COMP2610/6261 - Information Theory

Lecture 20: Joint-Typicality and the Noisy-Channel Coding Theorem

Robert C. Williamson

Research School of Computer Science

Australian National University

October 24th, 2018

Channel Capacity: Recap

The *largest possible* reduction in uncertainty achievable across a channel is its **capacity**

Channel Capacity

The capacity C of a channel Q is the largest mutual information between its input and output for any choice of input ensemble. That is,

$$C = \max_{\mathbf{p}_X} I(X; Y)$$

Block Codes: Recap

(N, K) Block Code

Given a channel Q with inputs $\mathcal{X}$ and outputs $\mathcal{Y}$, an integer $N > 0$, and $K > 0$, an (N, K) Block Code for Q is a list of $S = 2^K$ codewords

$$\mathcal{C} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(2^K)}\}$$

where each $\mathbf{x}^{(s)} \in \mathcal{X}^N$ consists of N symbols from $\mathcal{X}$.

Rate of a block code is $\frac{K}{N} = \frac{\log_2 S}{N}$

Reliability: Recap

$$\boxed{\mathbf{s}_{in}} \xrightarrow{\text{Encoder}} \boxed{\mathbf{x}} \xrightarrow{Q} \boxed{\mathbf{y}} \xrightarrow{\text{Decoder}} \boxed{\mathbf{s}_{out}}$$

Probability of (Block) Error

Given a channel Q the **probability of (block) error** for a code is

$$p_B = P(\mathbf{s}_{out} \neq \mathbf{s}_{in}) = \sum_{\mathbf{s}_{in}} P(\mathbf{s}_{out} \neq \mathbf{s}_{in} | \mathbf{s}_{in}) P(\mathbf{s}_{in})$$

and its **maximum probability of (block) error** is

$$p_{BM} = \max_{\mathbf{s}_{in}} P(\mathbf{s}_{out} \neq \mathbf{s}_{in} | \mathbf{s}_{in})$$

The Noisy-Channel Coding Theorem: Recap

Informal Statement

Recall that a rate R is achievable if there is a block code with this rate and arbitrarily small error probability

We highlighted the following remarkable result:

Noisy-Channel Coding Theorem (Informal)

If Q is a channel with capacity C then the rate R is *achievable* **if and only if** $R \leq C$, that is, the rate is no greater than the channel capacity.

The Noisy-Channel Coding Theorem: Recap

Informal Statement

Recall that a rate R is achievable if there is a block code with this rate and arbitrarily small error probability

We highlighted the following remarkable result:

Noisy-Channel Coding Theorem (Informal)

If Q is a channel with capacity C then the rate R is *achievable* **if and only if** $R \leq C$, that is, the rate is no greater than the channel capacity.

Ideally, we would like to know:

- Can we go above C if we allow some fixed probability of error?
- Is there a maximal rate for a fixed probability of error?

The Noisy-Channel Coding Theorem

Formal Statement

Recall: a rate is achievable if for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$

The Noisy-Channel Coding Theorem

Formal Statement

Recall: a rate is achievable if for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$

The Noisy-Channel Coding Theorem (Formal)

1. Any rate $R < C$ is *achievable* for Q

The Noisy-Channel Coding Theorem

Formal Statement

Recall: a rate is achievable if for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$

The Noisy-Channel Coding Theorem (Formal)

1. Any rate $R < C$ is *achievable* for Q
2. If probability of bit error $p_b := p_B/K$ is acceptable, (N, K) codes exists with rates

$$\frac{K}{N} \leq R(p_b) = \frac{C}{1 - H_2(p_b)}$$

The Noisy-Channel Coding Theorem

Formal Statement

Recall: a rate is achievable if for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$

The Noisy-Channel Coding Theorem (Formal)

1. Any rate $R < C$ is *achievable* for Q
2. If probability of bit error $p_b := p_B/K$ is acceptable, (N, K) codes exists with rates
$$\frac{K}{N} \leq R(p_b) = \frac{C}{1 - H_2(p_b)}$$
3. For any p, we **cannot** achieve a rate greater than $R(p)$ with probability of bit error p.

The Noisy-Channel Coding Theorem

Formal Statement

Recall: a rate is achievable if for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$

The Noisy-Channel Coding Theorem (Formal)

1. Any rate $R < C$ is *achievable* for Q
2. If probability of bit error $p_b := p_B/K$ is acceptable, (N, K) codes exists with rates
$$\frac{K}{N} \leq R(p_b) = \frac{C}{1 - H_2(p_b)}$$
3. For any p, we **cannot** achieve a rate greater than $R(p)$ with probability of bit error p.

The Noisy-Channel Coding Theorem

Formal Statement

Recall: a rate is achievable if for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$

The Noisy-Channel Coding Theorem (Formal)

1. Any rate $R < C$ is *achievable* for Q
2. If probability of bit error $p_b := p_B/K$ is acceptable, (N, K) codes exists with rates
$$\frac{K}{N} \leq R(p_b) = \frac{C}{1 - H_2(p_b)}$$
3. For any p, we **cannot** achieve a rate greater than $R(p)$ with probability of bit error p.

Note that as $p_b \to \frac{1}{2}$, $R(p_b) \to +\infty$, while as $p_b \to \{0, 1\}$, $R(p_b) \to C$, so we cannot achieve rate greater than C with probability of bit error arbitrarily small

Implications of NCCT

Suppose we know a channel has capacity 0.6 bits

Implications of NCCT

Suppose we know a channel has capacity 0.6 bits

We **cannot** achieve a rate of 0.8 with arbitrarily small error

Implications of NCCT

Suppose we know a channel has capacity 0.6 bits

We **cannot** achieve a rate of 0.8 with arbitrarily small error

We **can** achieve a rate of 0.8 with probability of bit error 5%, since $\frac{0.6}{1-H_2(0.05)} = 0.8408 > 0.8$

