerable distortion), and the overall encoded data (which is usually similar to host data) will be under data processing attacks (by introducing another level of distortion) that attempt to remove or degrade the message. The goal of the encoder-decoder party is to let the decoder correctly decodes the hidden message. Though the problem is related to cryptography, the secrecy of the message is not always required, e.g., in watermarking [3], [4] where the message is the personal identification for copyright protection.

Public watermarking [4] shares a similar setting with [1] that the host data is only available at the encoder. In compari-

This work was partially supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No.s: CUHK 24205621 (ECS), CUHK 14209823 (GRF)].

son, the *private* watermarking [3], [6] discusses the case where the host data is available at both the encoder and the decoder, and the capacity and error exponents are investigated in [3]. [1] has a questionable assumption that the attack scheme is always be learned by the decoder, which is dropped in [3], [4] (and is also the case in this paper). In [5], the capacity of a Gaussian watermarking game has been studied, and cases where the public and private games have the same capacity are discussed (which is not always true in general). Digital fingerprinting is another important application, which desires to embed a fingerprint into the host data that can uniquely identify the users for copyright protection or tracing illegal uses of the data. It is challenging due to the possible collusion between users [8], and the case of blockwise memoryless attacks has been discussed in [1]. See [22]–[26] for other related literature.

B. One-shot Information Theory

For all the literature on the information hiding problem above, the information-theoretic limits are investigated in the *asymptotic* regime, where the law of large numbers is employed to derive the asymptotic behavior of channels in the large blocklength limit. In the past decade, due to the fact that packets have bounded lengths in practice, which can even be very short in machine-type communications [27], *finite blocklength information theory* [11]–[13] and one-shot information theory [12]–[20] have been studied, both of which intend to provide nonasymptotic results of channels, which are expected to imply existing (first-order and second-order) asymptotic results. In the one-shot case, we assume the blocklength is 1, i.e., we consider a *single* use of the channel, with no assumption on memorylessness or ergodicity.

In this paper, we derive one-shot achievability for the information hiding problem by utilizing the *Poisson matching lemma* [19], which in turn is based on the Poisson functional representation [21]. It provides a unified framework for one-shot achievability results, which can improve upon previously known one-shot bounds in various settings with simpler analyses [19]. See Section III for details. Recently a refined version of the Poisson matching lemma has been used to provide one-shot bounds for (multi-hop) general noisy networks [20].

C. Information Hiding in Machine Learning

Information hiding systems for modern scenarios have recently attracted considerable attention. In the past decade, machine learning techniques have gained great success in a large number of areas. Software based on the generative models, e.g., Midjourney [28], Stable Diffusion [29] or ChatGPT [30] can produce contents as realistic as the original works by human creators used in training. However, the copyright issue becomes controversial since some generative models are possibly trained on publicly available data without obtaining permission from the authors. To protect the intellectual property of human creators, techniques to embed information (e.g., watermarks that are hard to be removed) into the original works have been proposed [31], [32]. Some AI-based algorithms for information hiding, watermarking and fingerprinting have also been proposed [33]–[35]. To design practical algorithms, it is crucial to understand the fundamental limits of any information hiding system, without impractical assumptions in such learning-based system designs, e.g., memorylessness, ergodicity, or the decoder being informed of the attack schemes. This also motivates us to provide a one-shot information-theoretic study on the information hiding problem.

Notations

We assume logarithm and entropy are to the base 2. For a statement S, we use $\mathbf{1}\{S\}$ to denote its indicator, i.e., $\mathbf{1}\{S\} = 1$ if S holds, and otherwise $\mathbf{1}\{S\} = 0$. We use δ_a to denote the degenerate distribution $\mathbf{P}\{X = a\} = 1$. For two random variables X, Y, the information density is defined as $\iota_{X;Y}(x;y) = \log((\mathrm{d}P_{X|Y}(\cdot|y)/\mathrm{d}P_X)(x))$, where $\mathrm{d}P_{X|Y}(\cdot|y)/\mathrm{d}P_X$ denotes the Radon-Nikodym derivative. We sometimes omit the subscript and write $\iota(x;y)$ if the random variables are clear from the context. The total variation (TV) distance between two distributions P, Q over $\mathcal{X}$ is $\|P-Q\|_{\mathrm{TV}} := \sup_{A\subseteq\mathcal{X}\text{ measurable}} |P(A) - Q(A)|$.

