Sparse Regression Codes

Сенин Игорь

Московский Физико-Технический Институт

senin.ia@phystech.edu

7 мая 2024 г.

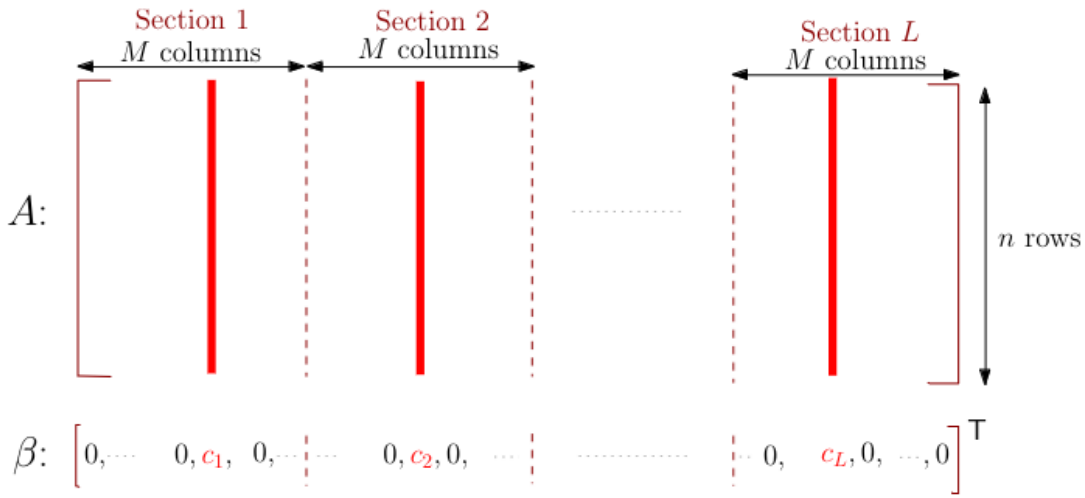
Теорема Шеннона-Хартли для AWGN канала

$\mathcal{C}$ - точная верхняя грань всех достижимых R.

Цель - придумать код с быстрыми алгоритмами кодирования и декодирования, который (асимптотически) достигал бы пропускной способности гауссовского канала $\frac{1}{2}\log(1+snr)$.

SPARCs

В SPARC'ах кодовые слова - вектора длины n вида $A\beta$, где A - матрица плана n x ML, β - разреженный вектор длины ML.
A и β разбиты на L секций длины M, причём в каждой из L секций β ровно одна ненулевая координата c_l.



Параметры M, L, элементы A, c_1, ..., c_L фиксируются и считаются известными.

SPARCs

- Элементы матрицы A выбираются i.i.d. $\mathcal{N}(0, \frac{1}{n})$.
- L выбирается $\Theta(n/\log n)$.
- $M = L^a$ для некоторой константы $a > 0$.

Благодаря такому выбору M и L размер матрицы A растёт полиномиально с n.

Общее число кодовых слов - M^L.
$M^L = 2^{Rn} \Leftrightarrow L \log M = Rn$

Outline

- Сжатие с потерями
- Применение SPARC к Lossy compression
- Конечный алфавит источника
- Successive и AMP энкодеры

Сжатие с потерями

Определения

Искажение (distortion) - функция $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$.
Среднеквадратичное искажение: $d(x, \hat{x}) = (x - \hat{x})^2$.
$(2^{nR}, n)$-**код**:

- функция-энкодер $f_n : \mathcal{X}^n \to \{1, 2, ..., 2^{nR}\}$
- функция-декодер $g_n : \{1, 2, ..., 2^{nR}\} \to \hat{\mathcal{X}}^n$

Искажение для $(2^{nR}, n)$-кода: $D(f_n, g_n) = \mathbb{E}d(X^n, g_n(f_n(X^n)))$.
Пара (R, D) называется **достижимой**, если существует последовательность $(2^{nR}, n)$-кодов (f_n, g_n) такая, что $\lim\limits_{n\to\infty} D(f_n, g_n) \leqslant D$.

Теорема Шеннона-Хартли для сжатия с потерями

Определение

Функция rate-distortion $R(D)$ - инфимум таких R, что пара (R, D) достижима.

Информационная rate-distortion функция: $R^I(D) = \inf\limits_{p(\hat{x}|x)} I(X; \hat{X})$,

где инфимум берётся по всем распределениям $p(\hat{x}|x)$, что $\sum\limits_{(x,\hat{x})} p(x, \hat{x}) \cdot d(x, \hat{x}) \leqslant D$.

Теорема Шеннона-Хартли

$R(D) = R^I(D)$.

Теорема

Для $\mathcal{X} = \mathbb{R}$ и источника $\mathcal{N}(0, \sigma^2)$ $R(D) = \frac{1}{2} \log \frac{\sigma^2}{D}$.

SPARC и Lossy compression

Оптимальный энкодер

Пусть $s = (s_1, ..., s_n)$ - набор вещественных чисел, которые хотим закодировать.

- Энкодер: $\hat{\beta} = \text{argmin}||s - A\beta||^2$.
- Декодер: $\hat{s} = A\hat{\beta}$.

Кодовое слово - набор индексов ненулевых элементов в векторе $\hat{\beta}$, то есть $L \log M$ бит. На практике количество бит уменьшается в $\log n$ раз.

