Статистический анализ корректирующих способностей различных методов при передаче данных в каналах со стираниями

A Preprint

Кожанов Илья Романович
Кафедра Математических Методов Прогнозирования
Факультет Вычислительной Математики и Кибернетики
Московский государственный университет имени М. В. Ломоносова
Научный руководитель: Гуров Сергей Исаевич
ilya.kozanov@yandex.ru

Abstract

В представленной работе осуществляется статистическое моделирование различных методов восстановления стертых пакетов при передаче данных в каналах типа «стирание». Основной целью данного исследования является сравнение методов посредством эмулирования их работы статистическими моделями и подсчетом их корректирующей способности с помощью этих моделей. Для некоторых методов восстановления пакетов аналитический расчет корректирующей способности невозможен или крайне затруднителен в виду сложной математической модели функционирования данных методов, а также различной эффективности при различных паттернах ошибок. Статистическое моделирование их работы позволяет рассчитать корректирующие способности таких методов и сравнить эффективность их работы. В данной рассматриваются метод добавления контрольного пакета четности, его модификация even/odd, коды Рида-Соломона, Фонтанные коды (в частности raptor-коды).

Keywords Канал со стираниями · Статистическое моделирование · Коды Рида-Соломона · Фонтанные коды

1 Введение

При передаче данных от источника к приемнику могут произойти сбои оборудования, приводящие к ошибкам различного вида. В данной работе рассматриваются часто наблюдаемые ошибки вида потеря пакета. Во многих протоколах при потере пакетов принимающая сторона запрашивает потерянные данные еще раз. Однако при потоковой передаче такая процедура занимает критически много времени. В таких случаях чтобы обнаружить и исправить ошибку, применяют помехоустойчивое кодирование, т.е. кодируют сообщение таким образом, чтобы принимающая сторона знала, произошла ошибка или нет, а так же исправить их в случае возникновения. В работе рассматриваются различные методы систематического кодирования сообщения (пакеты четности, even/odd и аналогичные модификации, коды Рида-Соломона) и анализируется их корректирующая способность на основании статистического моделирования их работы.

2 Постановка задачи

Для сравнительного анализа выбранных методов была выбрана часто используемая модель двоичного симметричного канала со стираниями (ДСКС) [В., 2011]. В данной модели на вход в канал подается последовательность пакетов данных. На выходе для каждого из пакетов существует одинаковая,

фиксированная вероятность q возникновения ошибки типа стирание, при которой теряется информация о содержимом пакета, но сохраняется информация о факте наличия ошибки и ее локализации в последовательности. Так же с фиксированной вероятностью p_0 при передаче очередного пакета возникает ошибка типа трансформация, при которой данные при передаче в силу несовершенности канала неверно интерпретируются на выходе из него, в результате чего пакет модифицируется. Однако, в отличие от ошибки типа стирание, такая ошибка не детектируется. На рисунке 1 представлен граф модели двоичного симметричного канала со стираниями для тривиального случая длины пакета, равной 1.

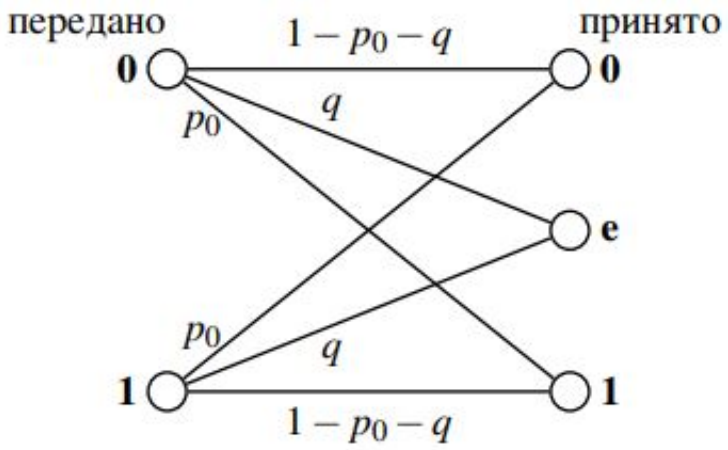

Рис. 1: Граф модели двоичного симметричного канала со стираниями

На практике, ошибки типа трансформация возникают значительно реже, чем ошибки типа стирание, поэтому в рамках данной работы рассматривается модель двоичного симметричного канала со стираниям с параметрами $p_0 = 0$ и $q \in [0.002; 0.05]$. На рисунке 2 представлен граф такой модели. Чтобы исправить