Joint Typicality

Recall that a random variable $\mathbf{z}$ from Z^N is typical for an ensemble Z whenever its average symbol information is within β of the entropy $H(Z)$

$$\left| -\frac{1}{N} \log_2 P(\mathbf{z}) - H(Z) \right| < \beta$$

Joint Typicality

Recall that a random variable **z** from Z^N is typical for an ensemble Z whenever its average symbol information is within β of the entropy $H(Z)$

$$\left| -\frac{1}{N} \log_2 P(\mathbf{z}) - H(Z) \right| < \beta$$

Joint Typicality

A pair of sequences $\mathbf{x} \in \mathcal{A}_X^N$ and $\mathbf{y} \in \mathcal{A}_Y^N$, each of length N, are **jointly typical** (to tolerance β) for distribution $P(x, y)$ if

1. $\mathbf{x}$ is typical of $P(\mathbf{x})$ [$\mathbf{z} = \mathbf{x}$ above]
2. $\mathbf{y}$ is typical of $P(\mathbf{y})$ [$\mathbf{z} = \mathbf{y}$ above]
3. $(\mathbf{x}, \mathbf{y})$ is typical of $P(\mathbf{x}, \mathbf{y})$ [$\mathbf{z} = (\mathbf{x}, \mathbf{y})$ above]

The **jointly typical set** of all such pairs is denoted $J_{N\beta}$.

Joint Typicality

Example

Example ($\mathbf{p}_X = (0.9, 0.1)$ and BSC with $f = 0.2$):

$\mathbf{x}$ 1111111111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

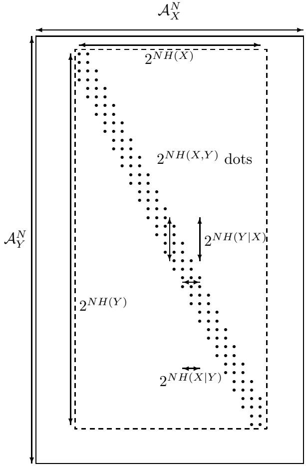
$\mathbf{y}$ 0011111111000000000000000000000000000000000000000000000000000000000000000000000000111111111111111111

Here:

- x has 10 1's (c.f. $p(X = 1) = 0.1$)
- y has 26 1's (c.f. $p(Y = 1) = (0.8)(0.1) + (0.2)(0.9) = 0.26$)
- x, y differ in 20 bits (c.f. $p(X \neq Y) = 0.2$)
    - This is essential in addition to the above two facts

Joint Typicality

Counts

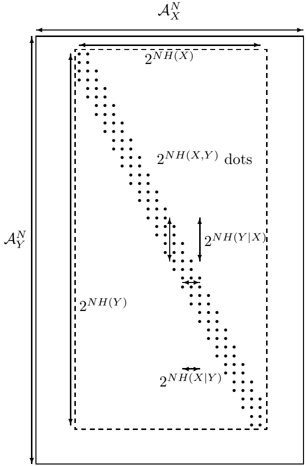


There are approximately:

- $2^{NH(X)}$ typical $\mathbf{x} \in \mathcal{A}_X^N$

Joint Typicality

Counts

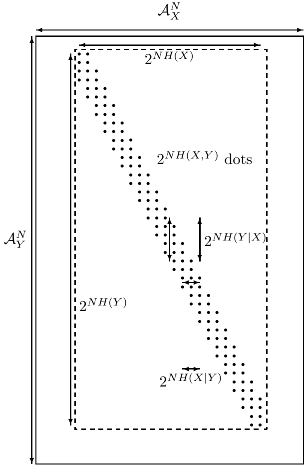


There are approximately:

- $2^{NH(X)}$ typical $\mathbf{x} \in \mathcal{A}_X^N$
- $2^{NH(Y)}$ typical $\mathbf{y} \in \mathcal{A}_Y^N$

Joint Typicality

Counts

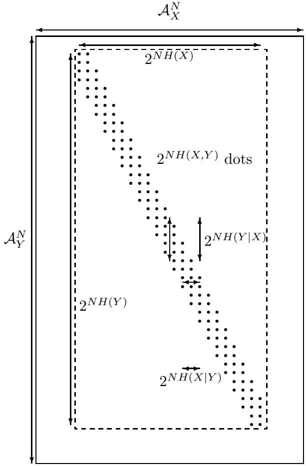


There are approximately:

- $2^{NH(X)}$ typical $\mathbf{x} \in \mathcal{A}_X^N$
- $2^{NH(Y)}$ typical $\mathbf{y} \in \mathcal{A}_Y^N$
- $2^{NH(X,Y)}$ typical $(\mathbf{x}, \mathbf{y}) \in \mathcal{A}_X^N \times \mathcal{A}_Y^N$

Joint Typicality

Counts

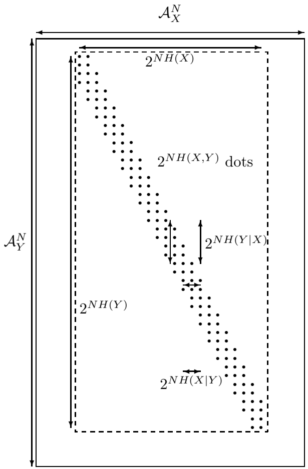


There are approximately:

- $2^{NH(X)}$ typical $\mathbf{x} \in \mathcal{A}_X^N$
- $2^{NH(Y)}$ typical $\mathbf{y} \in \mathcal{A}_Y^N$
- $2^{NH(X,Y)}$ typical $(\mathbf{x}, \mathbf{y}) \in \mathcal{A}_X^N \times \mathcal{A}_Y^N$
- $2^{NH(Y|X)}$ typical $\mathbf{y}$ given $\mathbf{x}$

Joint Typicality

Counts



There are approximately:

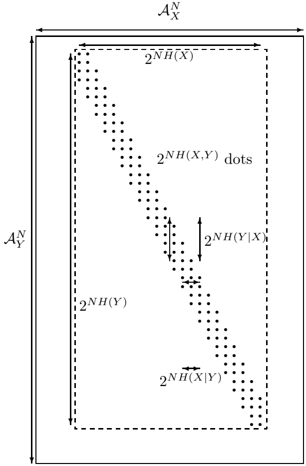
- $2^{NH(X)}$ typical $\mathbf{x} \in \mathcal{A}_X^N$
- $2^{NH(Y)}$ typical $\mathbf{y} \in \mathcal{A}_Y^N$
- $2^{NH(X,Y)}$ typical $(\mathbf{x}, \mathbf{y}) \in \mathcal{A}_X^N \times \mathcal{A}_Y^N$
- $2^{NH(Y|X)}$ typical $\mathbf{y}$ given $\mathbf{x}$

