On the rates of codes for high noise binary symmetric channels

Gábor P. Nagy

joint work with M. Maróti

University of Szeged (Hungary)

ALCOMA 2015, Kloster Banz
March 15-20, 2015

Basic concepts

Codes: linear codes of length n and dimension k over a field K (mostly $K = \mathbb{F}_2$)

Messages: random elements of K^k (pseudo-random, of course)

Channel: Binary Symmetric Channel with Bit Error Ratio p (I love these 3-letter acronyms: BSC, BER, TLA,...)

Decoding: hard decoding, nearest codeword (=maximum likelihood) (except when not)

Basic concepts

Codes: linear codes of length n and dimension k over a field K (mostly $K = \mathbb{F}_2$)

Messages: random elements of K^k (pseudo-random, of course)

Channel: Binary Symmetric Channel with Bit Error Ratio p (I love these 3-letter acronyms: BSC, BER, TLA,...)

Decoding: hard decoding, nearest codeword (=maximum likelihood) (except when not)

Basic concepts

Codes: linear codes of length n and dimension k over a field K (mostly $K = \mathbb{F}_2$)

Messages: random elements of K^k (pseudo-random, of course)

Channel: Binary Symmetric Channel with Bit Error Ratio p (I love these 3-letter acronyms: BSC, BER, TLA,...)

Decoding: hard decoding, nearest codeword (=maximum likelihood) (except when not)

Basic concepts

Codes: linear codes of length n and dimension k over a field K (mostly $K = \mathbb{F}_2$)

Messages: random elements of K^k (pseudo-random, of course)

Channel: Binary Symmetric Channel with Bit Error Ratio p (I love these 3-letter acronyms: BSC, BER, TLA,...)

Decoding: hard decoding, nearest codeword (=maximum likelihood) (except when not)

Cost-Benefit Analysis of codes

Cost: Expressed by the rate $R = \frac{k}{n}$ of the code

Benefit: Many definitions...

- Minimum distance d; the error correction ratio $\lfloor \frac{d-1}{2} \rfloor / n$
  Good theoretical tool for combinatorics and geometry
- Probability of wrong decoding of codewords

$$P_C = \frac{1}{|C|} \sum_{w \in C} P_{C,w}$$

  Good theoretical tool for probablity and information theory
  NB!!! Depends on p
- Maximum probability of wrong decoding of codewords
  Useful for engineers, can be estemated by p and d
- Improved Bit Error Ratio: bit errors after decoding

Cost-Benefit Analysis of codes

Cost: Expressed by the rate $R = \frac{k}{n}$ of the code

Benefit: Many definitions...

- Minimum distance d; the error correction ratio $\lfloor \frac{d-1}{2} \rfloor / n$
  Good theoretical tool for combinatorics and geometry
- Probability of wrong decoding of codewords

$$P_C = \frac{1}{|C|} \sum_{w \in C} P_{C,w}$$

  Good theoretical tool for probablity and information theory
  NB!!! Depends on p
- Maximum probability of wrong decoding of codewords
  Useful for engineers, can be estemated by p and d
- Improved Bit Error Ratio: bit errors after decoding

Cost-Benefit Analysis of codes

Cost: Expressed by the rate $R = \frac{k}{n}$ of the code

Benefit: Many definitions...

- Minimum distance d; the error correction ratio $\lfloor \frac{d-1}{2} \rfloor / n$
  Good theoretical tool for combinatorics and geometry
- Probability of wrong decoding of codewords

$$P_C = \frac{1}{|C|} \sum_{w \in C} P_{C,w}$$

  Good theoretical tool for probablity and information theory
  NB!!! Depends on p
- Maximum probability of wrong decoding of codewords
  Useful for engineers, can be estemated by p and d
- Improved Bit Error Ratio: bit errors after decoding

Cost-Benefit Analysis of codes

Cost: Expressed by the rate $R = \frac{k}{n}$ of the code

Benefit: Many definitions...

