# Exploring Features of Unitary Matrices in Kashin Representation Quantization

Зайнуллин Амир, Чирков Георгий, Богданов Азат, Черняков Севастьян, Нистюк Алексей

*Проект по методам оптимизации. МФТИ*

## Вступление

Проект основан на статье, в которой предлагается новый итерационный метод квантизации, основанный на представлении Кашина. В этом разложении один из основных объектов - ортогональная матрица. Мы проверим, какие способы получения ортогональной матрицы дают наискорейшую сходимость, исследуем сходимость метода от различной начальной инициализации и дадим теоретическое обоснование. Данный проект поможет в разработке более надежных методов сжатия LLM.

## Переопределенный базис и представление Кашина

Введем переопределенный базис размера $m = 2n$, где первые $n$ элементов из стандартного ОНБ базиса, а вторые $n$ - столбцы некоторой ортогональной матрицы $Q$. Пусть $B = [I \quad Q]$

$$x = u + Qv \quad \text{или} \quad x = \sum_{i=1}^{N} a_i b_i$$

Пусть имеется набор векторов $b_1, \ldots, b_N \in \mathbb{R}^n$. Мы называем представление вектора $x \in \mathbb{R}^n$ представлением Кашина уровня $c$, если

$$x = \sum_{i=1}^{N} a_i b_i \quad \text{и} \quad \max_{i=1,\ldots,N} |a_i| \le \frac{c}{\sqrt{N}} \|x\|_2$$

- Имеем вектор $x$ с большим разбросом значений. Мы хотим преобразовать этот вектор в новый с меньшим разбросом. Получим два вектора $u$ и $v$. Для этого существует жадный алгоритм.
- Оказывается, при некоторых обстоятельствах, элементы векторов $u$ и $v$ будут концентрироваться у нескольких пиков.

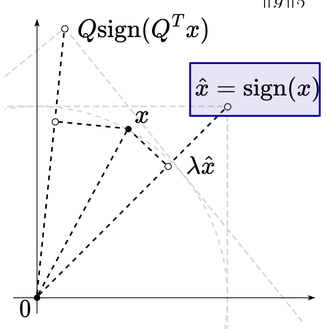
## Алгоритм

Псевдокод алгоритма **Vector Decomposition Kashin Algorithm**:

**Input:** Vector $x \in \mathbb{R}^n$, Orthogonal matrix $Q$, Tolerance $\varepsilon > 0$
**Output:** Vectors $u, \hat{v} \in \mathbb{R}^n$ such that $x \approx u + \hat{v} = u + Qv$, and both $u$ and $v$ have small infinity norm.

**Initialize:** $u \leftarrow 0^n, \hat{v} \leftarrow 0^n$ and **Define projection:** $\pi_x(y) \dfrac{x^\top y}{\|y\|_5^2} \cdot y$

**while** $\|x - u - \hat{v}\| \ge \varepsilon$ **do**
  **if** $\|x\|_1 > \|Q^T x\|_1$ **then**
    $\pi \leftarrow \pi_x(\text{Sign}(x))$
    $u \leftarrow u + \pi$
  **else**
    $\pi \leftarrow \pi_x(Q\text{Sign}(Q^T x))$
    $\hat{v} \leftarrow \hat{v} + \pi$
  **end if**
  $x \leftarrow x - \pi$
**end while**
**Return:** $x, u, \hat{v}$




## Теорема Кашина

**Theorem 1** *Для* $\forall x \in B_2^N = \{x \in \mathbb{R}^N : \|x\|_2 \le 1\}$ *жадный алгоритм за* $k$ *шагов строит вектора* $u_k$ *и* $v_k$*, такие что*

$$\|x - u_k - v_k\|_2 \le \eta^k$$

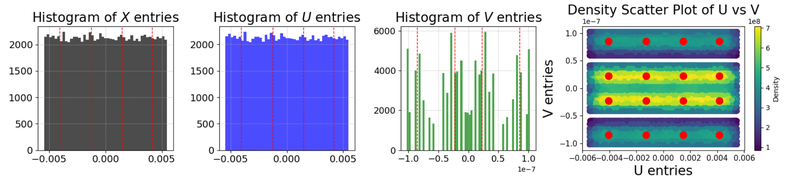
$$\|u_k\|_\infty \le \frac{c}{\sqrt{N}}, \quad \|Qv_k\|_\infty \le \frac{c}{\sqrt{N}}$$

