Information and Coding Theory **Winter 2021**

Lecture 11: February 16, 2021

Lecturer: Madhur Tulsiani

1 Matrix Scaling

We will consider an application of I-projections to a problem known as matrix scaling. Say we are given two nonnegative matrices $M, N \in \mathbb{R}_+^{n\times n}$ such that for all i, j, $M_{ij} = 0 \Leftrightarrow N_{ij} = 0$. The goal is to multiply (scale) each row i of M by a number r_i and each column j by c_j, such that the resulting matrix M' has the same row and column sums as the target matrix N. Another way of stating this is that we want to find diagonal matrices D_1 and D_2 such that for $M' = D_1 M D_2$, we have

$$\sum_j M'_{ij} \;=\; \sum_j N_{ij} \quad \forall i \in [n] \qquad \text{and} \qquad \sum_i M'_{ij} \;=\; \sum_i N_{ij} \quad \forall j \in [n]\,.$$

We will show a special case when the goal is to scale M so that the resulting matrix M' is doubly stochastic i.e.,

$$\sum_j M'_{ij} \;=\; \sum_i M'_{ij} \;=\; 1 \quad \forall i, j \in [n]\,.$$

First, note that by a *global* scaling of $1/\sum_{i,j} M_{ij}$, we can assume that $\sum_{i,j} M_{ij} = 1$, and the goal is instead to scale it to have row and column sums equal to $1/n$ i.e.,

$$\sum_j M'_{ij} \;=\; \sum_i M'_{ij} \;=\; \frac{1}{n} \quad \forall i, j \in [n]\,.$$

We can now think of this as a problem of going from one distribution to another. Assume that $M_{ij} > 0$ for all i, j, and think of the target matrix N with $N_{ij} = 1/n^2$ for all i, j. Since the entries of M are positive and sum to 1, we can think of it as a probability distribution Q with $\mathrm{Supp}(Q) = [n] \times [n]$ (where $q(i, j) = M_{ij}$). We consider the linear family of distributions on $[n] \times [n]$ (written as matrices) with the required row and column sums.

$$\mathcal{L} \;:=\; \left\{ P \;\middle|\; \sum_j p(i,j) \;=\; \sum_i p(i,j) \;=\; \frac{1}{n} \quad \forall i, j \in [n] \right\}$$

Note that the above is a linear family as defined in the previous lecture, since we can consider functions $f_1, \ldots, f_n$ and $g_1, \ldots, g_n$ defined as

$$f_{i_0}(i,j) \;=\; \begin{cases} 1 & \text{if } i = i_0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad g_{j_0}(i,j) \;=\; \begin{cases} 1 & \text{if } j = j_0 \\ 0 & \text{otherwise} \end{cases}.$$

Then, the above family can be written in terms of the expectations of the functions f_i and g_j for all $i, j \in [n]$. Moreover, we know from the previous lecture that the I-projection P^* of Q onto $\mathcal{L}$ is of the form

$$\begin{aligned} p^*(i,j) &= c_0 \cdot q(i,j) \cdot \exp\left(\sum_{i_0} \lambda_{i_0} \cdot f_{i_0}(i,j) + \sum_{j_0} \mu_{j_0} \cdot g_{j_0}(i,j)\right) \\ &= c_0 \cdot q(i,j) \cdot \exp\left(\lambda_i + \mu_j\right) \\ &= \left(\sqrt{c_0} \cdot \exp(\lambda_i)\right) \cdot M_{i,j} \cdot \left(\sqrt{c_0} \cdot \exp(\mu_j)\right) . \end{aligned}$$

Thus, we can define the diagonal matrices D_1 and D_2 as

$$(D_1)_{ii} = \sqrt{c_0} \cdot \exp(\lambda_i) \qquad \text{and} \qquad (D_2)_{jj} = \sqrt{c_0} \cdot \exp(\mu_j) .$$

We then have that the distribution p^* given by the resulting matrix $M' = D_1 M D_2$, belongs to the linear family $\mathcal{L}$. Thus, the row and column sums of M' are $1/n$. Combining this with another global scaling (replace $\sqrt{c_0}$ by $\sqrt{c_0 \cdot n}$) we can also get all the row and column sums to be 1 (i.e., make the matrix doubly stochastic).

