Научно-технологический университет Сириус

Теоретико-информационный анализ нейросетей

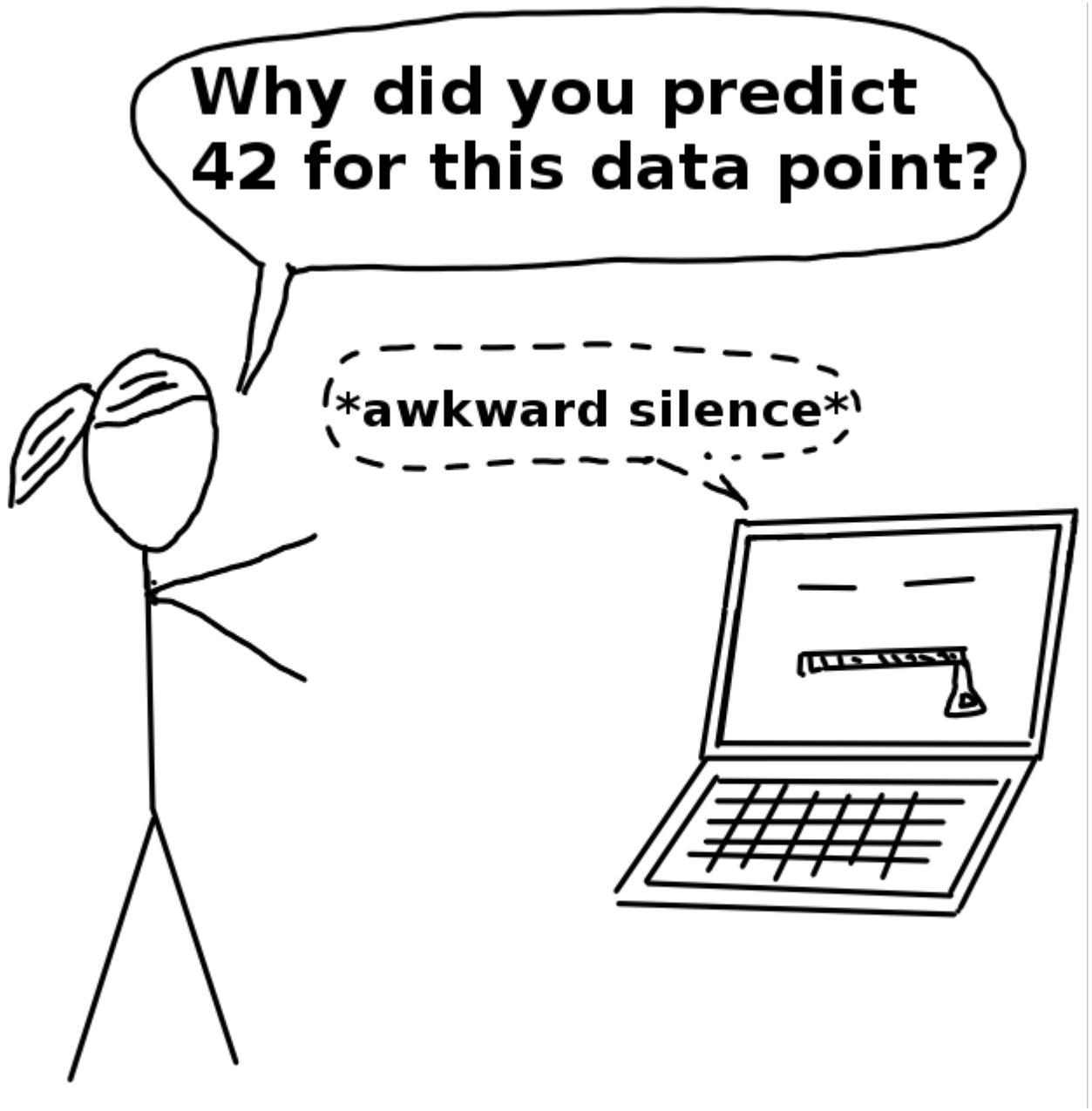
Бутаков Иван Дмитриевич
Маланчук София Владимировна
Неопрятная Анна Михайловна
Толмачев Александр Дмитриевич

Научно-технологический университет Сириус

Структура

1. Проблематика
2. Анализ подходов к решению проблемы
3. Цели и задачи
4. Эксперименты
5. Результаты
6. Выводы

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


Why did you predict 42 for this data point?
awkward silence

Научно-технологический университет Сириус

Проблематика

Нейронные сети являются «черными ящиками»: на данный момент методы исследования внутренней структуры сложных нейросетей слабо развиты.

Научно-технологический университет Сириус

Проблематика

Стоит задача нахождения способов оценки работы скрытой части нейронной сети, строгого анализа её архитектуры, поиска слабых частей и, по возможности, их улучшения.

Подход на основе теории информации

На данный момент перспективным кажется подход на основе теории информации.

В частности, Naftali Tishby совместно с Ravid Schwartz-Ziv предлагают исследовать структуру и процесс обучения нейронной сети при помощи взаимной информации.

$$I(X,Y) = H(X,Y) - H(X \mid Y) - H(Y \mid X)$$

Подход на основе теории информации

Предлагается измерять взаимную информацию между слоями, а также между слоем и входом, слоем и выходом.

Рассмотрим задачу классификации. **Гипотеза Tishby и Schwartz-Ziv** заключается в том, что нейронная сеть стремится сохранить максимум информации о метке, «отсеяв» при этом как можно больше ненужной информации о входе.

$$I(X, L_i) \to \min \qquad I(Y, L_i) \to \max$$

Научно-технологический университет Сириус

Подход на основе теории информации

Результаты экспериментов Tishby и Schwartz-Ziv были проанализированы в статье **«Estimating Information Flow in Deep Neural Networks»**

Estimating Information Flow in Deep Neural Networks

Ziv Goldfeld [1 2] **Ewout van den Berg** [2 3] **Kristjan Greenewald** [2 3] **Igor Melnyk** [2 3] **Nam Nguyen** [2 3]
Brian Kingsbury [2 3] **Yury Polyanskiy** [1 2]

Abstract

We study the estimation of the mutual information $I(X;T_\ell)$ between the input X to a deep neural network (DNN) and the output vector T_ℓ of its ℓth hidden layer (an "internal representation").

y 2019

et al., 2018; Achille & Soatto, 2018). Mutual information is an appealing measure due to its invariance to smooth, invertible transformations and the fact that it has meaningful units (bits or nats). But, these benefits come at a price: mutual information is often impossible to compute analytically, and its estimation from samples is difficult (Paninski, 2003). A

Подход на основе теории информации

Были указаны следующие принципиальные недочёты в методологии:

1) Детерминированность нейронной сети

2) Оценка энтропии дискретизацией распределения

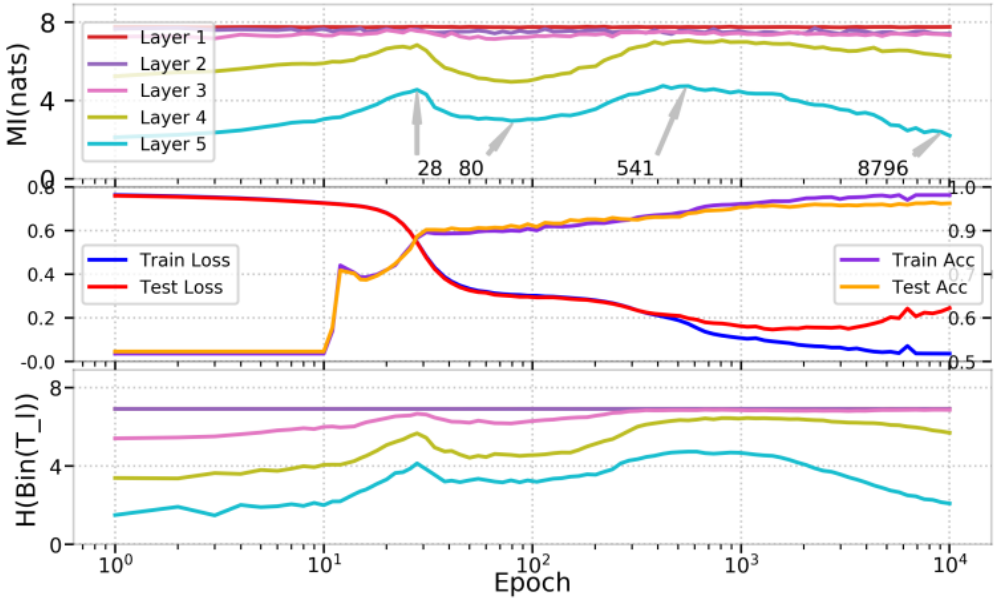
Предложенные решения:

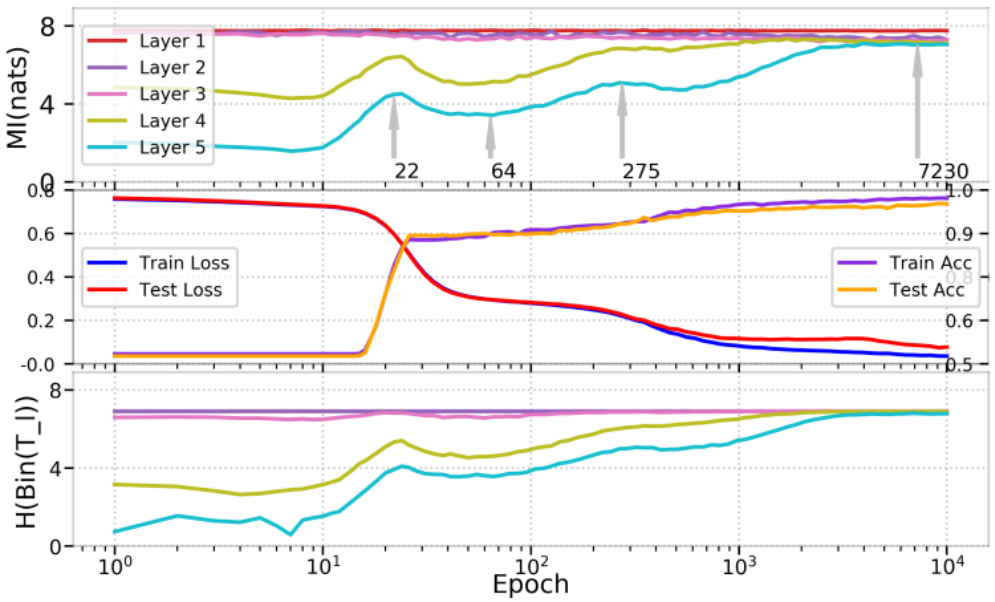
1) Добавление шума на каждом слое

2) Корректные и сходящиеся эстиматоры

Подход на основе теории информации

Тем не менее, авторы статьи утверждают, что в ходе своих экспериментов замечали явления, напоминающие **гипотезу** Tishby и Schwartz-Ziv. Однако не во всех случаях наблюдалось сжатие.



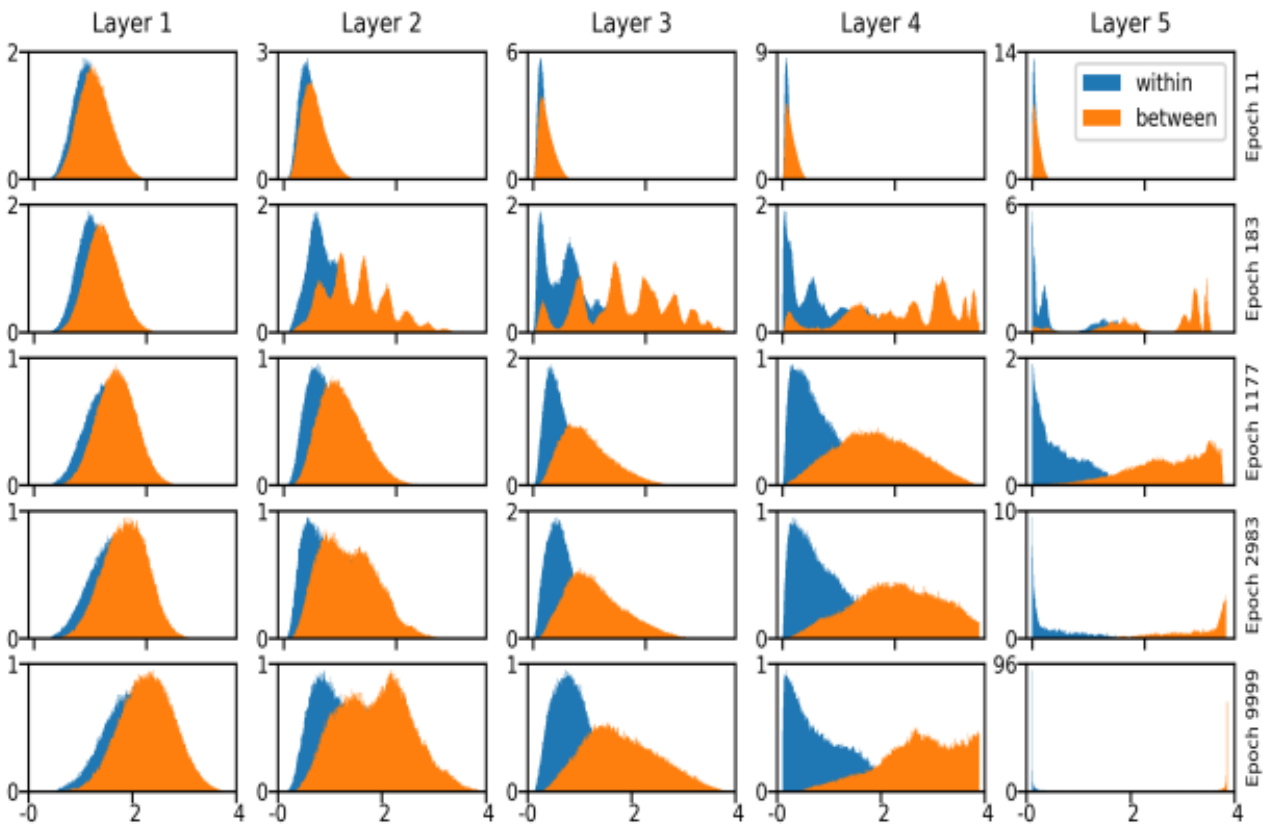


Научно-технологический университет Сириус

Подход на основе теории информации

Поэтому в статье предлагалось интерпретировать «сжатие» как кластеризацию внутреннего представления.

Новая гипотеза нашла экспериментальные подтверждения в виде гистограмм расстояний между образцами выходов слоёв.



Научно-технологический университет Сириус

Проблематика

Подсчёт энтропии напрямую по указанному способу невозможен или даёт большие отклонения для многомерных случайных величин.

Научно-технологический университет Сириус

Проблематика

Подсчёт энтропии напрямую по указанному способу невозможен или даёт большие отклонения для многомерных случайных величин.

Можно попробовать **сжать данные**

Научно-технологический университет Сириус

Цели

Разработать и протестировать метод оценки энтропии многомерных случайных величин. Попробовать связать оценку энтропии с внутренней размерностью случайной величины.

Научно-технологический университет Сириус

Задачи

1. Разработать метод оценки энтропии многомерных случайных величин, основанный на автоэнкодере.
2. Протестировать данный метод на синтетических наборах данных.
3. Использовать метод на наборе данных MNIST.
4. Оценить размерность многообразия цифр.

Научно-технологический университет Сириус

Методика проведения эксперимента

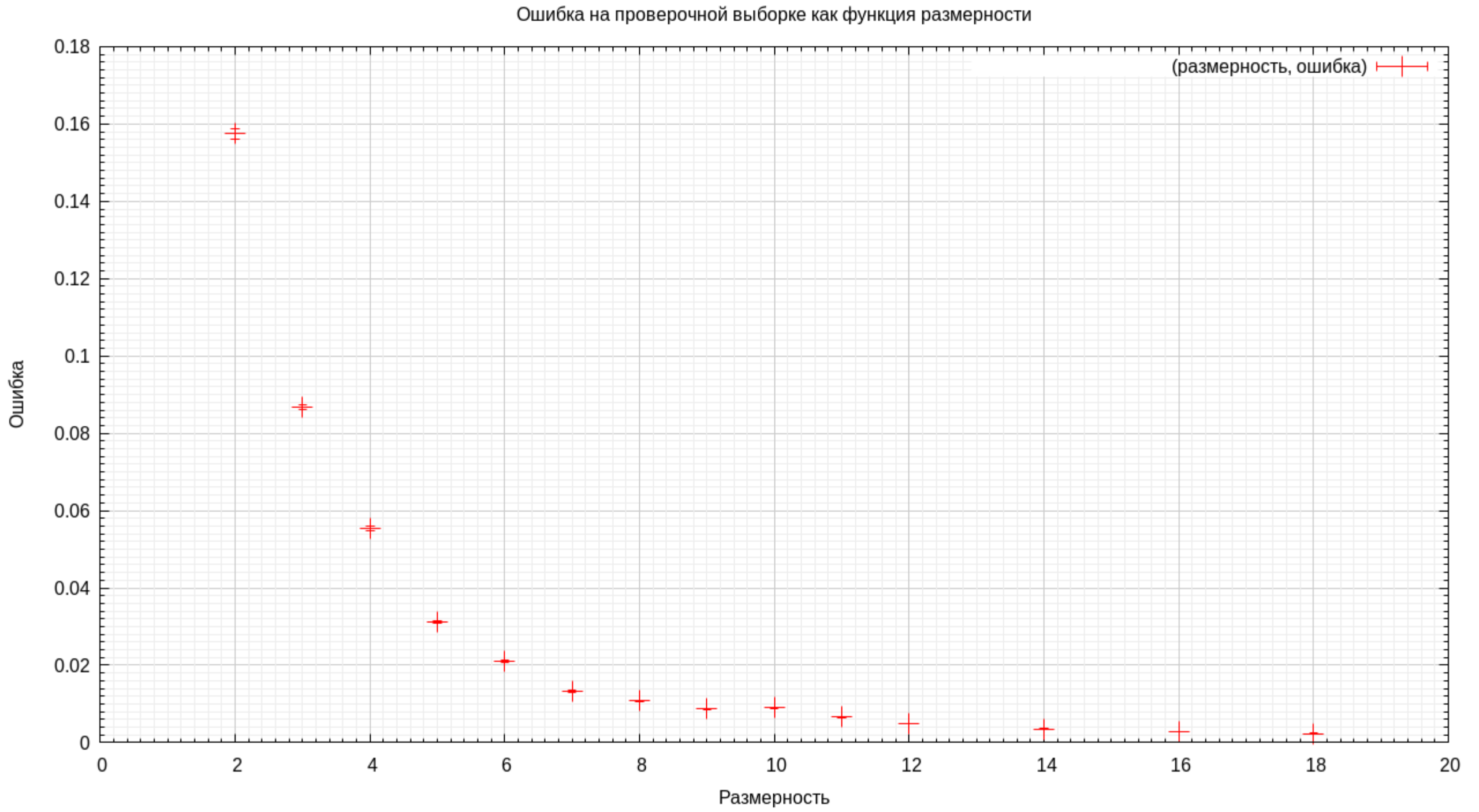
1. Построить и обучить автоэнкодер с заданной размерностью пространства кодов.
2. Получить набор кодов для всех входных данных.
3. Оценить функцию плотности вероятности при помощи KDE.
4. Оценить энтропию (метод Монте-Карло или leave one out).