*где* $c > 0$ *это константа. Для нашего алгоритма неравенство сходимости*

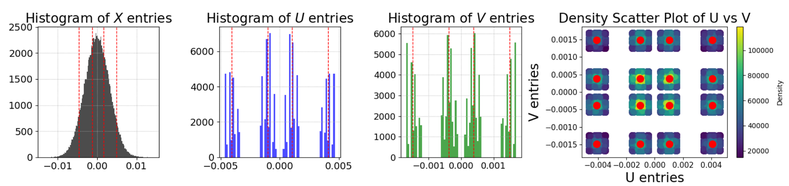
$$\|x - \hat{x}\|^2 \le 1 - \frac{1}{n}\left(\max(\|x\|_1, \|Q^T x\|_1)\right)^2 \tag{1}$$

## Генерация разных распределений компонент вектора X

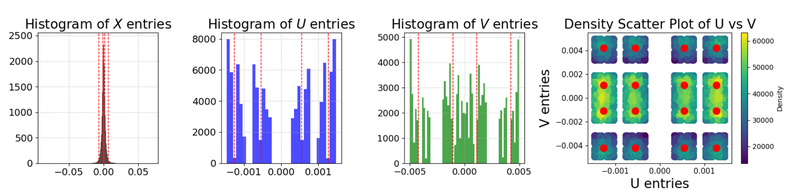
Равномерное распределение




Нормальное распределение



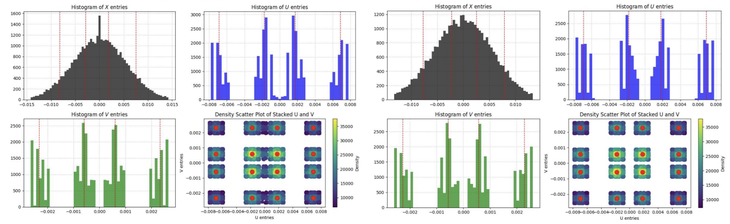

Распределение Стьюдента с двумя степенями свободы




- Как и было сказано, для неравномерных распределений наблюдается концентрирование у четырех пиков. Их количество зависит от размерности матрицы.
- Распределение с равномерным распределением (распределение с экстремально тяжелым хвостом) значений по компонентам дает зашумление по базису $U$
- Распределение с ярко выраженным пиком (Стьюдент) дает зашумление по базису $V$
- На худшее уменьшение невязки влияет меньшая повернутость гиперкуба, повернутого матрицей, относительно обычного гиперкуба

## Работа с outliers

Outliers - элементы матрицы, превышающие по абсолютному значению 99%(в нашем случае) остальных элементов матрицы (шумы). Исследуем, как влияет на кластеризацию удаление или зануление таких элементов.

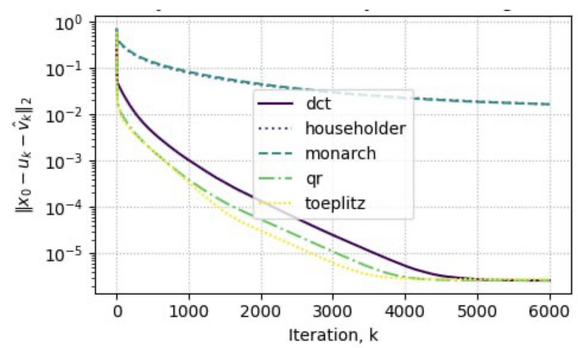



Левый график, который демонстрирует применение алгоритма в случае заполнения нулями компоненты outliers, отражает ухудшение квантизации. Тогда как правый график показывает состоятельность подхода зачистки outliers.

## Тест разных ортогональных матриц

В качестве метрики релевантности используется поперечник Колмогорова - нахождение входа, на котором происходит наихудшее уменьшение за итерацию (1). Нами дополнительно рассмотрены матрицы Теплица и Монарх, обе из которых крайне перспективны. Более тщательный анализ на слое модели показывает, что Монарх не так хорош, но вот матрица Теплица действительно может быть использована.

| $Q$ | $\min_{\|x\|_2=1} \max\{\|x\|_1, \|Q^T x\|_1\}$ |
|---|---|
| Random | 0.548 |
| Householder | 0.076 |
| DCT | 0.330 |
| Monarch | **0.686** |
| Toeplitz | **0.612** |




## Выводы

Наглядно продемонстрированы влияния экстремальных случаев распределения компонент входного вектора на квантизацию по базисам $U$ и $V$. Также выбран корректный подход работы с outliers. Исследованы новые ортогональные матрицы: Монарха и Теплица, и показано, что вторая матрица может давать улучшения в работе относительно случайной матрицы - подходе, используемом на данном этапе при квантизации.