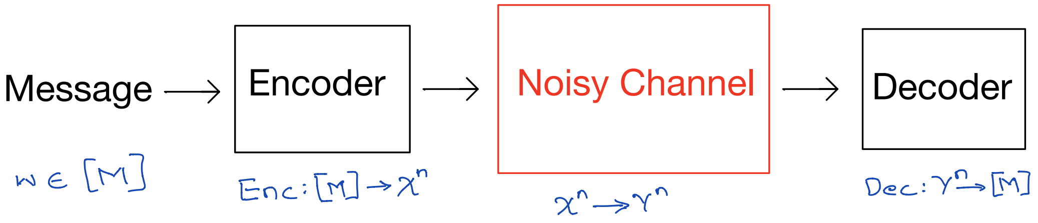
Exercise 1.1. *Where did we use the fact that $M_{ij} > 0$ for all $i, j \in [n]$?*

Exercise 1.2. *Use this the above techniques to solve the matrix scaling problem for an arbitrary target matrix N (assuming $M_{ij} = 0 \Leftrightarrow N_{ij} = 0$).*

Matrix scaling and its generalization, known as operator scaling have found a variety of applications in combinatorial optimization, complexity theory and analysis. Please take a look at the recent tutorial by Wigderson [Wig17] for an introduction to many of these connections.

2 Error-Correcting Codes

Over the next few lectures, we will switch to the "coding theory" part of the course and see how to construct (and work with) error correcting codes. We first define the model(s) we will be working with, and consider the tasks that we will aim for.


Message
Encoder
Noisy Channel
Decoder
$w \in [M]$
$\mathrm{Enc}:[M]\to\mathcal{X}^n$
$\mathcal{X}^n \to \mathcal{Y}^n$
$\mathrm{Dec}:\mathcal{Y}^n\to[M]$

As illustrated above, the goal is to communicate a message, which we will take to be one of M possible values in $[M] = \{1, \dots, M\}$. The problem is that the transmission goes through a "channel" which introduces some form of noise. The goal is then to "encode" the message as a length-n string in a finite alphabet (say) $\mathcal{X}$, in such a way that even after the noise, the received noisy transmission (say in $\mathcal{Y}^n$) can be decoded to the intended message (maybe with high probability). A code is thus specified by two (not necessarily efficiently computable) maps $\mathsf{Enc} : [M] \to \mathcal{X}^n$ and $\mathsf{Dec} : \mathcal{Y}^n \to [M]$. The specific requirements from these maps, depend on the error model we consider.

Shannon model. In this setting, we will think of the message being a random variable W chosen uniformly randomly in $[M]$, and the errors introduced by the channel as being the result of a probabilistic process. The goal will be design (deterministic) maps Enc and Dec such that

$$\mathbb{P}\left[\mathsf{Dec}(\text{Noisy-Transmission}(\mathsf{Enc}(W))) = W\right] \quad\longrightarrow\quad 1\,,$$

where the probability is over the choice of W, and the errors introduced by the noisy transmission through the channel. In particular, we will limit our discussion to *discrete memoryless channels* where the "discrete" part refers to $\mathcal{X}$ and $\mathcal{Y}$ being finite, and the "memoryless" property refers to the fact that the noise acts independently on each of the n symbols transmitted, and does not depend on what happens to the previous symbols. The channel is thus specified by a collection of probability distributions $P(Y|X)$ where X and Y refer to the input and output of the channel for a *single* transmission.