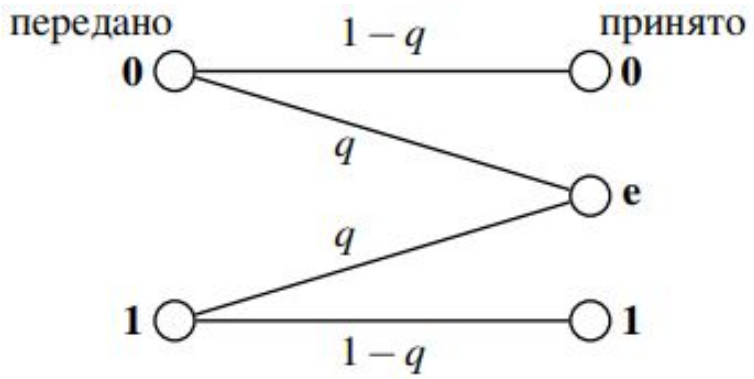

Рис. 2: Граф модели двоичного симметричного канала со стираниями для $p_0 = 0$

ошибку, применяют помехоустойчивое кодирование, т.е. кодируют сообщение таким образом, чтобы принимающая сторона знала, произошла ошибка или нет, и при могла исправить ошибки в случае их возникновения. По сути, кодирование — это добавление к исходной информации дополнительной, проверочной, информации. Для кодирования на передающей стороне используются энкодер, а на принимающей стороне — декодер для получения исходного сообщения. [Шинкаренко К. В., 2008]
Введем некоторые необходимые понятия:
Расстояние Хемминга - число позиций, в которых соответствующие символы двух слов одинаковой длины различны. В более общем случае расстояние Хэмминга применяется для строк одинаковой длины любых q-ичных алфавитов и служит метрикой различия (функцией, определяющей расстояние в метрическом пространстве) объектов одинаковой размерности.
Корректирующая способность — характеристика t кода C, описывающая возможность исправить ошибки в кодовых словах. Определяется как целое число, меньшее половины от минимального расстояния d_{min}

между кодовыми словами минус один в принятой метрике кода:

$$t = \left\lfloor \frac{(d_{min} - 1)}{2} \right\rfloor$$

Для Хеммингoвой метрики корректирующую способность кода можно определить как максимальный радиус сфер Хемминга, при котором для двух различных кодовых векторов сферы не пересекаются [P., 2006]:

$$t = \vec{v}_i, \vec{v}_j \in C \max \{l | S_l(\vec{v}_i) \cap S_l(\vec{v}_j) = \emptyset, \vec{v}_i \neq \vec{v}_j\}$$

Необходимо передать по двоичному симметричному каналу со стираниями сообщение длины K, которое состоит из K пакетов-символов. Каждый пакет имеет фиксированную вероятность возникновения ошибки типа стирание = q. Пусть после некоторого преобразования исходного сообщения, длина сообщения стала больше и теперь составляет $N = K + M$ пакетов. Тогда назовем M - избыточными пакетами, а M/N - избыточностью кода.

При этом, в результате кодирования, восстанавления и декодирования сообщения вероятность потери каждого пакета должна не превышать $\pi = 0.001$ (при вероятностной интерпретации полученного сообщения аналогичной исходному). В рамках данной работы рассматриваются систематические методы кодирования сообщения, то есть такие, в которых избыточные контрольные пакеты дописываются в конец сообщения.

Метрикой качества того или иного метода будем считать минимальную избыточность при кодировании сообщения, необходимую для снижения вероятности стирания пакета до нужной.

Выбор наилучшего кода будет зависеть от многих факторов, включая ограничения по используемому аппаратному обеспечению, тип используемых данных и требуемую производительность системы. В данной работе рассмотрены следующие методы - пакеты четности, even/odd и аналогичные модификации, коды Рида-Соломона.

3 Применение методов

3.1 Реализация моделирования работы канала

Статистическое моделирование работы двоичного канала со стираниями реализовано на языке Python3 с помощью встроенного генератора псевдослучайных чисел. Максимальная длина сообщения в пакетах вычислялась с помощью алгоритма бинарного поиска по значениям K и соответствующим им вероятностям стирания пакета после применения метода - π. Они, в свою очередь, расчитывались для каждого фиксированного K усреднением количества не восстановленных после применения метода пакетов по 10^6 (в случае кодов Рида-Соломона - $3 * 10^5$) независимым экспериментам.