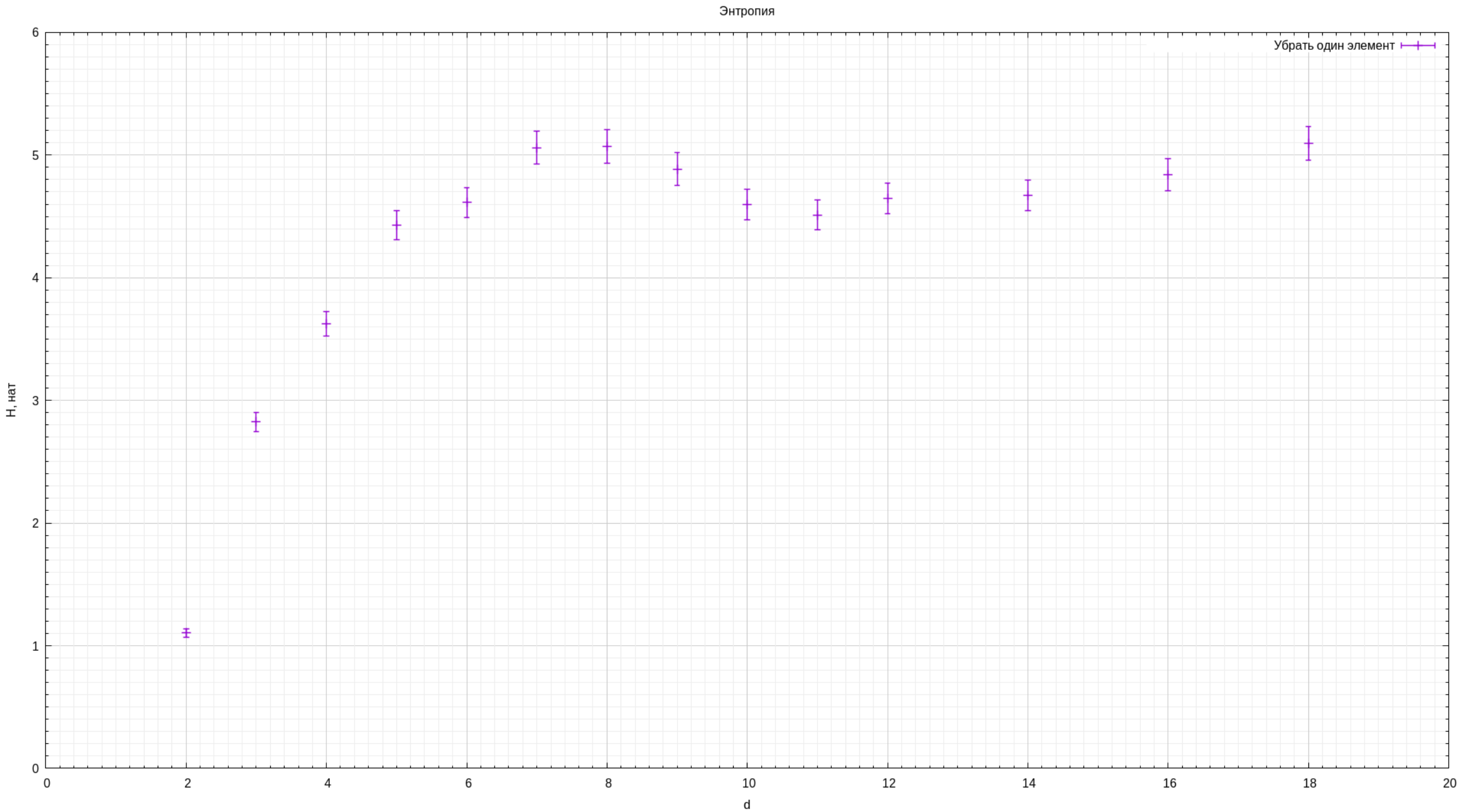
Научно-технологический университет Сириус

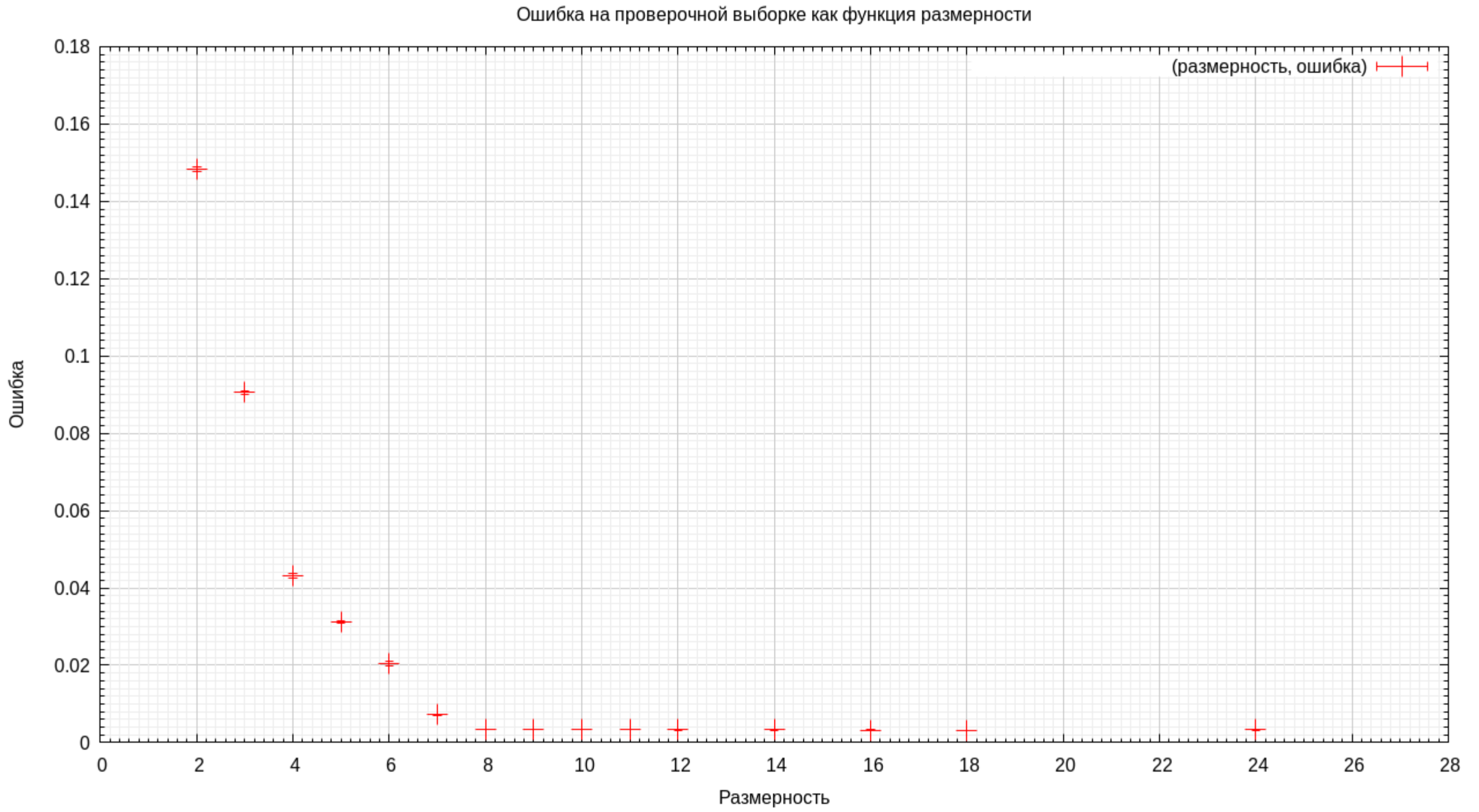
Гипотеза: график зависимости оценки энтропии от размерности кода имеет излом