- Minimum distance d; the error correction ratio $\lfloor \frac{d-1}{2} \rfloor / n$
  Good theoretical tool for combinatorics and geometry
- Probability of wrong decoding of codewords

$$P_C = \frac{1}{|C|} \sum_{w \in C} P_{C,w}$$

  Good theoretical tool for probablity and information theory
  NB!!! Depends on p
- Maximum probability of wrong decoding of codewords
  Useful for engineers, can be estemated by p and d
- Improved Bit Error Ratio: bit errors after decoding

Cost-Benefit Analysis of codes

Cost: Expressed by the rate $R = \frac{k}{n}$ of the code

Benefit: Many definitions...

- Minimum distance d; the error correction ratio $\lfloor \frac{d-1}{2} \rfloor / n$
  Good theoretical tool for combinatorics and geometry
- Probability of wrong decoding of codewords

$$P_C = \frac{1}{|C|} \sum_{w \in C} P_{C,w}$$

  Good theoretical tool for probablity and information theory
  NB!!! Depends on p
- Maximum probability of wrong decoding of codewords
  Useful for engineers, can be estemated by p and d
- Improved Bit Error Ratio: bit errors after decoding

Cost-Benefit Analysis of codes

Cost: Expressed by the rate $R = \frac{k}{n}$ of the code

Benefit: Many definitions...

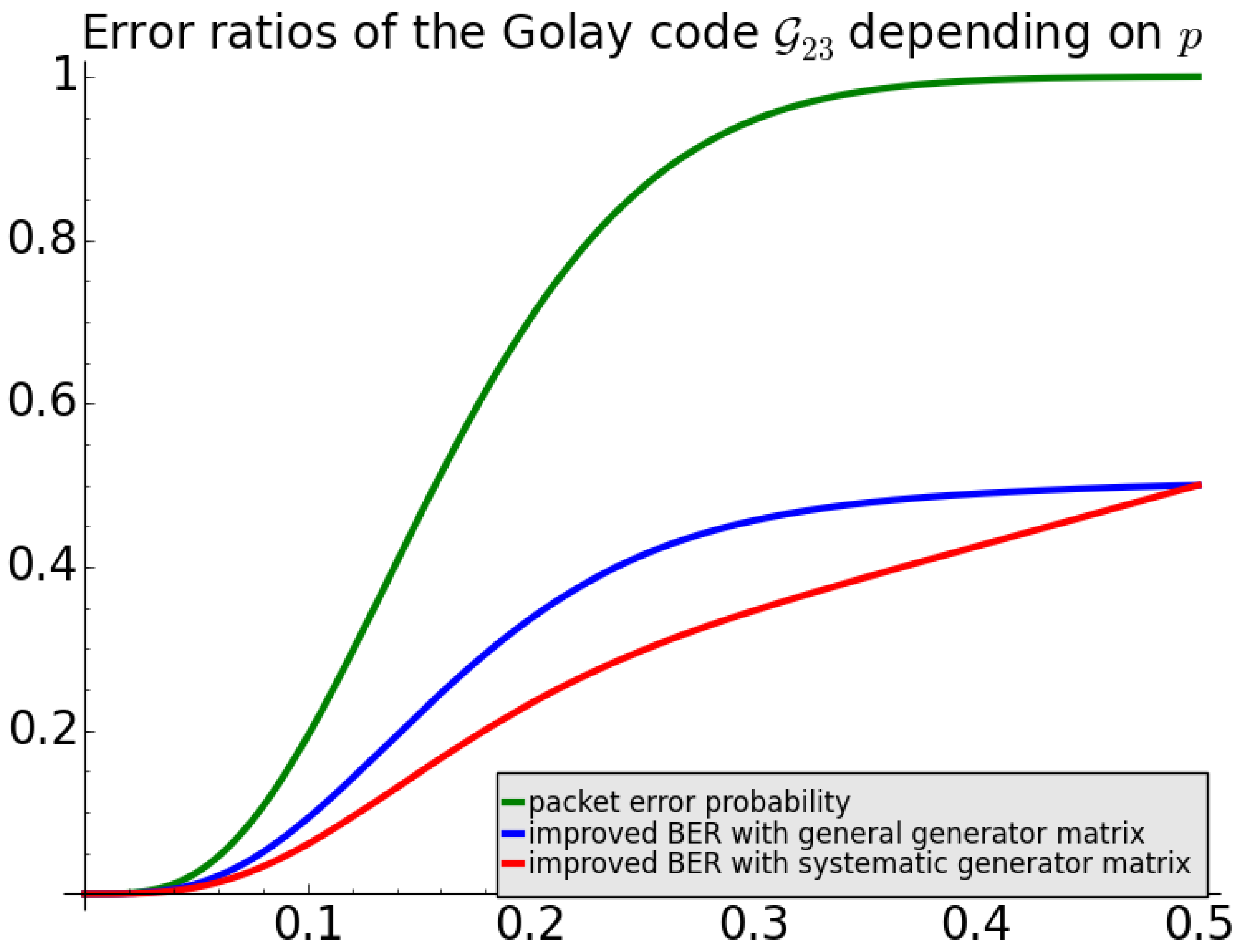
- Minimum distance d; the error correction ratio $\lfloor \frac{d-1}{2} \rfloor / n$
  Good theoretical tool for combinatorics and geometry
- Probability of wrong decoding of codewords

