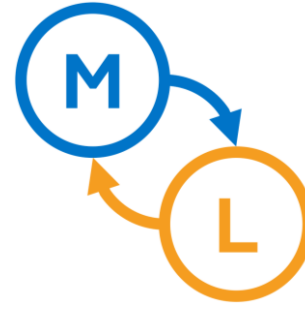




TurboQuant: Online Vector Quantization with Near-optimal Distortion Rate

по работе Zandieh et al., ICLR 2026

Максим Николаев
СПбГУ, 24 апреля 2026

Постановка задачи

Хотим искать пару таких **случайных** отображений $Q: \mathbb{R}^d \rightarrow \{0, 1\}^{b \cdot d}$ и $Q^{-1}: \{0, 1\}^{b \cdot d} \rightarrow \mathbb{R}^d$, что $Q^{-1}(Q(x)) \approx x$. Здесь b — число бит на одну координату.

Ясно, что Q не биекция, поэтому $Q^{-1} \circ Q$ имеет погрешность, которую мы хотим минимизировать.

Конкретнее, мы будем минимизировать

- MSE: $D_{mse} := \sup_x \mathbb{E}_Q \|x - Q^{-1}(Q(x))\|^2,$
- Ошибку скалярного произведения: $D_{prod} := \sup_{x,y} \mathbb{E}_Q |\langle y, x \rangle - \langle y, Q^{-1}(Q(x)) \rangle|^2,$

а также обеспечивать несмещенность скалярного произведения:

$$\mathbb{E}_Q \langle y, Q^{-1}(Q(x)) \rangle = \langle y, x \rangle.$$

TLDR

Авторами были предложены два алгоритма Q_{mse} и Q_{prod} со следующими характеристиками:

далее везде $\|x\| = 1$

- $D_{mse}(Q_{mse}) = \sup_x \mathbb{E} \|x - Q_{mse}^{-1}(Q_{mse}(x))\|^2 \leq \frac{\sqrt{3}\pi}{2} \cdot \frac{1}{4^b}$ для всех $b \geq 0$,
- Для $b = 1, 2, 3, 4$ имеется оценка $D_{mse}(Q_{mse}) \approx 0.36, 0.117, 0.03, 0.009$,
- $\mathbb{E} \left\langle y, Q_{prod}^{-1}(Q_{prod}(x)) \right\rangle = \langle y, x \rangle$,
- $D_{prod}(Q_{prod}) = \sup_{x,y} \mathbb{E} \left| \langle y, x \rangle - \left\langle y, Q_{prod}^{-1}(Q_{prod}(x)) \right\rangle \right|^2 \leq \frac{\sqrt{3}\pi^2 \cdot \|y\|^2}{d} \cdot \frac{1}{4^b}$ для всех $b \geq 0$.
- Для $b = 1, 2, 3, 4$ имеется оценка $D_{prod}(Q_{prod}) \cdot d \cdot \|y\|^{-2} \approx 1.57, 0.56, 0.18, 0.047$.

Кроме этого были получены нижние оценки:

- $D_{mse}(Q) \geq \frac{1}{4^b}$,
- $D_{prod}(Q) \geq \frac{\|y\|^2}{d} \cdot \frac{1}{4^b}$.

Алгоритм Q_{mse}

Algorithm 1 TURBOQUANT_{mse}: optimized for MSE

- 1: **input:** dimension d and bit-width b
// Global Parameters for Setting up TURBOQUANT_{mse}
 - 2: Generate a random rotation matrix $\mathbf{\Pi} \in \mathbb{R}^{d \times d}$
 - 3: Construct codebook by finding centroids $c_1, c_2, \dots, c_{2^b} \in [-1, 1]$ that minimize MSE cost in Eq. (4)
-

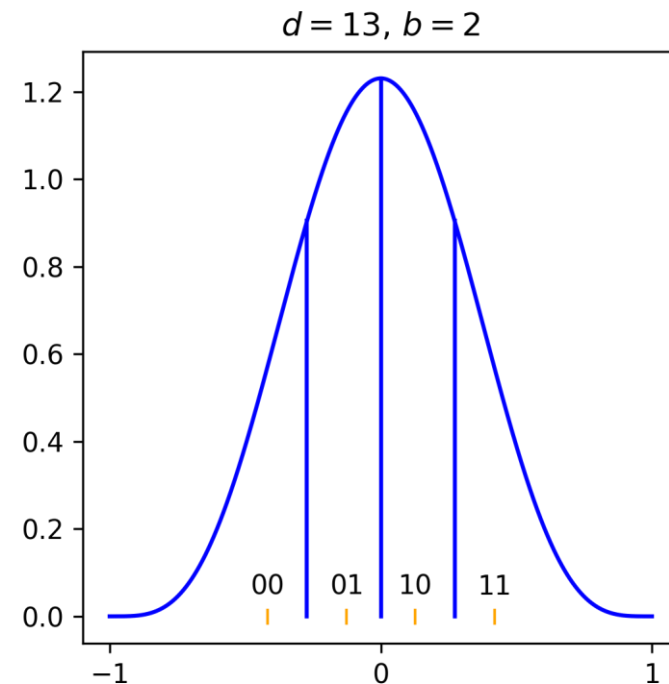
- 4: **Procedure** QUANT_{mse}(x) как и раньше, $\|x\| = 1$
 - 5: $y \leftarrow \mathbf{\Pi} \cdot x$
 - 6: $\text{idx}_j \leftarrow \arg \min_{k \in [2^b]} |y_j - c_k|$ for every $j \in [d]$ {idx_j's are b -bit integers}
 - 7: **output:** idx
-

