Lecture 1: Capacity of Finite State Channels

Jun Chen

1 Single-User Information Theory:

Definition 1.1. *A sequence $x^n \in \mathcal{X}^n$ is said to be ϵ-strongly typical with respect to a distribution $p(x)$ on $\mathcal{X}$ if*

1. *For all $a \in \mathcal{X}$ with $p(a) > 0$, we have*
$$\left|\frac{1}{n}N(a|x^n) - p(a)\right| < \frac{\epsilon}{|\mathcal{X}|},$$
2. *For all $a \in \mathcal{X}$ with $p(a) = 0$, $N(a|x^n) = 0$.*

$N(a|x^n)$ is the number of occurrences of the symbol a in the sequence x^n. The set of sequences $x^n \in \mathcal{X}^n$ such that x^n is strongly typical is call the strongly typical set and is denoted $A_\epsilon^n(X)$.

Theorem 1.1. *In the following, δ is a small positive quantity such that $\delta \to 0$ as $\epsilon \to 0$.*

1. *If $x^n \in A_\epsilon^n(X)$, then*
$$2^{-n(H(X)+\delta)} \le p(x^n) \le 2^{-n(H(X)-\delta)}.$$
2. *For n sufficiently large,*
$$Pr(X^n \in A_\epsilon^n(X)) > 1 - \epsilon.$$
3. *For n sufficiently large,*
$$(1-\epsilon)2^{n(H(X)-\delta)} \le |A_\epsilon^n(X)| \le 2^{n(H(X)+\delta)}.$$

Proof. To prove property 1, we write
$$p(x^n) = \prod_{a\in\mathcal{X}} p(a)^{N(a|x^n)}.$$
Then
$$\begin{aligned}
\log p(x^n) &= \sum_{a\in\mathcal{X}} N(a|x^n)\log p(a) \\
&= \sum_{a\in\mathcal{X}} (N(a|x^n) - np(a) + np(a))\log p(a) \\
&= n\sum_{a\in\mathcal{X}} p(a)\log p(a) - n\sum_{a\in\mathcal{X}}\left(\frac{1}{n}N(a|x^n) - p(a)\right)(-\log p(a)) \\
&= -n\left[H(X) + \sum_{a\in\mathcal{X}}\left(\frac{1}{n}N(a|x^n) - p(a)\right)(-\log p(a))\right].
\end{aligned}$$

If $x^n \in A_\epsilon^n(X)$, then

$$\sum_{a\in\mathcal{X}} \left|\frac{1}{n}N(a|x^n) - p(a)\right| \le \epsilon,$$

which implies

$$\begin{aligned}
& \left|\sum_{a\in\mathcal{X}} \left(\frac{1}{n}N(a|x^n) - p(a)\right)(-\log p(a))\right| \\
\le\ & \sum_{a\in\mathcal{X}} \left|\left(\frac{1}{n}N(a|x^n) - p(a)\right)\right|(-\log p(a)) \\
\le\ & -\log\left(\min_{a\in\mathcal{X}} p(a)\right)\sum_{a\in\mathcal{X}} \left|\left(\frac{1}{n}N(a|x^n) - p(a)\right)\right| \\
\le\ & -\log\left(\min_{a\in\mathcal{X}} p(a)\right)\epsilon \\
\triangleq\ & \delta.
\end{aligned}$$

Property 2 follows by the weak law of large numbers.

To prove property 3, we first write

$$\begin{aligned}
|A_\epsilon^n(X)|\, 2^{-n(H(X)+\delta)} &= \sum_{x^n\in A_\epsilon^n(X)} 2^{-n(H(X)+\delta)} \\
&\le \sum_{x^n\in A_\epsilon^n(X)} p(x^n) \\
&\le 1,
\end{aligned}$$

which implies

$$|A_\epsilon^n(X)| \le 2^{n(H(X)+\delta)}.$$

We also have

$$\begin{aligned}
|A_\epsilon^n(X)|\, 2^{-n(H(X)-\delta)} &= \sum_{x^n\in A_\epsilon^n(X)} 2^{-n(H(X)-\delta)} \\
&\ge \sum_{x^n\in A_\epsilon^n(X)} p(x^n) \\
&\ge 1-\epsilon,
\end{aligned}$$

which implies

$$|A_\epsilon^n(X)| \ge (1-\epsilon)2^{n(H(X)-\delta)}.$$

$\square$

Remark: Property 1 is actually the definition of weak typical sequences. So in this sense, strong typicality is indeed “stronger” than weak typicality.