$$P_C = \frac{1}{|C|} \sum_{w \in C} P_{C,w}$$

  Good theoretical tool for probablity and information theory
  NB!!! Depends on p
- Maximum probability of wrong decoding of codewords
  Useful for engineers, can be estemated by p and d
- Improved Bit Error Ratio: bit errors after decoding

Improved Bit Error Ratio

- Only for engineers!!! Depends on the *generator matrix*...
- Can be estimated by simulation.



The Challenge I: Fixing the benefit

- We have to transmit 3000 bits on a BSC with $p = 0.1$
- such that ≤ 3 incorrect bits are received
- with "some high probablity" for random streams of 3000 bits.
- **Notice:** This means an improved BER < 0.0005.

Definion: "Good code"

- Let C be a binary linear code given by its generator matrix.
- We make simulations for the improved BER with $p = 0.1$ and (pseudo-)random bit stream of length 3000, using error correction with C.
- We say that C is good, if the simulated BER value is ≤ 0.001 for at least 4 simulations out of 5.

- It is easy to show that the repetition code of length 11 is good.

The Challenge I: Fixing the benefit

- We have to transmit 3000 bits on a BSC with $p = 0.1$
- such that ≤ 3 incorrect bits are received
- with "some high probablity" for random streams of 3000 bits.
- **Notice:** This means an improved BER < 0.0005.

Definion: "Good code"

- Let C be a binary linear code given by its generator matrix.
- We make simulations for the improved BER with $p = 0.1$ and (pseudo-)random bit stream of length 3000, using error correction with C.
- We say that C is good, if the simulated BER value is ≤ 0.001 for at least 4 simulations out of 5.

- It is easy to show that the repetition code of length 11 is good.

The Challenge II: Minimizing the cost

The Challenge

Find good codes with high rate.

Remarks:

- The repetition code of length 11 has rate $R = 1/11 \approx 0.0909$.
- You must be able to **run the simulation** for your code in a reasonable amount of time!!!
- That is, the code must be **explicitly given** with implemented decoding algorithm.

The Team

- The supervisors: GN, M. Maróti (Szeged), P. Müller and F. Möller (Würzburg).
- Master and PhD students of the University of Szeged (Hungary) and the University of Potenza (Italy).

- Simulations were done in SageMath.
- SageMath uses Python: easy to program but slow.

