Fast JL Transform

Hu Fu @SHUFE

October 2022

Picture So Far

- Recall: JL-transform multiplies $\mathbf{x} \in \mathbb{R}^d$ with a $t \times d$ matrix with i.i.d. standard Gaussian entries
- AMS emulates JL-transform with a matrix with $\{0,1\}$ entries, each row generated by a 4-wise independent hash function
- This suggests that the matrix in JL-transform may be made simpler
  - Achlioptas (2003) gave a transform with matrix entries from $\{-1,0,1\}$, about 2/3 of them being 0
  - Count-Sketch in fact approximately preserves ℓ_2 norm (see homework)
    - Recall we had pairwise independent hash functions $h : [d] \to [w]$ and $g : [d] \to \{\pm 1\}$.
    - The operation of Count-Sketch can be seen as multiplying $\mathbf{x}$ first with a $d \times d$ diagonal matrix D with random entries from $\{\pm 1\}$, and then by a $w \times d$ matrix M with $M_{h(i),i} = g(i)$ and all other entries 0.

Making JL-transform Faster

- The original JL-transform takes time $\Omega(td)$ to multiply the matrix with $\mathbf{x}$
- $t = O(\log\left(\frac{1}{\delta}\right)\epsilon^{-2})$ if we would like to preserve the norm w.p. $1-\delta$
- Can we be faster asymptotically?
- Two approaches:
    - Sparse JL-Transforms: construct matrices satisfying the JL-property with few non-zero entries
    - Fast JL-Transforms: construct matrices satisfying the JL-property with structural properties which allow faster matrix multiplication [Ailon & Chazelle, 2006]

Today

First Attempt

- If we try with a $\{0,1\}$ matrix, what happens if we sample, for each entry, just one coordinate of $\mathbf{x}$?
- Consider $t \times d$ matrix S, where in each row a uniformly random entry is 1, and all other entries are 0.

$$\mathbb{E}\left[(S\mathbf{x})_i^2\right] = \sum_{j=1}^{d} \frac{x_j^2}{d} \quad \Rightarrow \quad \mathbb{E}\left[\|\sqrt{\frac{d}{t}}S\mathbf{x}\|_2^2\right] = \|\mathbf{x}\|_2^2$$

  - How well does $\|S\mathbf{x}\|_2^2$ concentrate around its expectation?
  - In the worst case, $\mathbf{x}$ has only one non-zero entry, then t needs to be $\Theta(d)$ for us to see that entry
  - Generally, this doesn't work well when $\dfrac{\|\mathbf{x}\|_\infty}{\|\mathbf{x}\|_2} \approx 1$

Rotating $\mathbf{x}$

- The ratio $\dfrac{\|\mathbf{x}\|_\infty}{\|\mathbf{x}\|_2}$ depends heavily on the bases in which we represent $\mathbf{x}$
  - E.g., the ratio is 1 for $(1,0,\cdots,0)$, but is only $1/\sqrt{d}$ if we rotate it to $(1/\sqrt{d},\cdots,1/\sqrt{d})$
  - The ratio is close to 1 if $\mathbf{x}$ "aligns well" with the axes, i.e., standard basis
  - Idea: first rotate $\mathbf{x}$ randomly — equivalent to multiplying it by a random orthogonal matrix M

In fact in the original JL paper, the matrix is a projection onto a random t-dimensional subspace of $\mathbf{R}^d$

Walsh-Hadamard Matrix

Def. A Hadamard matrix is an orthogonal $d \times d$ matrix with all entries from $\{1/\sqrt{d}, -1/\sqrt{d}\}$.

