1. Зависимость «ошибки кодирования» при обучении кодовой книги

Для вывода в аналитическом виде характера зависимости средней / максимальной нормы ошибки кодирования от размера кодовой книги, первым делом более внимательно посмотрим на формулы величин, рассчитываемых в процессе построения кодовой книги.

1) Начальный «центр» кодовой книги на первой итерации:

$$C^* = \frac{1}{M} \cdot \sum_{m=1}^{M} x_m \tag{1}$$

Данная формула представляет собой простое среднее всех векторов в обучающей выборке (своего рода «центр масс»)

2) Начальное значение squared error distortion measure (SEDM):

$$\begin{aligned} D^* &= \frac{1}{Mk} \cdot \sum_{m=1}^{M} \left\| x_m - C^* \right\|^2 = \\ &\frac{1}{k} \cdot \left(\frac{1}{M} \cdot \sum_{m=1}^{M} \sum_{n=1}^{k} [x_m^{(n)} - C^*_n]^2 \right) = \\ &\frac{1}{k} \cdot \left(\sum_{n=1}^{k} \frac{1}{M} \cdot \sum_{m=1}^{M} [x_m^{(n)} - C^*_n]^2 \right) = \\ &\frac{1}{k} \cdot \sum_{n=1}^{k} \sigma_n^2 \end{aligned} \tag{2}$$

Где σ_n^2 — дисперсия обучающей выборки по n-й координате нашего k-мерного пространства. Таким образом рассчитывается нормированный к размерности пространства след[1] ковариационной матрицы для множества x_m – нашей обучающей выборки.

3) На каждом шаге увеличения размера кодовой книги, для завершения шага должно выполниться условие:

1 Сумма элементов главной диагонали (http://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D0%B5%D0%B4_%D0%BC%D0%B0%D1%82%D1%80%D0%B8%D1%86%D1%8B)

$$\frac{(D_{i-1}-D_i)}{D_{i-1}} \le \epsilon \quad \rightarrow$$

$$(1-\epsilon) D_{i-1} \le D_i \quad \rightarrow \qquad (3)$$

$$D_i \ge (1-\epsilon)^i \cdot D^*$$

Поскольку на каждом шаге кодовая книга растет в 2 раза, то рассматривая лишь кодовые книги с размером size = 2^L (L – натуральное число), можно сказать, что номер шага p тождественен L. Другими словами, можно сказать, что p = log2(size).

Тем не менее, остро встает вопрос с количеством итераций на каждом шаге. Необходимо из каких-то соображений определить чему будет равно i в конце текущего p-го шага. В общем случае вывести аналитическую формулу для предела i пока не вышло. Однако эмпирически было замечено, что при обучении кодовой книги при заданном ε, количество итераций на каждом шаге в среднем примерно одинаковое. Таким образом, некое среднее i видится целесообразным оценивать по результатам обучения кодовой книги. Обозначим далее эту величину как i*.

ВЫВОДЫ по п.1:

Получаем, что в процессе обучения кодовой книги «ошибка кодирования», измеряемая с помощью SEDM, подчиняется правилу (4) и таким образом возможно получить оценку для нижней границы SEDM:

$$D(size) \ge (1-\epsilon)^{(log2(size) \cdot i^*)} \cdot D^* \qquad (4)$$

Где D*, в свою очередь, зависит от ковариационной матрицы обучающей выборки (см. (2)).

2. Ошибка кодирования входных данных

В качестве оценки ошибки кодирования входных данных предлагается применять норму разности y_m (элемент кодируемого множества) и соответствующего ему кода ($Q(y_m)$ в терминах предоставленного описания). Таким образом, норма m-ой ошибки выражается формулой:

$$norm_m = \sqrt{\sum_{n=1}^{k} [y_m^{(n)} - Q(y_m)]^2} \qquad (5)$$

Соответственно средняя норма будет выражаться как:

$$norm^* = \frac{1}{M} \sum_{m=1}^{M} norm_m \tag{6}$$

ВЫВОДЫ по п.2:

Для того, чтобы привести нашу «аналитическую» оценку для SEDM к размерности нормы ошибки, предлагается проделать следующие операции:

- убрать нормировку к размерности пространства в (2):

$$D^*{}' = \sum_{n=1}^{k} \sigma_n^2 \tag{7}$$

- взять квадратный корень из получившейся величины D*':

$$D^{**} = \sqrt{(D^*{}')} \tag{8}$$

- итоговое ожидание SEDM для некоторого размера size кодовой книги рассчитывать как:

$$D(size) \approx (1-\epsilon)^{(log2(size) \cdot i^*)} \cdot D^{**} \tag{9}$$

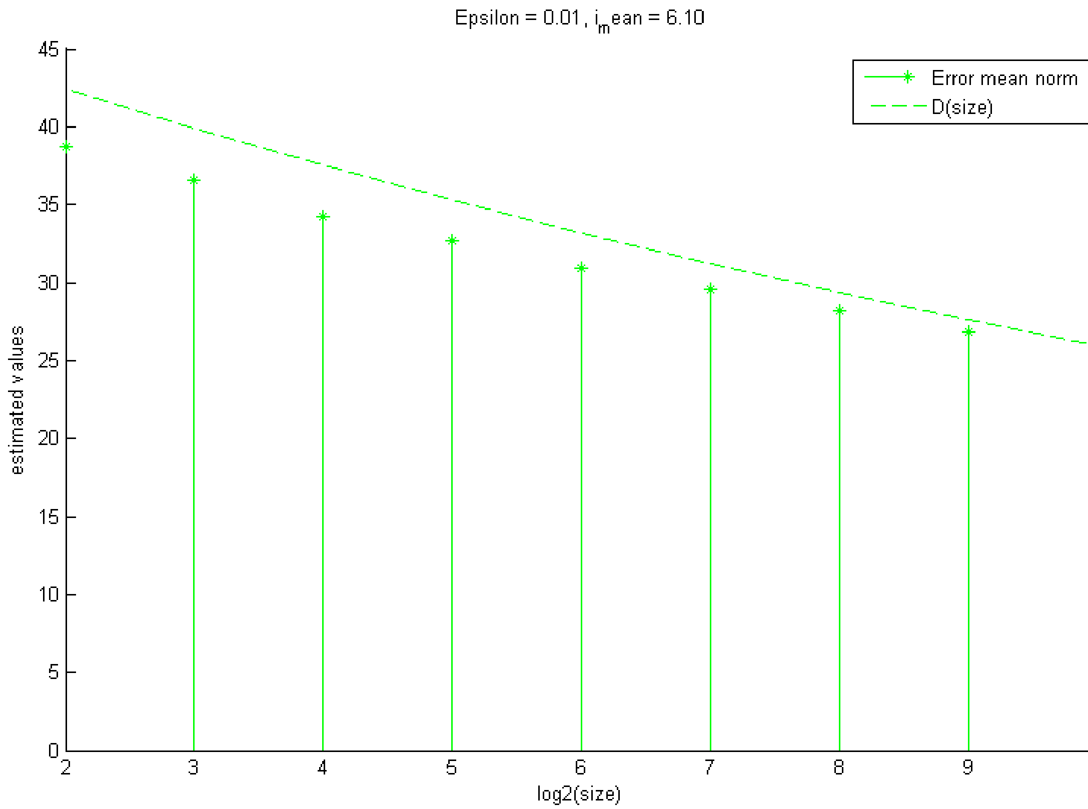
3. Итог

Проверим на практике наши предположения для значений ε = [0.01; 0.05; 0.1] и значений size = [4; 8; 16; 32; 64; 128; 256; 512; 1024].