- 8: **Procedure** DEQUANT_{mse}(idx)
 - 9: $\tilde{y}_j \leftarrow c_{\text{idx}_j}$ for every $j \in [d]$
 - 10: $\tilde{x} \leftarrow \mathbf{\Pi}^\top \cdot \tilde{y}$
 - 11: **output:** $\tilde{x}$
-

$$\mathcal{C}(f_X, b) := \min_{-1 \leq c_1 \leq c_2 \leq \dots \leq c_{2^b} \leq 1} \sum_{i=1}^{2^b} \int_{\frac{c_{i-1} + c_i}{2}}^{\frac{c_i + c_{i+1}}{2}} |x - c_i|^2 \cdot f_X(x) dx. \quad (4)$$

Ясно, что $D_{mse}(Q_{mse}) = d\mathcal{C}(f_X, b)$ и можно показать, что

$$\mathcal{C}(f_X, b) \leq \frac{1}{12} \cdot \left(\int f_X(x)^{13} dx \right)^3 \cdot \frac{1}{4^b} = \frac{\sqrt{3}\pi}{2d} \cdot 1/4^b$$



Пх это случайная точка на $\mathbb{S}^{d-1}$, ее координаты имеют распределение $2 \cdot \text{Beta}(\alpha, \alpha) - 1$, где $\alpha = (d - 3)/2$.

Почему это работает?

Если X это случайная точка на $\mathbb{S}^{d-1}$, то $X_i \sim 2 \cdot \text{Beta}(\alpha, \alpha) - 1$, где $\alpha = (d - 3)/2$.

В самом деле, плотность X_i в точке x пропорциональна площади $r\mathbb{S}^{d-2}$, деленной на r , где $r = (1 - x^2)^{1/2}$, что пропорционально $r^{d-2}/r = r^{d-3} = (1 - x^2)^{d-3} = (1 + x)^{(d-3)/2}(1 - x)^{(d-3)/2}$.

При $d \rightarrow \infty$ $\sqrt{d}X_i \xrightarrow{\text{dist}} \mathcal{N}(0,1)$, причем $\sqrt{d}X_i$ и $\sqrt{d}X_j$ асимптотически независимы. Это означает, что оптимальное кодирование x (асимптотически) эквивалентно конкатенации оптимальных кодов элементов.

Чтобы оптимально кодировать каждую координату с помощью b бит в терминах MSE, нужно решать задачу 2^b -средних для распределения координаты. Для каждого b и d решение свое, но мы можем предполагать, что d большое, а значит распределение почти нормальное, а значения центроидов предподсчитать для разных b .

Quantized Johnson—Lindenstrauss

Для данного $d \geq 1$ рассмотрим отображение $Q_{qjl}: \mathbb{S}^{d-1} \rightarrow \{-1, +1\}^d$:

$$Q_{qjl}(x) := \text{sign}(Sx),$$

где $S \in \mathbb{R}^{d \times d}$ и $S_{i,j}$ независимы и имеют распределение $\mathcal{N}(0,1)$, а sign применяется поэлементно.

Обратное преобразование определим как

$$Q_{qjl}^{-1}(z) = \frac{\sqrt{\pi/2}}{d} S^T z.$$

Фишка в том, что $\mathbb{E} \left\langle y, Q_{qjl}^{-1} \left(Q_{qjl}(x) \right) \right\rangle = \langle y, x \rangle$. Чтобы в этом убедиться, нужно проверить, что

$$\frac{\sqrt{\pi/2}}{d} \mathbb{E} \langle Sy, \text{sign}(Sx) \rangle = \sqrt{\pi/2} \mathbb{E} (S_1 y \cdot \text{sign}(S_1 x)) = \langle y, x \rangle.$$