Example. $H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} / \sqrt{2}$

Walsh-Hadamard Matrices. $H_d = \begin{pmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2} \end{pmatrix} / \sqrt{2}$

A recursive construction

Claim. H_d is a Hadamard matrix

Proof. By induction, each entry of $H_{d/2}$ is in $\{\pm 1/\sqrt{d/2}\}$, so each entry of H_d is in $\{\pm 1/\sqrt{d}\}$.

$$H_d^\top H_d = \begin{pmatrix} H_{d/2}^\top & H_{d/2}^\top \\ H_{d/2}^\top & -H_{d/2}^\top \end{pmatrix} \begin{pmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2} \end{pmatrix} /2 = \begin{pmatrix} 2H_{d/2}^\top H_{d/2} & 0 \\ 0 & 2H_{d/2}^\top H_{d/2} \end{pmatrix} /2 = I$$

Walsh-Hadamard Matrix

Def. A Hadamard matrix is an orthogonal $d \times d$ matrix with all entries from $\{1/\sqrt{d}, -1/\sqrt{d}\}$.

Example. $H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} / \sqrt{2}$

Walsh-Hadamard Matrices. $H_d = \begin{pmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2} \end{pmatrix} / \sqrt{2}$

A recursive construction

Claim. H_d is a Hadamard matrix

Claim. The product $H_d \mathbf{x}$ can be computed in time $O(d \log d)$.

similar to Fast Fourier Transform

Proof. Let $T(d)$ be the time to compute $H_d \mathbf{x}$, then by recursive calls we have $T(d) = O(d) + 2T(d/2)$

Randomized Hadamard Matrix

- Of course a "good" vector $\mathbf{x}$ can become a "bad" $H_d\mathbf{x}$.
- We need to introduce randomness to H_d
- The -1 in H_2 may as well be elsewhere
- Let D be a $d \times d$ diagonal matrix with diagonal entries randomly sampled from $\{-1,1\}$
- H_dD is still Hadamard: $(H_dD)^\top H_dD = D^\top H_d^\top H_dD = I$.
- Write $H = H_d$ henceforth

Thm. For nonzero $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{y} = HD\mathbf{x}$, then $\mathbb{P}\left[\frac{\|\mathbf{y}\|_\infty}{\|\mathbf{y}\|_2} \geq \sqrt{\frac{2\ln(4d/\delta)}{d}}\right] \leq \frac{\delta}{2}$.

Thm. For nonzero $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{y} = HD\mathbf{x}$, then $\mathbb{P}\left[\frac{\|\mathbf{y}\|_\infty}{\|\mathbf{y}\|_2} \geq \sqrt{\frac{2\ln(4d/\delta)}{d}}\right] \leq \frac{\delta}{2}$.

Proof. Without loss of generality, assume $\|\mathbf{x}\|_2 = 1$. Then $\|\mathbf{y}\|_2 = \|HD\mathbf{x}\|_2 = 1$ as well.
To bound $\|\mathbf{y}\|_\infty$, note that for each i, y_i has the same distribution as $\sum_j D_j x_j$ where D_j's are i.i.d. Rademacher variables. If we let $z_j := D_j x_j$, then $\mathbb{E}[z_j] = 0$. Chernoff bound does not apply…

Thm. (Hoeffding's Bound) If $X_1, \cdots, X_n$ are independent random variables where $X_i \in [a_i, b_i]$. Let $X = \sum_i X_i$.

Then $\mathbb{P}(|X - \mathbb{E}[X]| \geq s) \leq 2\exp\left(-\frac{2s^2}{\sum_i (b_i - a_i)^2}\right)$.

Proof idea similar to Chernoff bound. Use the following bound on $\mathbb{E}[e^{\lambda X_i}]$:

Lemma. (Hoeffding's Lemma) If random variable X is in $[a_i, b_i]$, then $\mathbb{E}[e^{\lambda(X - \mathbf{E}[X])}] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$.

Thm. For nonzero $\mathbf{x} \in \mathbb{R}^d$, let $\mathbf{y} = HD\mathbf{x}$, then $\mathbb{P}\left[\frac{\|\mathbf{y}\|_\infty}{\|\mathbf{y}\|_2} \geq \sqrt{\frac{2\ln(4d/\delta)}{d}}\right] \leq \frac{\delta}{2}$.