Thus, by selecting independent typical vectors, we arrive at a jointly typical vector with probability approximately

$$\frac{2^{NH(X,Y)}}{2^{NH(X)} \cdot 2^{NH(Y)}} = 2^{-NI(X;Y)}$$

Joint Typicality

Counts



There are approximately:

- $2^{NH(X)}$ typical $\mathbf{x} \in \mathcal{A}_X^N$
- $2^{NH(Y)}$ typical $\mathbf{y} \in \mathcal{A}_Y^N$
- $2^{NH(X,Y)}$ typical $(\mathbf{x}, \mathbf{y}) \in \mathcal{A}_X^N \times \mathcal{A}_Y^N$
- $2^{NH(Y|X)}$ typical $\mathbf{y}$ given $\mathbf{x}$

Thus, by selecting independent typical vectors, we arrive at a jointly typical vector with probability approximately

$$\frac{2^{NH(X,Y)}}{2^{NH(X)} \cdot 2^{NH(Y)}} = 2^{-NI(X;Y)}$$

Here we used

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

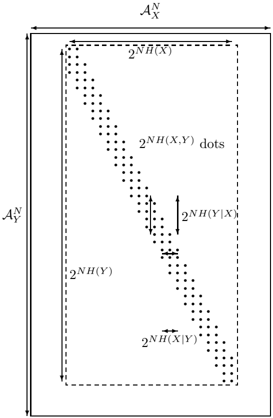
Joint Typicality Theorem

Let $\mathbf{x}, \mathbf{y}$ be drawn from $(XY)^N$ with $P(\mathbf{x}, \mathbf{y}) = \prod_n P(x_n, y_n)$.

Joint Typicality Theorem

For all tolerances $\beta > 0$

1. Almost every pair is eventually jointly typical
   $P((\mathbf{x}, \mathbf{y}) \in J_{N\beta}) \to 1$ as $N \to \infty$



Joint Typicality Theorem

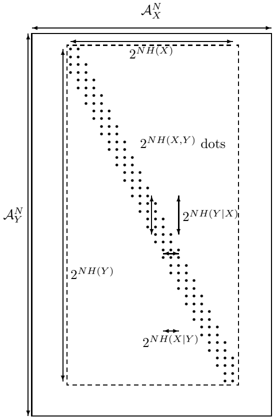
Let $\mathbf{x}, \mathbf{y}$ be drawn from $(XY)^N$ with $P(\mathbf{x}, \mathbf{y}) = \prod_n P(x_n, y_n)$.

Joint Typicality Theorem

For all tolerances $\beta > 0$

1. Almost every pair is eventually jointly typical
   $P((\mathbf{x}, \mathbf{y}) \in J_{N\beta}) \to 1$ as $N \to \infty$
2. The number of jointly typical sequences is roughly $2^{NH(X,Y)}$:

$$|J_{N\beta}| \leq 2^{N(H(X,Y)+\beta)}$$



Joint Typicality Theorem

Let $\mathbf{x}, \mathbf{y}$ be drawn from $(XY)^N$ with $P(\mathbf{x}, \mathbf{y}) = \prod_n P(x_n, y_n)$.

Joint Typicality Theorem

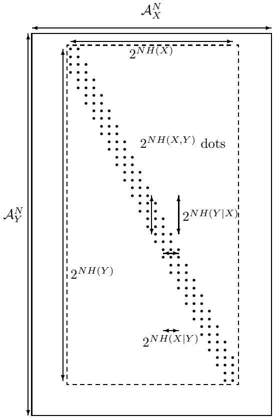
For all tolerances $\beta > 0$

1. Almost every pair is eventually jointly typical
   $P((\mathbf{x}, \mathbf{y}) \in J_{N\beta}) \to 1$ as $N \to \infty$
2. The number of jointly typical sequences is roughly $2^{NH(X,Y)}$:

$$|J_{N\beta}| \leq 2^{N(H(X,Y)+\beta)}$$

3. For $\mathbf{x}'$ and $\mathbf{y}'$ drawn independently from the marginals of $P(\mathbf{x}, \mathbf{y})$,

$$P((\mathbf{x}', \mathbf{y}') \in J_{N\beta}) \leq 2^{-N(I(X;Y)-3\beta)}$$



The Noisy-Channel Coding Theorem

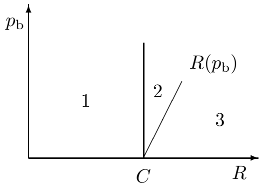
Let Q be a channel with inputs $\mathcal{A}_X$ and outputs $\mathcal{A}_Y$.
Let $C = \max_{\mathbf{p}_X} I(X; Y)$ be the capacity of Q and
$H_2(p) = -p \log_2 p - (1 - p) \log_2(1 - p)$.

The Noisy-Channel Coding Theorem

1. Any rate $R < C$ is *achievable* for Q (i.e., for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$)
2. If probability of bit error $p_b := p_B/K$ is acceptable, there exist (N, K) codes with rates
$$\frac{K}{N} \leq R(p_b) = \frac{C}{1 - H_2(p_b)}$$
3. For any p_b, rates greater than $R(p_b)$ are not achievable.

The Noisy-Channel Coding Theorem

Let Q be a channel with inputs $\mathcal{A}_X$ and outputs $\mathcal{A}_Y$.
Let $C = \max_{\mathbf{p}_X} I(X;Y)$ be the capacity of Q and
$H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$.

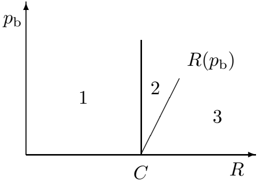


The Noisy-Channel Coding Theorem

1. Any rate $R < C$ is *achievable* for Q (i.e., for any tolerance $\epsilon > 0$, an (N,K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$)
2. If probability of bit error $p_b := p_B/K$ is acceptable, there exist (N,K) codes with rates
$$\frac{K}{N} \leq R(p_b) = \frac{C}{1 - H_2(p_b)}$$
3. For any p_b, rates greater than $R(p_b)$ are not achievable.