SPARC и Lossy compression

Successive-approximation энкодер

Идея: итеративно аппроксимировать остатки вида $\mathcal{R}_i = \mathcal{R}_{i-1} - c_i A_{m_i}$; $R_0 = s$.

Числа c_i - специального вида, $c_i = \sqrt{\frac{2R\sigma^2}{L}(1 - \frac{2R}{L})^{i-1}}$.

Утверждение (2014)

Пусть s приходит из нормального распределения $\mathcal{N}(0, \sigma^2)$. Тогда предложенный метод асимптотически достигает rate-distortion функции гауссовского источника $\sigma^2 e^{-2R}$, а вероятность эксцесса убывает экспоненциально относительно L.

SPARC и Lossy compression

AMP энкодер

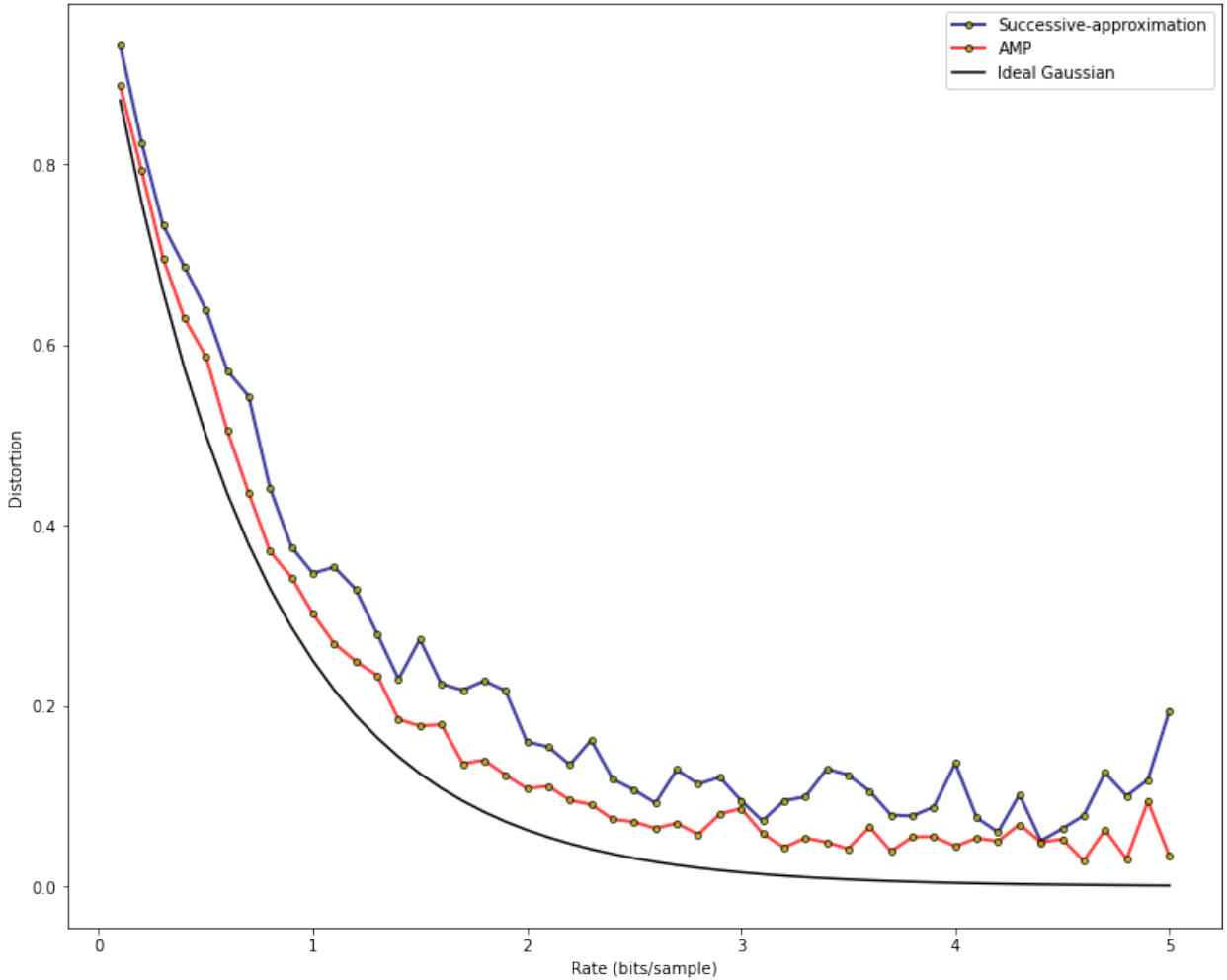
Алгоритм AMP - *Approximate Message Passing*, итеративно решает NP-трудную задачу построения байесовской оценки.

Конечный алфавит

Можно ли применить метод, например, к сжатию изображений?

Растровое изображение - набор чисел от 0 до 255 (например, 4 числа в случае RGBA). К самим наборам и большим блокам применять сжатие имеет мало смысла. Максимум - сжимать сами числа. Но это очень не эффективно.

Сравнение Successive и AMP энкодеров



Планы к отчётной конференции

- Доделать фреймворк для сжатия с потерями на основе SPARC
- Сделать выводы для случая конечного алфавита