Hamming model. Here we think of the message as an arbitrary element of $[M]$, and the errors introduced by the channel as adversarial. Of course, we need some assumption, as otherwise all transmissions can be mapped to a single output string with no hope of recovery. We will thus assume a bound on the *number of positions* where $\overline{\mathbf{x}} \in \mathcal{X}^n$ is corrupted. In particular, we take $\mathcal{Y} = \mathcal{X}$, but assume $\overline{\mathbf{x}} \in \mathcal{X}^n$ is corrupted in at most t positions, to give the output $\overline{\mathbf{y}} \in \mathcal{X}^n$. We will want codes which satisfy

$$\forall w \in [M] \quad \mathsf{Dec}(\text{Noisy-Transmission}(\mathsf{Enc}(w))) = w\,,$$

if the noisy transmission corrupts the string $\overline{\mathbf{x}} = \mathsf{Enc}(W)$ in at most t (arbitrary) positions.

We will first discuss the Shannon model of errors for a couple of lectures, before considering codes for the Hamming model.

2.1 Discrete memoryless channels

As discussed above, a discrete memoryless channel is specified by a collection of probability distributions, or a transition matrix, specifying the probabilities $\mathbb{P}[Y = y | X = x]$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and X and Y are random variables corresponding respectively to the input and the output. The limits on how efficiently we can transmit data through a channel, will be specified by a quantity called the channel capacity, defined as

$$C \;:=\; \max_{P(X)} I(X;Y)\,,$$

where we maximize over all the distributions $P(X)$ for the input X, and the joint distribution of (X,Y) is then given by the transition matrix corresponding to the channel. Consider the following examples.

Example 2.1 (Noiseless channel). *We take $\mathcal{X} = \mathcal{Y} = \{0,1\}$, and have $\mathbb{P}[Y = b_1 | X = b_2] = 1$ if $b_1 = b_2$ and 0 otherwise i.e., symbols are transmitted without any corruption. It is easy to see that $Y = X$ and*

$$C \;=\; \max_{P(X)} I(X;Y) \;=\; \max_{P(X)} H(X) \;=\; 1\,.$$

Example 2.2 (Noiseless random channel). *Let $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{a,b,c,d\}$ with $\mathbb{P}[Y = a|X = 0] = \mathbb{P}[Y = b|X = 0] = 1/2$, $\mathbb{P}[Y = c|X = 0] = \mathbb{P}[Y = d|X = 0] = 1/2$, and all other probabilities being 0. Thus, the value of Y still uniquely specifies X, and we have*

$$C \;=\; \max_{P(X)} I(X;Y) \;=\; \max_{P(X)} \{H(X) - H(X|Y)\} \;=\; \max_{P(X)} H(X) \;=\; 1\,.$$

Example 2.3 (Binary symmetric channel: $\mathsf{BSC}(p)$). *We take $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and*

$$\mathbb{P}[Y = b_1 | X = b_2] \;=\; \begin{cases} 1-p & \text{if } b_1 = b_2 \\ p & \text{if } b_1 \neq b_2 \end{cases}.$$

From the symmetry of the above transition matrix, we can say that $Y = X + Z \mod 2$ where $Z = 1$ with probability p and 0 with probability $1-p$. Using this, we get

$$I(X;Y) \;=\; H(Y) - H(Y|X) \;=\; H(X+Z) - H((X+Z)|X) \;=\; H(X+Z) - H_2(p)\,.$$

Since $H(X+Z) \leq 1$ and is maximized when $P(X)$ is uniform over $\{0,1\}$, we get that for the binary symmetric channel, $C = 1 - H_2(p)$.

Exercise 2.4 (Binary erasure channel)**.** *Take $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,1,\perp\}$ and define the transition probabilities as*

$$\mathbb{P}\left[Y = b_1 | X = b_2\right] \;=\; \begin{cases} 1-p & \text{if } b_1 = b_2 \\ p & \text{if } b_1 = \perp \end{cases} ,$$

i.e., each symbol in $\{0,1\}$ is transmitted without error with probability $1-p$, and is converted to $\perp$ (erasure) with probability p (but 0 is never converted to 1, and vice-versa). Prove that $C = 1-p$.