Lossless Source Coding: Since there are at most $2^{n(H(X)+\delta)}$ strongly typical sequences, we can use $2^{n(H(X)+\delta)}$ codewords, each of length $n(H(X)+\delta)$, to represent these sequences.

For those non-typical sequences, we don't compress them, thus the length is still n. For this coding scheme, the average codeword length is upper bounded by

$$n(H(X)+\delta)+\Pr(X^n \notin A_\epsilon^n(X))n \leq n(H(X)+\delta)+n\epsilon,$$

and the rate is upper bounded by $H(X)+\epsilon+\delta$.

Definition 1.2. *A pair of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is said to be ϵ-strongly typical with respect to a distribution $p(x,y)$ on $\mathcal{X} \times \mathcal{Y}$ if*

1. *For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) > 0$, we have*

$$\left|\frac{1}{n}N(a,b|x^n,y^n)-p(a,b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|},$$

2. *For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) = 0$, $N(a,b|x^n,y^n) = 0$.*

$N(a,b|x^n,y^n)$ is the number of occurrences of the pair (a,b) in the pair of sequences (x^n,y^n). The set of sequences $(x^n,y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that (x^n,y^n) is strongly typical is called the strongly typical set and is denoted $A_\epsilon^n(X,Y)$.

Theorem 1.2. *Let (X^n,Y^n) be sequences of length n drawn i.i.d. according to $p(x^n,y^n) = \prod_{i=1}^n p(x_i,x_i)$. Then*

1. *For n sufficiently large,*

$$Pr((X^n,Y^n) \in A_\epsilon^n(X,Y)) > 1-\epsilon.$$

2. *For n sufficiently large,*

$$(1-\epsilon)2^{n(H(X,Y)-\lambda)} \leq |A_\epsilon^n(X,Y)| \leq 2^{n(H(X,Y)+\lambda)}.$$

3. *If $(\widetilde{X}^n,\widetilde{Y}^n) \sim p(x^n)p(y^n)$, i.e., $\widetilde{X}^n$ and $\widetilde{Y}^n$ are independent with the same marginals as $p(x^n,y^n)$, then for n sufficiently large, we have*

$$(1-\epsilon)2^{-n(I(X;Y)+\lambda)} \leq Pr((\widetilde{X}^n,\widetilde{Y}^n) \in A_\epsilon^n(X,Y)) \leq 2^{-n(I(X;Y)-\lambda)}$$

Here λ is a small positive quantity such that $\lambda \to 0$ as $\epsilon \to 0$.

Theorem 1.3. *For any $x^n \in A_\epsilon^n(X)$, define*

$$A_\epsilon^n(Y|X)(x^n) = \left\{y^n : (x^n,y^n) \in A_\epsilon^n(X,Y)\right\}.$$

If $|A_\epsilon^n(Y|X)(x^n)| \geq 1$, then

$$2^{n(H(Y|X)-\nu)} \leq |A_\epsilon^n(Y|X)(x^n)| \leq 2^{n(H(Y|X)+\nu)},$$

where $\nu \to 0$ as $n \to \infty$ and $\epsilon \to 0$.

Channel Coding Theorem:

$$C = \max_{p(x)} I(X;Y).$$

Independently generate 2^{nR} codewords $\{X^n(i)\}_{i=1}^{2^{nR}}$ according to $\prod_{j=1}^{n} p(x_j)$. Suppose $X^n(1)$ is transmitted and Y^n is the channel output. $X^n(1)$ and Y^n are strongly typical with high probability. For codeword $X^n(i)$, $i \neq 1$, the probability that $X^n(i)$ with strongly typical with Y^n is upper bounded by $2^{-n(I(X;Y)-\lambda)}$. Thus by the union bound, the probability of existing some codeword $X^n(i)$, $i \neq 1$, such that $X^n(i)$ and Y^n are strongly typical is upper bounded by $2^{nR}2^{-n(I(X;Y)-\lambda)}$, which goes to 0 if $R < I(X;Y) - \lambda$. Since λ can be arbitrarily close to 0 as $n \to \infty$, the channel coding theorem follows by optimizing the input distribution.