Алгоритм Q_{prod}

Algorithm 2 TURBOQUANT_{prod}: optimized for inner product

1: **input:** dimension d and bit-width b

 // Global Parameters for Setting up TURBOQUANT_{prod}

2: Instantiate a TURBOQUANT_{mse} with bit-width $b - 1$ as per Algorithm 1

3: Generate a random projection matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ with i.i.d. entries $\mathbf{S}_{i,j} \sim \mathcal{N}(0, 1)$

4: **Procedure** QUANT_{prod}($\mathbf{x}$)

5: $\text{idx} \leftarrow \text{QUANT}_{\text{mse}}(\mathbf{x})$

6: $\mathbf{r} \leftarrow \mathbf{x} - \text{DEQUANT}_{\text{mse}}(\text{idx})$

{residual vector}

7: $\text{qjl} \leftarrow \text{sign}(\mathbf{S} \cdot \mathbf{r})$

{QJL on residual vector}

8: **output:** ($\text{idx}, \text{qjl}, \|\mathbf{r}\|_2$)

9: **Procedure** DEQUANT_{prod}($\text{idx}, \text{qjl}, \gamma$)

10: $\tilde{\mathbf{x}}_{\text{mse}} \leftarrow \text{DEQUANT}_{\text{mse}}(\text{idx})$

11: $\tilde{\mathbf{x}}_{\text{qjl}} \leftarrow \frac{\sqrt{\pi/2}}{d} \cdot \gamma \cdot \mathbf{S}^\top \cdot \text{qjl}$

12: **output:** $\tilde{\mathbf{x}}_{\text{mse}} + \tilde{\mathbf{x}}_{\text{qjl}}$

Качество работы алгоритма

Theorem 2 (performance guarantee: $\text{TURBOQUANT}_{\text{prod}}$). *For any bit-width $b \geq 1$ and any vector $\mathbf{x} \in \mathbb{S}^{d-1}$, the procedure $\text{QUANT}_{\text{prod}}(\mathbf{x})$ in Algorithm 2 outputs an index vector $\text{idx} \in [2^{b-1}]^d$ along with a sign vector $\mathbf{qjl} \in \{-1, 1\}^d$ and a positive number $\gamma \geq 0$. When these vectors and the scalar value are passed to the primitive $\text{DEQUANT}_{\text{prod}}(\text{idx}, \mathbf{qjl}, \gamma)$, it produces a reconstructed vector $\tilde{\mathbf{x}} \in \mathbb{R}^d$ that for any vector $\mathbf{y} \in \mathbb{R}^d$ satisfies the following properties:*

- *Expected inner-product $\mathbb{E}_{\tilde{\mathbf{x}}} [\langle \mathbf{y}, \tilde{\mathbf{x}} \rangle] = \langle \mathbf{y}, \mathbf{x} \rangle$*
- *Inner-product distortion defined as $D_{\text{prod}} := \mathbb{E}_{\tilde{\mathbf{x}}} \left[|\langle \mathbf{y}, \mathbf{x} \rangle - \langle \mathbf{y}, \tilde{\mathbf{x}} \rangle|^2 \right]$ is bounded by $D_{\text{prod}} \leq \frac{\sqrt{3}\pi^2 \cdot \|\mathbf{y}\|_2^2}{d} \cdot \frac{1}{4^b}$ for any $b \geq 0$.*
- *For small bit-widths, specifically $b = 1, 2, 3, 4$, D_{prod} exhibits finer-grained distortion values: $D_{\text{prod}} \approx \frac{1.57}{d}, \frac{0.56}{d}, \frac{0.18}{d}, \frac{0.047}{d}$, respectively.*

Качество работы алгоритма

Proof. First we compute the conditional expectation of the inner product estimate $\langle \mathbf{y}, \tilde{\mathbf{x}} \rangle$ conditioned on $\tilde{\mathbf{x}}_{\text{mse}}$ as follows:

$$\begin{aligned}\mathbb{E}[\langle \mathbf{y}, \tilde{\mathbf{x}} \rangle | \tilde{\mathbf{x}}_{\text{mse}}] &= \mathbb{E}_{\tilde{\mathbf{x}}_{\text{qj1}}} [\langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{mse}} + \tilde{\mathbf{x}}_{\text{qj1}} \rangle | \tilde{\mathbf{x}}_{\text{mse}}] \\ &= \langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{mse}} \rangle + \mathbb{E}_{\tilde{\mathbf{x}}_{\text{qj1}}} [\langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{qj1}} \rangle | \tilde{\mathbf{x}}_{\text{mse}}] \\ &= \langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{mse}} \rangle + \langle \mathbf{y}, \mathbf{r} \rangle \\ &= \langle \mathbf{y}, \mathbf{x} \rangle,\end{aligned}$$