2.2 Channel coding theorem

We next discuss Shannon's channel coding theorem which says that the maximum achievable efficiency for any code is given by the channel capacity. To measure the efficiency of a code, we need a few definitions.

Recall that a code is specified by maps $\mathsf{Enc} : [M] \to \mathcal{X}^n$ and $\mathsf{Dec} : \mathcal{Y}^n \to [M]$. We will also take the map Enc to be injective so that no two messages have the same encoding. We can think of this process as the following Markov chain

$$W \;\to\; \left(\mathsf{Enc}(W) = \overline{\mathbf{X}}\right) \;\to\; \overline{\mathbf{Y}} \;\to\; \left(\widehat{W} = \mathsf{Dec}(\overline{\mathbf{Y}})\right) ,$$

where $\overline{\mathbf{X}} \in \mathcal{X}^n$ and $\overline{\mathbf{Y}} \in \mathcal{Y}^n$ denote the input and output sequences for the channel. We take the probability of error for a code to be $p_e = \mathbb{P}\left[W \neq \widehat{W}\right]$, where the probability is over the choice of the message, and the noise in the channel.

Definition 2.5. *We define the* rate *of a code as above to be*

$$R \;:=\; \frac{\log M}{n} \qquad (\textit{bits per transmission})\,.$$

We say that a rate R is achievable *for a channel, if there exists a sequence of codes for $n \geq n_0$ with rates at least R and error probabilities $p_e^{(n)}$ such that $p_e^{(n)} \to 0$ as $n \to \infty$. We define the maximum achievable rate for a channel as $R^* = \sup\{R \mid R \text{ is achievable}\}$.*

Remark 2.6. *Often achievable rates are defined with respect to the* maximum *probability of error $\lambda_e = \max_{w \in [M]} \mathbb{P}\left[\widehat{W} \neq W \mid W = w\right]$, where the probability is taken only over the noise in the channel. However, it will be slightly more convenient to work with the average error probability p_e for us. Moreover, you can check that R^* does not change according to the two notions. In particular, check that if there exists a sequence of codes with rate at least R such that $p_e^{(n)} \to 0$, then for every $\varepsilon > 0$, there exists a sequence of codes with rate at least $R - \varepsilon$ such that $\lambda_e^{(n)} \to 0$ (simply discard messages for which probability of error larger than $2p_e$).*

We can now state Shannon's channel coding theorem for discrete memoryless channels.

Theorem 2.7. *For any discrete memoryless channel, we have $R^* = C$.*

While we will prove that $R^* \leq C$ for every channel, we will only prove $R^* \geq C$ for the binary symmetric channel. The proof idea for the case of general channels is similar and (using a random code) but the analysis a bit more cumbersome.

2.3 Channel capacity as an upper bound on achievable rates

We first prove the following.

Proposition 2.8. *Let R be any achievable rate for a given channel with capacity C. Then, $R \leq C$.*

Proof: Since R is achievable, there exists a sequence of codes with encoding lengths (also called block-lengths) $n \geq n_0$ and rate at least R, such that $p_e^{(n)} \to 0$. Consider any such code with block-length n and rate $\log(M)/n \geq R \Rightarrow M \geq 2^{nR}$. Recall that we think of the transmission process as the Markov chain

$$W \;\to\; \left(\mathsf{Enc}(W) = \overline{\mathbf{X}}\right) \;\to\; \overline{\mathbf{Y}} \;\to\; \left(\widehat{W} = \mathsf{Dec}(\overline{\mathbf{Y}})\right) .$$

Applying Fano's inequality gives that for $p_e^{(n)} = \mathbb{P}\left[\widehat{W} \neq W\right]$, we have

$$1 + p_e^{(n)} \cdot \log(M) \;\;\geq\;\; H_2(p_e^{(n)}) + p_e^{(n)} \cdot \log(\mathrm{Supp}(W) - 1) \;\;\geq\;\; H\left(W|\widehat{W}\right)$$

