On Secure Source Coding with Side Information at the Encoder

Yeow-Khiang Chia
Modulation and Coding Dept.
Institute for Infocomm Research
Email: yeowkhiang@gmail.com

Kittipong Kittichokechai
ACCESS Linnaeus Center
KTH Royal Institute of Technology, Sweden
Email: kki@kth.se

***Abstract*—We consider a secure source coding problem with side informations at the decoder and the eavesdropper. The encoder has a source that it wishes to describe with limited distortion through a rate-limited link to a legitimate decoder. The message sent is also observed by the eavesdropper. The encoder aims to minimize both the distortion incurred by the legitimate decoder; and the information leakage rate at the eavesdropper. When the encoder has access to the side information (S.I.) at the decoder, we characterize the rate-distortion-information leakage rate (R.D.I.) region under a Markov chain assumption and when S.I. at the encoder does not improve the rate-distortion region as compared to the case when S.I. is absent. We then extend our setting to consider the case where the encoder and decoder obtain coded S.I. through a rate-limited helper, and characterize the R.D.I. region for several special cases under logarithmic loss distortion (log-loss). Finally, we consider the case of list or entropy constraints at the decoder and show that the R.D.I. region coincides with R.D.I. region under log-loss.**

I. Introduction

Consider the secure lossy source coding problem with side informations at the decoders in Figure 1. The encoder has source X^n that it wishes to describe through a rate-limited link to a legitimate decoder (denoted simply as a decoder in Figure 1). The message sent is also observed by another decoder, which we denote as an eavesdropper in our setup (see Figure 1). The encoder aims to minimize the distortion incurred by the legitimate decoder in reconstructing the source sequence, while at the same time, minimize the information leakage rate at the eavesdropper given its side information (S.I.) and the common message M: $I(X^n; M, Z^n)/n$.

The problem of source coding with security constraints has received attention in recent years, [1]–[4], due to potential applications in areas such as privacy in sensor networks and databases (see for e.g. [5]). In [3], the authors considered our setting when S.I. Y^n is unavailable at the encoder and gave the full characterization of the rate-distortion-information leakage rate (R.D.I.) region for discrete memoryless sources and arbitrary distortion measures. [4] considered both the case when S.I. Y^n is available at the encoder and the case when S.I. Y^n is unavailable at encoder. However, the authors were interested in the information leakage rate for the S.I., $I(Y^n; M, Z^n)/n$, instead of $I(X^n; M, Z^n)/n$. As we will discuss in the sequel, the differences give rise to a new role, that of generating secret key from common randomness [6], for the S.I. observed at the encoder and decoder.

A particular distortion measure that we will focus on in this paper is the *logarithmic loss* (log-loss) distortion measure, first proposed in [7]. Log-loss has the interesting property that S.I. at the encoder does not improve the rate-distortion region, with respect to the Wyner-Ziv setting [8] where S.I. is absent at the encoder. This property will be key in establishing the results in this paper. Following [9] and [10], we will also extend our work to consider source amplification measures for our setting. We consider the amplification measures of list constraint [11], and the entropy constraint, $H(X^n|M, Y^n)/n$, at the decoder. Interestingly, we find, for our settings, that the R.D.I. region is the same regardless of whether one uses symbol by symbol log-loss or the above amplification measures.

The rest of this paper is as follow. We provide formal definitions in Section II. Section III considers our setting in Figure 1. General inner and outer bounds are given for this setup and the R.D.I. region is characterized when these conditions hold: (i) a Markov Chain $X - Y - Z$ between the source and the side informations; and (ii) S.I. at the encoder does not improve the rate distortion region. Section IV considers the setting in Figure 2, where the encoder and decoder obtain *coded* S.I. sent by a helper via a rate-limited link. Section V deals with the amplification measures listed in the previous paragraph. Several proofs and other results omitted from this paper are deferred to an extended version to be posted online [12].

II. Definitions

We will follow the notation in [13]. Throughout this paper, source and side informations (X^n, Y^n, Z^n, W^n) are assumed to be i.i.d.; i.e. $p(x^n, y^n, z^n, w^n) = \prod_{i=1}^n p(x_i, y_i, z_i, w_i)$.