III. Poisson Matching Lemma

The techniques in [1], [4] are not suitable for the one-shot setting (see Section IV-B for details). In this paper, we utilize the Poisson matching lemma [19] to prove one-shot achievabilities of the information hiding game, which is rooted in the Poisson functional representation [21] that is reviewed as follows. Fix a probability distribution $\bar{P}$ over $\mathcal{U}$. Let $(T_i)_{i=1,2,\ldots}$ be a Poisson process with rate 1, i.e., $T_1, T_2 - T_1, T_3 - T_2 \overset{iid}{\sim} \mathrm{Exp}(1)$. Let $(\bar{U}_i)_i$ be an independent i.i.d. sequence with distribution $\bar{P}$. This "marked" Poisson process $(\bar{U}_i, T_i)_i$ supports a "query operation" given by the Poisson functional representation, where one can input a distribution P over $\mathcal{U}$, and obtain one sample $\tilde{U}_P$ with distribution P. The Poisson functional representation is given by

$$\tilde{U}_P := \bar{U}_K, \text{ where } K := \arg\min_i T_i \cdot \Big(\frac{\mathrm{d}P}{\mathrm{d}\bar{P}}(\bar{U}_i)\Big)^{-1}.$$

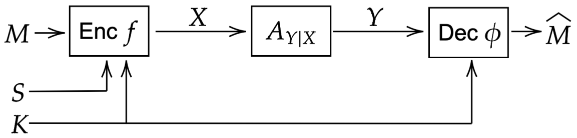
The way this Poisson process is used in communication settings (e.g., in [19], [20]) is that the encoder would query the process using the prior distribution of the signal to obtain the signal to be sent, and the decoder would query using the posterior distribution of the signal given the noisy observation to obtain the message. There is no error in the communication if the two queries return the same sample. The probability of error can be bounded by the Poisson matching lemma in [19].

Lemma 1 (Poisson matching lemma [19]). *Consider two distributions $P, Q \ll \bar{P}$. Almost surely, we have*

$$\mathbf{P}\big(\tilde{U}_Q \neq \tilde{U}_P \,\big|\, \tilde{U}_P\big) \leq 1 - \Big(1 + \frac{\mathrm{d}P}{\mathrm{d}Q}(\tilde{U}_P)\Big)^{-1}.$$

IV. One-shot Information Hiding Game

In this section, we formulate the one-shot information hiding problem and provide the main results.



A. Problem Formulation

The one-shot information hiding game is described in the figure above. A host source $S \in \mathcal{S}$ and a side information source $K \in \mathcal{K}$ (available to both the encoder and the decoder) are distributed according to the joint distribution $P_{S,K}$. A message M is uniformly chosen from the set $[1:\mathsf{L}]$, where L is the *message size*. An encoder produces X as a function of S, K, M and sends X through an attack channel $A_{Y|X}$. The attack channel $A_{Y|X}$ is chosen by an attacker, who attempts to destroy the embedded message under some distortion constraint, which will be discussed later. The decoder observes Y, K and desires to decode M correctly.

The roles of random variables are briefly described below.

- **Message** M: A message that is desired to be transmitted reliably to the decoder through a noisy channel (attacks). To protect M from the attacker, the encoder hides M into a host data source S to produce X.
- **Host data source** S: A host data set from text, image or video, which is allowed to suffer some tolerable level of distortions (from both the encoder and the attacker).
- **Side information** K: Common randomness available at both the encoder and the decoder, but not the attacker. It reveals information about S to the decoder, where the dependency are from the joint distribution $P_{S,K}$.

Given the random variables, the information hiding problem can be viewed as a game between two parties: the first party consists of the encoder (information hider) and the decoder, who are cooperatively transmitting the message M; the second party is an attacker, who is trying to destroy or degrade the hidden message M in S so that the decoder cannot correctly decode. Their roles and assumptions are elaborated as follows.

- **Encoder**: The goal of the encoder is to hide the message M into S. Given S, K, M, the encoder outputs $X = f(S, K, M)$, where $f : \mathcal{S} \times \mathcal{K} \times [1:\mathsf{L}] \to \mathcal{X}$. The encoder wants X to be close to S, in the sense that the distortion $d_1(S, X)$ is small, where $d_1 : \mathcal{S} \times \mathcal{X} \to [0, \infty)$ is a distortion measure. We want $d_1(S, X) \le \mathsf{D}_1$ with high probability. This will be elaborated later.
- **Attacker**: The attacker is formulated as a noisy channel with input X and output Y, called the *attack channel* $A_{Y|X}$. It performs data processing attacks on the received X and produces Y, a corrupted version of X. Its objective is to (partially) remove or degrade the message M so that the decoder cannot correctly find the original message M from Y. We assume the attack channel must be chosen from a class of channels $\mathcal{A}$, for example, the class of channels satisfying some distortion constraint between X and Y, or the class of memoryless channels in case X and Y are sequences. Different attack strategies could be performed, e.g., deterministic attacks that X is mapped by a deterministic function, or a randomized strategy. We assume the attacker has knowledge of the distributions (but not the values) of S, M, K, and also knows the code that the encoder-decoder team uses.
- **Decoder** ϕ: Upon observing the attacker's output Y, the decoder wishes to recover the message M. It outputs $\hat{M} = \phi(K, Y)$, where $\phi : \mathcal{K} \times \mathcal{Y} \to [1:\mathsf{L}]$. The decoder is uninformed of the attacker's strategy. We require the encoder-decoder team's worst case failure probability