We can write the RHS of the inequality above as

$$H\left(W|\widehat{W}\right) \;\;=\;\; H(W) - I\left(W;\widehat{W}\right) \;\;\geq\;\; \log(M) - I\left(W;\overline{\mathbf{Y}}\right) \;\;=\;\; nR - I\left(\overline{\mathbf{X}};\overline{\mathbf{Y}}\right) .$$

Finally, we can analyze the mutual information term as follows (using $Y_{<i}$ as a shorthand for $Y_1, \ldots, Y_{i-1}$)

$$\begin{aligned} I(\overline{\mathbf{X}};\overline{\mathbf{Y}}) &= H(\overline{\mathbf{Y}}) - H(\overline{\mathbf{Y}}|\overline{\mathbf{X}}) \\ &= \sum_{i=1}^{n} \left(H(Y_i \mid Y_{<i}) - H(Y_i \mid Y_{<i}, \overline{\mathbf{X}})\right) \\ &\leq \sum_{i=1}^{n} \left(H(Y_i) - H(Y_i \mid X_i)\right) \\ &= \sum_{i=1}^{n} I(X_i;Y_i) \\ &\leq n \cdot C \end{aligned}$$

Note that we used above that $H(Y_i \mid Y_{<i}, \overline{\mathbf{X}}) = H(Y_i|X_i)$ by the memoryless property of the channel, and that for all i, $I(X_i; Y_i) \leq C$. Combining the above bounds, we get

$$1 + p_e^{(n)} \cdot \log(M) \;\geq\; \log(M) - nC \quad \Rightarrow \quad R \cdot (1 - p_e^{(n)}) \;\leq\; C + \frac{1}{n} \,.$$

Taking the limit $n \to \infty$ then gives $R \leq C$ as desired. ■

2.4 Achieving capacity for the binary symmetric channel

We will show next that a random collection of codewords (called *codebook* or simply *code*) can achieve capacity for the binary symmetric channel $\mathrm{BSC}(p)$. Recall that the capacity for the channel is $1 - H_2(p)$. We will show that for every $\varepsilon > 0$, there is a sequence of codes with rate at least $1 - H_2(p) - \varepsilon$ and vanishing probability of error. We can assume that $p < 1/2$ (why?)

The code construction. For parameters R to be chosen later, let $M = 2^{nR}$. We define the codewords, the maps Enc and Dec as below. We will use $\Delta(\overline{\mathbf{x}}, \overline{\mathbf{y}})$ for $\overline{\mathbf{x}}, \overline{\mathbf{y}} \in \{0,1\}$ to denote the Hamming distance, i.e., the number of positions in which the two strings differ.

- **Codewords:** Select M independent random codewords $\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_M \in \{0,1\}^n$ with each bit of each codeword chosen independently and uniformly at random in $\{0,1\}$. Here we are using the fact that for $\mathrm{BSC}(p)$, the distribution $P(X)$ maximizing the mutual information is uniform on $\{0,1\}$. For the case of general channels and alphabet $\mathcal{X}$, each symbol is chosen independently from the distribution $P(X)$ maximizing the mutual information $I(X;Y)$.
- **Encoding:** For each $w \in [M]$, define $\mathsf{Enc}(w) = \overline{\mathbf{x}}_w$ (the w-th codeword).
- **Decoding:** Given $\overline{\mathbf{y}} \in \{0,1\}^n$, define $\mathsf{Dec}(\overline{\mathbf{y}})$ as

$$\mathsf{Dec}(\overline{\mathbf{y}}) \;=\; \begin{cases} w & \text{if } \exists \text{ unique } w \in [M] \text{ s.t. } \Delta(\overline{\mathbf{x}}_w, \overline{\mathbf{y}}) \leq (p+\delta) \cdot n \\ \text{arbitrary} & \text{otherwise} \end{cases} .$$

  Note that we will always count the second case towards the error probability, so we don't care how the decoding is defined there.