A. Uncoded S.I. case (Figure 1)

An $(n, 2^{nR})$ code for this setup consists of

- A *stochastic* encoder F_e that takes (X^n, Y^n) as input and generates $M \in [1 : 2^{nR}]$ according to a conditional pmf $p(m|x^n, y^n)$; and
- A decoder $f_D : M \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$.

The *expected distortion* incurred by the code is given by $\mathrm{E}\, d(X^n, \hat{X}^n) := \sum_{i=1}^n Ed(X_i, \hat{X}_i)/n$, and the *information leakage rate* at the eavesdropper is given by $I(X^n; M, Z^n)/n$. A (R, D, Δ) tuple is said to be achievable if there exists a

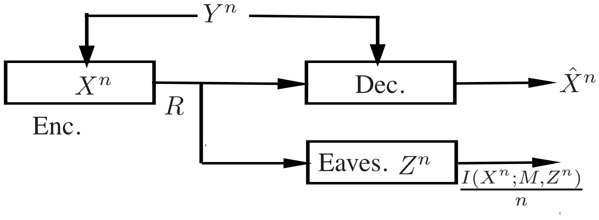

Fig. 1: Uncoded S.I. at the encoder and decoder

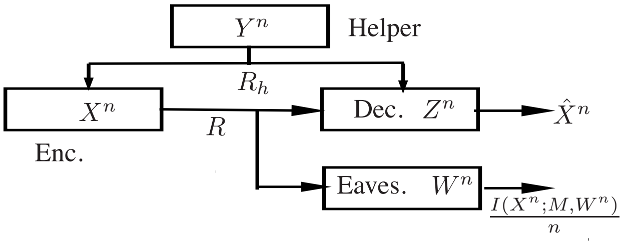

Fig. 2: Rate-limited helper (coded S.I. at encoder and decoder)

sequence of $(n, 2^{nR})$ codes such that

$$\limsup_{n\to\infty} \mathrm{E}\, d(X^n, \hat{X}^n) \le D, \tag{1}$$

$$\limsup_{n\to\infty} \frac{I(X^n; Z^n, M)}{n} \le \Delta. \tag{2}$$

The *rate-distortion-information leakage rate* (R.D.I.) region is then defined as the closure of all achievable (R, D, Δ) tuples.

B. Rate-limited helper case (Figure 2)

The rate-limited helper setting is shown in Figure 2. An $(n, 2^{nR}, 2^{nR_h})$ code for this setup consists of

- A stochastic *helper* encoder F_h that takes Y^n as input and outputs $M_h \in [1 : 2^{nR_h}]$ according to the conditional pmf $p(m_h|y^n)$;
- A stochastic encoder F_e that takes (X^n, M_h) as input and generates $M \in [1 : 2^{nR}]$ according to a conditional pmf $p(m|x^n, m_h)$;
- A decoder $f_D : M \times M_h \times \mathcal{Z}^n \to \hat{\mathcal{X}}^n$.

The definitions of expected distortion incurred by the decoder and information leakage rate at the eavesdropper are the same as previous setting, with Z^n in (2) replaced by W^n for the information leakage rate. A (R, R_h, D, Δ) tuple is said to be achievable if there exists a sequence of $(n, 2^{nR}, 2^{nR_h})$ codes such that (1) and (2) are satisfied. The R.D.I. region is then defined as the closure of all achievable (R, R_h, D, Δ) tuples.

Remark 2.1: We note that the helper encoder is a stochastic encoder. Hence, it can choose to send a combination of *independent randomness* and description of S.I. Y^n to the encoder and decoder. Therefore, the previous setting (Figure 1) is not recovered by simply setting a large enough R_h.

C. Side information and rate distortion region

Let $R_{\mathrm{WZ}}(D)$ be the rate-distortion function for the Wyner-Ziv setting (see [13, Chapter 11]) where S.I. Y^n is available at the decoder only. Let $R_{\mathrm{SI-Enc}}(D)$ be the rate-distortion function when Y^n is also available at the encoder. We say that *S.I. at the encoder does not improve the rate-distortion region* if $R_{\mathrm{WZ}}(D) = R_{\mathrm{SI-Enc}}(D)$ for all $D \ge 0$. We denote this condition by $\mathcal{R}_{\mathrm{WZ}} = \mathcal{R}_{\mathrm{SI-Enc}}$.