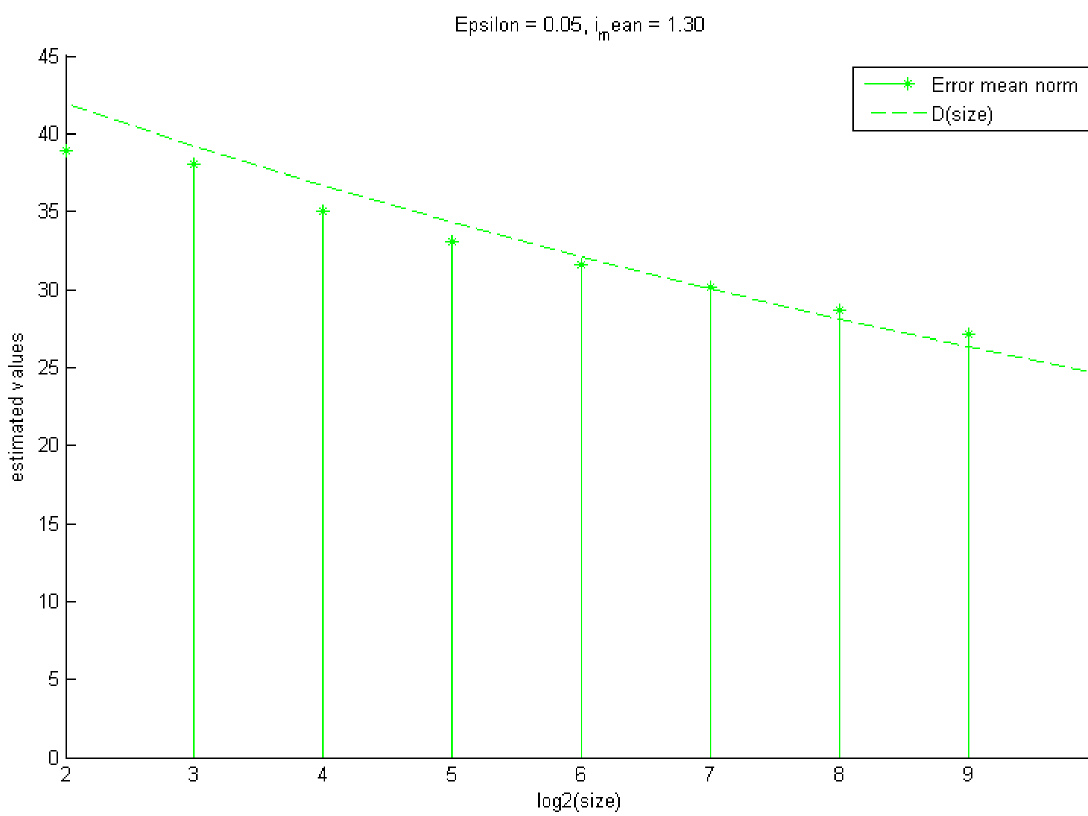
Значения i*= [6.1; 1.3; 1] для соответствующих ε были оценены в процессе обучения кодовых книг.

Графики средней нормы ошибки кодирования и соответствующих им оценок D(size) из формулы (9) представлены ниже.

