Алимагадов Курбан, ПМ-11М

Отчётное домашнее задание №2

Задание 1

MATLAB. Для источника с 4-мя состояниями:

p(A) = 0,2; p(B) = 0,5; p(C) = 0,2; p(D) = 0,1

закодировать и декодировать по полученному кодовому числу сообщение «CBAABD», используя регистры «бесконечной» разрядности. (Для этого модифицировать task1.m, task2.m). Привести полученный скрипт и код сообщения. В качестве сообщения вместо CBAABD использовать по вариантам: вариант № 1 – CBABD.

Сообщению «CBABD» при арифметическом кодировании соответствует отрезок [0.753; 0.754], в качестве кодового выберем число 0.753. При декодировании получаем сообщение «CBABD».

Задание 2

Изучить исходные коды неадаптивного арифметического кодера arcode.c. При кодировании в данной программе используется входной двоичный файл, в каждом байте которого содержится символ, подлежащий кодированию. Таким образом, максимальная длина алфавита кодера равна 256 символам.

Задание 3

Изменить arcode.c, реализовав адаптивную модель для арифметического кодера, которая обновляет гистограмму частот после кодирования/декодирования каждого обработанного символа сообщения. В отчет включить модифицированный программный код соответствующей процедуры update_model(int symbol).

Код процедуры update_model(int symbol):

```
inline void update_model(int symbol)
{
	if (cum_freq[NO_OF_SYMBOLS] == MAX_FREQUENCY)
	{
		int cum = 0, freq = 0;
		for (int i = 0; i < NO_OF_SYMBOLS; i++)
		{
			freq = (cum_freq[i + 1] - cum_freq[i] + 1) >> 1;
			cum_freq[i] = cum;
			cum += freq;
		}
		cum_freq[NO_OF_SYMBOLS] = cum;
	}
	for (int i = symbol + 1; i <= NO_OF_SYMBOLS; i++)
		cum_freq[i]++;
}
```

Задание 4

Сравнить эффективность сжатия данных неадаптивного (с фиксированной гистограммой, найденной в результате предварительного статистического анализа обрабатываемых данных) и адаптивного кодера (полученного в п.3) на различных типах данных, полученных как естественным (текстовые, исполняемые файлы, исходные коды, зашумленные и незашумленные проквантованные изображения из прошлого ДЗ), так и искусственным путем. В последнем случае необходимо сгенерировать входные данные с различной статистикой символов, а также входные данные с резко изменяющейся на протяжении файла (нестационарной) статистикой: Например, создав наборы данных $x_1, x_2, \ldots, x_N$ и $y_1, y_2, \ldots, y_N$ с существенно различающимися статистиками, изучить характеристики адаптивного кодера при обработке последовательностей $x_1, x_2, \ldots, x_N, y_1, y_2, \ldots, y_N$ и $x_1, y_1, x_2, y_2, \ldots, x_N, y_N$. Исследование должно быть выполнено в общей сложности не менее чем на 5 различных файлах размера ~100 кБ.

Для сравнения эффективности сжатия рассмотрим PDF-файл с домашним заданием №2, изображения, полученные в ДЗ №1, и файлы с числовыми последовательностями, сгенерированными с помощью скрипта task4.m.

При сжатии PDF-файла (test_PDF.pdf, исходный размер 535335 байта) адаптивным кодером его размер становится равен 523752 байта, при сжатии этого же файла неадаптивным кодером его размер становится равен 569266 байт. Отсюда можно заключить, что в данном случае неадаптивное кодирование приводит к б**о**льшим битовым затратам, а значит применять его для сжатия такого файла не следует.

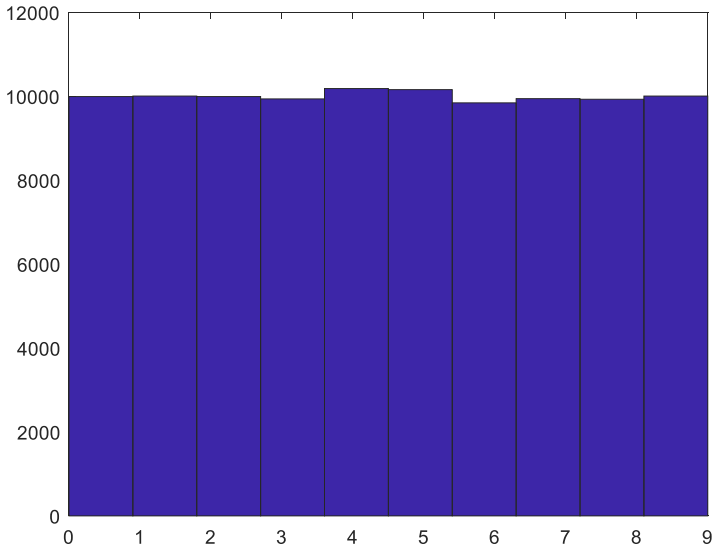
При сжатии проквантованного (без добавления шума) изображения (quantized_without_noise.bmp, исходный размер 263222 байта) адаптивным кодером его размер становится равен 125685 байт, при сжатии этого же файла неадаптивным кодером его размер становится равен 142149 байт.