III. Uncoded S.I. at encoder and decoder

In this section, we present results for the setting in Figure 1.

A. General inner and outer bounds

Proposition 1: An outer bound to the R.D.I. region for the setting in Figure 1 is given by

$$R \ge I(X; U, V|Y),$$
$$\Delta \ge \max \left\{ \begin{array}{l} I(X; Z), \\ I(X; Z, V, U) + I(V; Z|U) \\ -I(V; Y|U) - H(Y|U, V, X, Z) \end{array} \right\},$$

for some $p(x, y, z)p(u, v|x, y)$ and reconstruction function $\hat{x}(Y, U, V)$ satisfying $\mathrm{E}\, d(X, \hat{x}(Y, U, V)) \le D$. The cardinalities of U and V may be upper bounded by $|\mathcal{U}| \le |\mathcal{X}||\mathcal{Y}| + 2$ and $|\mathcal{V}| \le |\mathcal{X}||\mathcal{Y}| + 2$.
Proof of this proposition is given in [12]. We now present an inner bound (achievability scheme) for this setting.

Proposition 2: An inner bound to the R.D.I. region for the setting in Figure 1 is given by

$$R > I(X; U, V|Y),$$
$$\Delta > I(X; Z, U) + I(V; X|U, Y) - R_K,$$

where $R_K = \min\{I(V; X|U, Y), H(Y|U, V, X, Z)\}$ for $p(u, v, x, y, z) = p(x, y)p(u, v|x, y)p(z|x, y)$ and reconstruction function $\hat{x}(Y, U, V)$ satisfying $\mathrm{E}\, d(X, \hat{x}(Y, U, V)) \le D$.

Before we present a proof sketch for a special case of Proposition 2, we first give some intuition behind the general achievability scheme. The encoder sends two layers of descriptions U^n and V^n to the decoder, which decodes by successive decoding. This results in rates of $I(X; U|Y)$ for the first U^n layer and $I(V; X|U, Y)$ for the second layer. We assume that the eavesdropper is able to decode the U^n codeword, resulting in side information (Z^n, U^n) at the eavesdropper. S.I. Y^n is binned to 2^{nR_K} bins to generated a secret key. This key can be kept secret from the eavesdropper if $R_K < H(Y|U, V, X, Z)$, and it is then used to scramble the message sent to the decoder about the V^n layer of codewords. This operation increases the uncertainty that the eavesdropper has about the V^n codewords. The information leakage rate is then upper bounded by $I(X; Z, U)$ plus $I(V; X|U, Y) - R_K$. $I(V; X|U, Y)$ is an upper bound on the leakage rate due to the V^n codeword if no scrambling was done, while $-R_K$ represents the reduction in the leakage rate due to the secret key scrambling operation.

Note that we did not scramble the first layer of codewords, but a straightforward way of scrambling the first layer of codewords as well as the second layer is to define in the inner bound $U = \emptyset$ and $V' = (V, U)$. Such a scheme leads to the following R.D.I. trade-off.

$$R > I(V'; X|Y)$$
$$= I(V, U; X|Y),$$

$$\begin{aligned}\Delta &> I(X;Z) + I(V';X|Y) - R_K \\ &= I(X;Z) + I(V,U;X|Y) - R_K,\end{aligned}$$

where $R_K = \min\{I(V,U;X|Y), H(Y|U,V,X,Z)\}$.

Sketch of achievability: For simplicity of exposition, we give the proof sketch for only one layer of codewords, V^n (by setting $U = \emptyset$), and defer analysis of the general case, and proofs of some technical bounds to [12].

Codebook generation

We generate two codebooks, the rate-distortion codebook and the key generation codebook. We first start with the rate distortion codebook, $\mathcal{C}_{\mathrm{RD}}$.

- Generate $2^{n(I(V;X,Y)+\delta(\epsilon))}$ $V^n(l_1)$ sequences according to $\prod_{i=1}^n p(v_i)$, $l_1 \in [1:2^{n(I(V;X,Y)+\delta(\epsilon)}]$.
- Partition the set of V^n sequences to $2^{n(I(V;X|Y)+3\delta(\epsilon))}$ bins, $\mathcal{B}_{\mathrm{RD}}(m_1)$, $m_1 \in [1:2^{n(I(V;X|Y)+3\delta(\epsilon))}]$.