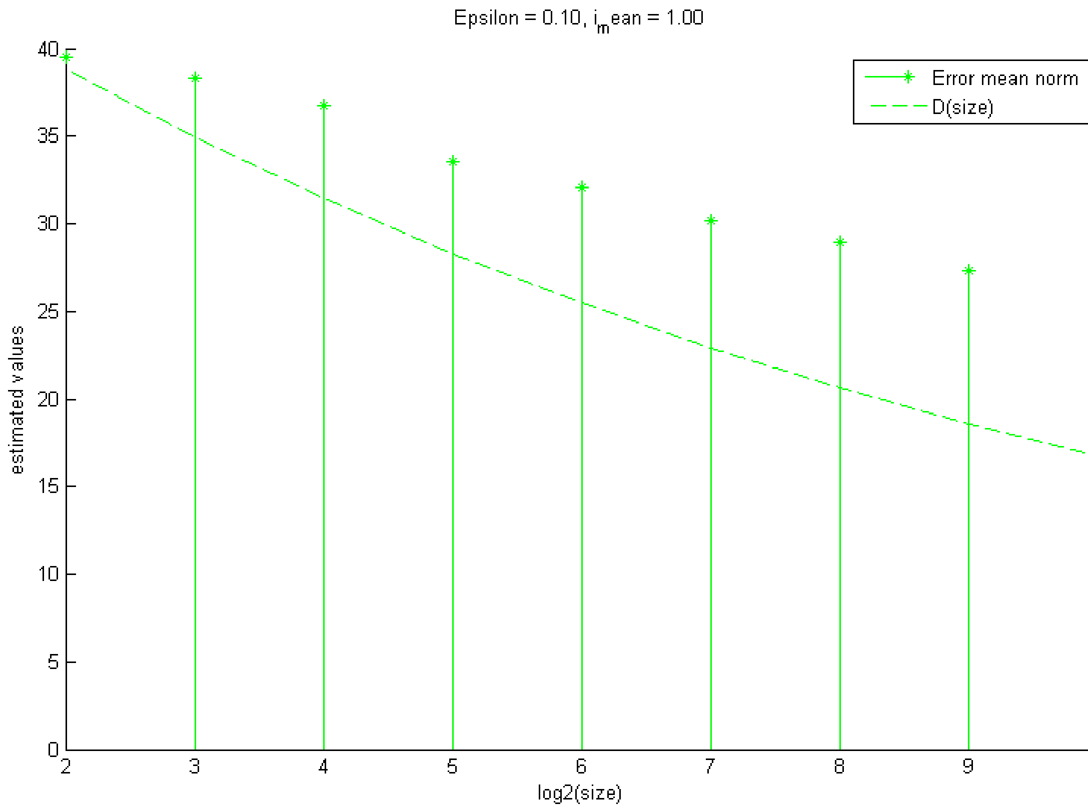
ε = 0.01 :


Epsilon = 0.01, i_mean = 6.10
Error mean norm
D(size)
estimated values
log2(size)

ε = 0.05 :


Epsilon = 0.05, i_mean = 1.30
Error mean norm
D(size)
estimated values
log2(size)

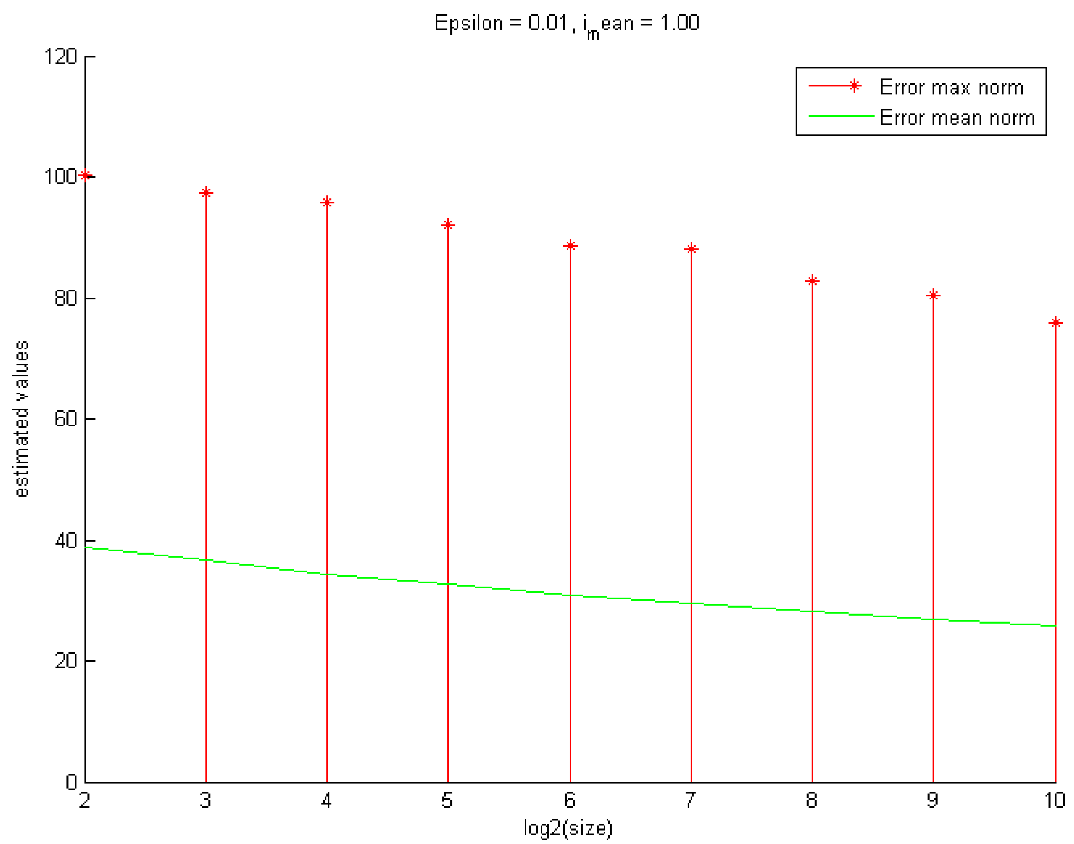
ε = 0.1 :



