Single-Shot Lossy Compression for Joint Inference and Reconstruction

Oguzhan Kubilay Ulger ou2007@nyu.edu, Elza Erkip elza@nyu.edu

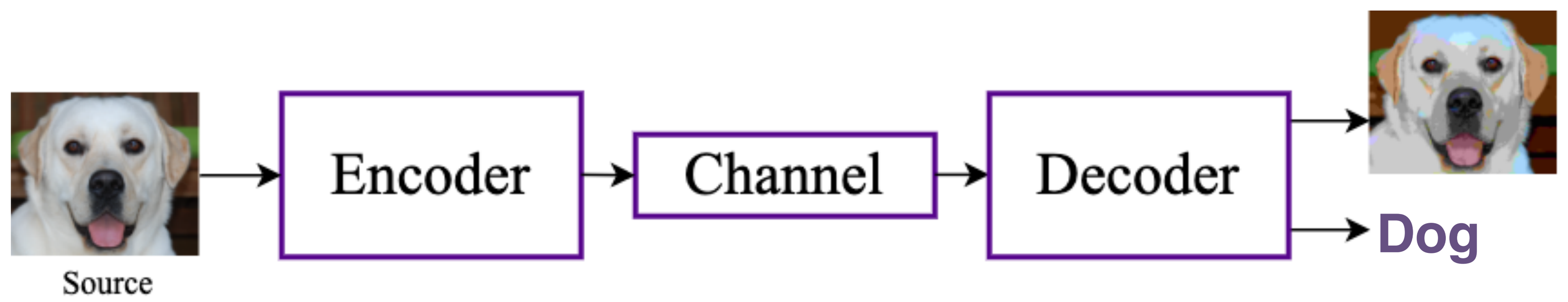
NYU TANDON SCHOOL OF ENGINEERING — NYU WIRELESS

I. Introduction

Motivation: Compression scenarios where we are interested not only reconstructing the source but also making inferences from it.

- Image compression with classification
- Speech compression with underlying text

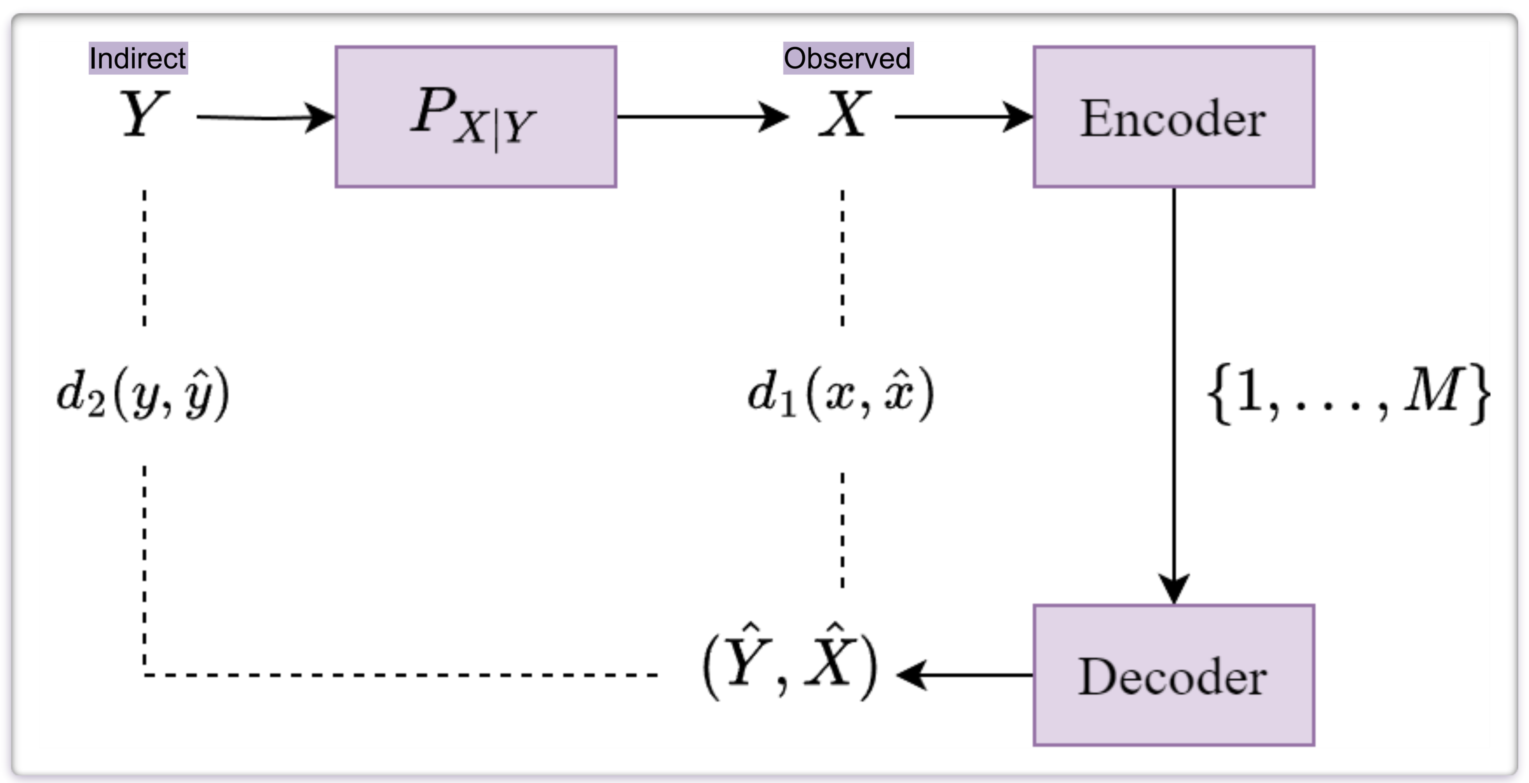
In both examples, **both reconstruction and the inference** tasks are desirable.