This completes the codebook generation for $\mathcal{C}_{\mathrm{RD}}$. We now turn to the key generation codebook, $\mathcal{C}_{\mathrm{K}}$, which has only a single step.

- Randomly and uniformly bin the set of Y^n sequences to 2^{nR_K} bins, $\mathcal{B}_{\mathrm{K}}(m_k)$, where $R_K := \min\{H(Y|V,X,Z), I(V;X|Y)\}$ and $m_k \in [1:2^{nR_K}]$.

We use $\mathcal{C} := \{\mathcal{C}_{\mathrm{RD}}, \mathcal{C}_{\mathrm{K}}\}$ to denote the combined codebook.

Encoding

- Given sequences (x^n, y^n), the encoder first looks for a sequence $v^n(l_1)$ such that $(v^n(l_1), x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}$. If there is more than one such sequence, the encoder selects one sequence uniformly at randomly from the set of jointly typical v^n sequences. If there is none, the encoder randomly and uniformly selects a sequence v^n from the set of all sequences.
- The encoder then looks for the index m_1 such that $v^n(l_1) \in \mathcal{B}_{\mathrm{RD}}(m_1)$.
- Next, it splits the index m_1 into two parts, $m_{1s} \in [1:2^{nR_K}]$ and $m_{1o} \in [1:2^{n(I(V;X|Y)+3\delta(\epsilon)-R_K)}]$.
- The encoder then looks for the index m_k such that $y^n \in \mathcal{B}_{\mathrm{K}}(m_k)$.
- Finally, the encoder sends out the indices m_0, m_{1o} and $m_{1s} \oplus m_k := (m_{1s} + m_{1o})_{2^{nR_K}}$, resulting in a rate of $I(X;V|Y) + 3\delta(\epsilon)$.

Decoding and analysis of distortion

Since the decoder has the sequence y^n, it first finds m_k to unscramble $m_{1s} \oplus m_k$. The rest of the decoding and analysis of expected distortion are fairly standard and we defer them to [12].

Analysis of information leakage rate

We analyze the information leakage averaged over codebooks generated. For notational convenience, we will use $o(\epsilon)$ to represent terms that go to zero as $\epsilon \to 0$.

$$\begin{aligned}n\Delta &= I(X^n;Z^n,M_{1o},M_{1s}\oplus M_K|\mathcal{C}) \\ &= I(X^n,Y^n;Z^n,M_{1o},M_{1s}\oplus M_K|\mathcal{C}) \\ &\quad - I(Y^n;Z^n,M_{1o},M_{1s}\oplus M_K|X^n,\mathcal{C}) \\ &= nI(X,Y;Z) + I(X^n,Y^n;M_{1o},M_{1s}\oplus M_K|Z^n,\mathcal{C}) \\ &\quad - I(Y^n;Z^n,M_{1o},M_{1s}\oplus M_K|X^n,\mathcal{C}) \\ &\overset{(a)}{\le} nI(X,Y;Z) + nI(V;X|Y) + no(\epsilon) \\ &\quad - I(Y^n;Z^n,M_{1o},M_{1s}\oplus M_K|X^n,\mathcal{C}), \qquad (3)\end{aligned}$$

where (a) follows from the size of M_{1o} and $M_{1s} \oplus M_K$. For the remaining term, we have