The Noisy-Channel Coding Theorem

Let Q be a channel with inputs $\mathcal{A}_X$ and outputs $\mathcal{A}_Y$.
Let $C = \max_{\mathbf{p}_X} I(X;Y)$ be the capacity of Q and
$H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$.



The Noisy-Channel Coding Theorem

1. Any rate $R < C$ is *achievable* for Q (i.e., for any tolerance $\epsilon > 0$, an (N, K) code with rate $K/N \geq R$ exists with max. block error $p_{BM} < \epsilon$)
2. If probability of bit error $p_b := p_B/K$ is acceptable, there exist (N, K) codes with rates
$$\frac{K}{N} \leq R(p_b) = \frac{C}{1 - H_2(p_b)}$$
3. For any p_b, rates greater than $R(p_b)$ are not achievable.

Some Intuition for the NCCT

The proof of the NCCT is based on the following observations:

- Each choice of input distribution $\mathbf{p}_X$ induces an output distibution $\mathbf{p}_Y$

Some Intuition for the NCCT

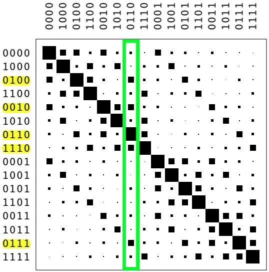
The proof of the NCCT is based on the following observations:

- Each choice of input distribution $\mathbf{p}_X$ induces an output distibution $\mathbf{p}_Y$
- There are $2^{NH(Y)}$ typical $\mathbf{y}$ (i.e., with prob. per symbol $\approx H(Y)$)

Some Intuition for the NCCT

The proof of the NCCT is based on the following observations:

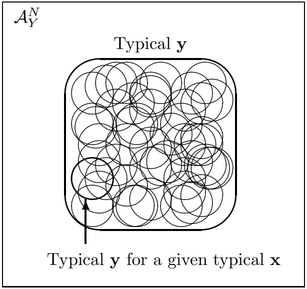
- Each choice of input distribution $\mathbf{p}_X$ induces an output distibution $\mathbf{p}_Y$
- There are $2^{NH(Y)}$ typical $\mathbf{y}$ (i.e., with prob. per symbol $\approx H(Y)$)
- For each $\mathbf{x}$ there are $2^{NH(Y|X)}$ typical $\mathbf{y}$ for $\mathbf{x}$

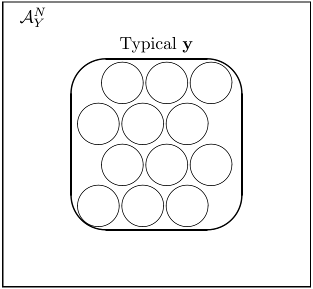


Some Intuition for the NCCT

The proof of the NCCT is based on the following observations:

- Each choice of input distribution $\mathbf{p}_X$ induces an output distibution $\mathbf{p}_Y$
- There are $2^{NH(Y)}$ typical $\mathbf{y}$ (i.e., with prob. per symbol $\approx H(Y)$)
- For each $\mathbf{x}$ there are $2^{NH(Y|X)}$ typical $\mathbf{y}$ for $\mathbf{x}$
- *At most* there are $\frac{2^{NH(Y)}}{2^{NH(Y|X)}} = 2^{N(H(Y)-H(Y|X))} = 2^{NI(X;Y)}$ $\mathbf{x}$ with disjoint typical $\mathbf{y}$. Coding with these $\mathbf{x}$ minimises error

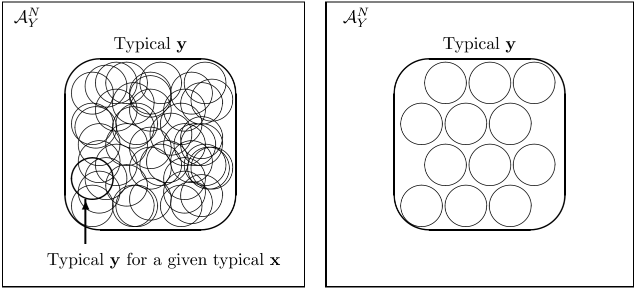




Some Intuition for the NCCT

The proof of the NCCT is based on the following observations:

- Each choice of input distribution $\mathbf{p}_X$ induces an output distibution $\mathbf{p}_Y$
- There are $2^{NH(Y)}$ typical $\mathbf{y}$ (i.e., with prob. per symbol $\approx H(Y)$)
- For each $\mathbf{x}$ there are $2^{NH(Y|X)}$ typical $\mathbf{y}$ for $\mathbf{x}$
- *At most* there are $\frac{2^{NH(Y)}}{2^{NH(Y|X)}} = 2^{N(H(Y)-H(Y|X))} = 2^{NI(X;Y)}$ $\mathbf{x}$ with disjoint typical $\mathbf{y}$. Coding with these $\mathbf{x}$ minimises error
- Best rate K/N achieved when number of such $\mathbf{x}$ (i.e., 2^K) is maximised: $2^K \leq \max_{\mathbf{p}_X} 2^{NI(X;Y)} = 2^{N \max_{\mathbf{p}_X} I(X;Y)} = 2^{NC}$



Proof Sketch of NCCT Part 1

We can:

- define a family of random codes, which rely on joint typicality, and which achieve the given rate

Proof Sketch of NCCT Part 1

We can:

- define a family of random codes, which rely on joint typicality, and which achieve the given rate
- show that on average, such a code has a low probability of block error

Proof Sketch of NCCT Part 1

We can:

- define a family of random codes, which rely on joint typicality, and which achieve the given rate
- show that on average, such a code has a low probability of block error
- deduce that at least one such code must have a low probability of block error

Proof Sketch of NCCT Part 1

We can:

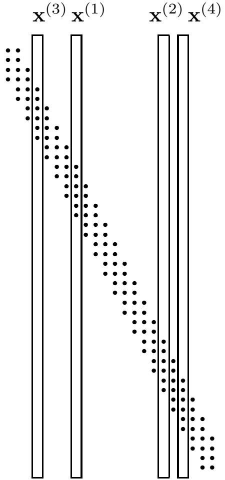
- define a family of random codes, which rely on joint typicality, and which achieve the given rate
- show that on average, such a code has a low probability of block error
- deduce that at least one such code must have a low probability of block error
- “expurgate” the above code so that it has low maximal probability of error

This will establish that the final code achieves low maximal probability of error, while achieving the given rate!