Our approach:

- Model the setting as a combination of **direct and indirect source coding** problem.
- Assume that the source has two parts: **a direct part** that is directly observed by the encoder and **an indirect part** that has to be inferred.
- **Single-shot source coding** approach with **excess distortion** probability constraint. Single-shot setting is relevant in settings requiring low latency (sensor networks in **autonomous cars**) and the **modern neural compressors** work in a single shot manner.

II. System Model



III. Objectives

Excess distortion probability: The probability of exceeding either distortion levels.

$$\mathbb{P}\left[d_1(X,\hat{X}) > D_1 \cup d_2(Y,\hat{Y}) > D_2\right] \le \epsilon$$

Find **achievability** and **converse** bounds for minimizing the excess distortion probability under fixed length **single-shot rate constraint**.

IV. Large Blocklength

- (X^n, Y^n) sampled i.i.d. from $P_{X,Y}$ and **jointly compress block of** n with $n \to \infty$. expected distortion constraint. The rate distortion function is given by [1]:

$$R(D_1,D_2) = \min_{p(\hat{x},\hat{y}|x)} I(X;\hat{X},\hat{Y})$$
$$\text{s.t. } \mathbb{E}[d_1(X,\hat{X})] \le D_1$$
$$\mathbb{E}[\hat{d}_2(X,\hat{Y})] \le D_2$$
$$\text{where } \hat{d}_2(x,\hat{y}) = \mathbb{E}[d_2(Y,\hat{y})\,|\,x]$$

V. Single-Shot Bounds

- We can extend the single-shot achievability and converse results from [2],[3] to obtain general bounds for our problem.

A. Achievability: The optimal excess distortion probability is less than:

$$\epsilon^*(M,D_1,D_2) \le \inf_{P_{\hat{X}\hat{Y}}} \int_0^1 \mathbb{E}\left[\mathbb{P}\left[\pi(X,\hat{X},\hat{Y}) > t \,\middle|\, X\right]^M\right] dt$$

- The result is based on random coding
- Generate a codebook of M symbols
- For every x, the encoder picks the codeword that gives lowest probability of exceeding either distortion levels

B. Converse: Excess distortion probability for any encoder/decoder pair has to be greater than:

$$\epsilon^*(M,D_1,D_2) \ge \inf_{P_{\hat{X}\hat{Y}|X}} \sup_{\gamma\ge 0} \left\{\mathbb{P}\left[J_{X;\hat{X}^*,\hat{Y}^*}(X,Y,\hat{X},\hat{Y},D_1,D_2) \ge \log M + \gamma\right] - 2^{-\gamma}\right\}$$

- These general bounds are **hard to compute** due to optimization over probability distributions on $\mathcal{X} \times \mathcal{Y}$. Especially when support of X is large.