$$P_e := \sup_{A_{Y|X} \in \mathcal{A}} \mathbf{P}\big(d_1(S, X) > \mathsf{D}_1 \ \text{ OR } \ M \neq \hat{M}\big) \quad (1)$$

to be small, where we assume $(S, K, M) \sim P_{S,K} \times \mathrm{Unif}[1:\mathsf{L}]$, $X = f(S, K, M)$, $Y|X \sim A_{Y|X}$ and $\hat{M} = \phi(K, Y)$ in the probability.[1]

B. Discussions

In [1], it is assumed that the attack channel must be memoryless, and hence the decoder can obtain full knowledge about the attack channel, justified by the large blocklength of signals. In this paper, we drop this assumption, and consider a one-shot setting where the set of possible attack channels $\mathcal{A}$ can be *any* set of channels. Also, we do not assume that the decoder knows the attack channel, which is unrealistic in the one-shot setting where the attacker can be arbitrary. In [4], the memoryless assumption is also dropped, and an asymptotic hiding capacity expressed as the limit of a sequence of single-letter expressions has been derived using constant composition codes. The key difference between [4] and our setting (and also [1]) is that the side information K in [4] is a shared key of unlimited size independent of M, S that can be chosen as a part of the coding scheme, whereas in our paper and [1] the K is a given side information that may be correlated with S (where the dependence is from the joint distribution $P_{S,K}$), and cannot be changed. In some watermarking problems [7], [26] certain components can be further constrained, e.g., there may exist a mapping from the message M to a codeword $V(M)$ which is independent of S, and then composite data are obtained by a mapping from S, K and $V(M)$.

The information hiding can be regarded as a variant of Gelfand-Pinsker coding for channels with side information at the encoder [9], [10], where the channel is fixed and not chosen by the attacker, and there is no K shared between the encoder and the decoder. Since the encoder and the decoder have to account for all possible attack channels, this can be regarded as a combination of Gelfand-Pinsker coding and compound channel [36]–[38]. The analyses in [1], [4] utilize techniques such as random binning, joint typicality decoding and constant composition codes, which are also commonly utilized in the asymptotic analyses of Gelfand-Pinsker coding [9], [39]. These techniques may not be suitable for our one-shot

[1] Note that [1] imposes a constraint on the expected distortion $\mathbf{E}[d_1(S, X)]$, which is reasonable in the context of [1] because the memoryless assumption and the law of large numbers ensure that the actual distortion is close to the expected distortion. Since we are considering a one-shot setting, we consider $d_1(S, X) > \mathsf{D}_1$ a failure event and bound the probability of failure instead.

setting. Strong typicality and constant composition codes are inapplicable when the blocklength is 1. While random binning can be applied to one-shot Gelfand-Pinsker coding [14]–[16], it produces weaker results compared to the Poisson matching lemma [19]. To obtain tight one-shot bounds for information hiding, we utilize the Poisson matching lemma instead.

V. One-shot Achievability of Information Hiding

Since the encoder-decoder team has to account for all possible attack channels in $\mathcal{A}$, it suffers a penalty depending on the "size" of $\mathcal{A}$. Though the cardinality of $\mathcal{A}$ is often infinite, we can often find a finite subset $\tilde{\mathcal{A}}$ such that every attack channel $A \in \mathcal{A}$ is close enough to some $\tilde{A} \in \tilde{\mathcal{A}}$. This notion of size is captured by the ϵ-covering number defined below. Similar covering arguments have been used in [1], [36].