Theorem 1.4. *Let $Y_1, Y_2, \cdots, Y_n$ be drawn i.i.d. $\sim \prod p(y)$. For $x^n \in A_\epsilon^n(X)$, the probability that $(x^n, Y^n) \in A_\epsilon^n(X,Y)$ is bounded by*

$$2^{-n(I(X;Y)+\epsilon_1)} \leq \Pr((x^n, Y^n) \in A_\epsilon^n(X,Y)) \leq 2^{-n(I(X;Y)-\epsilon_1)},$$

where ϵ_1 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Rate Distortion Theorem:

$$R(D) = \min_{p(y|x):\mathbb{E}d(X,Y)\leq D} I(X;Y).$$

Independently generate 2^{nR} codewords $\{Y^n(i)\}_{i=1}^{2^{nR}}$ according to $\prod_{j=1}^{n} p(y_j)$. Given $X^n = x^n$, if it is not strongly typical (its probability is no greater than ϵ), then we declare an error and incur the maximum distortion $d_{\max}$. Otherwise we search for a codeword $Y^n(i)$ that is strongly typical with x^n. The probability that such a codeword does not exist can be bounded as

$$\begin{aligned}
\Pr((x^n, Y^n(i)) \notin A_\epsilon^n(X,Y), \forall i) &= \prod_{i=1}^{2^{nR}} (\Pr((x^n, Y^n(i)) \notin A_\epsilon^n(X,Y))) \\
&\leq \left(1 - 2^{-n(I(X;Y)+\epsilon_1)}\right)^{2^{nR}} \\
&\to 0
\end{aligned}$$

if $R > I(X;Y) + \epsilon_1$. Note: ϵ_1 can be arbitrarily small as $n \to \infty$. Furthermore, if x^n and $Y^n(i)$ are strongly joint typical, we have

$$\frac{1}{n}\sum_{j=1}^{n} d(x_j, Y_j(i)) = \frac{1}{n}\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}}^{n} N(a,b|x^n, Y^n(i))d(a,b)$$

and

$$\mathbb{E}d(X;Y) - \epsilon d_{\max} \leq \frac{1}{n}\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}}^{n} N(a,b|x^n, Y^n(i))d(a,b) \leq \mathbb{E}d(X;Y) + \epsilon d_{\max}.$$

Therefore, as $n \to \infty$, the average distortion is upper bounded by $\mathbb{E}d(X,Y) + 2\epsilon d_{\max}$, which can be made arbitrarily to $\mathbb{E}d(X,Y)$. Now the rate distortion theorem follows by optimizing $I(X;Y)$ over all $p(y|x)$ such that $\mathbb{E}d(X,Y) \leq D$.

Remark: For the case $d_{\max} = \infty$, we need a more careful analysis of the tail distribution of $p(x)$.

2 Channel Capacities with Different Knowledge of Channel State Information

Channel model: $p(Y(t)|X^t, S^t) = p(Y(t)|X(t), S(t))$. The state process $\{S(t)\}_{t=-\infty}^{\infty}$ is stationary and ergodic with stationary distribution $\{\pi(s)\}_{s\in\mathcal{S}}$, and is unaffected by the input and output processes, i.e., $p(S(t)|S^{t-1}, X^{t-1}, Y^{t-1}) = p(S(t)|S^{t-1})$.

1. No Knowledge of Channel State Information:

   The channel capacity is given by

   $$C = \limsup_{n\to\infty} \frac{1}{n} \max_{p(x^n)} I(X^n; Y^n).$$

   When $\{S(t)\}_{t=-\infty}^{\infty}$ is an i.i.d. process, then the channel becomes memoryless, and the capacity formula can be simplified to

   $$C = \max_{p(x)} I(X; Y).$$

   Here the channel transition probability is given by

   $$p(y|x) = \sum_{s\in\mathcal{S}} \pi(s)p(y|x, s).$$

   $\mathcal{S}$ is the state space.