Random Coding and Typical Set Decoding

Make **random code** $\mathcal{C}$ with rate R':

- Fix $\mathbf{p}_X$ and choose $S = 2^{NR'}$ codewords, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$, each with $P(\mathbf{x}) = \prod_n P(x_n)$



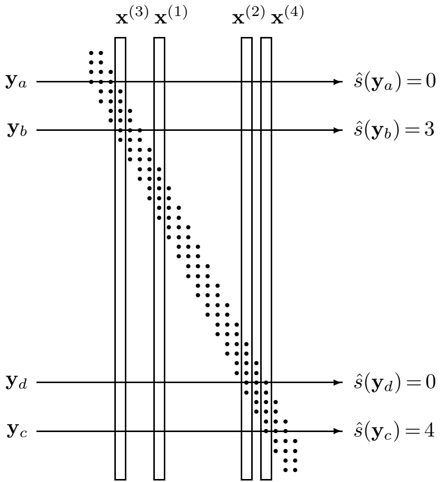
Random Coding and Typical Set Decoding

Make **random code** $\mathcal{C}$ with rate R':

- Fix $\mathbf{p}_X$ and choose $S = 2^{NR'}$ codewords, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$, each with $P(\mathbf{x}) = \prod_n P(x_n)$

Decode y via typical sets:

- If there is *exactly one* $\hat{s}$ so that $(\mathbf{x}^{\hat{s}}, \mathbf{y})$ are jointly typical then decode $\mathbf{y}$ as $\hat{s}$
- Otherwise, fail ($\hat{s} = 0$)



Random Coding and Typical Set Decoding

Make **random code** $\mathcal{C}$ with rate R':

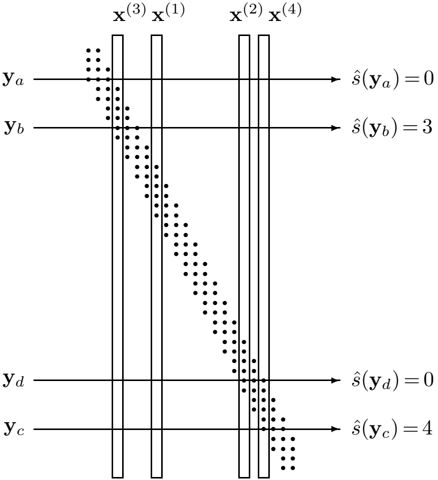
- Fix $\mathbf{p}_X$ and choose $S = 2^{NR'}$ codewords, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$, each with $P(\mathbf{x}) = \prod_n P(x_n)$

Decode y via typical sets:

- If there is *exactly one* $\hat{s}$ so that $(\mathbf{x}^{\hat{s}}, \mathbf{y})$ are jointly typical then decode $\mathbf{y}$ as $\hat{s}$
- Otherwise, fail ($\hat{s} = 0$)

Errors:

- $p_B(\mathcal{C}) = P(\hat{s} \neq s|\mathcal{C})$
- $\langle p_B \rangle = \sum_{\mathcal{C}} P(\hat{s} \neq s|\mathcal{C})P(\mathcal{C})$
- $p_{BM}(\mathcal{C}) = \max_s P(\hat{s} \neq s|s, \mathcal{C})$
  (Aim: $\exists\mathcal{C}$ s.t. $p_{BM}(\mathcal{C})$ small)



Average Error Over All Codes

Let's consider the average error over random codes:

$$\langle p_B \rangle = \sum_{\mathcal{C}} P(\hat{s} \neq s | \mathcal{C}) P(\mathcal{C})$$

A bound on the average $\langle f \rangle$ of some function f of random variables $z \in \mathcal{Z}$ with probabilities $P(z)$ *guarantees* there is at least one $z^* \in \mathcal{Z}$ such that $f(z^*)$ is smaller than the bound.[1]

[1] If $\langle f \rangle < \delta$ but $f(z) \geq \delta$ for all z, $\langle f \rangle = \sum_z f(z) P(z) \geq \sum_z \delta P(z) = \delta$!!

Average Error Over All Codes

Let's consider the average error over random codes:

$$\langle p_B \rangle = \sum_{\mathcal{C}} P(\hat{s} \neq s | \mathcal{C}) P(\mathcal{C})$$

A bound on the average $\langle f \rangle$ of some function f of random variables $z \in \mathcal{Z}$ with probabilities $P(z)$ *guarantees* there is at least one $z^* \in \mathcal{Z}$ such that $f(z^*)$ is smaller than the bound.[1]

$$\text{So } \langle p_B \rangle < \delta \implies p_B(\mathcal{C}^*) < \delta \text{ for some } \mathcal{C}^*.$$

Analogy: Suppose the average height of class is not more than 160 cm. Then one of you *must* be shorter than 160 cm.

[1] If $\langle f \rangle < \delta$ but $f(z) \geq \delta$ for all z, $\langle f \rangle = \sum_z f(z) P(z) \geq \sum_z \delta P(z) = \delta$!!

Proof Sketch of NCCT Part 1

Want to prove

> Any rate $R < C$ is *achievable* for Q (i.e., an (N, K) code with rate $N/K \geq R$ exists with max. block error $p_{BM} < \epsilon$ for any tolerance ϵ)

Let us thus bound $\langle p_B \rangle$ for our random code

Choose some $\delta > 0$

1. Part one of the Joint Typicality Theorem says we can find an $N(\delta)$ such that the probability $(\mathbf{x}, \mathbf{y})$ are not jointly typical is less than δ.