Definition 1. Given a set of channels $\mathcal{A}$ from $\mathcal{X}$ to $\mathcal{Y}$, its *ϵ-covering number* is defined as

$$N_\epsilon(\mathcal{A}) := \min \Big\{ |\tilde{\mathcal{A}}| : \, \tilde{\mathcal{A}} \subseteq \mathcal{A}, \\ \sup_{A \in \mathcal{A}} \min_{\tilde{A} \in \tilde{\mathcal{A}}} \sup_{x \in \mathcal{X}} \|A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x)\|_{\mathrm{TV}} \le \epsilon \Big\},$$

where $\|A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x)\|_{\mathrm{TV}} \in [0,1]$ denotes the total variation distance between $A_{Y|X}(\cdot|x)$ (the distribution of Y if $X = x$ and Y follows $A_{Y|X}$) and $\tilde{A}_{Y|X}(\cdot|x)$.

We now present the main result, which is a one-shot achievability result with a bound on the error probability in terms of $N_\epsilon(\mathcal{A})$ and information density terms.

Theorem 2. *Fix any $P_{U,X|S,K}$ and channel $\hat{A}_{Y|X}$. Then for any $\epsilon \ge 0$, there exists an information hiding scheme satisfying*

$$P_e \le N_\epsilon(\mathcal{A}) \sup_{A_{Y|X} \in \mathcal{A}} \mathbf{E}_{Y|X \sim A_{Y|X}} \bigg[1 - \mathbf{1}\{d_1(S,X) \le \mathsf{D}_1\} \\ \cdot \Big(1 + \mathsf{L}2^{-\hat{\iota}(U;Y|K) + \iota(U;S|K)}\Big)^{-1} \bigg] + \epsilon,$$

where we assume $(S,K,U,X,Y) \sim P_{S,K}P_{U,X|S,K}A_{Y|X}$ in the expectation, and $\hat{\iota}(U;Y|K)$ is the information density computed by the joint distribution $P_{S,K}P_{U,X|S,K}\hat{A}_{Y|X}$ (instead of $A_{Y|X}$), assuming that $\iota(U;S|K), \hat{\iota}(U;Y|K)$ are almost surely finite for every $A_{Y|X} \in \mathcal{A}$.

Proof. The idea is that we design the decoder assuming that the attack channel is fixed to $\hat{A}_{Y|X}$, and hope that this decoder works for every attack channel $A_{Y|X}$. Let $\mathcal{C} := ((\bar{U}_i, \bar{M}_i), T_i)_i$ where $(T_i)_i$ is a Poisson process, $\bar{U}_i \overset{iid}{\sim} P_U$, and $\bar{M}_i \overset{iid}{\sim} P_M$ (where $P_M = \mathrm{Unif}[1 : \mathsf{L}]$). This will act as a random codebook shared between the encoder and the decoder (we will fix the codebook later). The encoder observes the message $M \sim P_M$, the host signal S and side information K, by the Poisson functional representation [19], [21] on the distribution $P_{U|S,K}(\cdot|S,K) \times \delta_M$ over $\mathcal{U} \times [1 : \mathsf{L}]$ it produces $U = \tilde{U}_{P_{U|S,K}(\cdot|S,K) \times \delta_M}$,[2] and sends the generated $X|(S,K,U) \sim P_{X|S,K,U}$. The decoder observes Y, K and outputs $\hat{M} = \tilde{M}_{\hat{P}_{U|Y,K}(\cdot|Y,K) \times P_M}$ by the Poisson functional representation, where $\hat{P}_{U|Y,K}$ is the conditional distribution computed by the joint distribution $P_{S,K}P_{U,X|S,K}\hat{A}_{Y|X}$. When the attack channel is $A_{Y|X} \in \mathcal{A}$, the error probability is

$$\begin{aligned}
&P_e(A) := 1 - \mathbf{P}_{Y|X \sim A_{Y|X}}\big(d_1(S,X) \le \mathsf{D}_1 \ \text{ AND } \ M = \hat{M}\big) \\
&= \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S,X) \le \mathsf{D}_1\} \cdot \mathbf{1}\{M = \hat{M}\}\Big] \\
&= \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S,X) \le \mathsf{D}_1\}\mathbf{P}\big(M = \hat{M} | M,S,U,Y,K\big)\Big] \\
&\le \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S,X) \le \mathsf{D}_1\} \\
&\qquad \cdot \mathbf{P}\big((U,M) = (\tilde{U},\tilde{M})_{\hat{P}_{U|Y,K}(\cdot|Y,K) \times P_M} | M,S,U,Y,K\big)\Big] \\
&\overset{(a)}{\le} \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S,X) \le \mathsf{D}_1\} \\
&\qquad \cdot \Big(1 + \frac{\mathrm{d}P_{U|S,K}(\cdot|S,K) \times \delta_M}{\mathrm{d}\hat{P}_{U|Y,K}(\cdot|Y,K) \times P_M}(U,M)\Big)^{-1}\Big] \\
&= \mathbf{E}\Big[1 - \mathbf{1}\{d_1(S,X) \le \mathsf{D}_1\}\Big(1 + \mathsf{L}2^{-\hat{\iota}(U;Y|K) + \iota(U;S|K)}\Big)^{-1}\Big] \\
&\le \sup_{A_{Y|X} \in \mathcal{A}} \mathbf{E}_{Y|X \sim A_{Y|X}}\Big[1 - \mathbf{1}\{d_1(S,X) \le \mathsf{D}_1\} \\
&\qquad \cdot \Big(1 + \mathsf{L}2^{-\hat{\iota}(U;Y|K) + \iota(U;S|K)}\Big)^{-1}\Big] \quad =: \overline{P_e},
\end{aligned}$$