   Proof. For the general stationary and ergodic channels, we need to invoke the asymptotic equipartition property (AEP) for stationary and ergodic processes, which in turn relies on the ergodic theorem. For memoryless channels, the AEP for i.i.d. processes suffices, which only requires the law of large numbers. □

   Remark: For non-stationary (or non-asymptotically-stationary) channels, the AEP does not apply anymore since the law of large numbers or more generally the ergodic theorem do not hold for non-stationary processes. For this case, the notion of channel capacity is still well-defined. Verdu and Han has found a general capacity formula for arbitrarily channels based on a new method called information spectrum. We also want to point out that the renowned Gallager's bounding technique (which is mainly used to derive the error exponent) also does not use the asymptotic equipartition property.

2. Perfect Channel State Information at Receiver:

   Firstly, we assume the channel is memoryless, i.e., $\{S(t)\}_{t=\infty}^{\infty}$ is an i.i.d. process. In this case, we can view $\{Y(t), S(t)\}$ as the channel output, and thus the capacity formula is given by

   $$C = \max_{p(x)} I(X; Y, S).$$

   Since the transmitter does not have the channel state information, it follows that $I(X; S) = 0$, which yields

   $$I(X; Y, S) = I(X; Y|S) + I(X; S) = I(X; Y|S).$$

Thus the capacity formula can be simplified to

$$C = \max_{p(x)} I(X;Y|S).$$

It turns out that even for the general stationary and ergodic channels, this capacity formula is still true.

Proof. Let $\{X(t)\}$ be an i.i.d. process. We have

$$\begin{aligned}
I(X^n;Y^n,S^n) &= I(X^n;Y^n|S^n) \\
&= H(Y^n|S^n) - H(Y^n|S^n,X^n) \\
&= \sum_{t=1}^{n} H(Y(t)|S(t)) - H(Y(t)|S(t),X(t)) \\
&= \sum_{t=1}^{n} I(X(t);Y(t)|S(t)) \\
&= nI(X;Y|S).
\end{aligned}$$

Therefore,

$$\begin{aligned}
C &= \limsup_{n\to\infty} \frac{1}{n} \max_{p(x^n)} I(X^n;Y^n,S^n) \\
&\geq \max_{p(x)} I(X;Y|S).
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
I(X^n;Y^n,S^n) &= I(X^n;Y^n|S^n) \\
&= H(Y^n|S^n) - H(Y^n|S^n,X^n) \\
&= \sum_{t=1}^{n} [H(Y(t)|Y_1^{t-1},S^n) - H(Y(t)|Y_1^{t-1},S^n,X^n)] \\
&\leq \sum_{t=1}^{n} [H(Y(t)|S(t)) - H(Y(t)|S(t),X(t))] \\
&= \sum_{t=1}^{n} I(X(t);Y(t)|S(t)) \\
&\leq n \max_{p(x)} I(X;Y|S),
\end{aligned}$$

which completes the proof. □

It is interesting to see that even when the channel has memory, it still possesses a single-letter characterization of its capacity when the receiver has the knowledge of channel state channel information. The single-letter characterization implies that the channel capacity depends only on the stationary distribution of the state process $\{\pi(s)\}_{s\in\mathcal{S}}$ (over the state space) but not on the joint distribution of the state process (over time). It is because the channel becomes memoryless conditioned on the realization of channel state process. This should be contrasted with the capacity result of case 1 where we need to know the joint distribution of the state process to carry out the maximization.

Remark: There is no difference as to the channel capacity whether the receiver knows the state process in a causal or non-causal way. This is because the receiver can always wait until it receives all the channel outputs before starting to decode.

3. Perfect Causal Channel State Information at Transmitter:

   To simplify the analysis, we assume $\{S(t)\}$ is an i.i.d. process.