Proof Sketch of NCCT Part 1

Want to prove

> Any rate $R < C$ is *achievable* for Q (i.e., an (N, K) code with rate $N/K \geq R$ exists with max. block error $p_{BM} < \epsilon$ for any tolerance ϵ)

Let us thus bound $\langle p_B \rangle$ for our random code

Choose some $\delta > 0$

1. Part one of the Joint Typicality Theorem says we can find an $N(\delta)$ such that the probability $(\mathbf{x}, \mathbf{y})$ are not jointly typical is less than δ.
2. Thus, the average probability of error satisfies (by Part 3 of JCT)

$$\langle p_B \rangle = \sum_{\text{atypical } (\mathbf{x},\mathbf{y})} P(\hat{s} \neq s|\cdot) + \sum_{\text{typical } (\mathbf{x},\mathbf{y})} P(\hat{s} \neq s|\cdot)$$

Proof Sketch of NCCT Part 1

Want to prove

> Any rate $R < C$ is *achievable* for Q (i.e., an (N, K) code with rate $N/K \geq R$ exists with max. block error $p_{BM} < \epsilon$ for any tolerance ϵ)

Let us thus bound $\langle p_B \rangle$ for our random code

Choose some $\delta > 0$

1. Part one of the Joint Typicality Theorem says we can find an $N(\delta)$ such that the probability $(\mathbf{x}, \mathbf{y})$ are not jointly typical is less than δ.
2. Thus, the average probability of error satisfies (by Part 3 of JCT)

$$\langle p_B \rangle \leq \delta + \sum_{s'=2}^{2^{NR'}} 2^{-N(I(X;Y)-3\beta)}$$

Proof Sketch of NCCT Part 1

Want to prove

> Any rate $R < C$ is *achievable* for Q (i.e., an (N, K) code with rate $N/K \geq R$ exists with max. block error $p_{BM} < \epsilon$ for any tolerance ϵ)

Let us thus bound $\langle p_B \rangle$ for our random code

Choose some $\delta > 0$

1. Part one of the Joint Typicality Theorem says we can find an $N(\delta)$ such that the probability $(\mathbf{x}, \mathbf{y})$ are not jointly typical is less than δ.
2. Thus, the average probability of error satisfies (by Part 3 of JCT)

$$\langle p_B \rangle \leq \delta + 2^{-N(I(X;Y) - R' - 3\beta)}$$

Proof Sketch of NCCT Part 1

Want to prove

> Any rate $R < C$ is *achievable* for Q (i.e., an (N, K) code with rate $N/K \geq R$ exists with max. block error $p_{BM} < \epsilon$ for any tolerance ϵ)

Let us thus bound $\langle p_B \rangle$ for our random code

Choose some $\delta > 0$

1. Part one of the Joint Typicality Theorem says we can find an $N(\delta)$ such that the probability $(\mathbf{x}, \mathbf{y})$ are not jointly typical is less than δ.
2. Thus, the average probability of error satisfies (by Part 3 of JCT)

$$\langle p_B \rangle \leq \delta + 2^{-N(I(X;Y) - R' - 3\beta)}$$

3. Increasing N will make $\langle p_B \rangle < 2\delta$ if $R' < I(X; Y) - 3\beta$

Proof Sketch of NCCT Part 1

Want to prove

> Any rate $R < C$ is *achievable* for Q (i.e., an (N, K) code with rate $N/K \geq R$ exists with max. block error $p_{BM} < \epsilon$ for any tolerance ϵ)

Let us thus bound $\langle p_B \rangle$ for our random code

Choose some $\delta > 0$

1. Part one of the Joint Typicality Theorem says we can find an $N(\delta)$ such that the probability $(\mathbf{x}, \mathbf{y})$ are not jointly typical is less than δ.
2. Thus, the average probability of error satisfies (by Part 3 of JCT)
$$\langle p_B \rangle \leq \delta + 2^{-N(I(X;Y)-R'-3\beta)}$$
3. Increasing N will make $\langle p_B \rangle < 2\delta$ if $R' < I(X;Y) - 3\beta$
4. Choosing maximal $P(x)$ makes required condition $R' < C - 3\beta$

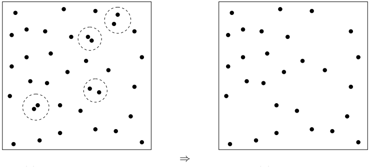
Code Expurgation

The last main "trick" is to show that if there is an (N, K) code with rate R' and $p_B(\mathcal{C}) < \delta$ we can construct a new (N, K') code $\mathcal{C}'$ with rate $R' - \frac{1}{N}$ and maximum probability of error $p_{BM}(\mathcal{C}') < 2\delta$.

Code Expurgation

The last main "trick" is to show that if there is an (N, K) code with rate R' and $p_B(\mathcal{C}) < \delta$ we can construct a new (N, K') code $\mathcal{C}'$ with rate $R' - \frac{1}{N}$ and maximum probability of error $p_{BM}(\mathcal{C}') < 2\delta$.

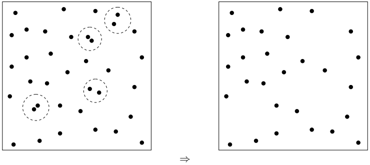
We create $\mathcal{C}'$ by **expurgating** (throwing out) half the codewords from $\mathcal{C}$, specifically the half with the largest *conditional* probability of error.



Code Expurgation

The last main “trick” is to show that if there is an (N, K) code with rate R' and $p_B(\mathcal{C}) < \delta$ we can construct a new (N, K') code $\mathcal{C}'$ with rate $R' - \frac{1}{N}$ and maximum probability of error $p_{BM}(\mathcal{C}') < 2\delta$.

We create $\mathcal{C}'$ by **expurgating** (throwing out) half the codewords from $\mathcal{C}$, specifically the half with the largest *conditional* probability of error.



Proof:

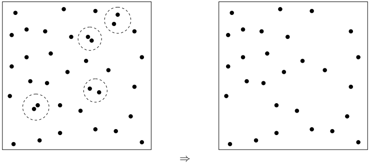
- Code $\mathcal{C}'$ has $2^{NR'}/2 = 2^{NR'-1}$ messages, so rate of $K'/N = R' - \frac{1}{N}$.
- Suppose $p_{BM}(\mathcal{C}') = \max_s P(\hat{s} \neq s | s, \mathcal{C}') \geq 2\delta$, then every $s \in \mathcal{C}$ that was thrown out must have conditional probability $P(\hat{s} \neq s | s, \mathcal{C}) \geq 2\delta$
- But then

$$p_B(\mathcal{C}) = \sum_s P(\hat{s} \neq s | s, \mathcal{C}) P(s) \geq \frac{1}{2} \sum_{s \notin \mathcal{C}'} 2\delta + \frac{1}{2} \sum_{s \in \mathcal{C}'} P(\hat{s} \neq s | s, \mathcal{C}) \geq \delta$$

Code Expurgation

The last main "trick" is to show that if there is an (N, K) code with rate R' and $p_B(\mathcal{C}) < \delta$ we can construct a new (N, K') code $\mathcal{C}'$ with rate $R' - \frac{1}{N}$ and maximum probability of error $p_{BM}(\mathcal{C}') < 2\delta$.