3.2 Добавление контрольного пакета четности

Принцип работы данного метода заключается в том, что к сообщению добавляется 1 проверочный пакет, который формируется побитовым применением логической операции XOR между всеми пакетами исходного сообщения. Таким образом, в случае стирания одного пакета, его можно будет восстановить путем применения логической операции XOR между проверочным пакетом и всеми пакетами исходного сообщения, кроме стертого. Данный метод позволяет гарантированно восстанавливать 1 пакет в случае единичного стирания, но в случае возникновения большего количества ошибок или в случае стирания проверочного пакета при передаче он не позволяет восстановить ни одного из стертых пакетов. Избыточность метода составляет $\frac{1}{K+1}$.

Зафиксируем некоторое значение q. Аналитическая формула максимальной длины сообщения, при которой в результате применения метода добавления контрольного пакета четности вероятность стирания пакета $\leq \pi$.

$$K = \left\lfloor \frac{ln(\pi) - ln(q)}{ln(1 - q)} \right\rfloor$$

В результате статистического моделирования работы метода были получены результаты, представленные в таблице 1.

На рисунке 3 представлен график зависимости вероятности стирания пакета после применения метода - π (вероятность того, что отдельно взятый пакет окажется невосстановленным) от исходной длины сообщения для исходной вероятности стирания пакета $q = 0.002$

Вероятность стирания пакета до применения метода (q)	Число информационных пакетов (K)	Вероятность стирания пакета после применения метода (pi)	Вероятность стирания пакета для сообщения длины K+1
0.002	346	0.000999	0.001002
0.003	134	0.000997	0.001001
0.005	44	0.000995	0.001015
0.007	21	0.000962	0.001009
0.01	10	0.000953	0.001051
0.02	2	0.000791	0.001181
0.03	1	0.000899	0.001772
0.05	-	-	-