$$\begin{aligned}&- I(Y^n;Z^n,M_{1o},M_{1s}\oplus M_K|X^n,\mathcal{C}) \\ &= -H(Y^n|X^n,\mathcal{C}) + H(Y^n|X^n,Z^n,M_{1o},M_{1s}\oplus M_K,\mathcal{C}) \\ &\le -nH(Y|X) + H(Y^n,L_1|X^n,Z^n,M_{1o},M_{1s}\oplus M_K,\mathcal{C}) \\ &= -nH(Y|X) + H(L_1|X^n,Z^n,M_{1o},M_{1s}\oplus M_K,\mathcal{C}) \\ &\quad + H(Y^n|X^n,Z^n,L_1,M_k,\mathcal{C}) \\ &\overset{(a)}{\le} -nH(Y|X) + H(L_1|X^n,Z^n,M_{1o},M_{1s}\oplus M_K,\mathcal{C}) \\ &\quad + nH(Y|V,X,Z) - nR_K + no(\epsilon) \\ &\le -nH(Y|X) + H(L_1|X^n,Z^n,\mathcal{C}) \\ &\quad + nH(Y|V,X,Z) - nR_K + no(\epsilon) \\ &\overset{(b)}{\le} -nH(Y|X) + nI(V;Y|X,Z) \\ &\quad + nH(Y|V,X,Z) - nR_K + no(\epsilon) \\ &\le -nI(X;Y|Z) - nR_K + no(\epsilon). \qquad (4)\end{aligned}$$

In (a), we used the bound that if $R_K \le H(Y|V,X,Z)$ and $\mathrm{P}(V^n(L_1),X^n,Y^n,Z^n) \in \mathcal{T}_\epsilon^{(n)}) \to 1$ as $n \to \infty$, then $H(Y^n|X^n,Z^n,L_1,M_k,\mathcal{C}) \le nH(Y|V,X,Z) - nR_K + no(\epsilon)$. Proof of this bound is given in [12]. The condition $\mathrm{P}(V^n(L_1),X^n,Y^n,Z^n) \in \mathcal{T}_\epsilon^{(n)}) \to 1$ as $n \to \infty$ follows from the codebook generation and encoding process. In (b), we used the bound $H(L_1|X^n,Y^n) \le nI(V;Y|X,Z) + no(\epsilon)$. Proof of this bound is given in [12]. Using (4) in (3) then lead us to $\Delta \le n(I(X;Z) + I(V;X|Y) - R_K + o(\epsilon))$. Hence, any $\Delta' > \Delta$ is achievable. ∎

B. R.D.I. regions

Proposition 3: For the setting in Figure 1, if $X - Y - Z$ and $\mathcal{R}_{\mathrm{SI-Enc}} = \mathcal{R}_{\mathrm{WZ}}$, the R.D.I. region is given by

$$\begin{aligned}&R \ge R_{\mathrm{SI-Enc}}(D), \\ &\Delta \ge \max\{I(X;Z), I(X;Z) + R_{\mathrm{SI-Enc}}(D) - H(Y|X,Z)\}.\end{aligned}$$

Proof of this Proposition follows from tightening the outer bound in Proposition 1 using the two conditions and showing achievability using Proposition 2. It is given in [12].

C. Examples

We now provide two examples involving canonical sources and distortion measures in information theory that satisfy the two assumptions stated in the previous subsection.

Corollary 1: Let $X - Y - Z$ and Y be an erased version of X. That is $Y = X$ with probability $1 - p_e$, and e with probability p_e. Let $|\hat{\mathcal{X}}| = |\mathcal{X}|$ and the distortion measure be the Hamming distance:

$$d(X,\hat{X}) = \left\{\begin{array}{ll} 0 & \text{if } \hat{X} = X \\ 1 & \text{if } \hat{X} \neq X \end{array}\right. .$$

Then, the R.D.I. region is given by

$$R \geq p_e I(X;\hat{X}),$$
$$\Delta \geq \max\{I(X;Z), I(X;Z) + p_e I(X;\hat{X}) - H(Y|X,Z)\}$$

for $0 \leq D \leq p_e$, $p(\hat{x}|x)$ such that $\mathrm{E}\, d(X,\hat{X}) \leq D/p_e$.

Proof: The proof follows from an application of Proposition 3 and a result in [14, Theorem 6]. Since $X - Y - Z$ by assumption, it remains to check that $\mathcal{R}_{\mathrm{SI-Enc}} = \mathcal{R}_{\mathrm{WZ}}$, which follows from [14, Theorem 6]. Further, [14, Theorem 6] states that $R_{\mathrm{SI-Enc}}(D) = p_e \min_{p(\hat{x}|x):\mathrm{E}\, d(X,\hat{X}) \leq D/p_e} I(X;\hat{X})$. ■