Proof. Without loss of generality, assume $\|\mathbf{x}\|_2 = 1$. Then $\|\mathbf{y}\|_2 = \|HD\mathbf{x}\|_2 = 1$ as well.
To bound $\|\mathbf{y}\|_\infty$, note that for each i, y_i has the same distribution as $\sum_j D_j x_j$ where D_j's are i.i.d. Rademacher variables. If we let $z_j := D_j x_j$, then $\mathbb{E}[z_j] = 0$. Chernoff bound does not apply since $z_j \in [-x_j/\sqrt{d}, x_j/\sqrt{d}]$
To apply Hoeffding's bound, note that $\sum_j 4x_j^2/d^2 = 4/d$.

$$\mathbb{P}\left[|z_j| \geq \sqrt{\frac{2\ln(4d/\delta)}{d}}\right] \leq 2\exp\left(-\frac{d}{2}\frac{2\ln(4d/\delta)}{d}\right) = 2 \cdot \frac{\delta}{4d} = \frac{\delta}{2d}.$$

The theorem follows from a union bound over z_j's.

Thm. (Hoeffding's Bound) If $X_1, \cdots, X_n$ are independent random variables where $X_i \in [a_i, b_i]$. Let $X = \sum_i X_i$.
Then $\mathbb{P}(|X - \mathbb{E}[X]| \geq s) \leq 2\exp\left(-\frac{2s^2}{\sum_i (b_i - a_i)^2}\right)$.

The Pieces We Have...

- For any $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| = 1$, $\mathbb{P}[\|\sqrt{\frac{d}{t}}SHD\mathbf{x}\| \in [1-\epsilon, 1+\epsilon]] = ?$
    - $D \in \{-1,0,1\}^{d\times d}$: diagonal matrix with Rademacher entries
    - $H \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^{d\times d}$: Walsh Hadamard matrix
    - $S \in \{0,1\}^{t\times d}$: Sampling matrix, with exactly one 1 in each row
- $\|HD\mathbf{x}\|_2 = 1$. With probability $\geq 1 - \frac{\delta}{2}$, $\|HD\mathbf{x}\|_\infty \leq \sqrt{\frac{2\ln(4d/\delta)}{d}}$
- Let $\mathbf{z} := SHD\mathbf{x}$. Then each $\mathbb{E}[z_i^2] = \frac{1}{d}$, $\mathbb{E}[\|\mathbf{z}\|_2^2] = \frac{t}{d}$. With probability $\geq 1 - \frac{\delta}{2}$, $z_i^2 \leq \frac{2\ln(4d/\delta)}{d}$.

Claim. If $z_i^2 \leq \frac{2\ln(4d/\delta)}{d}$, for $t = \frac{2\ln^2(4d/\delta)\ln(4/\delta)}{\epsilon^2}$, $\mathbb{P}\left[\|\sqrt{\frac{d}{t}}S\mathbf{z}\|_2^2 \notin [1-\epsilon, 1+\epsilon]\right] \leq \frac{\delta}{2}$

Directly applying Hoeffding's bound yields $t = \Omega(\sqrt{d})$. To get the better bound, we need to make use of the variance of $S\mathbf{z}$ and another concentration bound called Berstein's inequality.

Fast JL-Transform

Thm. For any $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| = 1$, for $t \geq \dfrac{2\ln^2(4d/\delta)\ln(4/\delta)}{\epsilon^2}$,

$$\mathbb{P}\left[\|\sqrt{\frac{d}{t}}SHD\mathbf{x}\| \in [1-\epsilon, 1+\epsilon]\right] \geq 1-\delta.$$

- $D \in \{-1,0,1\}^{d\times d}$: diagonal matrix with Rademacher entries
- $H \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^{d\times d}$: Walsh Hadamard matrix
- $S \in \{0,1\}^{t\times d}$: Sampling matrix, with exactly one 1 in each row