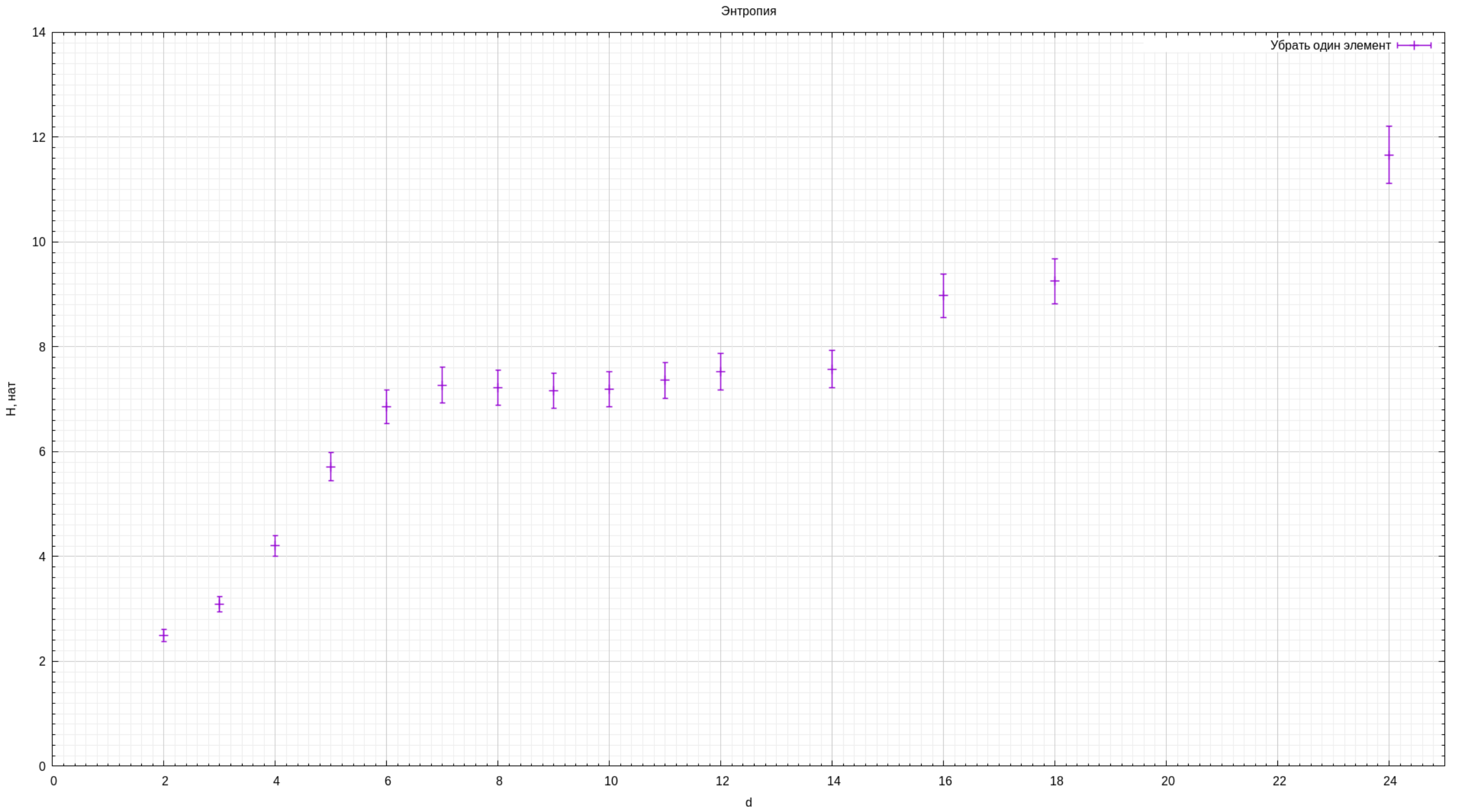
Научно-технологический университет Сириус

Эксперименты на синтетических данных


Ошибка на проверочной выборке как функция размерности
(размерность, ошибка)
Ошибка
Размерность
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
2
4
6
8
10
12
14
16
18
20

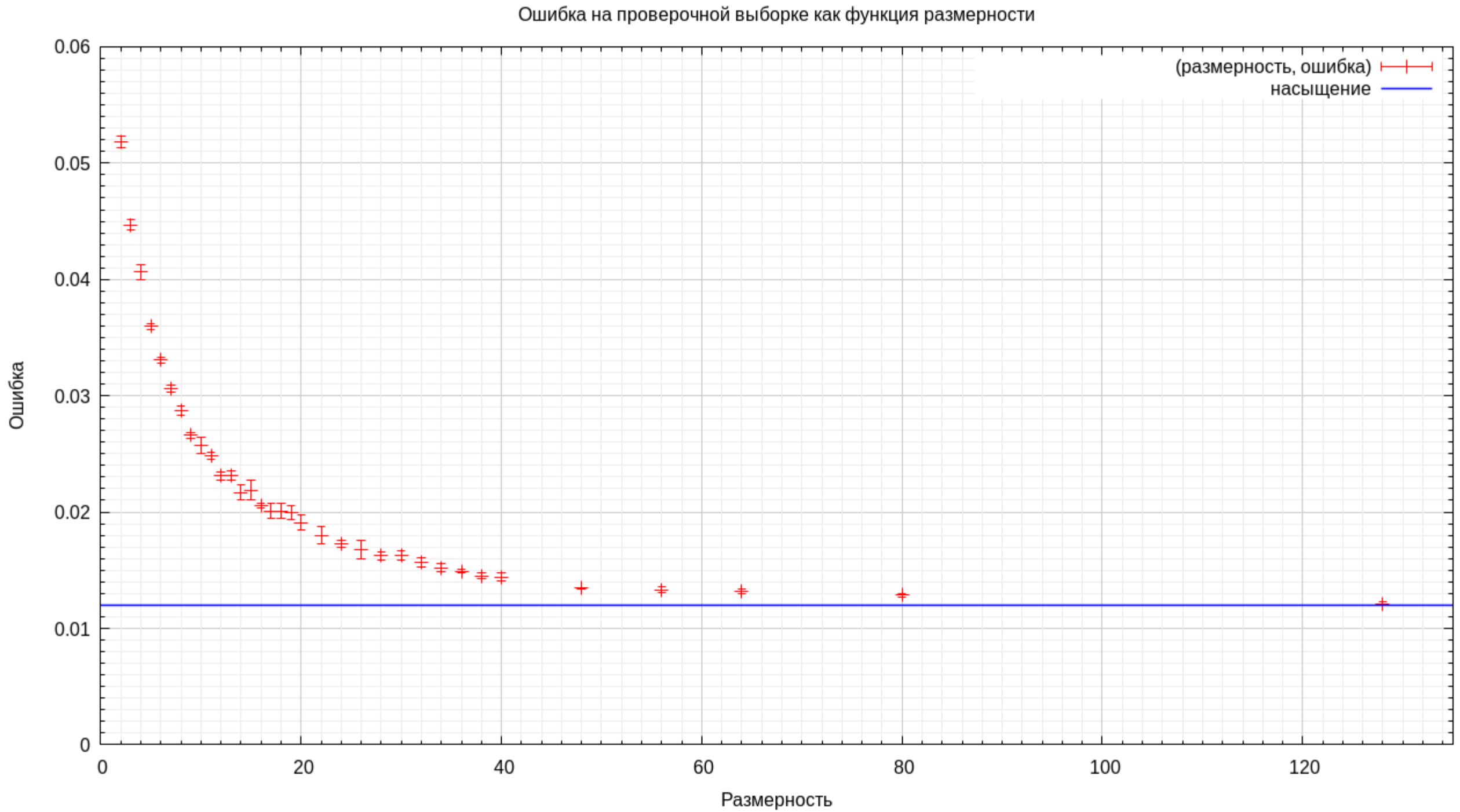

Энтропия
Убрать один элемент
H, нат
d
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18
20