Corollary 2: Let $X - Y - Z$ and let the distortion measure be given by the log-loss distortion [7]. That is, the reconstruction alphabet is a vector representing the set of probability distributions of the source X. Thus, $\hat{x}(x)$, $1 \leq x \leq |\mathcal{X}|$, represents the x component of the vector $\hat{x}$ that gives the estimated probability of $X = x$. Then, the log-loss measure is defined by

$$d(x,\hat{x}) = \log \frac{1}{\hat{x}(x)}.$$

With this distortion measure, the R.D.I. region is given by

$$R \geq [H(X|Y) - D]^+,$$
$$\Delta \geq \max\{I(X;Z), I(X;Z) + H(X|Y) - D - H(Y|X,Z)\}.$$

Here, $[x]^+ := \max\{0, x\}$.

Proof: This result follows again from a straightforward application of Proposition 3. The fact that $R_{\mathrm{SI-Enc}}(D) = R_{\mathrm{WZ}}(D)$ for arbitrary discrete memoryless X, Y, Z under logarithmic loss follows from results in [7]. Further, [7] showed that $R_{\mathrm{SI-Enc}}(D) = H(X|Y) - D$. ■

IV. Rate-limited helper setting

In this section, we consider the rate-limited helper setting in Figure 2. We begin with a general inner bound for the R.D.I. region.

Proposition 4: An inner bound for the R.D.I. region for the rate-limited helper case in Figure 2 is given by

$$\begin{aligned} R_h &> \max\{I(U_h;Y|Z), I(U_h;Y|X)\},\\ R &> I(X;V|Z,U_h),\\ \Delta &> I(X;W) + I(X;V|Z,U_h)\\ &\quad + I(V;U_h|Y,X) - R_K - R'_K \end{aligned}$$

for some $U_h - Y - (X,Z,W)$, $V - (X,U_h) - (Y,Z,W)$, reconstruction function $\hat{x}(u_h,v,z)$, R_K and R'_K such that $\mathrm{E}\, d(X,\hat{X}) \leq D$, $R_K \leq I(U_h;Y) - I(U_h;X,W,V)$, $R'_K \leq R_h - \max\{I(U_h;Y|Z), I(U_h;Y|X)\}$ and $R_K + R'_K \leq I(X;V|Z,U_h)$.

The proof follows similar lines to that in Proposition 2, with the main differences being the actions of the helper and how the secret key is being generated. To reduce R, the helper sends a description U_h^n to both the encoder and the decoder. To ensure that both the encoder and the decoder can decode U_h^n, we require $R_h > \max\{I(U_h;Y|Z), I(U_h;Y|X)\}$. The secret key is generated in two parts. The first part of the secret key comes from the codeword U_h^n. A secret key of rate R_K can be generated by random binning of the U_h^n codewords if $R_K < I(U_h;Y) - I(U_h;X,W,V)$. Next, the helper can also use its own randomness and the remaining rate to send to the encoder and the decoder a uniform random variable of size $2^{nR'_K}$ as a second secret key. Hence, $R'_K \leq R_h - \max\{I(U_h;Y|Z), I(U_h;Y|X)\}$. These two keys are then used to scramble the message sent on the rate-limited link, which is of rate $I(X;V|U_h,Z)$.

Proof of a more general version of this inner bound, with two layers of codewords, U^n and V^n, instead of just one layer V^n, is given in [12].

In this achievability scheme, there is a tradeoff between the amount of secret key generated and the quality of the description that the helper sends to reduce the rate required by the encoder. The independent randomness sent on the helper link reduces the amount of information leakage through secret key scrambling, but does not help to reduce the distortion at the decoder. While we can generate another secret key using the helper codeword, U_h^n, the rate of the key that can be generated is usually not as large as it would be if uniform randomness is used. In some cases such as those in the next subsection, the tradeoff is tight.

A. Special cases under log-loss

Proposition 5: For the setting in Figure 2, if $Y - X - Z - W$ and the distortion measure is log-loss distortion, then the R.D.I. region is given by

$$\begin{aligned} R_h &\geq I(U_h;Y|Z),\\ R &\geq [H(X|U_h,Z) - D]^+,\\ \Delta &\geq \max\{I(X;W), I(X;W) + H(X|Z,U_h) - D\\ &\qquad - (R_h - I(U_h;X|Z))\} \end{aligned}$$

for some $U_h - Y - X - Z - W$, with $|\mathcal{U}_h| \leq |\mathcal{Y}| + 2$.