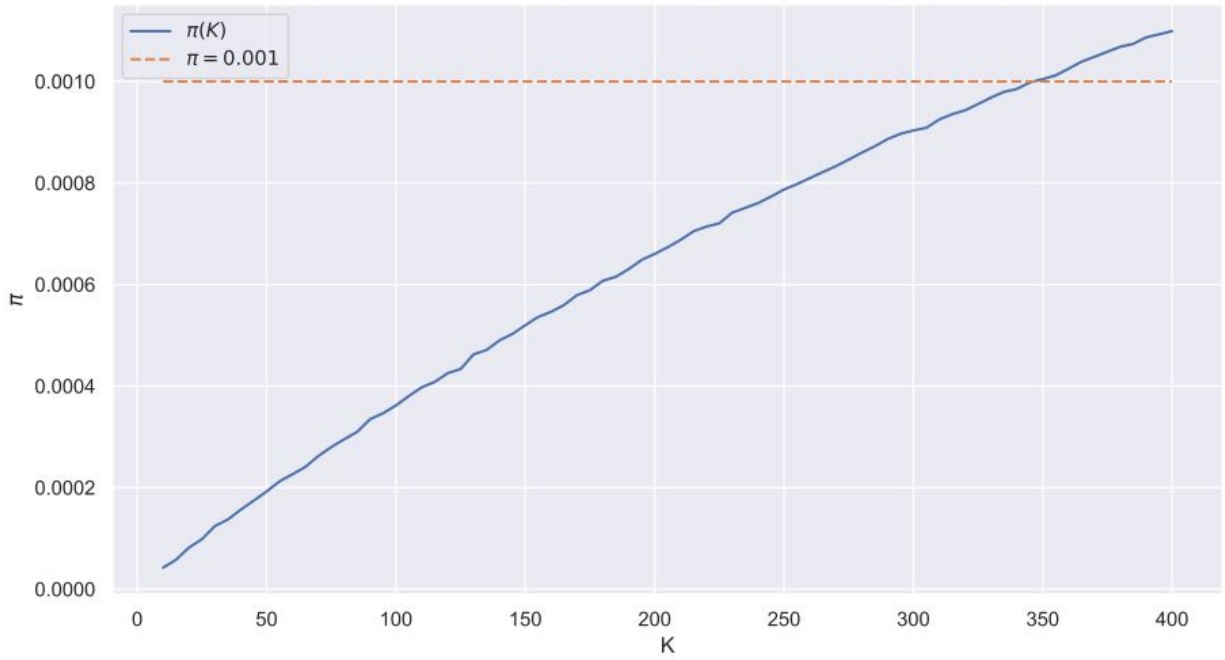

Рис. 3: График зависимости π от K

Среднеквадратическая ошибка с аналитической формулой - $2 * 10^{-9}$

3.3 Модификация even/odd

Преимуществом метода добавления контрольного пакета четности является простота масштабирования. Модификация even/odd подразумаевает добавление в конец сообщения не одного, а двух контрольных пакетов четности, один из которых формируется побитовым применением логической операции XOR между пакетами исходного сообщения, стоящими на четных позициях, а другой - между пакетами исходного сообщения, стоящими на нечетных позициях. Таким образом, данный метод позволяет гарантированно восстанавливать 1 стертый пакет и с некоторой вероятностью 2 пакета (в том случае, если за них отвечают разные проверочные пакеты). Избыточность метода составляет $\frac{2}{K+1}$.

Аналитическая формула максимальной длины сообщения, при которой в результате применения метода добавления контрольного пакета четности вероятность стирания пакета $\leq \pi$.

$$K = \left\lfloor 2 \cdot \frac{ln(\pi) - ln(q)}{ln(1-q)} \right\rfloor$$

В результате статистического моделирования работы метода были получены результаты, представленные в таблице 2:

Таким образом, метод even/odd является логичным развитием метода добавления контрольного пакета четности. Дальнейшее увеличение количества контрольных пакетов позволяет линейно увеличить максимальную длину передаваемого сообщения.

Вероятность стирания пакета до применения метода (q)	Число информационных пакетов (K)	Вероятность стирания пакета после применения метода (pi)	Вероятность стирания пакета для сообщения длины K+1
0.002	692	0.000998	0.001008
0.003	268	0.000985	0.001005
0.005	88	0.000970	0.001014
0.007	42	0.000996	0.001030
0.01	20	0.000941	0.001112
0.02	4	0.000802	0.001101
0.03	2	0.000780	0.001078
0.05	-	-	-

На рисунке 4 представлен график зависимости вероятности стирания пакета после применения метода - π (вероятность того, что отдельно взятый пакет окажется невосстановленным) от исходной длины сообщения для исходной вероятности стирания пакета $q = 0.002$

Среднеквадратическая ошибка с аналитической формулой - $1.7 * 10^{-9}$

3.4 Коды Рида-Соломона

Рида – Соломона коды (РС-коды) можно интерпретировать как недвоичные коды БЧХ (Боуза – Чоудхури – Хоквингема), значения кодовых символов которых взяты из поля $GF(2^r)$, т. е. r информационных символов отображаются отдельным элементом поля. Коды Рида – Соломона – это линейные недвоичные систематические циклические коды, символы которых представляют собой r-битовые последовательности, где r – целое положительное число, большее 1.

Коды Рида – Соломона (n, k) определены на r-битовых символах при всех n и k, для которых: $0 < k < n < 2r + 2$, где k – число информационных символов, подлежащих кодированию, n – число кодовых символов в кодируемом блоке. [Питерсон У., 1972]

РС-коды обладают наибольшим минимальным расстоянием (числом символов, которыми отличаются последовательности), возможным для линейного кода, определяемым следующим образом: $d_{min} = n - k + 1$. [К., 2003]

Замечательным свойством данного метода является тот факт, что Код Рида — Соломона, исправляющий t ошибок типа стирание, требует t проверочных пакетов, и с его помощью исправляются произвольные t ошибок и меньше. Данное свойство делает РС-коды лучшими среди систематических кодов с точки зрения избыточности в данной задаче. На практике применение данного метода усложняется необходимостью использования достаточно большого кодового алфавита. Размер алфавита должен быть не меньше длины кодовых слов, то есть не меньше длины пакета. Более того, отдельной задачей является оптимальный алгоритм декодирования сообщения - несмотря на теоритическую полиномиальную сложность, алгоритм достаточно медленный и вычислительно емкий. Данные проблемы решаются различными модификациями и улучшениями РС-кодов (например, каскадные коды), однако в рамках данной работы исследуется исключительно корректирующая способность методов. [Ромащенко А. Е., 2011]

Решение остановиться именно на РС-кодах среди методов и модификаций с аналогичной корректирующей способностью было продиктовано хорошей изученностью данного метода и наличием большого количества прикладных реализаций. Для декодирования используется алгоритм Берлекэмпа – Мэсси.[Е., 1983]

На рисунке 5 представлен график зависимости вероятности стирания пакета после применения метода - π (вероятность того, что отдельно взятый пакет окажется невосстановленным) от исходной длины сообщения для исходной вероятности стирания пакета $q = 0.005$. Для сравнения были выбраны несколько вариантов РС - кодов, исправляющих различное количество ошибок.