We create $\mathcal{C}'$ by **expurgating** (throwing out) half the codewords from $\mathcal{C}$, specifically the half with the largest *conditional* probability of error.



Proof:

- Code $\mathcal{C}'$ has $2^{NR'}/2 = 2^{NR'-1}$ messages, so rate of $K'/N = R' - \frac{1}{N}$.
- Suppose $p_{BM}(\mathcal{C}') = \max_s P(\hat{s} \neq s | s, \mathcal{C}') \geq 2\delta$, then every $s \in \mathcal{C}$ that was thrown out must have conditional probability $P(\hat{s} \neq s | s, \mathcal{C}) \geq 2\delta$
- But then

$$p_B(\mathcal{C}) = \sum_s P(\hat{s} \neq s | s, \mathcal{C}) P(s) \geq \frac{1}{2} \sum_{s \notin \mathcal{C}'} 2\delta + \frac{1}{2} \sum_{s \in \mathcal{C}'} P(\hat{s} \neq s | s, \mathcal{C}) \geq \delta$$

Wrapping It All Up

From the previous slide, $\langle p_B \rangle < 2\delta \implies$ some $\mathcal{C}'$ such that $p_{BM}(\mathcal{C}') < 4\delta$ with rate $R' - \frac{1}{N}$

Setting $R' = (R + C)/2, \delta = \epsilon/4, \beta < (C - R')/3$ gives the result!

NCCT Part 1: Comments

NCCT shows the existence of good codes; actually constructing practical codes is another matter

In principle, one could try the coding scheme outlined in the proof

- However, it would require a lookup in an exponential sized table (for the typical set decoding)!

Over the past few decades, some codes (e.g. Turbo codes) have been shown to achieve rate close to the Shannon capacity

- Beyond the scope of this course!

NCCT Converse: Comments

One can in fact make a stronger statement about

$$p_{B,\text{avg}} = \frac{1}{2^K} \sum_{\mathbf{s}_{\text{in}}} P(\mathbf{s}_{\text{out}} \neq \mathbf{s}_{\text{in}} \mid \mathbf{s}_{\text{in}}),$$

the probability of block error assuming a uniform distribution over inputs

We have:

$$p_{B,\text{avg}} \geq 1 - O(e^{-N(R-C)})$$

Thus, if $R > C$, the probability of block error shoots to 1 as N increases!

- We have a “phase transition” around C between perfectly reliable and perfectly unreliable communication!

Theory and Practice

> *The difference between theory and practice is that, in theory, there is no difference between theory and practice but, in practice, there is.*
>
> — *Jan L. A. van de Snepscheut*

Theory and Practice

> *The difference between theory and practice is that, in theory, there is no difference between theory and practice but, in practice, there is.*
>
> *— Jan L. A. van de Snepscheut*

Theory vs. Practice

- The NCCT theorem tells us that good block codes exist for any noisy channel (in fact, most random codes are good)
- However, the theorem is non-constructive: it does not tell us **how** to create *practical* codes for a given noisy channel
- The construction of practical codes that achieve rates up to the capacity for general channels is ongoing research

Types of Codes

When we talk about types of codes we will be referring to schemes for creating (N, K) codes for any size N. MacKay makes the following distinctions:

- **Bad**: Cannot achieve arbitrarily small error, or only achieve it if the rate goes to zero (i.e., either $p_{BM} \to a > 0$ as $N \to \infty$ or $p_{BM} \to 0 \implies K/N \to 0$)

Types of Codes

When we talk about types of codes we will be referring to schemes for creating (N, K) codes for any size N. MacKay makes the following distinctions:

- **Bad**: Cannot achieve arbitrarily small error, or only achieve it if the rate goes to zero (i.e., either $p_{BM} \to a > 0$ as $N \to \infty$ or $p_{BM} \to 0 \implies K/N \to 0$)
- **Good**: Can achieve arbitrarily small error up to some maximum rate strictly less than the channel capacity (i.e, for any ϵ a good coding scheme can make a code with $K/N = R_{max} < C$ and $p_{BM} < \epsilon$)

Types of Codes

When we talk about types of codes we will be referring to schemes for creating (N, K) codes for any size N. MacKay makes the following distinctions:

- **Bad**: Cannot achieve arbitrarily small error, or only achieve it if the rate goes to zero (i.e., either $p_{BM} \to a > 0$ as $N \to \infty$ or $p_{BM} \to 0 \implies K/N \to 0$)
- **Good**: Can achieve arbitrarily small error up to some maximum rate strictly less than the channel capacity (i.e, for any ϵ a good coding scheme can make a code with $K/N = R_{max} < C$ and $p_{BM} < \epsilon$)
- **Very Good**: Can achieve arbitrarily small error at any rate up to the channel capacity (i.e., for any $\epsilon > 0$ a very good coding scheme can make a code with $K/N = C$ and $p_{BM} < \epsilon$)

Types of Codes

When we talk about types of codes we will be referring to schemes for creating (N, K) codes for any size N. MacKay makes the following distinctions:

- **Bad**: Cannot achieve arbitrarily small error, or only achieve it if the rate goes to zero (i.e., either $p_{BM} \to a > 0$ as $N \to \infty$ or $p_{BM} \to 0 \implies K/N \to 0$)
- **Good**: Can achieve arbitrarily small error up to some maximum rate strictly less than the channel capacity (i.e, for any ϵ a good coding scheme can make a code with $K/N = R_{max} < C$ and $p_{BM} < \epsilon$)
- **Very Good**: Can achieve arbitrarily small error at any rate up to the channel capacity (i.e., for any $\epsilon > 0$ a very good coding scheme can make a code with $K/N = C$ and $p_{BM} < \epsilon$)
- **Practical**: Can be coded and decoded in time that is polynomial in the block length N.