This result generalizes some of the results found in [2]. By setting $W = \emptyset$ and $D = 0$[1], we recover [2, Theorem 4] and by setting $Z = \emptyset$ as well, we recover [2, Theorem 2]. Achievability of the R.D.I. region in Proposition 5 for $D \leq H(X|U_h,Z)$ follows from Proposition 4 by setting V to be the following random variable

$$V = \begin{cases} X & \text{with probability } 1 - \frac{D}{H(X|U_h,Z)} \\ \emptyset & \text{otherwise} \end{cases}.$$

The reconstruction function is given by $\hat{x}(u_h,v,z) := p(x|u_h,v,z)$ and it can be verified that this reconstruction function achieves $\mathrm{E}\, d(X,\hat{X}) \leq H(X|U_h,V,Z) = D$.

Next, we note now that the definition of V results in the Markov Chain $V - X - (U_h,Y,Z,W)$. Further, since $U_h - Y - X - Z - W$, we have $I(U_h;Y|Z) \geq I(U_h;Y|X)$. The achievable leakage rate is then given by

$$\Delta > I(X;W) + H(X|Z,U_h) - D - R_K - R'_K$$

[1] Please see Proposition 7, with more details in [12], for proof of the claim that the lossless case in [2] is recovered when $D = 0$.

for $R_K \leq I(U_h;Y) - I(U_h;X)$, $R'_K \leq R_h - I(U_h;Y|Z)$ and $R_K + R'_K \leq I(V;X|Z,U)$. Hence, the achievable Δ is either $I(W;Z)$, or if $R_h - I(U_h;X|Z) < H(X|Z,U_h) - D$, $I(X;W) + H(X|Z,U_h) - D - (R_h - I(U_h;X|Z))$.

Proof of the converse is given in [12]. The identification of the auxiliary random variable U_h and lower bounds for the rates R and R_h follow steps similar to those in [15].

Proposition 6: For the setting in Figure 2, if $W = Z$, $Y - Z - X$ and the distortion measure is log-loss, then the R.D.I. region is given by

$$\begin{aligned}
&R_h \geq 0,\\
&R \geq [H(X|Z) - D]^+,\\
&\Delta \geq \max\{I(X;Z), I(X;Z) + H(X|Z) - D - R_h\}.
\end{aligned}$$

In this setting, side information at the decoder is of higher quality than the side information at the encoder. Any side information sent by the helper does not help to reduce the rate required to achieve a required distortion at the decoder. The helper's only role is to generate a secret key to reduce the information leakage rate. Hence, in this case, there is no tradeoff in the role of the helper between sending a higher quality description versus sending a secret key to reduce the information leakage rate. Proof is given in [12].

It may be of interest to note that the achievability scheme in this proposition relies on a helper with enough independent randomness to generate a secret key of size 2^{nR_h}. The side information Y^n is completely ignored. If, however, the helper is stochastically constrained [16], then Y^n may be used to generate an additional secret key. A complete characterization of in the stochastically constrained case is an open question.

V. Amplification Measures

We now turn our attention to source amplification measures at the decoder. Let U_{dec} be the overall information at the decoder. Instead of symbol by symbol distortion measures like those considered in the previous sections, we consider the following two amplification measures.

1) *List constraint*: Based on the decoder's information, it forms a list, $\mathcal{L}(U_{\text{dec}})$, of x^n sequences such that $|\mathcal{L}(U_{\text{dec}})| \leq 2^{nD_{\text{list}}}$ and $\limsup_{n\to\infty} \mathrm{P}(X^n \notin \mathcal{L}(U_{\text{dec}})) = 0$. The list constraint is a straightforward generalization of lossless source coding, with $D = 0$ corresponding to the lossless case.