On rates of good codes: Shannon's Theorems

- Define the entropy function

$$h(p) = -p \log_2 p - (1-p)\log_2(1-p), \qquad 0 \le p \le 1.$$

Shannon's Theorems

1. Let $0 < R < 1 - h(p)$ and $\mathcal{F}_n$ be a balanced family of linear codes with codewords of length n and dimension $k = \lfloor Rn \rfloor$. Then
$$\min_{C \in \mathcal{F}_n} P_C \to 0, \qquad n \to \infty.$$
2. If $C_n \subseteq \mathbb{F}_2^n$ is a sequence of codes such that for some fixed $K > 1 - h(p)$
$$K \le R_{C_n} \le 1$$
holds, then $\lim_{n\to\infty} P_{C_n} = 1$.

- We have the upper bound $1 - h(0.1) = 0.531$ for the rates of good codes.

NP-completeness of decoding of binary codes

Theorem (Berlekamp, McEliece, van Tilborg 1978)

The following problem is NP-complete:
Given a linear subspace $C \leq \mathbb{F}_2^n$, a vector $y \in \mathbb{F}_2^n$ and a positive integer w.
Does there exist an element $x \in C$ such that $d_H(x, y) \leq w$?

- **Straightforward implementations** of maximum likelihood decoding stop working at $k \approx 20$, $n \approx 60$.
- Good **random codes** with rate ≈ 0.2 are found easily.

Some classes of binary codes

- Binary Golay codes **fail badly.........** for bit error ratio $p > 0.05$.
- Good binary BCH codes with rate > 0.2 are **hard to find.**
  Algebraic decoding only up to the designed minimum distance
- Product codes are **good!!!**
  (Extended Golay) $*$ (Extended Golay) has rate $R = 0.25$.
- **Good** convolution codes with parameters $n = 100$, $k = 30$ give rates $R = 0.3$.

Some classes of binary codes

- Binary Golay codes **fail badly.........** for bit error ratio $p > 0.05$.
- Good binary BCH codes with rate > 0.2 are **hard to find.**
  Algebraic decoding only up to the designed minimum distance
- Product codes are **good!!!**
  (Extended Golay) $*$ (Extended Golay) has rate $R = 0.25$.
- **Good** convolution codes with parameters $n = 100$, $k = 30$ give rates $R = 0.3$.

Some classes of binary codes

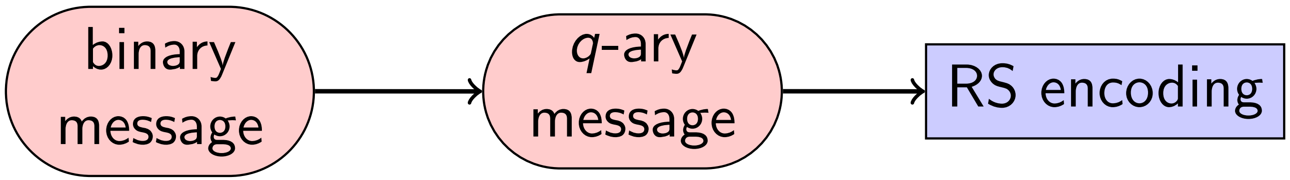
- Binary Golay codes **fail badly.........** for bit error ratio $p > 0.05$.
- Good binary BCH codes with rate > 0.2 are **hard to find.**
  Algebraic decoding only up to the designed minimum distance
- Product codes are **good!!!**
  (Extended Golay) $*$ (Extended Golay) has rate $R = 0.25$.
- **Good** convolution codes with parameters $n = 100$, $k = 30$ give rates $R = 0.3$.

Some classes of binary codes

- Binary Golay codes **fail badly.........** for bit error ratio $p > 0.05$.
- Good binary BCH codes with rate > 0.2 are **hard to find.**
  Algebraic decoding only up to the designed minimum distance
- Product codes are **good!!!**
  (Extended Golay) $*$ (Extended Golay) has rate $R = 0.25$.
- **Good** convolution codes with parameters $n = 100$, $k = 30$ give rates $R = 0.3$.