where the first equality follows from the definition of $\tilde{\mathbf{x}}$ in line 12 of the algorithm. The third equality above follows from Lemma 4 and last line follows from definition of the residual vector $\mathbf{r} = \mathbf{x} - \tilde{\mathbf{x}}_{\text{mse}}$ in line 6. Now we can compute the unconditional expectation using the law of total expectation: $\mathbb{E}_{\tilde{\mathbf{x}}} [\langle \mathbf{y}, \tilde{\mathbf{x}} \rangle] = \mathbb{E}_{\tilde{\mathbf{x}}_{\text{mse}}} [\mathbb{E}[\langle \mathbf{y}, \tilde{\mathbf{x}} \rangle | \tilde{\mathbf{x}}_{\text{mse}}]] = \mathbb{E}[\langle \mathbf{y}, \mathbf{x} \rangle] = \langle \mathbf{y}, \mathbf{x} \rangle$, which proves the first claim of the theorem.

Качество работы алгоритма

We apply the same conditioning on $\tilde{\mathbf{x}}_{\text{mse}}$, when computing the distortion, and then compute the resulting conditional distortion:

$$\begin{aligned}\mathbb{E} \left[|\langle \mathbf{y}, \mathbf{x} \rangle - \langle \mathbf{y}, \tilde{\mathbf{x}} \rangle|^2 \middle| \tilde{\mathbf{x}}_{\text{mse}} \right] &= \mathbb{E}_{\tilde{\mathbf{x}}_{\text{qj1}}} \left[|\langle \mathbf{y}, \mathbf{x} \rangle - \langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{mse}} + \tilde{\mathbf{x}}_{\text{qj1}} \rangle|^2 \middle| \tilde{\mathbf{x}}_{\text{mse}} \right] \\ &= \mathbb{E}_{\tilde{\mathbf{x}}_{\text{qj1}}} \left[|\langle \mathbf{y}, \mathbf{r} \rangle - \langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{qj1}} \rangle|^2 \middle| \tilde{\mathbf{x}}_{\text{mse}} \right] \\ &= \text{Var} \left(\langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{qj1}} \rangle \middle| \tilde{\mathbf{x}}_{\text{mse}} \right) \\ &\leq \frac{\pi}{2d} \cdot \|\mathbf{r}\|_2^2 \|\mathbf{y}\|_2^2,\end{aligned}$$

where the second equality above follows from the definitions of $\mathbf{r}$ and $\tilde{\mathbf{x}}_{\text{mse}}$ in lines 6 and 10 of Algorithm 2. The third line above follows because $\mathbb{E}[\langle \mathbf{y}, \tilde{\mathbf{x}}_{\text{qj1}} \rangle] = \langle \mathbf{y}, \mathbf{r} \rangle$, by Lemma 4. The last line follows from the variance bound of QJL estimator shown in Lemma 4 and using the fact that $\tilde{\mathbf{x}}_{\text{qj1}}$ in line 11 is re-scaled by $\gamma = \|\mathbf{r}\|$.

Нижние оценки

Shannon Lower Bound

Пусть $X \in \mathbb{R}^d$ это случайный вектор с плотностью $p(x)$ с конечной дифференциальной энтропией $H(X)$. Пусть

$$D(p, B) := \inf\{\mathbb{E}\|X - Y\|^2 \mid I(X, Y) \leq B\},$$

где инфимум берется по всем совместным распределениям (X, Y) , в которых взаимная информация $I(X, Y) = H(Y) - H(Y|X)$ не больше B .

Тогда $D(p, B) \geq \frac{d}{2\pi e} \cdot 2^{(2/d)(H(x)-B)}$.

В случае $X \sim U(\mathbb{S}^{d-1})$ получаем $D(B) \geq 2^{-2B/d} = \frac{1}{4^b}$.

Нижние оценки

Для любого рандомизированного алгоритма $Q: \mathbb{S}^{d-1} \rightarrow \{0, 1\}^{b \cdot d}$ и любого $Q^{-1}: \{0, 1\}^{b \cdot d} \rightarrow \mathbb{R}^d$ существует такой $x \in \mathbb{S}^{d-1}$, что

$$\mathbb{E} \|x - Q^{-1}(Q(x))\|^2 \geq \frac{1}{4^d}.$$

Кроме того, существует $y \in \mathbb{S}^{d-1}$ такой, что

$$\mathbb{E} (\langle y, x \rangle - \langle y, Q^{-1}(Q(x)) \rangle)^2 \geq \frac{1}{d} \cdot \frac{1}{4^d}.$$

Yao's minimax principle

Ожидаемое время работы вероятностного алгоритма в наихудшем случае не меньше, чем ожидаемое время работы на наихудшем распределении на входных данных детерминированного алгоритма, который лучше всего подходит для этого распределения.

Shannon Lower Bound и Yao's principle $\Rightarrow$ первая оценка, принцип Дирихле $\Rightarrow$ вторая.