2) *Entropy constraint*: Here, we wish to ensure that $\limsup_{n\to\infty} \frac{1}{n} H(X^n|U_{\text{dec}}) \leq D$. The entropy constraint can be shown to be equivalent to *block log-loss* constraint [9], [10]. That is, the decoder's reconstruction vector is the set of all probability distributions over $|\mathcal{X}|^n$, and the loss function is defined as $\log \frac{1}{\hat{x}^n(x^n)}$. Block log-loss is a strengthening of the symbol-by-symbol log-loss distortion measure defined in Corollary 2 since it allows more general probability distributions over $|\mathcal{X}|^n$ instead of only product distributions (in the case of symbol by symbol log loss).

We now consider how the R.D.I. regions change when amplification measures are used.

Proposition 7: For the settings in Corollary 2, and Propositions 5 and 6, the R.D.I. regions remain unchanged if the log-loss measure at the decoder is replaced by a list or entropy constraint.

For the case of entropy constraint, Proposition 7 states that even if we allow more general probability distributions than the product distributions for symbol-by-symbol log-loss, there is no gain in the R.D.I. regions for our settings in restricting attention to product distributions. In the case of list constraint, it relates achievable distortion under log-loss to the exponent of the achievable list size.

Proof of this proposition is given in [12]. A key property used, one that also appears in symbol-by-symbol log-loss, is that $H(X^n|U_{\text{dec}})/n$ is upper bounded by the achievable distortion [10], or the exponent of the list size [11].

Acknowledgment

We thank Prof. Tsachy Weissman of Stanford University, Profs. Mikael Skoglund and Tobias Oechtering of KTH Sweden for helpful discussions. This work was partially supported by the Swedish Research Council.

References

[1] D. Gündüz, E. Erkip, and H. V. Poor, “Lossless compression with security constraints,” in *Proc. IEEE International Symposium on Information Theory*, Toronto, ON, Canada, July 2008, pp. 111–115.

[2] R. Tandon, S. Ulukus, and K. Ramachandran, “Secure source coding with a helper,” *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2178 –2187, June 2013.

[3] J. Villard and P. Piantanida, “Secure lossy source coding with side information at the decoders,” in *48th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, USA, September 2010, pp. 733 –739.

[4] R. Tandon, L. Sankar, and H. V. Poor, “Discriminatory lossy source coding: Side information privacy,” submitted to *IEEE Trans. Inf. Theory*.

[5] L. Sankar, S. R. Rajagopalan, and H. V. Poor, “Utility-privacy tradeoff in databases: An information-theoretic approach,” submitted to IEEE Trans. on Information Forensics and Security. Online: http://arxiv.org/abs/1102.3751.

[6] R. Ahlswede and I. Csiszár, “Common randomness in information theory and cryptography—I: Secret sharing,” *IEEE Trans. Inf. Theory*, vol. 39, no. 4, 1993.

[7] T. Courtade and R. Wesel, “Multiterminal source coding with an entropy- based distortion measure,” in *Proc. IEEE International Symposium on Information Theory*, St. Petersburg, Russia, Aug 2011, pp. 2040–2044.

[8] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[9] T. Courtade, “Information masking and amplification: The source coding setting,” in *Proc. IEEE International Symposium on Information Theory*, Boston, MA, USA, July 2012, pp. 189–193.

[10] T. Courtade and T. Weissman, “Multiterminal source coding under logarithmic loss,” *IEEE Trans. Inf. Theory*, to appear.

[11] Y. H. Kim, A. Sutivong, and T. Cover, “State amplification,” *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1850–1859, May 2008.

[12] Y. K. Chia and K. Kittichokechai, “On secure source coding with side information at the encoder,” 2013, in preparation. To be posted online at ArXiv.

[13] A. El Gamal and Y. H. Kim, *Network Information Theory*, 1st ed. Cambridge University Press, 2011.

[14] E. Perron, S. Diggavi, and E. Teletar, “The kaspi rate-distortion problem with encoder side-information: Binary erasure case, licos-report-2006-004,” École polytechnique fédérale de Lausanne, Tech. Rep., 2007.

[15] H. Permuter, Y. Steinberg, and T. Weissman, “Two-way source coding with a helper,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2905 –2919, June 2010.

[16] S. Watanabe and Y. Oohama, “Broadcast channels with confidential messages by randomness constrained stochastic encoder,” in *Proc. IEEE International Symposium on Information Theory*, Boston, MA, USA, July 2012, pp. 61 –65, extended version available online at ArXiv.