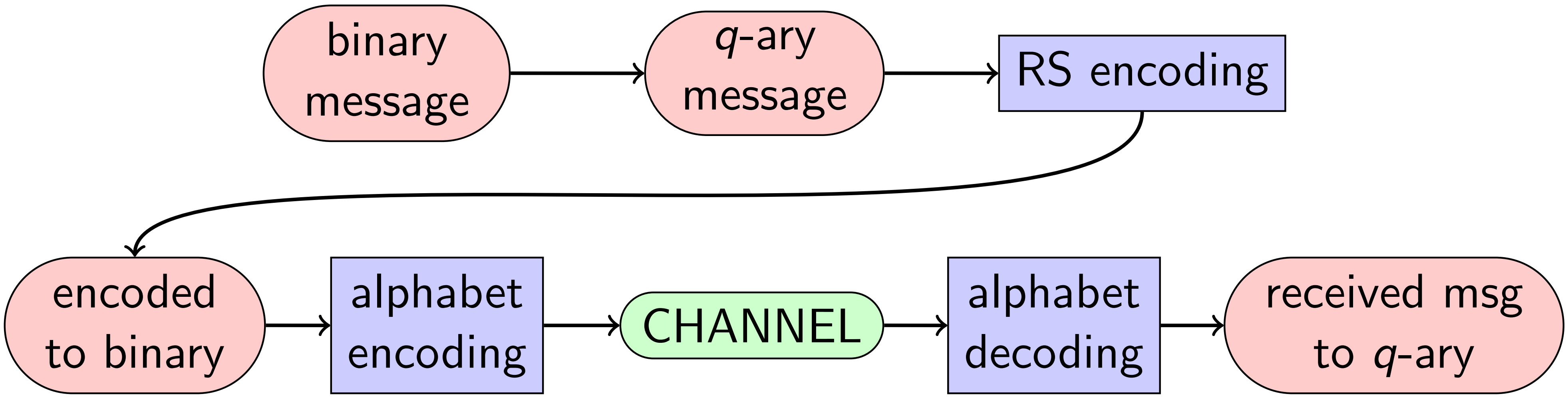
Reed-Solomon codes over $\mathbb{F}_q$, $q = 2^f$