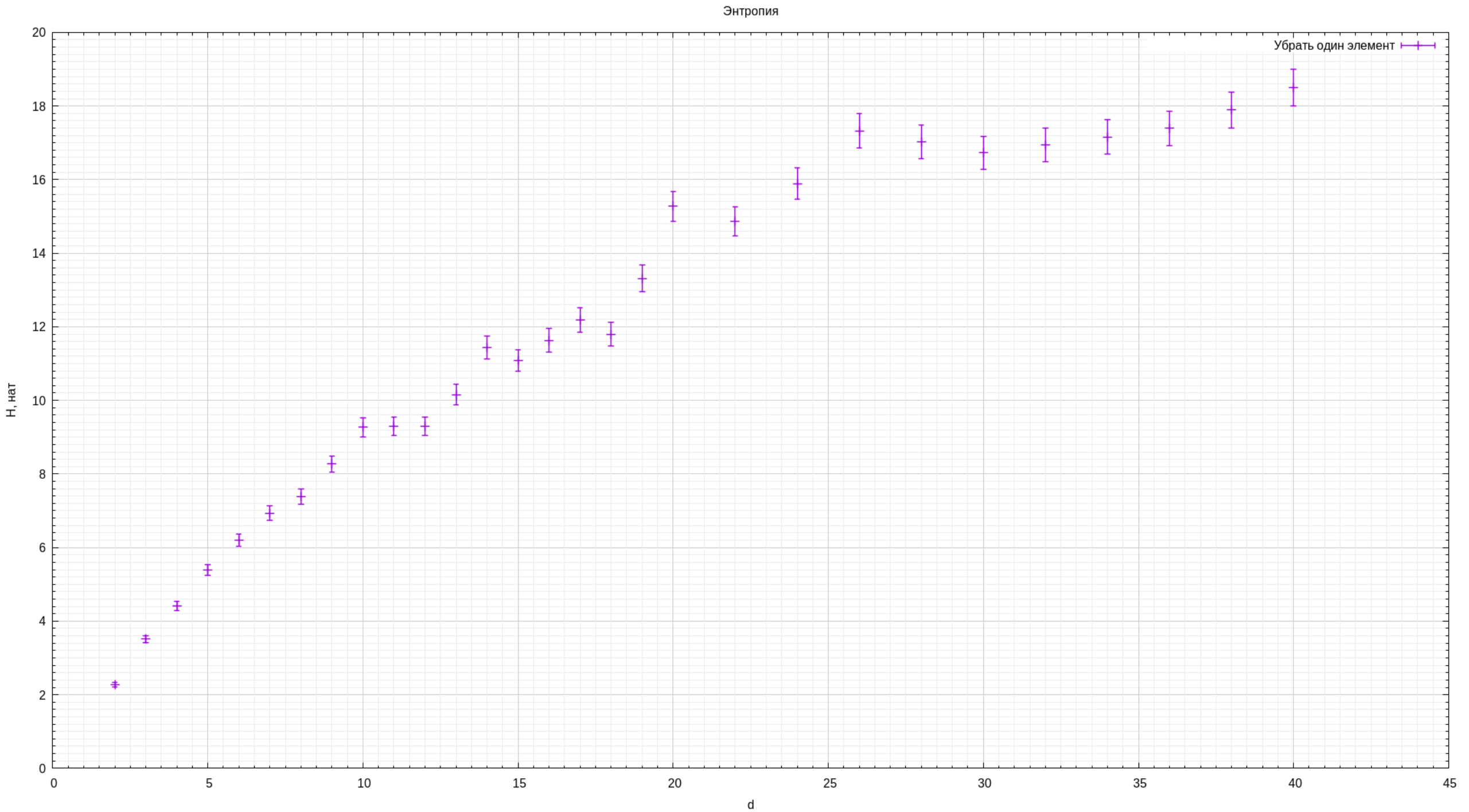

Ошибка на проверочной выборке как функция размерности
(размерность, ошибка)
Ошибка
Размерность
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28

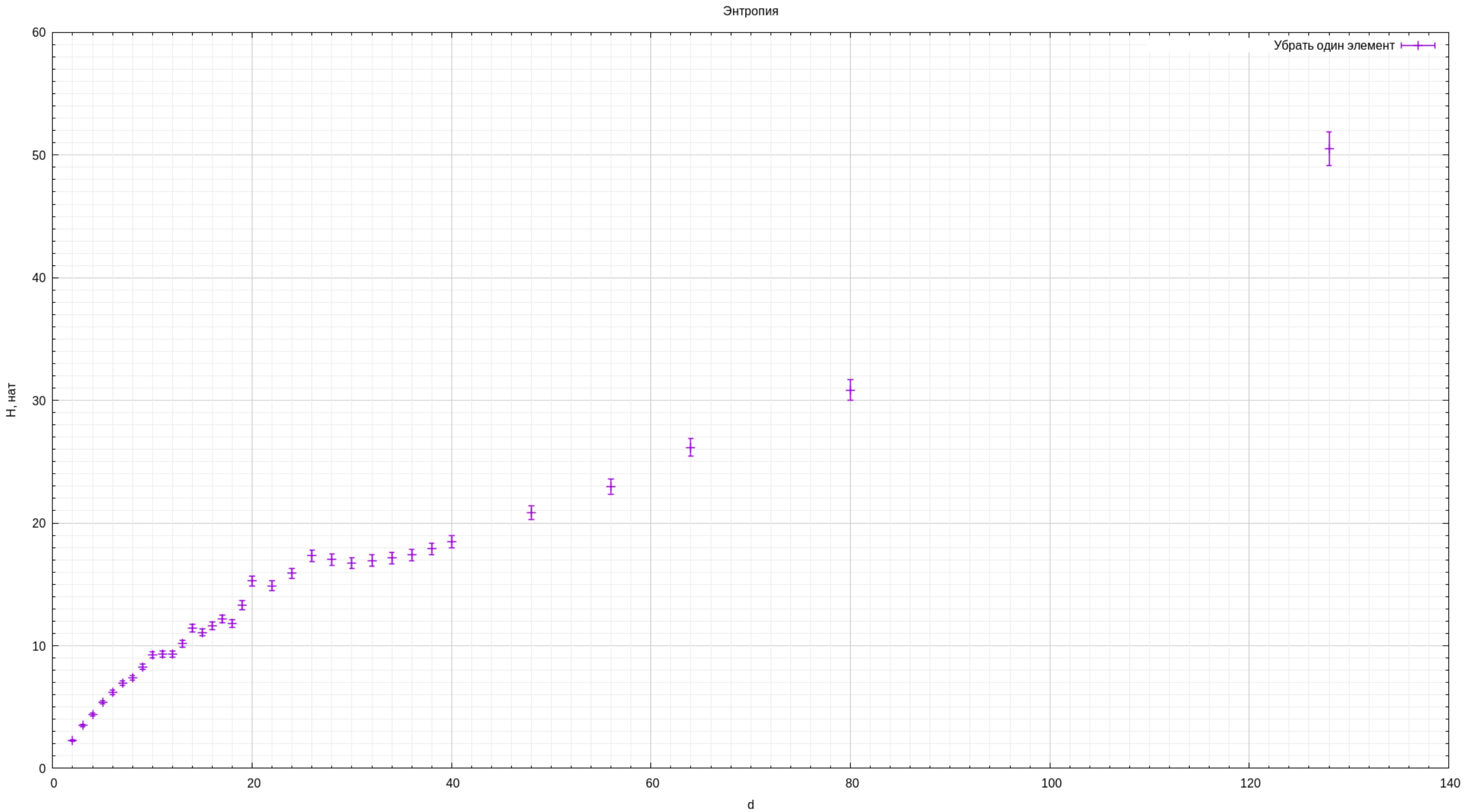

Энтропия
Убрать один элемент
H, нат
d
0
2
4
6
8
10
12
14
16
18
20
22
24

Научно-технологический университет Сириус

Эксперименты на наборе данных MNIST


Ошибка на проверочной выборке как функция размерности
(размерность, ошибка)
насыщение
Ошибка
Размерность
0.06
0.05
0.04
0.03
0.02
0.01
0
0
20
40
60
80
100
120


Энтропия
Убрать один элемент
H, нат
d
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30
35
40
45


Энтропия
Убрать один элемент
H, нат
d
0
10
20
30
40
50
60
0
20
40
60
80
100
120
140