Как видно из графика, РС-коды значительно превосходят предыдущий метод при аналогичной избыточности. Ниже приведена таблица, в которой приведено требуемое количество пакетов, которые необходимо восстановить, чтобы понизить вероятность потери пакета до $\pi = 0.001$. Заметим, что в случае РС-кодов количество пакетов, которые необходимо восстановить, равно числу избыточных пакетов.

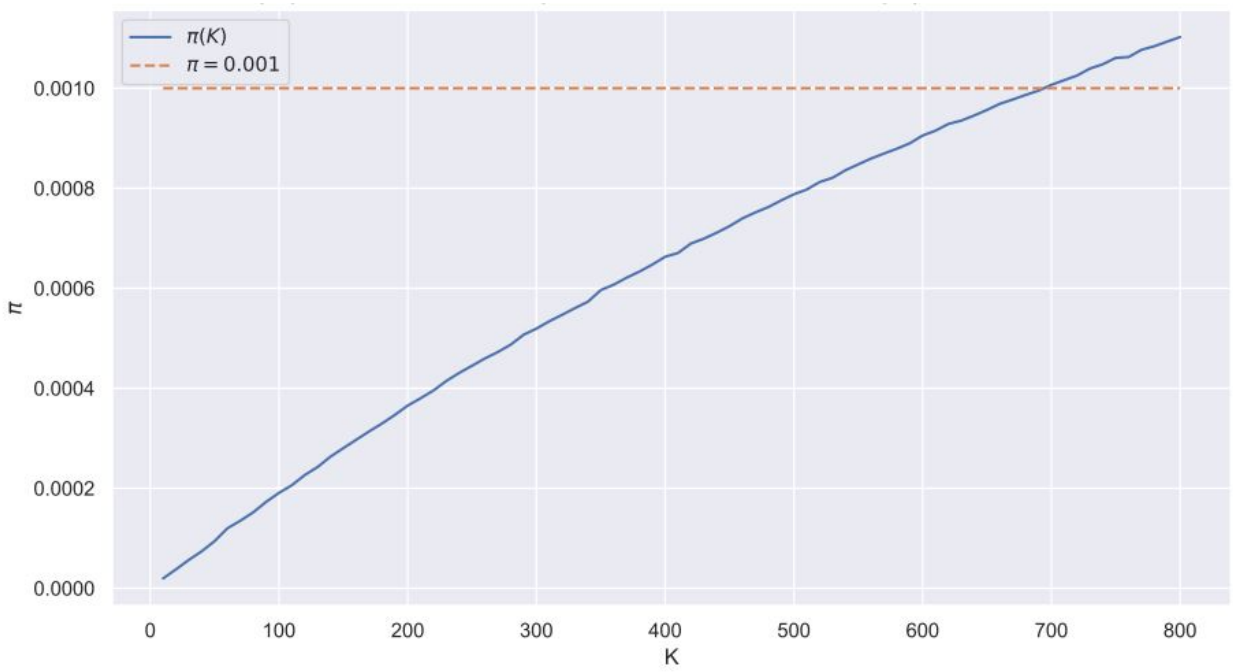

Рис. 4: График зависимости π от K

Исходная длина сообщения - K	Вероятность одного потери пакета в канале %	Кол-во пакетов, которые необходимо восстановить
5*5	0.2	1
	0.5	1
	0.7	1
	1	1
	1.5	2
5*10	0.2	1
	0.5	1
	0.7	1
	1	1
	1.5	2
5*50	0.2	1
	0.5	2
	0.7	2
	1	2
	1.5	3
5*100	0.2	1
	0.5	2
	0.7	3
	1	3
	1.5	5
5*300	0.2	1
	0.5	3
	0.7	5
	1	6
	1.5	9
5*500	0.2	2
	0.5	5
	0.7	7
	1	9
	1.5	13
5*1000	0.2	3
	0.5	8
	0.7	11
	1	15
	1.5	22