where (a) is by the Poisson matching lemma.[3] If we allow the encoder and the decoder to share unlimited additional common randomness, we can assume the codebook $\mathcal{C} = ((\bar{U}_i, \bar{M}_i), T_i)_i$ is actually shared, and conclude that $P_e = \sup_{A \in \mathcal{A}} P_e(A) \le \overline{P_e}$. Nevertheless, the only actual common randomness between the encoder and the decoder is K, which we cannot control. Therefore, we have to fix the codebook.

Let $P_e(A,c)$ be the probability of error when the attack channel is A and the codebook is $\mathcal{C} = c$. We have $P_e(A) = \mathbf{E}_{\mathcal{C}}[P_e(A,\mathcal{C})]$ Let $\tilde{\mathcal{A}} \subseteq \mathcal{A}$ attain the minimum in $N_\epsilon(\mathcal{A})$. Consider any $A \in \mathcal{A}$, and let $\tilde{A} \in \tilde{\mathcal{A}}$ satisfy $\sup_{x \in \mathcal{X}} \|A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x)\|_{\mathrm{TV}} \le \epsilon$. The TV distance between the joint distribution of M,S,K,U,X,Y under the attack channel A conditional on $\mathcal{C} = c$ and the joint distribution under the attack channel $\tilde{A}$ conditional on $\mathcal{C} = c$ is also bounded by ϵ. Hence $|P_e(A,c) - P_e(\tilde{A},c)| \le \epsilon$ and

$$P_e(A,c) \le P_e(\tilde{A},c) + \epsilon \le \sum_{\tilde{A} \in \tilde{\mathcal{A}}} P_e(\tilde{A},c) + \epsilon.$$

Therefore,

$$\begin{aligned}
\mathbf{E}_{\mathcal{C}}\Big[\sup_{A \in \mathcal{A}} P_e(A,\mathcal{C})\Big] &\le \mathbf{E}_{\mathcal{C}}\Big[\sum_{\tilde{A} \in \tilde{\mathcal{A}}} P_e(\tilde{A},\mathcal{C}) + \epsilon\Big] \\
&= \sum_{\tilde{A} \in \tilde{\mathcal{A}}} P_e(\tilde{A}) + \epsilon \ \le |\tilde{\mathcal{A}}| \cdot \overline{P_e} + \epsilon.
\end{aligned}$$

[2] The Poisson functional representation produces a pair $(\tilde{U}, \tilde{M})$, and U is set to the first component of the pair.

[3] The Poisson matching lemma is applied on the conditional distributions given M,S,U,Y,K. Also see the conditional Poisson matching lemma [19].

The proof is completed by the existence of a codebook c such that $\sup_{A \in \mathcal{A}} P_e(A, c) \le |\tilde{\mathcal{A}}| \cdot \overline{P_e} + \epsilon$. □

Note that when $K = \emptyset$, $d_1(s, x) = 0$, and $\mathcal{A} = \{A_{Y|X}\}$ is a singleton set, taking $\hat{A}_{Y|X} = A_{Y|X}$, Theorem 2 reduces to the one-shot Gelfand-Pinsker coding result in [19].

VI. Recovering the Asymptotic Result

We first give a simple bound on the ϵ-covering number in the case that X and Y are discrete and finite.