- Non-binary linear code $\Longrightarrow$ We need **alphabet coding.**



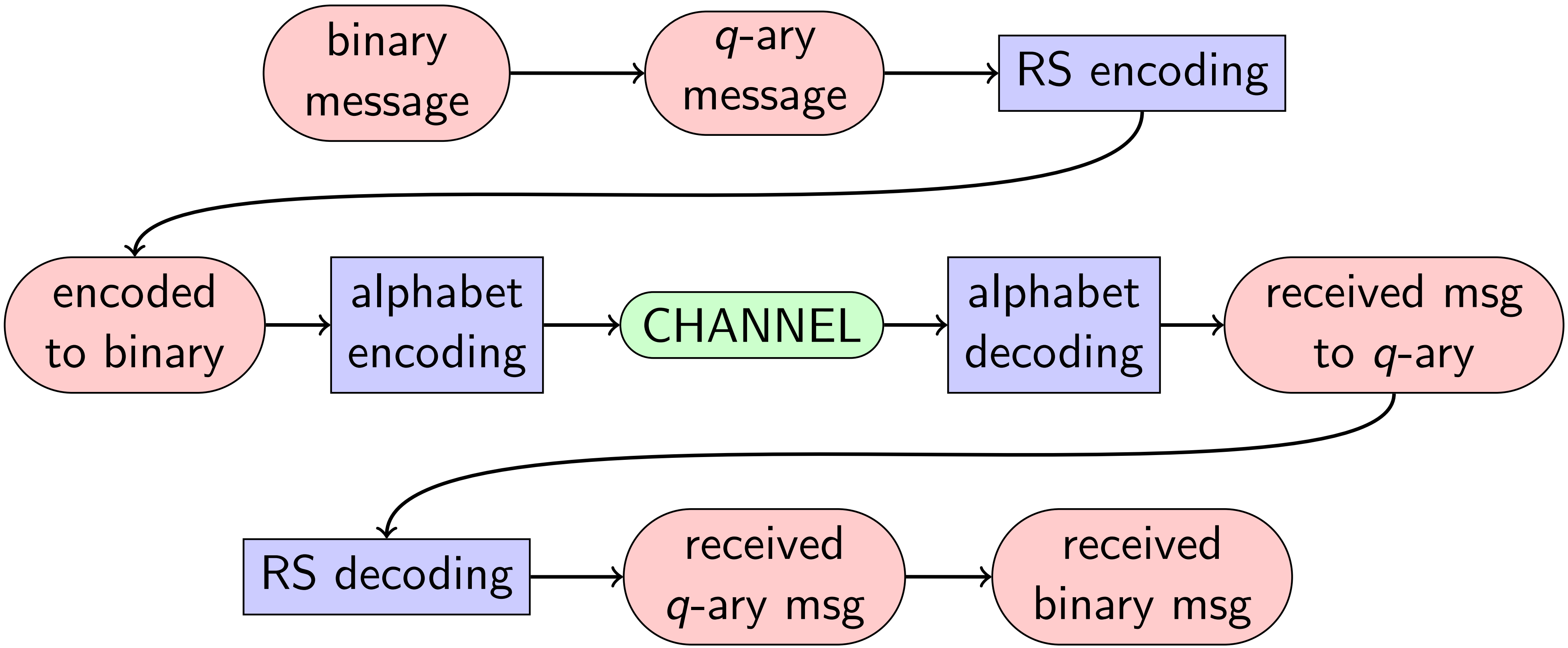
Reed-Solomon codes over $\mathbb{F}_q$, $q = 2^f$

- Non-binary linear code $\Longrightarrow$ We need **alphabet coding.**



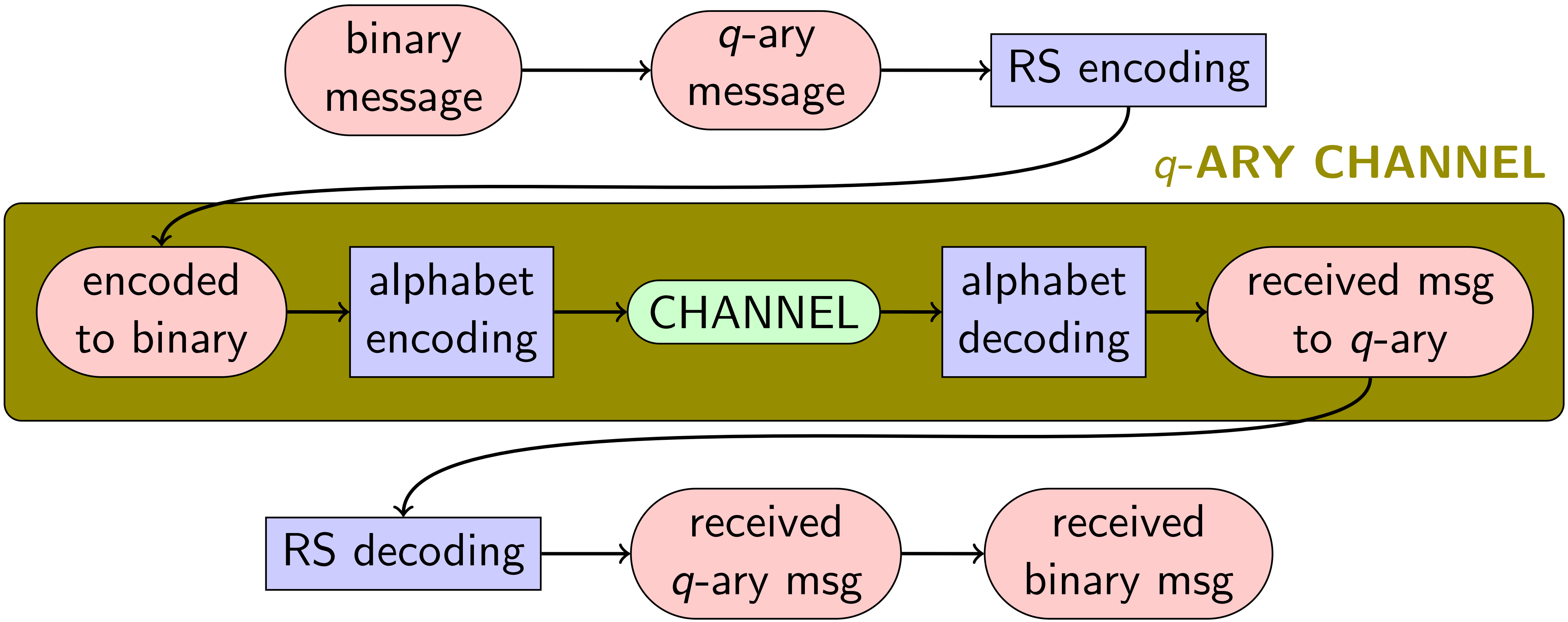
Reed-Solomon codes over $\mathbb{F}_q$, $q = 2^f$

