Lossy Compression Theory (Achievability)

John MacLaren Walsh, Ph.D.

ECET 602, Winter Quarter, 2013

1 References

- *Elements of Information Theory*, T. M. Cover and J. A. Thomas, John Wiley and Sons, 1991.
- *Information Theory and Reliable Communication*, R. G. Gallager, John Wiley and Sons, 1968.
- *Information Theory: Coding Theorems For Discrete Memoryless Systems.*, I. Csiszar, J. Koerner, Academic Press, 1981.

These proof technique in these notes follows extremely closely that in Chapter 13 in the first reference (Chapter 10 in the second edition).

2 Overview

This lecture covers:

- achievability of the information rate distortion function via a random coding argument employing distortion typical sequences.

3 Ingredients

Consider two sequences $X^N = (X_1, \ldots, X_N)$, and $\hat{X}^N = (X_1, \ldots, X_N)$, with $\{(X_i, \hat{X}_i)\}$ i.i.d. from a joint distribution $\mathsf{p}_{X,\hat{X}}$. Recall that the motivation for the set of jointly typical sequences was based on the convergence in probability of

$$-\frac{1}{N}\log(\mathsf{p}(X^N)) \to \mathcal{H}(X)$$
$$-\frac{1}{N}\log(\mathsf{p}(\hat{X}^N)) \to \mathcal{H}(\hat{X})$$
$$-\frac{1}{N}\log(\mathsf{p}(\hat{X}^N, X^N)) \to \mathcal{H}(\hat{X}, X)$$

for which we provided an easy proof using the weak law of large numbers. The jointly typical set was then defined to be the set of deterministic realizations $x^n, \hat{x}^N$ for $\hat{X}^N, X^N$, for which each of these three convergences had happened within a common ϵ.

In order to prove the achievability for rate distortion, we observed another convergence in probability, namely if $\hat{X}^N, X^N$ are specified as above, then additionally, we must have

$$\frac{1}{N}\sum_{i=1}^{N} d(X_i, \hat{X}_i) \to \mathbb{E}\left[d(X, \hat{X})\right] \tag{1}$$

in probability as $N \to \infty$. This is clear through the WLLN because the i.i.d. nature of $X_i, \hat{X}_i$ implies that $d(X_i, \hat{X}_i)$ must also be i.i.d.

We can then consider a subset of the jointly typical sequences, called the *distortion jointly typical sequences* $\mathcal{A}^N_{\epsilon,d}$ for which the convergence (1) has also occurred within the same ϵ.

Naturally, then, the convergences in probability we have mentioned make that the probability $\Pr\left[(X^N, \hat{X}^N) \in \mathcal{A}_{\epsilon,d}^N\right] \to 1$ as $N \to \infty$. Another interesting property is for $(\hat{x}^N, x^N) \in \mathcal{A}_{\epsilon,d}^N$ that we can bound $\mathsf{p}(\hat{x}^N)$ using $\mathsf{p}(\hat{x}^N|x^N)$ according to

$$\mathsf{p}(\hat{x}^N) \geq \mathsf{p}(\hat{x}^N|x^N)2^{-N(\mathcal{I}(X;\hat{X})+3\epsilon)} \tag{2}$$

We proved this with the chain of inequalities

$$\mathsf{p}(\hat{x}^N|x^N) = \mathsf{p}(\hat{x}^N)\frac{\mathsf{p}(x^N,\hat{x}^N)}{\mathsf{p}(x^N)\mathsf{p}(\hat{x}^N)} \leq \mathsf{p}(\hat{x}^N)\frac{2^{-N(\mathcal{H}(X,\hat{X})-\epsilon)}}{2^{-N(\mathcal{H}(\hat{X})+\epsilon)}2^{-N(\mathcal{H}(X)+\epsilon)}} = \mathsf{p}(\hat{x}^N)2^{N\left(\mathcal{I}(X;\hat{X})+3\epsilon\right)}$$

which proves the proposition.

We presently use these two new tools to prove the achievability of rate distortion codes.

4 Rate Distortion Theory: Achievability

Here, we construct a lossy source code with rate $R = R^I(D)$ and distortion arbitrarily close to D as the block length $N \to \infty$. Consider a source X_i i.i.d. according to p_X, with a bounded distortion metric $d \leq d_{max}$ and information rate distortion function

$$R^I(D) := \max_{\mathsf{p}(\hat{X}|X):\sum \mathsf{p}(\hat{X}|X)\mathsf{p}(X)d(\hat{X},X)\leq D} \mathcal{I}(\hat{X};X) \tag{3}$$

The first step in our construction is to choose a $\mathsf{p}_{\hat{X}|X}$ achieving the maximum in (3) for a distortion D. This induces (together with the distribution for the source symbols) a distribution for $\hat{X}$ that is $\mathsf{p}_{\hat{X}}(\hat{x}) = \sum_{x\in\mathcal{X}} \mathsf{p}_{\hat{X}|X}(\hat{x}|x)\mathsf{p}_X(x)$.