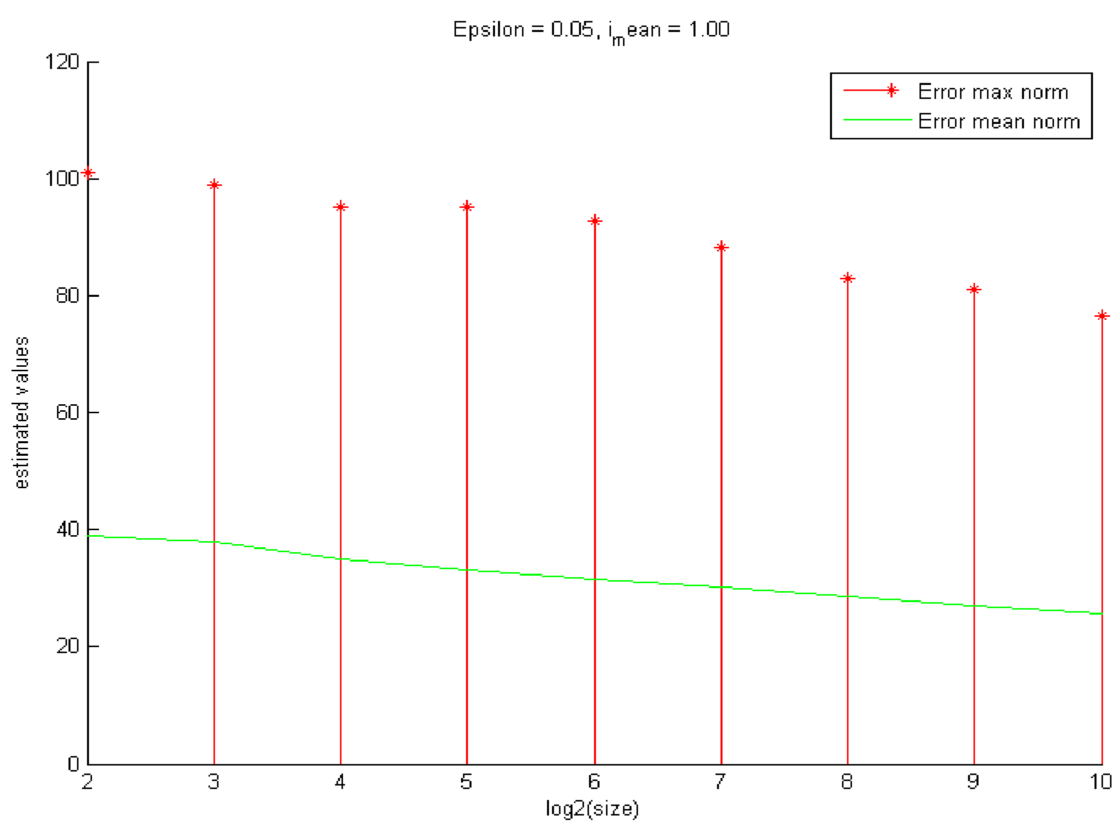
Для последнего графика видно, что значение ε было выбрано слишком большим, т. к. условие оптимизации SEDM всегда выполнялось с 1-го раза. Однако меньше одной итерации на шаге сделать не представляется возможным.