- Non-binary linear code $\Longrightarrow$ We need **alphabet coding.**



Reed-Solomon codes over $\mathbb{F}_q$, $q = 2^f$

- Non-binary linear code $\Longrightarrow$ We need **alphabet coding.**



- The middle layer can be seen as a q-ary channel with erasure.
- For $q = 2^6$ and an $(17, 6)$ alphabet code we reached $R = 0.27$.

Decoding by solvers – preliminary impressions

Different NP-complete problems have good performing solver software:

- **INTEGER PROGRAMMING** (GLPK, SCIP, GUROBI, etc.)
  works for $k \approx 40$, $n \approx 80$
  performs better with sparse parity check matrix.
- **SAT-SOLVER** (MiniSAT, Glucose, etc.)
  works for $k \approx 30$, $n \approx 70$.
  performs better with sparse parity check matrix.
- **GROEBNER BASIS** (approach by M. Borges-Quintana, M. A. Borges-Trenard, P. Fitzpatrick, E. Martínez-Moro 2008)
  not usable in practice,
  the Groebner basis is larger than the standard array.

Decoding by solvers – preliminary impressions

Different NP-complete problems have good performing solver software:

- **INTEGER PROGRAMMING** (GLPK, SCIP, GUROBI, etc.)
  works for $k \approx 40$, $n \approx 80$
  performs better with sparse parity check matrix.
- **SAT-SOLVER** (MiniSAT, Glucose, etc.)
  works for $k \approx 30$, $n \approx 70$.
  performs better with sparse parity check matrix.
- **GROEBNER BASIS** (approach by M. Borges-Quintana, M. A. Borges-Trenard, P. Fitzpatrick, E. Martínez-Moro 2008)
  not usable in practice,
  the Groebner basis is larger than the standard array.

Decoding by solvers – preliminary impressions

Different NP-complete problems have good performing solver software:

- **INTEGER PROGRAMMING** (GLPK, SCIP, GUROBI, etc.)
  works for $k \approx 40$, $n \approx 80$
  performs better with sparse parity check matrix.
- **SAT-SOLVER** (MiniSAT, Glucose, etc.)
  works for $k \approx 30$, $n \approx 70$.
  performs better with sparse parity check matrix.
- **GROEBNER BASIS** (approach by M. Borges-Quintana, M. A. Borges-Trenard, P. Fitzpatrick, E. Martínez-Moro 2008)
  not usable in practice,
  the Groebner basis is larger than the standard array.

Open problem

The Challenge (Beer+Pizza)

Find good codes with rate > 0.3.

THANK YOU FOR YOUR ATTENTION!