При сжатии проквантованного (с добавлением шума) изображения (quantized_with_noise.bmp, исходный размер 263222 байта) адаптивным кодером его размер становится равен 126277 байт, при сжатии этого же файла неадаптивным кодером его размер становится равен 141411 байт.

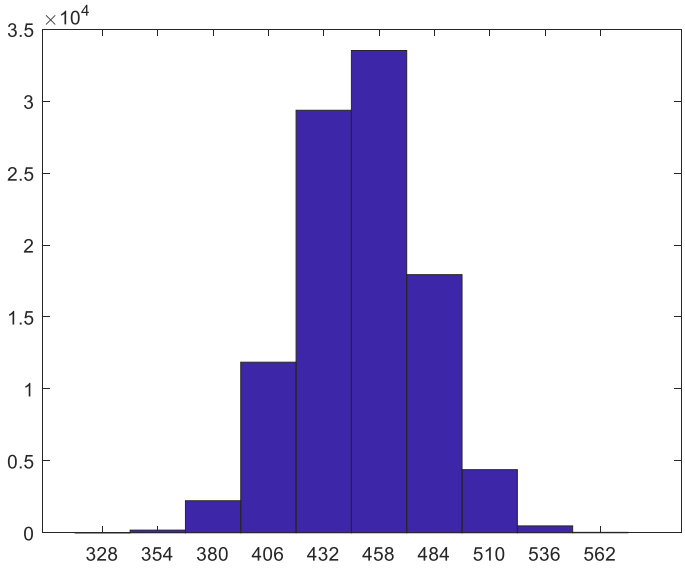
Выбор адаптивного/неадаптивного способа кодирования для данной реализации арифметического кодера несущественно влияет на эффективность сжатия рассмотренных изображений.

При помощи скрипта task4.m были сгенерированы три одномерных массива случайных целых чисел, сохранённые в файлах row1.bin (массив x_1, x_2, …, x_M; M = 100000), row2.bin (массив x_1, x_2, …, x_N, y_1, y_2, …, y_N; N = 50000) и row3.bin (массив x_1, y_1, x_2, y_2, …, x_N, y_N; N = 50000), случайные величины, соответствующие выборкам $\{x_i\}$ и $\{y_i\}$, имеют равномерный и нормальный законы распределения вероятностей соответственно.

Гистограмма $\{x_i\}$ (i = 1, …, N; N = 100000).



Гистограмма $\{y_i\}$ (i = 1, …, M; M = 100000).



При сжатии файла row1.bin (исходный размер 100000 байт) адаптивным кодером его размер становится равен 41918 байт, при сжатии этого же файла неадаптивным кодером его размер становится равен 41763 байт.

При сжатии файла row2.bin (исходный размер 100000 байт) адаптивным кодером его размер становится равен 27260 байт, при сжатии этого же файла неадаптивным кодером его размер становится равен 41465 байт.

При сжатии файла row3.bin (исходный размер 100000 байт) адаптивным кодером его размер становится равен 33654 байта, при сжатии этого же файла неадаптивным кодером его размер становится равен 33500 байт.

Кодирование последовательности случайных чисел, подчиняющихся одному закону распределения (row1.bin), с помощью адаптивного и неадаптивного кодера даёт приблизительно одинаковые результаты эффективности сжатия, так как гистограмма частот в обоих случаях имеет приблизительно одинаковое распределение по своим частотам. Кодирование последовательности случайных чисел с изменяющимся законом распределения (row2.bin) с использованием адаптивного кодера демонстрирует большую эффективность сжатия, по сравнению с неадаптивным кодированием (отношение размеров полученных сжатых файлов, большего к меньшему: > 1.5), так как в процессе адаптивного кодирования гистограмма частот настраивается под изменяющуюся выборку, в отличии от гистограммы при неадаптивном кодировании. Эффективность кодирования перемешанных последовательностей случайных чисел с разными законами распределения (row3.bin) несущественно зависит от выбора адаптивного/неадаптивного кодера, так как в данном случае гистограмме частот сложно настроиться под конкретное распределение вероятностей.

Выводы:

В работе был рассмотрен алгоритм арифметического кодирования в двух вариантах: и использованием адаптивного кодера и без него. Установлено, что применение адаптивного арифметического кодирования целесообразно лишь в тех случаях, когда гистограмма частот может настроиться под изменяющийся закон распределения вероятностей кодируемой последовательности. В случаях, когда гистограмма не изменяется в процессе кодирования (например, когда кодируемая последовательность подчиняется одному закону распределения вероятностей), эффективность сжатия адаптивного кодера приближается к эффективности неадаптивного.

Код программ представлен в файлах adaptive.cpp и notadaptive.cpp.