Научно-технологический университет Сириус

Фрактальная размерность кодов

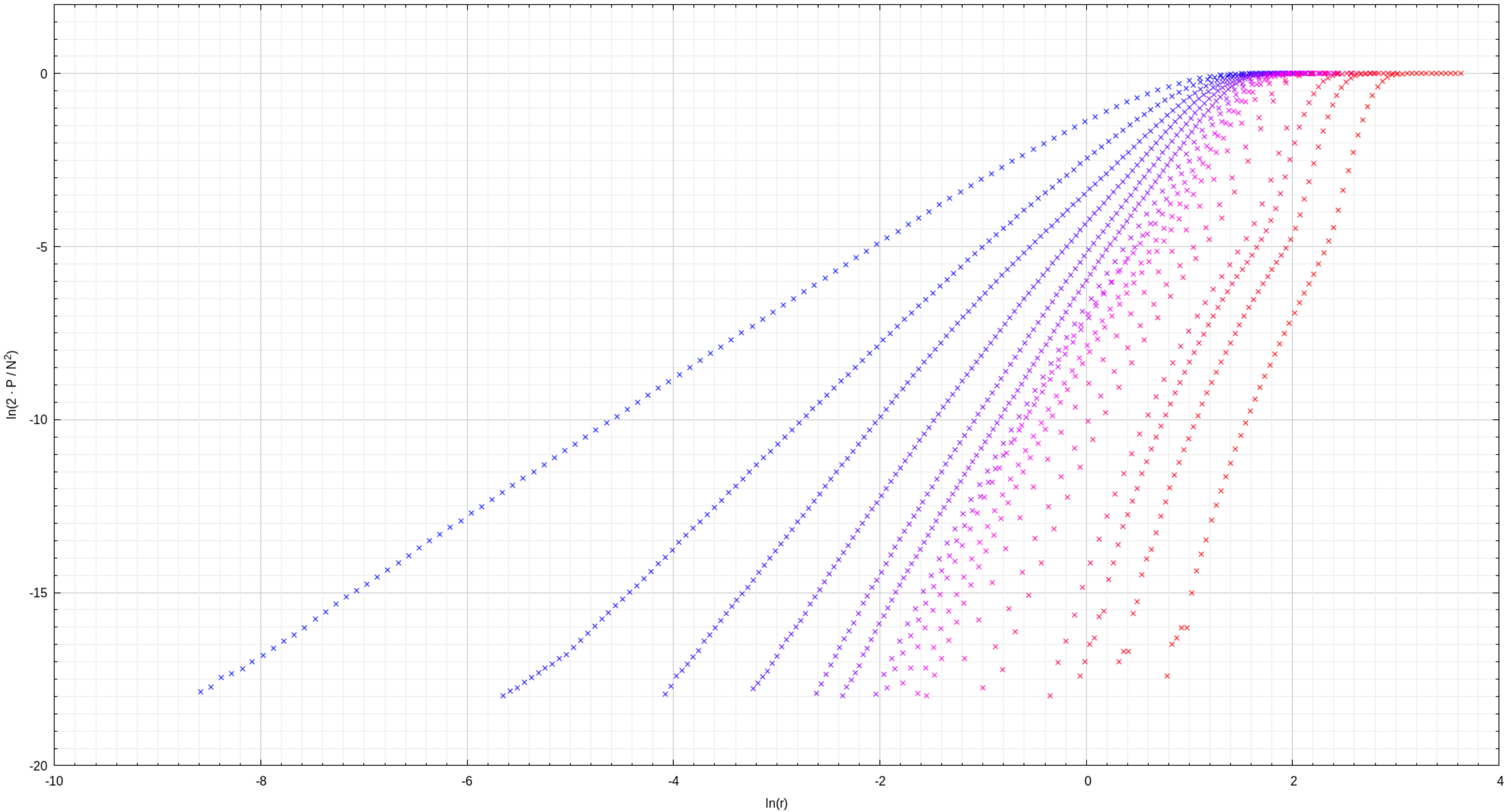
Научно-технологический университет Сириус

Фрактальная размерность

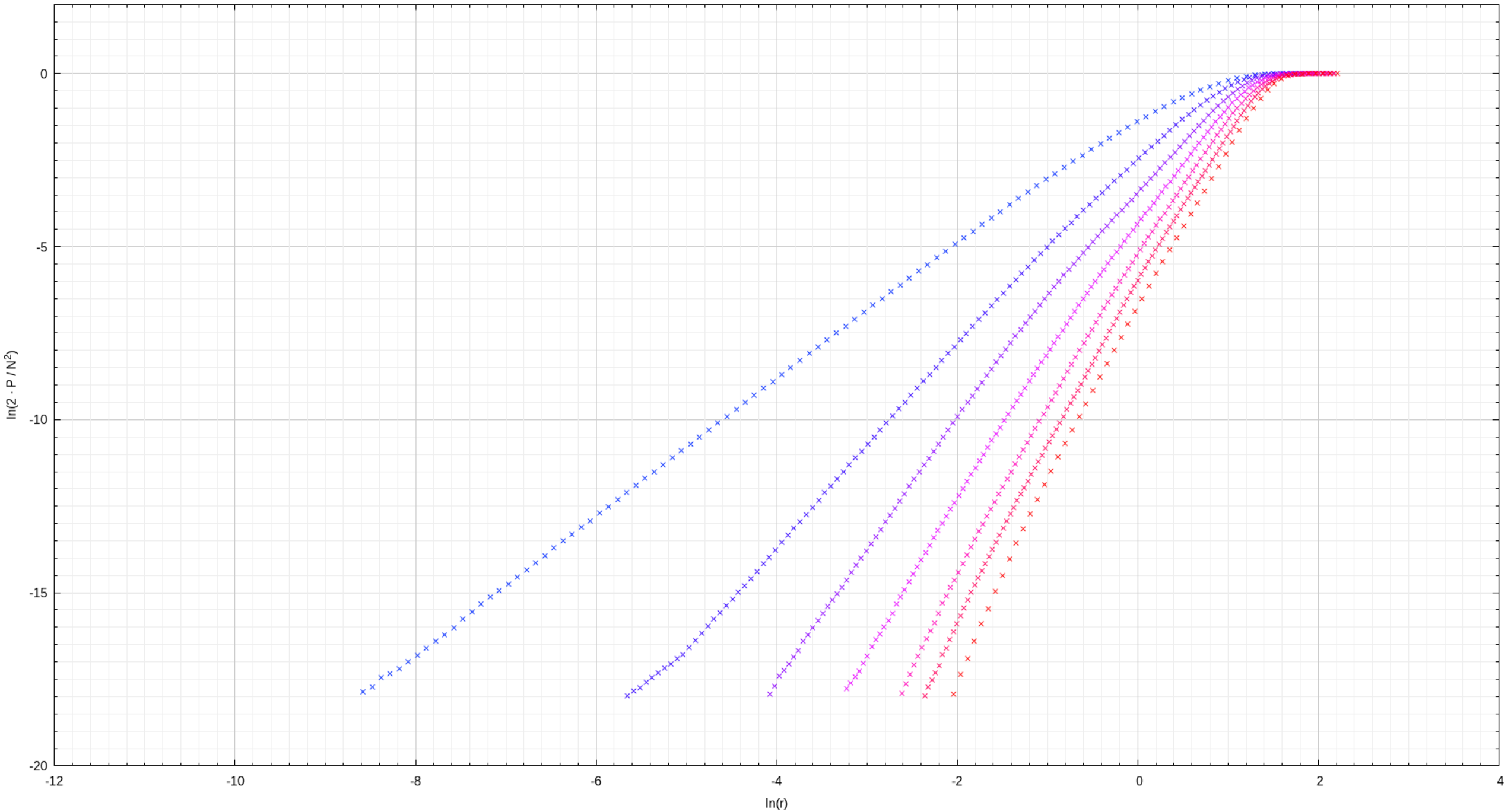
Исследуем **корреляционную размерность** для множеств кодов, а также для выборки из множества изображений.

$$d_{corr} = \lim_{r \to 0,\ N \to \infty} \frac{\ln\left(\frac{2 \cdot P(r)}{N^2}\right)}{\ln r}$$