Before analyzing the error probability, we note that the noise in $\mathrm{BSC}(p)$ can be written a nice form. For input and output sequences $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$, we can write $\overline{\mathbf{y}} = \overline{\mathbf{x}} + \overline{\mathbf{z}} \mod 2$, where $\overline{\mathbf{z}} \in \{0,1\}^n$ is a sequence each bit independently 1 with probability p and 0 with probability $1 - p$. We will refer to this distribution for each bit of $\overline{\mathbf{z}}$ as the Bernoulli distribution with parameter p, denoted $\mathsf{Bern}(p)$. We thus have $\overline{\mathbf{z}} \sim (\mathsf{Bern}(p))^n$.

We now analyze the *expected* probability of error for a random collection of codewords C, chosen as above. Obtaining a bound on the error probability (for each n) will show that there *exists* a good collection of codewords for each n, although we don't explicitly know what this code is. We will discuss explicit constructions in the next lecture. We now prove the following.

Claim 2.9. *Let C be random code constructed as above. Then*

$$\mathop{\mathbb{E}}_{C}\left[p_e\right] \;\leq\; n \cdot 2^{-n \cdot D(p+\delta\|p)} \;+\; 2^{nR} \cdot n \cdot 2^{-n \cdot D(p+\delta\|\frac{1}{2})},$$

where $D(p\|q)$ denotes $D(\mathsf{Bern}(p)\|\mathsf{Bern}(q))$ as usual.

Proof: We get

$$\mathop{\mathbb{E}}_{C}\left[p_e\right] \;=\; \mathop{\mathbb{E}}_{C}\left[\mathbb{P}\left[\widehat{W} \neq W\right]\right] \;=\; \mathop{\mathbb{E}}_{C}\left[\sum_{w\in[M]} \frac{1}{M} \cdot \mathbb{P}\left[\widehat{W} \neq w | W = w\right]\right].$$

By symmetry in the code construction, we can say that $\mathbb{E}_C\left[\mathbb{P}\left[\widehat{W} \neq w | W = w\right]\right]$ is the same for all $w \in [M]$. Replacing all these by the case for $w = 1$, we get

$$\mathop{\mathbb{E}}_{C}\left[p_e\right] \;=\; \mathop{\mathbb{E}}_{C}\left[\mathbb{P}\left[\widehat{W} \neq 1 | W = 1\right]\right].$$

We consider two cases in which we can have an error: either the output $\overline{\mathbf{y}}$ of the channel was too far from the input $\overline{\mathbf{x}}_1$, or $\Delta(\overline{\mathbf{x}}_w, \overline{\mathbf{y}}) \leq (p+\delta) \cdot n$ for some other $w > 1$. Thus, we have

$$\mathop{\mathbb{E}}_{C}\left[p_e\right] \;\leq\; \mathop{\mathbb{E}}_{C}\left[\mathbb{P}\left[\Delta(\overline{\mathbf{x}}_1, \overline{\mathbf{y}}) > (p+\delta) \cdot n\right]\right] + \sum_{w>1} \mathop{\mathbb{E}}_{C}\left[\mathbb{P}\left[\Delta(\overline{\mathbf{x}}_w, \overline{\mathbf{y}}) \leq (p+\delta) \cdot n\right]\right]$$

For a fixed $\overline{\mathbf{x}}_1$, let $\overline{\mathbf{y}} = \overline{\mathbf{x}}_1 + \overline{\mathbf{z}} \mod 2$, where $\overline{\mathbf{z}} \sim \mathsf{Bern}(p)$ is independent of $\overline{\mathbf{x}}_1$. The event $\Delta(\overline{\mathbf{x}}_1, \overline{\mathbf{y}}) > (p+\delta) \cdot n$ can be written in terms of the "type" $P_{\overline{\mathbf{z}}}$ of $\overline{\mathbf{z}}$ as $P_{\overline{\mathbf{z}}} \in \Pi$, where $\Pi = \{\mathsf{Bern}(p') \mid p' > p + \delta\}$. By Sanov's theorem, we then have that for each fixed $\overline{\mathbf{x}}_1$

$$\mathop{\mathbb{P}}_{\overline{\mathbf{y}}}\left[\Delta(\overline{\mathbf{x}}_1, \overline{\mathbf{y}}) > (p+\delta) \cdot n\right] \;\leq\; n \cdot 2^{-n \cdot D(p+\delta\|p)}.$$