Proposition 3. *If $\mathcal{X}$ and $\mathcal{Y}$ are finite, then*

$$N_\epsilon(\mathcal{A}) \le \Big(\frac{1}{2\epsilon} + \frac{|\mathcal{Y}|+1}{2}\Big)^{|\mathcal{X}|\cdot|\mathcal{Y}|}.$$

Proof. Write $d(A, \tilde{A}) := \sup_{x \in \mathcal{X}} \|A_{Y|X}(\cdot|x) - \tilde{A}_{Y|X}(\cdot|x)\|_{\mathrm{TV}}$. We use the standard method to bound the covering number, where we start with $\tilde{\mathcal{A}} = \emptyset$, and add $A \in \mathcal{A}$ not currently covered by $\tilde{\mathcal{A}}$ (i.e., $\min_{\tilde{A} \in \tilde{\mathcal{A}}} d(A, \tilde{A}) > \epsilon$) to $\tilde{\mathcal{A}}$ one by one until all of $\mathcal{A}$ is covered. Note that every two different $\tilde{A}, \tilde{A}' \in \tilde{\mathcal{A}}$ produced this way must satisfy $d(\tilde{A}, \tilde{A}') > \epsilon$, and hence the $(\epsilon/2)$-balls $\{A : d(A, \tilde{A}) \le \epsilon/2\}$ must be disjoint for $\tilde{A} \in \tilde{\mathcal{A}}$.

We now treat $A_{Y|X}$ as a transition probability matrix $A \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$. We have $d(A, \tilde{A}) = (1/2)\|A - \tilde{A}\|_1 = (1/2)\max_x \sum_y |A_{y,x} - \tilde{A}_{y,x}|$. The volume of the ball $\{A \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|} : d(A, \tilde{A}) \le \epsilon/2\}$ (i.e., its Lebesgue measure in the space $\mathbb{R}^{|\mathcal{Y}|\cdot|\mathcal{X}|}$) is $((2\epsilon)^{|\mathcal{Y}|}/(|\mathcal{Y}|!))^{|\mathcal{X}|}$, and all these balls are subsets of $\{A \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|} : \min_{x,y} A_{y,x} \ge -\epsilon,\ \max_x \sum_y A_{y,x} \le 1+\epsilon\}$, which has a volume $((1+(|\mathcal{Y}|+1)\epsilon)^{|\mathcal{Y}|}/(|\mathcal{Y}|!))^{|\mathcal{X}|}$. Hence, the size of $\tilde{\mathcal{A}}$ is upper-bounded by

$$\frac{((1+(|\mathcal{Y}|+1)\epsilon)^{|\mathcal{Y}|}/(|\mathcal{Y}|!))^{|\mathcal{X}|}}{((2\epsilon)^{|\mathcal{Y}|}/(|\mathcal{Y}|!))^{|\mathcal{X}|}} = \Big(\frac{1}{2\epsilon} + \frac{|\mathcal{Y}|+1}{2}\Big)^{|\mathcal{X}|\cdot|\mathcal{Y}|}.$$

□

We now show that Theorem 2 recovers the asymptotic result in [1] when S, K, X, Y are finite and discrete, and the attack channel must be memoryless and is subject to a distortion constraint, and hence giving a simple alternative proof to [1]. Consider sequences $S^n = (S_1, \ldots, S_n)$, K^n, X^n, Y^n where $(S_i, K_i) \overset{iid}{\sim} P_{S,K}$. Consider a channel input distribution P_X. The class of attack channels $\mathcal{A}_n = \mathcal{A}_n(P_X)$ (which depends on P_X) is taken to be

$$\mathcal{A}_n(P_X) := \big\{A^n_{Y|X} : A_{Y|X} \in \mathcal{A}(P_X)\big\},$$

$$\mathcal{A}(P_X) := \big\{A_{Y|X} : \mathbf{E}_{(X,Y) \sim P_X A_{Y|X}}[d_2(X, Y)] \le \mathsf{D}_2\big\},$$

and $d_2 : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is a distortion measure, and D_2 is the allowed distortion level. In other words, the attacker can only use memoryless channels $A^n_{Y|X}$ that satisfy the expected distortion constraint $\mathbf{E}[d_2(X, Y)] \le \mathsf{D}_2$. The asymptotic hiding capacity given in [1] is

$$C = \max_{P_{U,X|S,K}} \min_{A_{Y|X} : \mathbf{E}[d_2(X,Y)] \le \mathsf{D}_2} \big(I(U; Y|K) - I(U; S|K)\big).$$

where the maximum is over $P_{U,X|S,K}$ with $\mathbf{E}[d_1(S, X)] \le \mathsf{D}_1$.