VI. Logarithmic Loss Case

We can obtain special bounds for the case when the distortion metric of X is logarithmic loss $d_1(x,\hat{x}) = -\log \hat{x}(x)$ where reconstructions are probability distributions (soft reconstruction).

A. Achievability: Using properties of logarithmic loss we can obtain a special bound that removes dependency on $P_{\hat{X}}$

$$\epsilon^*(M,D_1,D_2) \le \inf_{P_{\hat{Y}}} \inf_{\gamma\ge 0} \inf_{0\le\epsilon'\le 1} \Big\{\epsilon'\left(1 - \mathbb{E}\left[\eta(\epsilon')^M\right]\right)$$
$$+\mathbb{E}\left[\eta(\epsilon')^M\right](1 + 2^{1-\gamma})$$
$$+2^{1-\gamma}\sum_{k=1}^{M}\binom{M}{k}\frac{M}{k}\mathbb{E}[\eta(\epsilon')^{M-k}(1-\eta(\epsilon'))^k]$$
$$+\mathbb{P}[\imath_X(X) > D_1 + \log M - \gamma]\Big\}$$

$$J_{X;\hat{X}^*\hat{Y}^*}(x,y,\hat{x},\hat{y},D_1,D_2) = \imath_{X;\hat{X}^*\hat{Y}^*}(x;\hat{x},\hat{y}) + \lambda_1(d_1(x,\hat{x}) - D_1) + \lambda_2(d_2(y,\hat{y}) - D_2)$$
$$\pi(x,\hat{x},\hat{y}) = \mathbb{P}\left[\{d_1(X,\hat{x}) > D_1\} \cup \{d_2(Y,\hat{y}) > D_2\}\,|\,X = x\right]$$
$$\eta(\epsilon') = \mathbb{P}[\pi'(X,\hat{Y}) > \epsilon'\,|\,X]$$