The next step is our construction is to generate a $2^{NR} \times N$ matrix $\mathbf{C}$ of codewords, with the wth row denoted by $\hat{X}^N(w)$, $w \in \{1, \ldots, 2^{NR}\}$, by sampling each element of the matrix i.i.d. from the distribution $\mathsf{p}_{\hat{X}}(\hat{x})$.

Encoding of an observation $X^N = x^N$ is accomplished by sending w if the wth row of the codebook $\hat{x}^N(w)$ is distortion jointly typical with x^N. If there is more than one such w we transmit the smallest, and if there is no such w, we transmit $w = 1$.

Decoding is accomplished at the decoder by mapping w to $\hat{X}^N(w)$ (via knowledge of the codebook).

We presently prove that this code has expected (over both word to encode X^N and codebook $\mathbf{C}$) distortion $D+\delta$ with δ vanishing as $N \to \infty$. To do this, we consider the following two possible situations

- There exists a w such that $\hat{x}^N(w)$ and x^N are in the jointly distortion typical set $\mathcal{A}_{\epsilon,d}^N$. This set of sequences x^N and codebooks $\mathbf{C}$ can not have more probability than one, and furthermore, by definition of the jointly distortion typical set the distortion $d(x^n, \hat{x}^N(w)) < D + \epsilon$, thus these sequences contribute no more that $D + \epsilon$ to the overall distortion.
- The sequence x^N and codebook $\mathbf{C}$ are such that there does not exist such a w with the wth codeword jointly distortion typical with x^N. Let the collective probability of these sequences and codebook pairs be denoted by P_e, then these sequences can contribute no more than $P_e d_{max}$ to the overall expected distortion.

Thus, having established that the expected distortion is bounded above by $D+\epsilon+P_e d_{max}$, all that remains to show is that P_e can be made arbitrarily small as $N \to \infty$. To do this, we obtain an upper bound on P_e. First off all, note that

$$P_e = \sum_{\mathbf{C}} \mathsf{p}(\mathbf{C}) \sum_{x^N|\nexists \hat{x}^N\in\mathbf{C} \text{ s.t.}(x^N,\hat{x}^N)\in\mathcal{A}_{\epsilon,d}^N} \mathsf{p}_{X^N}(x^N)$$

we can rewrite this as

$$P_e = \sum_{x^N\in\mathcal{X}^N} \mathsf{p}_{X^N}(x^N) \sum_{\mathbf{C}|\nexists \hat{x}^N\in\mathbf{C} \text{ s.t.}(x^N,\hat{x}^N)\in\mathcal{A}_{\epsilon,d}^N} \mathsf{p}(\mathbf{C})$$

Now, considering the fact that the rows of the code matrix (i.e. the codewords) are selected independently of one another in an identically distributed, we can write the result of the latter sum as the product over

the 2^{NR} different code words of the probability that one selected code word is not jointly typical with the given x^N. Thus, we have

$$P_e = \sum_{x^N \in \mathcal{X}^N} \mathsf{p}_{X^N}(x^N) \left(1 - \sum_{\hat{x}^N | (x^N, \hat{x}^N) \in \mathcal{A}^N_{\epsilon,d}} \mathsf{p}_{\hat{X}^N}(\hat{x}^N)\right)^{2^{NR}}$$

We can bound the probability in the parenthesis using (2) to get

$$P_e \leq \sum_{x^N \in \mathcal{X}^N} \mathsf{p}_{X^N}(x^N) \left(1 - \sum_{\hat{x}^N | (x^N, \hat{x}^N) \in \mathcal{A}^N_{\epsilon,d}} \mathsf{p}(\hat{x}^N | x^N) 2^{-N(\mathcal{I}(X;\hat{X})+3\epsilon)}\right)^{2^{NR}}$$

We then use the deterministic bound that for $0 \leq x, y \leq 1, N > 0$ $(1-xy)^N \leq 1 - x + e^{-yN}$ to get

$$\begin{aligned} P_e &\leq \sum_{x^N \in \mathcal{X}^N} \mathsf{p}_{X^N}(x^N) \left(1 - \sum_{\hat{x}^N | (x^N, \hat{x}^N) \in \mathcal{A}^N_{\epsilon,d}} \mathsf{p}(\hat{x}^N | x^N) + e^{-2^{-N(\mathcal{I}(X;\hat{X})+3\epsilon)+NR}}\right) \quad (4) \\ &= 1 - \sum_{(x^N, \hat{x}^N) \in \mathcal{A}^N_{\epsilon,d}} \mathsf{p}(\hat{x}^N, x^N) + e^{-2^{-N(\mathcal{I}(X;\hat{X})+3\epsilon)+NR} \quad (5)} \end{aligned}$$

The first two terms are then recognized as one minus the probability of the distortion typical set, which we know to be $\leq \epsilon$. Finally, remembering that we selected $\mathsf{p}(\hat{X}|X)$ to achieve the maximum in $R^I(D)$ for D, we observe that the last term may be rewritten to yield

$$P_e \leq \epsilon + e^{-2^{N(R-R^I(D)-3\epsilon)}}$$

which goes to zero as $N \to \infty$ if $R \geq R^I(D) + 3\epsilon$. We have thus established achievability of the rate distortion function.