We now show the achievability of the above asymptotic rate as a direct corollary of Theorem 2. Fix $P_{U,X|S,K}$ which achieves the above maximum subject to $\mathbf{E}[d_1(S, X)] \le \mathsf{D}_1'$ where $\mathsf{D}_1' < \mathsf{D}_1$. Take $\hat{A}_{Y|X}$ to be the minimizer of the rate-distortion function $\min_{A_{Y|X} : \mathbf{E}[d_2(X,Y)] \le \mathsf{D}_2} I(U; Y|K)$, and assume $(S, K, U, X, Y) \sim P_{S,K} P_{U,X|S,K} \hat{A}_{Y|X}$. Write the information density and mutual information obtained from this distribution as $\hat{\iota}_{U;Y|K}$ and $\hat{I}(U; Y|K)$, respectively. Fix a coding rate $R < \hat{I}(U; Y|K) - I(U; S|K)$. We want to show that this rate is achievable.

Consider any attack channel $A_{Y|X}$ with $\mathbf{E}[d_2(X, Y)] \le \mathsf{D}_2$. Let $A^\lambda_{Y|X} := (1 - \lambda)\hat{A}_{Y|X} + \lambda A_{Y|X}$ for $0 \le \lambda \le 1$. Write $I_\lambda(U; Y|K)$ for the mutual information computed assuming $Y|X \sim A^\lambda_{Y|X}$. It is straightforward to check that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} I_\lambda(U; Y|K)\Big|_{\lambda=0} = \mathbf{E}_{Y|X \sim A_{Y|X}}[\hat{\iota}(U; Y|K)] - \hat{I}(U; Y|K).$$

By the optimality of $\hat{A}$, the above derivative is nonnegative, and hence $\mathbf{E}_{Y|X \sim A_{Y|X}}[\hat{\iota}(U; Y|K)] \ge \hat{I}(U; Y|K)$. Therefore, when we have i.i.d. sequences $(S^n, K^n, U^n, X^n, Y^n) \sim P^n_{S,K} P^n_{U,X|S,K} A^n_{Y|X}$ and $\mathsf{L} = \lfloor 2^{nR} \rfloor$, by law of large numbers,

$$\begin{aligned} &\mathsf{L} 2^{-\hat{\iota}(U^n; Y^n|K^n) + \iota(U^n; S^n|K^n)} \\ &\le 2^{nR - \sum_{i=1}^n (\hat{\iota}(U_i; Y_i|K_i) - \iota(U_i; S_i|K_i))} \to 0 \end{aligned}$$

exponentially as $n \to \infty$ since $\mathbf{E}[\hat{\iota}(U_i; Y_i|K_i) - \iota(U_i; S_i|K_i))] \ge \hat{I}(U; Y|K) - I(U; S|K) > R$. We also have $d_1(S^n, X^n) = \sum_{i=1}^n d_1(S_i, X_i) > n\mathsf{D}_1$ with probability approaching 0 exponentially since $\mathsf{D}_1' < \mathsf{D}_1$. These convergences are uniform over all such attack channels $A_{Y|X}$ since the random variables are discrete and finite.

Therefore, to bound P_e using Theorem 2, it is left to bound the ϵ-covering number $N_\epsilon(\mathcal{A}_n(P_X))$. Note that $\|A^n_{Y|X}(\cdot|x^n) - \tilde{A}^n_{Y|X}(\cdot|x^n)\|_{\mathrm{TV}} \le \sum_{i=1}^n \|A_{Y|X}(\cdot|x_i) - \tilde{A}_{Y|X}(\cdot|x_i)\|_{\mathrm{TV}}$, and hence we can construct a ϵ-cover of $\mathcal{A}_n(P_X)$ using an (ϵ/n)-cover of $\mathcal{A}(P_X)$. Therefore, $N_\epsilon(\mathcal{A}_n(P_X)) \le N_{\epsilon/n}(\mathcal{A}(P_X)) = O((n/\epsilon)^{|\mathcal{X}|\cdot|\mathcal{Y}|})$ by Proposition 3, which grows much slower than the exponential decrease of the expectation in Theorem 2. Therefore, taking $\epsilon = 1/n$, we have $P_e \to 0$ as $n \to \infty$. Taking $\mathsf{D}_1' \to \mathsf{D}_1$ completes the proof.

It is straightforward to convert this to a finite blocklength result where n is a fixed number using the Berry-Esseen theorem [40], [41]. This is left for future studies.

VII. Conclusion

In this paper, we presented a one-shot information-theoretic analysis of the information hiding problem, proved by utilizing the Poisson matching lemma. Compared with the existing asymptotic results, our result applies to any distribution of the host data, and any class of attack channels (not necessarily memoryless or ergodic), and the decoder is uninformed of the attack channel. We showed that our one-shot achievability result recovers the asymptotic result in [1], hence giving a simple alternative proof to [1] where X, K, X, Y are finite and discrete, and the attack channel is memoryless and subject to a distortion constraint.