   The channel capacity is given by

$$C = \max_{p(u)} I(U; Y)$$

   where $U \in \mathcal{X}^{|\mathcal{S}|}$ is a random vector of length $|\mathcal{S}|$ with elements in $\mathcal{X}$ and probability distribution $p(u)$. A code for this channel is a set of $|\mathcal{W}|$ sequences of length n of vectors $u \in \mathcal{X}^{|\mathcal{S}|}$. For a given source message $w \in \mathcal{W}$, the codeword $u^n(w)$ is selected. At each time t, the channel input is given by $x(t) = u_t(w, s(t))$, where $u_t(w, s)$ denotes the sth element of the vector $u_t(w)$. A codeword $u^n(w)$ defines a sequence of n functions $\mathcal{S} \to \mathcal{X}$. This coding scheme essentially converts a memoryless channel with state information at transmitter to a memoryless vector channel without state information. Now the direct coding theorem follows immediately.

   Now we proceed to prove the converse. By Fano's inequality, in order to find an upper bound on the channel capacity, we just need to bound $\frac{1}{n} I(W; Y^n)$. We have

$$\begin{aligned}
& \frac{1}{n} I(W; Y^n) \\
= \ & \frac{1}{n} [H(Y^n) - H(Y^n|W)] \\
\leq \ & \frac{1}{n} \left[\sum_{t=1}^{n} H(Y(t)) - H(Y^n|W) \right] \\
\leq \ & \frac{1}{n} \left[\sum_{t=1}^{n} H(Y(t)) - \sum_{t=1}^{n} H(Y(t)|W, Y^{t-1}, S^{t-1}) \right].
\end{aligned}$$

   Let $U(t) = (f_t(W, S^{t-1}, S(t) = s))_{s \in \mathcal{S}} \in \mathcal{X}^{|\mathcal{S}|}$. Note that $U(t)$ is a function of W and S^{t-1}. Moreover, $(W, Y^{t-1}, S^{t-1}) - U(t) - Y(t)$ form a Markov chain. Therefore, we have

$$\begin{aligned}
& \frac{1}{n} \left[\sum_{t=1}^{n} H(Y(t)) - \sum_{t=1}^{n} H(Y(t)|W, Y^{t-1}, S^{t-1}) \right] \\
= \ & \frac{1}{n} \sum_{t=1}^{n} [H(Y(t)) - H(Y(t)|U(t))] \\
= \ & \frac{1}{n} \sum_{t=1}^{n} I(U(t); Y(t)) \\
\leq \ & \max_{p(u)} I(U; Y).
\end{aligned}$$

4. Perfect Non-causal Channel State Information at Transmitter: Gelfand-Pinsker Theorem.

   We will discuss this after introducing the random binning technique.

5. Perfect Channel State Information at both Transmitter and Receiver:

   The channel capacity is given by

   $$C = \sum_{s\in\mathcal{S}} \pi(s) \max_{p(x|s)} I(X;Y|S=s).$$

   Again, the capacity has a single-letter characterization even when the channel has memory.

   This capacity has an intuitive interpretation. Since both the transmitter and the receiver know the state realization, they can use the multiplexing technique to decompose the channel to $|\mathcal{S}|$ memoryless channels, each corresponding to a state $s \in \mathcal{S}$. For each state s, the optimal coding scheme is used to achieve the memoryless channel capacity with respect to state s, i.e.,

   $$C_s = \max_{p(x|s)} I(X;Y|S=s).$$

   Since the state process is stationary and ergodic, given n time slots, the number of times that the state has realization s is approximately $n\pi(s)$ as n is large enough. Therefore, the average rate is

   $$\sum_{s\in\mathcal{S}} \pi(s) C_s = \sum_{s\in\mathcal{S}} \pi(s) \max_{p(x|s)} I(X;Y|S=s).$$

   Note: Although the multiplexing method gives us a simple way to achieve that channel capacity, it is not optimal in the sense of error exponent since the dependency in the state process is not fully exploited. Furthermore, this multiplexing method requires extremely long block coding when the state space is large, so it may cause long delay in decoding.