Графики максимальных норм ошибок кодирования представлены ниже. Указанное автором задания замечание, что максимальная норма ошибки как правило получается примерно в 3 раза больше, нежели соответствующая ей средняя норма, может быть условно объяснено «правилом 3-х сигм». Хотя в данном случае стоит уточнить, что сигма для каждой ошибки у нас k-мерная (как и пространство кодируемых значений) и взятие квадратного корня от суммы дисперсий по каждой координате — несколько необычная операция. Хотя возможно тут есть простое объяснение, для которого нужны более глубокие знания теории вероятности.

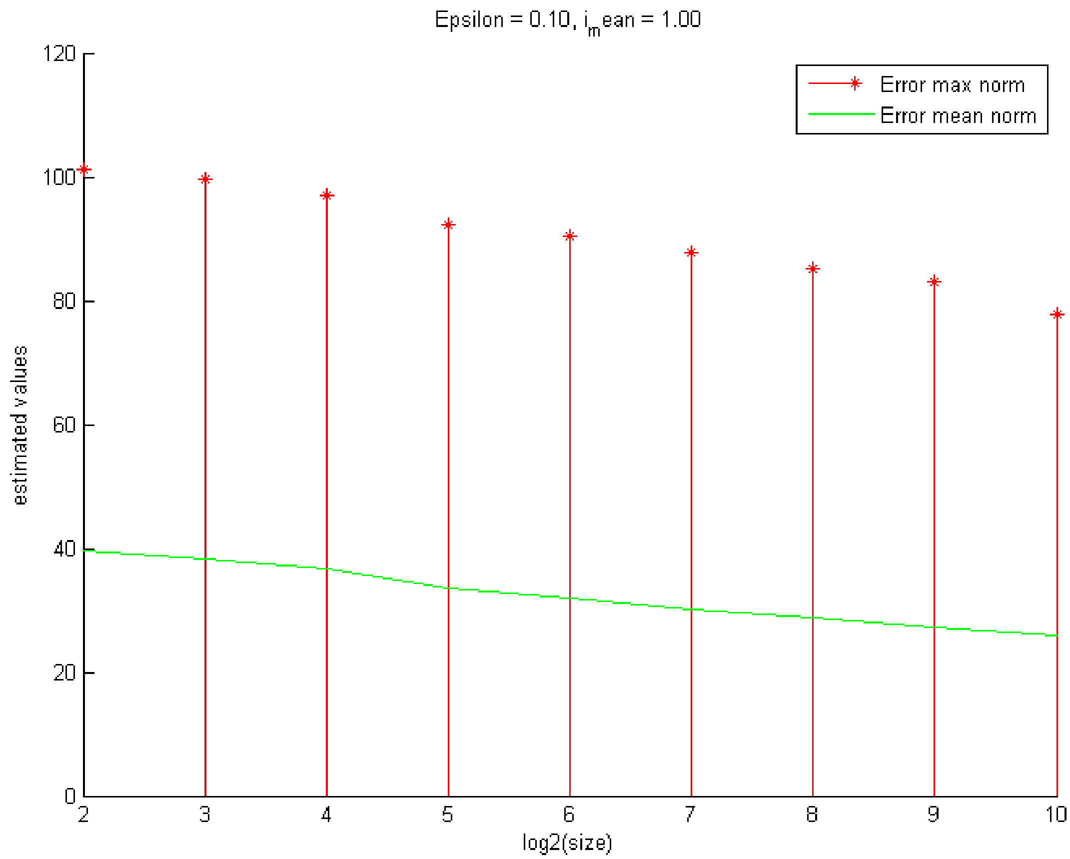
ε = 0.01 :


Epsilon = 0.01, i_mean = 1.00
Error max norm
Error mean norm
estimated values
log2(size)
0
20
40
60
80
100
120
2
3
4
5
6
7
8
9
10

ε = 0.05 :


Epsilon = 0.05, i_mean = 1.00
Error max norm
Error mean norm
estimated values
log2(size)
0
20
40
60
80
100
120
2
3
4
5
6
7
8
9
10

ε = 0.1 :


Epsilon = 0.10, i_mean = 1.00
Error max norm
Error mean norm
estimated values
log2(size)
120
100
80
60
40
20
0
2
3
4
5
6
7
8
9
10