References

[1] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Transactions on information theory*, vol. 49, no. 3, pp. 563–593, 2003.

[2] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1064–1087, 1998.

[3] A. Somekh-Baruch and N. Merhav, "On the error exponent and capacity games of private watermarking systems," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 537–562, 2003.

[4] ——, "On the capacity game of public watermarking systems," *IEEE Transactions on Information Theory*, vol. 50, no. 3, pp. 511–524, 2004.

[5] A. S. Cohen and A. Lapidoth, "The gaussian watermarking game," *IEEE transactions on Information Theory*, vol. 48, no. 6, pp. 1639–1667, 2002.

[6] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, "Perceptual watermarks for digital images and video," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1108–1126, 1999.

[7] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE transactions on image processing*, vol. 6, no. 12, pp. 1673–1687, 1997.

[8] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1897–1905, 1998.

[9] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. and Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[10] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE transactions on Information theory*, vol. 29, no. 5, pp. 731–739, 1983.

[11] V. Y. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 881–903, 2013.

[12] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[13] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4947–4966, 2009.

[14] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 1–8.

[15] M. H. Yassaee, M. R. Aref, and A. Gohari, "A technique for deriving one-shot achievability results in network information theory," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 1287–1291.

[16] S. Watanabe, S. Kuzuoka, and V. Y. F. Tan, "Nonasymptotic and second-order achievability bounds for coding with side-information," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1574–1605, April 2015.

[17] J. Liu, P. Cuff, and S. Verdú, "One-shot mutual covering lemma and marton's inner bound with a common message," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1457–1461.

[18] E. C. Song, P. Cuff, and H. V. Poor, "The likelihood encoder for lossy compression," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1836–1849, 2016.

[19] C. T. Li and V. Anantharam, "A unified framework for one-shot achievability via the poisson matching lemma," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2624–2651, 2021.

[20] Y. Liu and C. T. Li, "One-shot coding over general noisy networks," in *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024.

[21] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, 2018.

[22] F. M. Willems, "An information theoretical approach to information embedding," in *2000 Symposium on Information Theory in the Benelux, SITB 2000*. Werkgemeenschap voor Informatie-en Communicatietheorie (WIC), 2000, pp. 255–260.

[23] Y. Steinberg and N. Merhav, "Identification in the presence of side information with application to watermarking," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1410–1422, 2001.

[24] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 420–430, 2000.

[25] O. Evsutin, A. Melman, and R. Meshcheryakov, "Digital steganography and watermarking for digital images: A review of current research directions," *IEEE Access*, vol. 8, pp. 166 589–166 611, 2020.

[26] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1079–1107, 1999.

[27] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna rayleigh-fading channels," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 618–629, 2015.

[28] Midjourney, "Midjourney (V5)," https://www.midjourney.com/, 2023, text-to-image model.

[29] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[30] OpenAI, "ChatGPT (Feb 13 version)," https://chat.openai.com, 2023, large language model.

[31] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," *arXiv preprint arXiv:2302.04222*, 2023.

[32] L. Ditria and T. Drummond, "Hey that's mine imperceptible watermarks are preserved in diffusion generated outputs," *arXiv preprint arXiv:2308.11123*, 2023.

[33] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 657–672.

[34] O. Byrnes, W. La, H. Wang, C. Ma, M. Xue, and Q. Wu, "Data hiding with deep learning: A survey unifying digital watermarking and steganography," *arXiv preprint arXiv:2107.09287*, 2021.

[35] C. Zhang, C. Lin, P. Benz, K. Chen, W. Zhang, and I. S. Kweon, "A brief survey on deep learning based data hiding," *arXiv preprint arXiv:2103.01607*, 2021.

[36] D. Blackwell, L. Breiman, A. Thomasian *et al.*, "The capacity of a class of channels," *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1229–1241, 1959.

[37] R. Dobrushin, "Optimum information transmission through a channel with unknown parameters," *Radio Eng. Electron*, vol. 4, no. 12, pp. 1–8, 1959.

[38] J. Wolfowitz, *Simultaneous Channels*. New York: Springer-Verlag, 1980.

[39] J. Scarlett, "On the dispersions of the gel'fand–pinsker channel and dirty paper coding," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4569–4586, 2015.

[40] A. C. Berry, "The accuracy of the Gaussian approximation to the sum of independent variates," *Transactions of the American Mathematical Society*, vol. 49, no. 1, pp. 122–136, 1941.

[41] C.-G. Esseen, "On the Liapunov limit error in the theory of probability," *Ark. Mat. Astr. Fys.*, vol. 28, pp. 1–19, 1942.