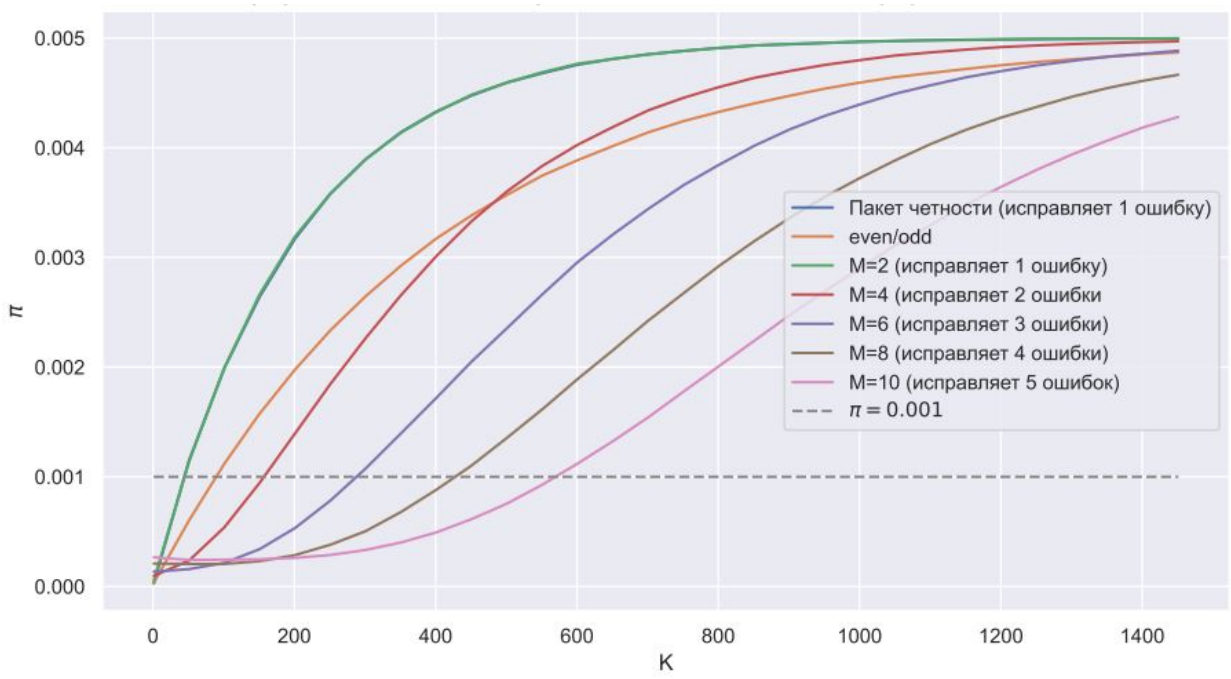

Рис. 5: График зависимости π от K

4 Вывод

В результате проведения статистического моделирования работы различных методов восстановления потерянных пакетов при передаче по двоичному каналу со стираниями было выявлено, что среди систематических кодов наилучший в метрике избыточности кода результат показали коды Рида - Соломона. Результаты, приведенные в таблице, позволяют оценить, какую конфигурацию кода Рида - Соломона (или аналогичных в смысле корректирующей способности) выбрать. Однако выбор конкретного метода и реализации в любом случае будет зависеть от специфики задачи. Стоит отметить, что сравнение производилось только среди кодов представляющих собой систематическое изменение исходного сообщения, то есть при кодировании не изменялисть исходные сообщения.

Список литературы

Марков М. В. Модели дискретных каналов связи. pages 143–156, 2011.

Кориков А. М. Шинкаренко К. В. Восстановление потерь пакетов в компьютерных сетях. pages 105–109, 2008.

Морелос-Сарагоса Р. Искусство помехоустойчивого кодирования. – Техносфера. pages 20—23, 2006.

Уэлдон Э. Питерсон У. Коды, исправляющие ошибки/пер. с англ. под ред. Р. Добрушина. 1972.

Касперски К. Могущество кодов Рида-Соломона, или Информация, воскресшая из пепла. pages 88–94, 2003.

Шень А. Ромащенко А. Е., Румянцев А. Ю. Заметки по теории кодирования. pages 17–20, 2011.

Blahut R. E. Theory and practice of error control codes. – reading : Addison-wesley. t 126. 1983.