Random Codes

During the discussion of the Noisy-Channel Coding Theorem we saw how to construct very good **random codes** via typical set decoding

Properties:

- Very Good: Rates up to C are achievable with arbitrarily small error
- Construction is easy
- Not Practical:
  - The 2^K codewords have no structure and must be "memorised"
  - Typical set decoding is expensive

Linear Codes

(N, K) Block Code

An (N, K) **block code** is a list of $S = 2^K$ codewords $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}\}$, each of length N. A message $s \in \{1, 2, \ldots, 2^K\}$ is encoded as $\mathbf{x}^{(s)}$.

Linear Codes

(N, K) Block Code

An (N, K) **block code** is a list of $S = 2^K$ codewords $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}\}$, each of length N. A message $s \in \{1, 2, \ldots, 2^K\}$ is encoded as $\mathbf{x}^{(s)}$.

Linear (N, K) Block Code

A **linear** (N, K) **block code** is an (N, K) block code where s is first represented as a K-bit binary vector $\mathbf{s} \in \{0, 1\}^K$ and then encoded via multiplication by an $N \times K$ binary matrix $\mathbf{G}^\top$ to form $\mathbf{t} = \mathbf{G}^\top\mathbf{s}$ modulo 2.

Here linear means all $S = 2^K$ messages can be obtained by "adding" different combinations of the K codewords $\mathbf{t}_i = \mathbf{G}^\top\mathbf{e}_i$ where $\mathbf{e}_i$ is K-bit string with single 1 in position i.

Linear Codes

(N, K) Block Code

An (N, K) **block code** is a list of $S = 2^K$ codewords $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}\}$, each of length N. A message $s \in \{1, 2, \ldots, 2^K\}$ is encoded as $\mathbf{x}^{(s)}$.

Linear (N, K) Block Code

A **linear** (N, K) **block code** is an (N, K) block code where s is first represented as a K-bit binary vector $\mathbf{s} \in \{0, 1\}^K$ and then encoded via multiplication by an $N \times K$ binary matrix $\mathbf{G}^\top$ to form $\mathbf{t} = \mathbf{G}^\top \mathbf{s}$ modulo 2.

Here linear means all $S = 2^K$ messages can be obtained by "adding" different combinations of the K codewords $\mathbf{t}_i = \mathbf{G}^\top \mathbf{e}_i$ where $\mathbf{e}_i$ is K-bit string with single 1 in position i.

Example: Suppose $(N, K) = (7, 4)$. To send $s = 3$, first create $\mathbf{s} = 0011$ and send $\mathbf{t} = \mathbf{G}^\top \mathbf{s} = \mathbf{G}^\top(\mathbf{e}_0 + \mathbf{e}_1) = \mathbf{G}^\top \mathbf{e}_0 + \mathbf{G}^\top \mathbf{e}_1 = \mathbf{t}_0 + \mathbf{t}_1$ where $\mathbf{e}_0 = 0001$ and $\mathbf{e}_1 = 0010$.

Types of Linear Code

Many commonly used codes are linear:

- Repetition Codes: e.g., $0 \rightarrow 000$; $1 \rightarrow 111$
- Convolution Codes: Linear coding plus bit shifts
- Concatenation Codes: Two or more levels of error correction
- Hamming Codes: Parity checking
- Low-Density Parity-Check Codes: Semi-random construction

Types of Linear Code

Many commonly used codes are linear:

- Repetition Codes: e.g., 0 $\rightarrow$ 000 ; 1 $\rightarrow$ 111
- Convolution Codes: Linear coding plus bit shifts
- Concatenation Codes: Two or more levels of error correction
- Hamming Codes: Parity checking
- Low-Density Parity-Check Codes: Semi-random construction

An NCCT can be proved for linear codes (i.e., “there exists a linear code” replacing “there exists a code”) but the proof is still non-constructive.

Types of Linear Code

Many commonly used codes are linear:

- Repetition Codes: e.g., 0 $\rightarrow$ 000 ; 1 $\rightarrow$ 111
- Convolution Codes: Linear coding plus bit shifts
- Concatenation Codes: Two or more levels of error correction
- Hamming Codes: Parity checking
- Low-Density Parity-Check Codes: Semi-random construction

An NCCT can be proved for linear codes (i.e., "there exists a linear code" replacing "there exists a code") but the proof is still non-constructive.

Practical linear codes:

- Use very large block sizes N
- Based on semi-random code constructions
- Apply probabilistic decoding techniques
- Used in wireless and satellite communication

Linear Codes: Examples

(7,4) Hamming Code

$$\mathbf{G}^\top = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

For $\mathbf{s} = 0011$,

$$\mathbf{G}^\top \mathbf{s} (\bmod 2) = [0\ 0\ 1\ 1\ 1\ 0\ 0]^\top$$

(6,3) Repetition Code

$$\mathbf{G}^\top = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

For $\mathbf{s} = 010$,

$$\mathbf{G}^\top \mathbf{s} (\bmod 2) = [0\ 1\ 0\ 0\ 1\ 0]^\top$$

Decoding

We can construct codes with a relatively simple encoding but how do we decode them? That is, given the input distribution and channel model Q how do we find the posterior distribution over $\mathbf{x}$ given we received $\mathbf{y}$?

Decoding

We can construct codes with a relatively simple encoding but how do we decode them? That is, given the input distribution and channel model Q how do we find the posterior distribution over $\mathbf{x}$ given we received $\mathbf{y}$?

Simple? Just compute

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{C}} P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}$$

But:

- the number of codes $\mathbf{x} \in \mathcal{C}$ is 2^K so, naively, the sum is expensive
- linear codes provide structure that the above method doesn't exploit

Summary and Reading

Main Points:

- Joint Typicality and the Joint Typicality Theorem
- The (Longer) Noisy Channel Coding Theorem
- Proof Ideas
  - Random Coding & Typical Set Decoding
  - Average Error Over Random Codes
  - Code Expurgation

Reading:

- MacKay §9.7, §10.1-§10.5