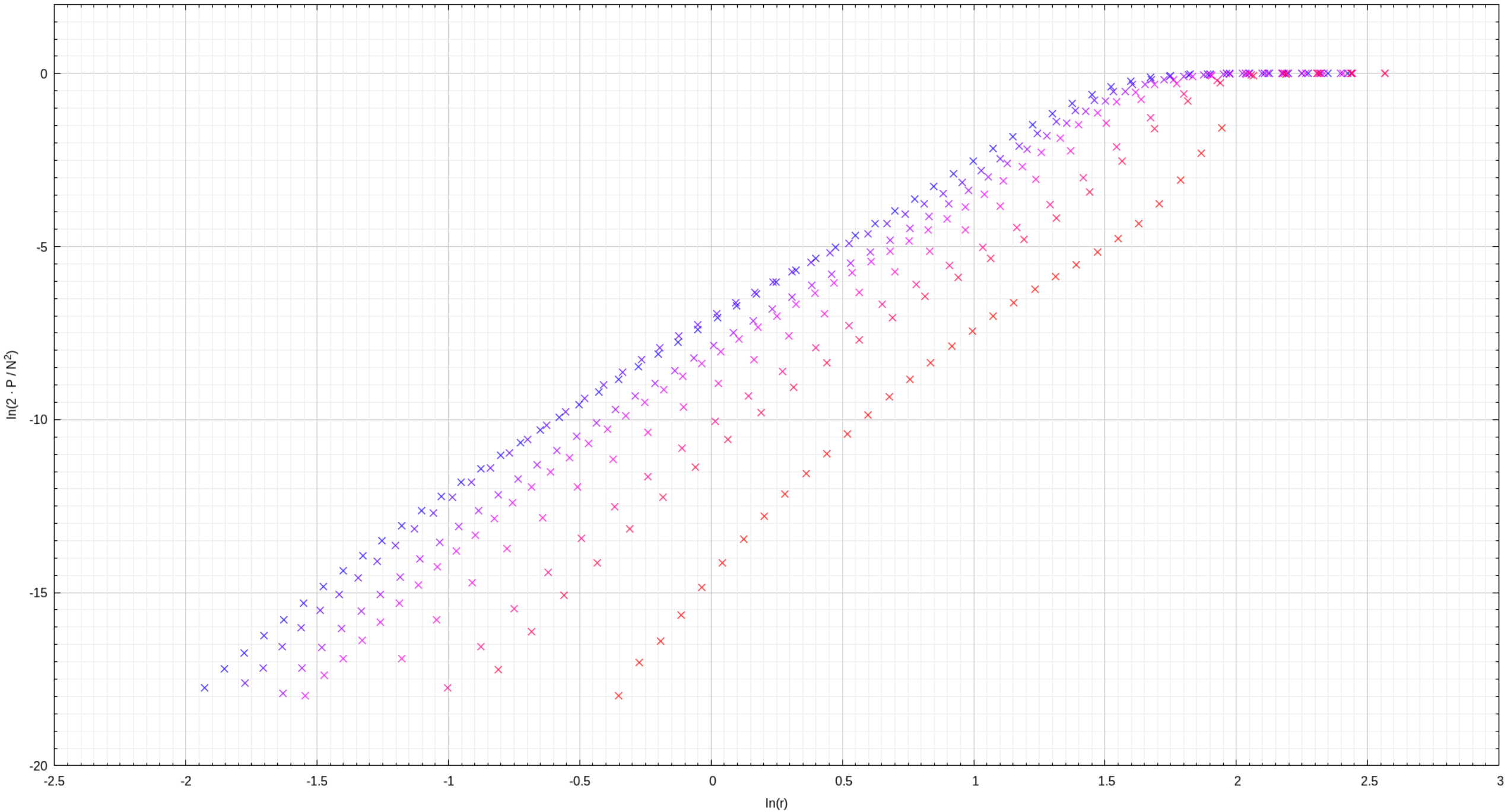
Корреляционная размерность



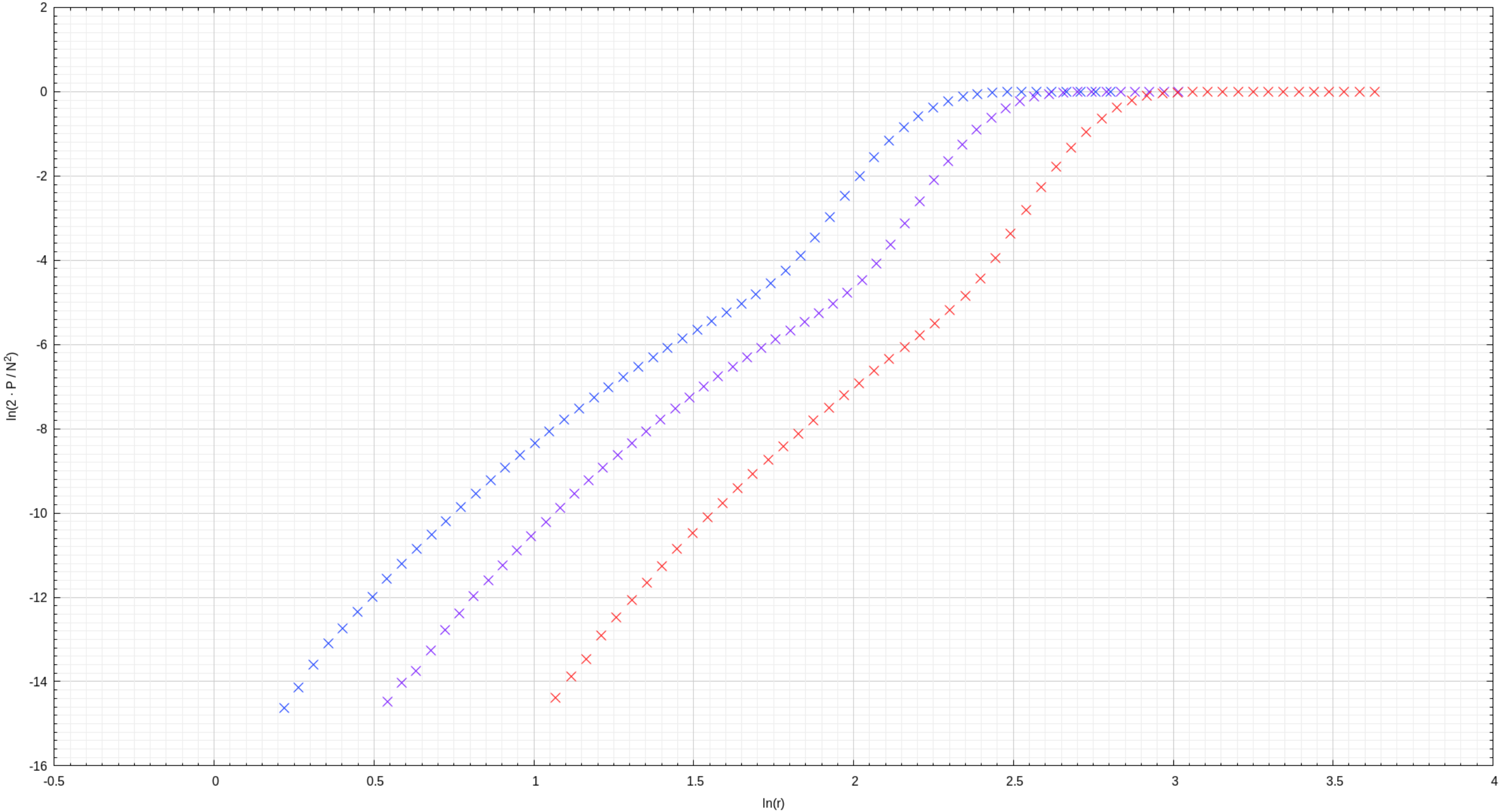
Корреляционная размерность



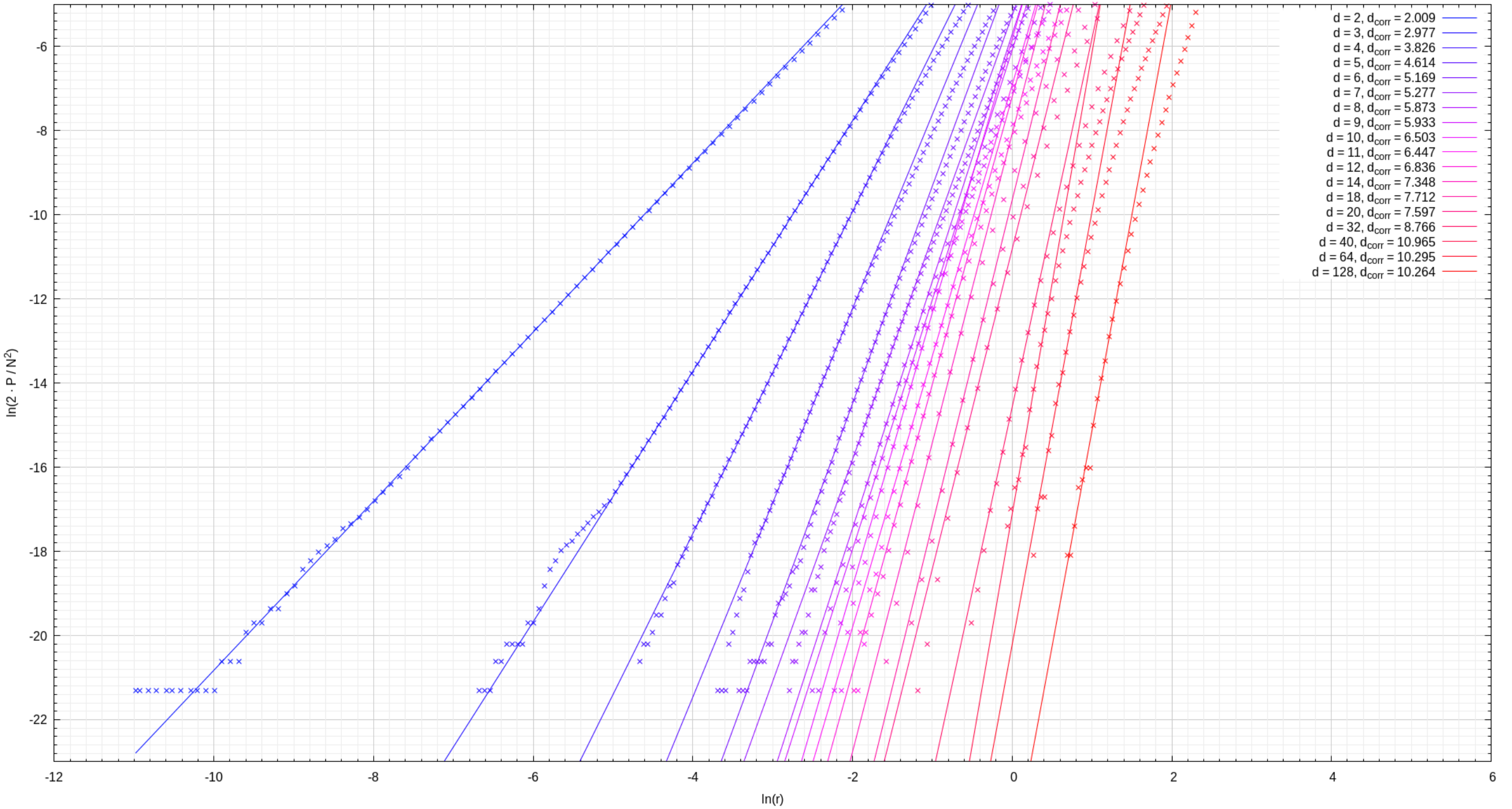
Корреляционная размерность

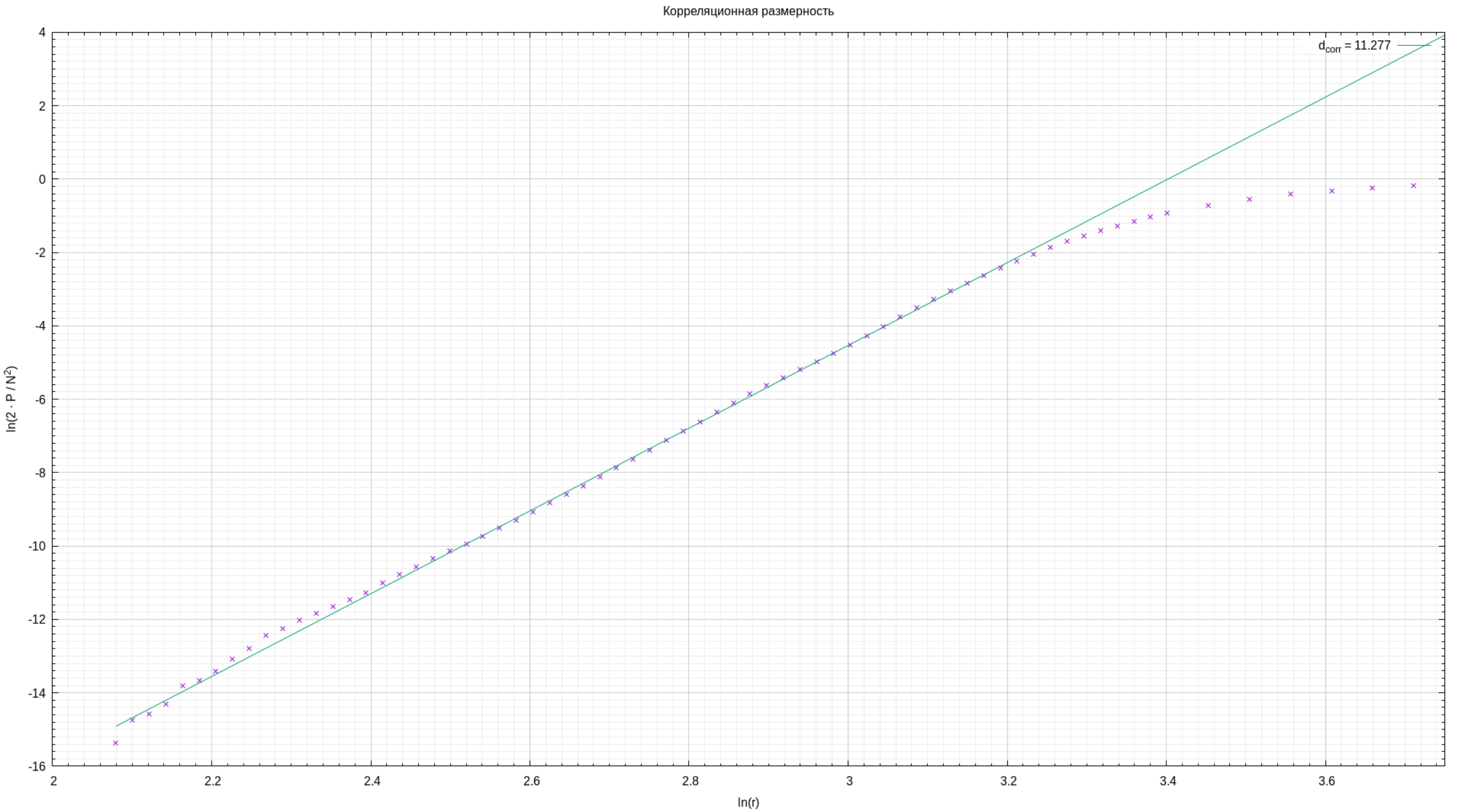


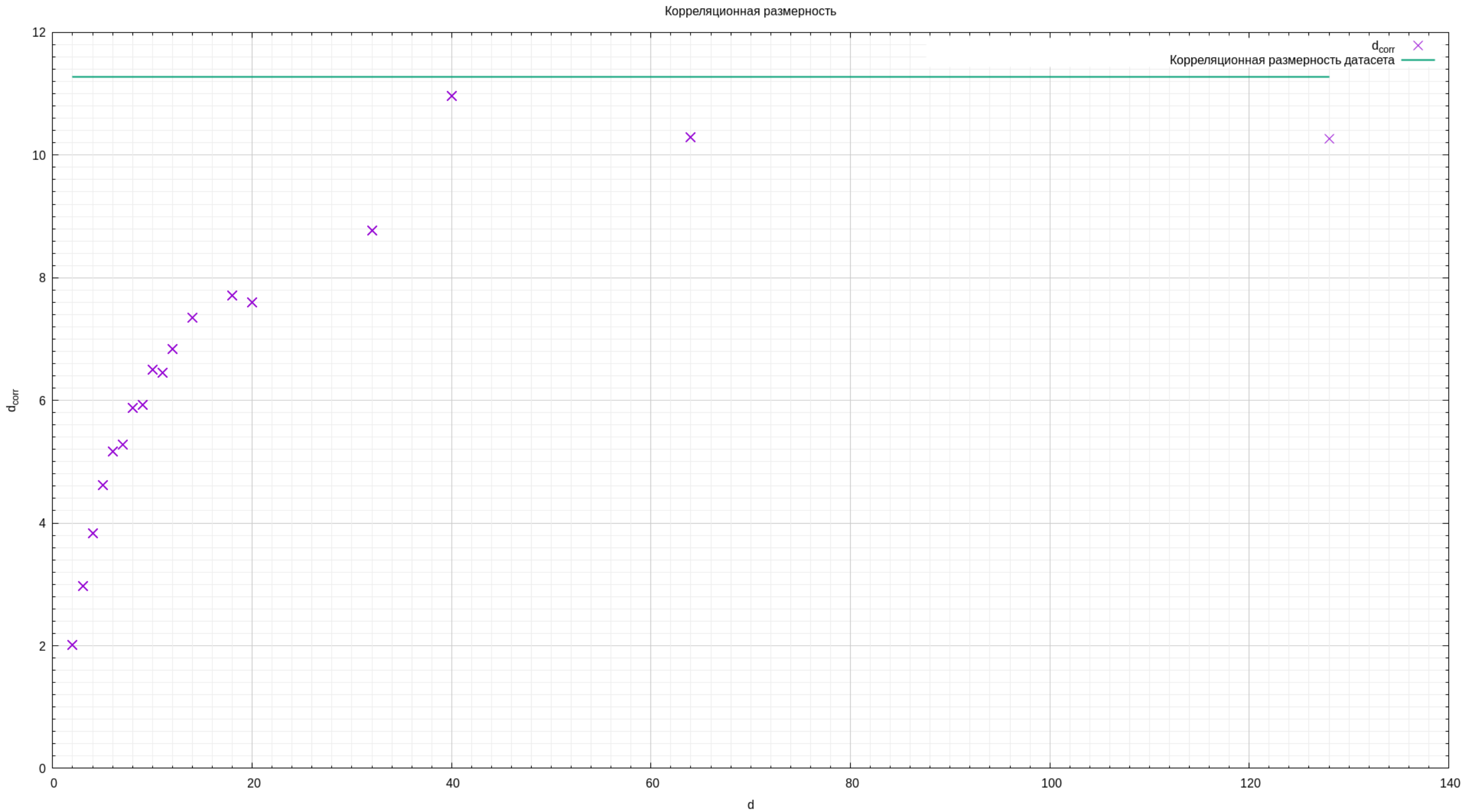
Корреляционная размерность



Корреляционная размерность




Корреляционная размерность
d_{corr} = 11.277
ln(2 · P / N²)
ln(r)
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
2
2.2
2.4
2.6
2.8
3
3.2
3.4
3.6


Корреляционная размерность
d_{corr}
Корреляционная размерность датасета
d_{corr}
d
0
2
4
6
8
10
12
0
20
40
60
80
100
120
140

Научно-технологический университет Сириус

Выводы

1. Графики *H(d)* для всех экспериментов соответствуют исходному предположению об изломе.
2. Излом для графиков экспериментов с синтетическими данными располагается в окрестности истинной размерности.
3. На всех графиках *H(d)* так или иначе наблюдается некоторый рост энтропии после плато.
4. Фрактальная структура кодов с ростом размерности, в целом, меняется непрерывно.
5. Корреляционная размерность множества кодов насыщается корреляционной размерностью множества изображений

Научно-технологический университет Сириус

Задачи

1. Разработать метод оценки энтропии многомерных случайных величин, основанный на автоэнкодере.
2. Протестировать данный метод на синтетических наборах данных.
3. Использовать метод на наборе данных MNIST.
4. Оценить размерность многообразия цифр.

Научно-технологический университет Сириус

Наша команда

Бутаков Иван Дмитриевич — написание кода для проведения экспериментов с MNIST.
Маланчук София Владимировна — проведение экспериментов с MNIST, измерение фрактальной размерности кодов.
Неопрятная Анна Михайловна — анализ альтернативных подходов к измерению размерностей, проведение экспериментов с синтетическими данными.
Толмачев Александр Дмитриевич — написание кода для проведения экспериментов с синтетическими данными.