6. Perfect Channel State Information at Receiver and Delayed Channel State Information at Transmitter:

   Suppose at time t, the transmitter has the knowledge of $\cdots, S(t-d-1), S(t-d)$. In order to get a single-letter characterization of channel capacity, we assume $\{S(t)\}$ is an irreducible Markov process. For this case, the capacity is given by

   $$C = \sum_{s'\in\mathcal{S}} \pi(s') \max_{p(x|s')} I(X(t);Y(t)|S(t), S(t-d)=s'),$$

   where

   $$I(X(t);Y(t)|S(t), S(t-d)=s') = \sum_{s\in\mathcal{S}} K^d(s',s) I(X(t);Y(t)|S(t)=s, S(t-d)=s')$$

   and K is the one-step transition matrix of the state process.

   The multiplexing technique still works. But now we shall multiplex the codewords over the delayed state process. The intuitive interpretation of this capacity formula is that if $S(t-d)=s'$, then $P(S(t)=s|S(t-d)=s') = K^d(s',s)$ for $s \in \mathcal{S}$. Now we imagine the situation that state process is stationary with stationary distribution $\{K^d(s',s)\}_{s\in\mathcal{S}}$. Since the receiver knows the state process (see case 2), the capacity, conditioned on $S(t-d)=s'$, is thus given by

   $$\begin{aligned} C_{s'} &= \max_{p(x|s')} I(X(t);Y(t)|S(t), S(t-d)=s') \\ &= \max_{p(x|s')} \sum_{s\in\mathcal{S}} K^d(s',s) I(X(t);Y(t)|S(t)=s, S(t-d)=s'). \end{aligned}$$

Now average over $\pi(s')$, we get

$$C = \sum_{s' \in \mathcal{S}} \pi(s') C_{s'} = \sum_{s' \in \mathcal{S}} \pi(s') \max_{p(x|s')} I(X(t); Y(t)|S(t), S(t-d) = s').$$

If the Markov chain is irreducible and aperiodic, then by the Markov convergence theorem $K^d(s', s) \to \pi(s)$ as $d \to \infty$ and becomes independent of s'. So

$$C_{s'} \to \max_{p(x)} I(X; Y|S) \text{ as } d \to \infty, \quad \forall s' \in \mathcal{S},$$

and consequently $C \to \max_{p(x)} I(X; Y|S)$ too. This is to say, if the delay is too large, then the capacity converges to the case where only the receiver has the state information. Of course, this is not necessarily true if the state process is periodic. For example, $\mathcal{S} = \{0, 1\}$ and $p(S(t) \neq S(t-1)) = 1$. Now as long as we know $S(t)$ for some t, we can predict all the future state realizations. Thus for this case, the delay of the state information at the transmitter does not affect the channel capacity, i.e., the capacity is the same for all finite d. But when $d = \infty$, i.e., the transmitter does not have the state information, then the capacity reduces, why? (comparing the capacity formulas of case 2 and case 5).

References

[1] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: Wiley, 1991.

[2] R. Yeung, A First Course in Information Theory, Kluwer Academic/ Plenum Publishers, 2002.

[3] C. E. Shannon, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, pp.289-293, Oct. 1958.

[4] G. Caire and S. Shamai, "On the capacity of some channels with channel state information," *IEEE Tran. Inform. Theory*, vol. 45, pp. 2007-2019, Sept. 1999.

[5] J. Wolfowitz, Coding Theorems of Information Theory. Berlin, Prentice-Hall, 1961.

[6] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Tran. Inform. Theory*, vol. 43, pp.1986-1992, Nov. 1997.

[7] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Tran. Inform. Theory*, vol. 45, pp. 761-771, Mar. 1999.

[8] S. C. Tatikonda, "Control Under Communication Constraints," PhD dissertation, MIT, 2000.

[9] D. Blackwell, L. Breiman, and A. J. Thomansian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Stat.*, 29, pp. 1209-1220, 1958.

[10] R. G. Gallager, Information Theory and Reliable Communication. New York, Wiley, 1968.

[11] S. Verdu and T. S. Han, "A general formula for channel capacity," *IEEE Tran. Inform. Theory*, vol. 40, pp. 1147-1157, July 1994.