B. Converse: The general bounds divides into two parts depending on which distortion constraint more restricting

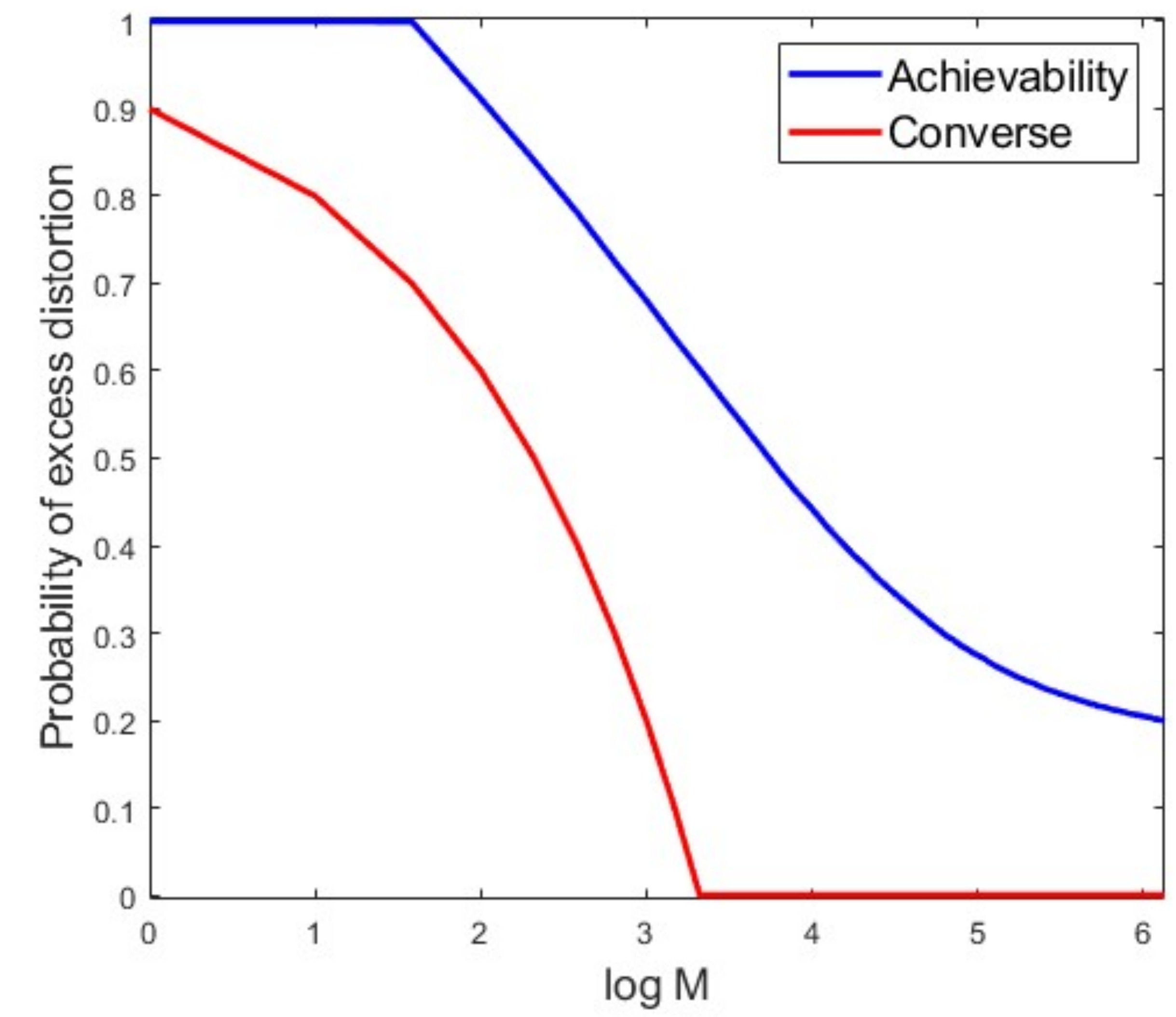
- When D_1 is small compared to D_2

$$\epsilon^*(M,D_1,D_2) \ge \sup_{\gamma\ge 0}\left\{\mathbb{P}\left[\imath_X(X) \ge D_1 + \log M + \gamma\right] - 2^{-\gamma}\right\}$$

- When D_2 is small compared to D_1

$$\epsilon^*(M,D_1,D_2) \ge \inf_{P_{\hat{Y}|X}} \sup_{\gamma\ge 0} \left\{\mathbb{P}\left[J_{X;\hat{Y}^*}(X,Y,\hat{Y},D_2) \ge \log M + \gamma\right] - 2^{-\gamma}\right\}$$

C. Example: Numerical example with $|\mathcal{X}| = 70$ and $|\mathcal{Y}| = 10$. Specialized bounds allows us to calculate the bounds when alphabet of X is large.



$Y \sim \text{Uniform}\{0,\ldots,9\}$ and $X \sim \text{Binom}\{n=7, p=0.1\}$ given Y with non-overlapping alphabets. $d_1(x,\hat{x})$ is logarithmic loss and $d_2(y,\hat{y})$ is Hamming distortion.

VII. Conclusion and References

- We study the joint inference and reconstruction problem and characterized upper and lower bounds to excess distortion probability.
- We obtain specialized achievability bound for the case where direct distortion metric is log-loss.
- Possible future direction is looking at the case where inference task is unknown and comes from a class of tasks.

[1] J. Liu, S. Shao, W. Zhang, and H. Vincent Poor, "An indirect rate-distortion characterization for semantic sources: General model and the case of gaussian observation," IEEE Transactions on Communications, pp. 1–1, 2022.

[2] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," IEEE Transactions on Information Theory, vol. 58, no. 6, pp. 3309–3338, 2012.

[3] V. Kostina and S. Verdú, "Nonasymptotic noisy lossy source coding," IEEE Transactions on Information Theory, vol. 62, no. 11, pp. 6111–6123, 2016.