For the second term, we use the fact that for each $\overline{\mathbf{y}}$ (which may depend on $\overline{\mathbf{x}}_1$), $\overline{\mathbf{x}}_w$ is independent of $\overline{\mathbf{y}}$ for all $w > 1$ (since codewords are chosen independently). Now defining $\overline{\mathbf{z}}$ so that $\overline{\mathbf{y}} + \overline{\mathbf{x}}_w = \overline{\mathbf{z}} \mod 2$, we get that $\overline{\mathbf{z}} \sim (\mathsf{Bern}(1/2))^n$ (why?) For this $\overline{\mathbf{z}}$, we can now write the even $\Delta(\overline{\mathbf{x}}_w, \overline{\mathbf{y}}) \leq (p+\delta) \cdot n$ as $P_{\overline{\mathbf{z}}} \in \Pi'$, where $\Pi' = \{\mathsf{Bern}(p') | p' \leq p + \delta\}$. Applying Sanov's theorem again, we get that

$$\mathop{\mathbb{P}}_{\overline{\mathbf{x}}_w}\left[\Delta(\overline{\mathbf{x}}_w, \overline{\mathbf{y}}) \leq (p+\delta) \cdot n\right] \;\leq\; n \cdot 2^{-n \cdot D(p+\delta\|\frac{1}{2})}.$$

Combining the above bounds, we get

$$\mathop{\mathbb{E}}_{C}[p_e] \;\le\; n \cdot 2^{-n \cdot D(p+\delta\|p)} + \sum_{w>1} n \cdot 2^{-n \cdot D(p+\delta\|\frac{1}{2})} \;\le\; n \cdot 2^{-n \cdot D(p+\delta\|p)} \;+\; 2^{nR} \cdot n \cdot 2^{-n \cdot D(p+\delta\|\frac{1}{2})},$$

as claimed. ■

To analyze the bound, and compare it to the channel capacity $1 - H_2(p)$, we note that $D(p+\delta\|\frac{1}{2}) = 1 - H_2(p+\delta)$. Check that $\forall \varepsilon > 0$, there exists $\delta > 0$ such that $H_2(p+\delta) \le H_2(p) + \varepsilon$. Using a δ such that $H_2(p+\delta) \le H_2(p) + \varepsilon/2$, we get that

$$\mathop{\mathbb{E}}_{C}[p_e] \;\le\; n \cdot 2^{-n \cdot D(p+\delta\|p)} \;+\; 2^{nR} \cdot n \cdot 2^{-n \cdot (1 - H_2(p) - \varepsilon/2)},$$

which tends to zero for $R = (1 - H_2(p) - \varepsilon)$. Thus, for every $\varepsilon > 0$, we have a sequence of codes (as $n \to \infty$) with rate at least $(1 - H_2(p) - \varepsilon)$, and $p_e^{(n)} \to 0$.

Exercise 2.10. *For $R = 1 - H_2(p) - \varepsilon$ in the above proof, let $n_0(\varepsilon)$ be the smallest n (block-length) such that the probability of error $p_e^{(n)} \to 0$ for $n \ge n_0(\varepsilon)$. Check that $n_0(\varepsilon) = O(1/\varepsilon^2)$ suffices in the above proof.*

References

[Wig17] Avi Wigderson, *Operator scaling: theory, applications and connections*, 2017, Tutorial given at CCC 2017. 2