Vector Search **Class #XXXXX Fall 2023**

Lecture number XX: Fast random projection

Lectures: Edo Liberty and Matthijs Douse

Warning: *Please do not cite this note as a peer reviewed source. If you find mistakes, please inform the authors.*

We discussed in class the fact that random projection matrices cannot be made sparse in general. That is because projecting sparse vectors and preserving their norm requires the projecting matrix is almost fully dense see also [1] and [2].

But, the question is, can we actively make sure that x is not sparse? If so, can we achieve a sparse random projection for non sparse vectors? These two questions received a positive answer in the seminal work by Ailon and Chazelle [3]. The results of [3] were improved and simplified over the years. See [4] for the latest result and an overview.

In this lesson we will produce a very simple algorithm based on the ideas in [3]. This algorithm will require a target dimension of $O(\log^2(n)/\varepsilon^2)$ instead of $O(\log(n)/\varepsilon^2)$ but will be much simpler to prove.

0.1 Fast vector ℓ_4 norm reduction

The goal of this subsection is to devise a linear mapping which preserves vector's ℓ_2 norms but reduces their ℓ_4 norms with high probability. This will work to our advantage because, intuitively, vectors whose ℓ_4 norm is small cannot be too sparse. For this we will need to learn what Hadamard matrices are.

Hadamard matrices are commonly used in coding theory and are conceptually close for Fourier matrices. We assume for convenience that d is a power of 2 (otherwise we can pad out vectors with zeros). The Walsh Hadamard transform of a vector $x \in \mathbb{R}^d$ is the result of the matrix-vector multiplication Hx where H is a $d \times d$ matrix whose entries are $H(i,j) = \frac{1}{\sqrt{d}}(-1)^{\langle i,j \rangle}$. Here $\langle i,j \rangle$ means the dot product over F_2 of the bit representation of i and j as binary vectors of length $\log(d)$. Another way to view this is to define Hadamard Matrices recursively.

$$H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad H_d = \frac{1}{\sqrt{2}} \left(\begin{array}{c:c} H_{d/2} & H_{d/2} \\ \hdashline H_{d/2} & -H_{d/2} \end{array}\right)$$

Here are a few interesting (and easy to show) facts about Hadamard matrices.

1. H_d is a unitary matrix $\|Hx\| = \|x\|$ for any vector $x \in \mathbb{R}^d$.
2. Computing $x \mapsto Hx$ requires $O(d\log(d))$ operations.

We also define a diagonal matrix D to be such that $D(i,i) \in \{1,-1\}$ uniformly. Clearly, we have that $\|HDx\|_2 = \|x\|_2$ since both H and D are isotropies. Let us now bound $\|HDx\|_\infty$. $(HDx)(1) = \sum_{i=1}^{d} H(1,i)D(i,i)x_i = \sum_{i=1}^{d} \frac{x_i}{\sqrt{d}} s_i$ where $s_i \in \{-1,1\}$ uniformly. To bound this we recap Hoeffding's inequality.

Fact 0.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables s.t. $X_i \in [a_i, b_i]$. Let $X = \sum_{i=1}^{n} X_i$.*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}} \tag{1}$$

Invoking Hoeffding's inequality and then the union bound we get that if $\|HDx\|_\infty \leq \sqrt{\frac{c\log(n)}{d}}$ for all points x. Remark, for this we assumed $\log(d) = O(\log(n))$ otherwise we should have had $\log(nd)$ in the bound. The situation, however, that the dimension is super polynomial in the number of points is unlikely. Usually it is common to have $n > d$.

Lemma 0.1. *Let $x \in \mathbb{R}^d$ by such that $\|x\| = 1$. Then:*

$$\|HDx\|_4^4 = O(log(n)/d)$$

with probability at least $1 - 1/\text{poly}(n)$

Proof. Let us define $y = HDx$ and $z_i = y_i^2$. From the above we have that $z_i \leq \frac{c\log(n)}{d} = \eta$ with probability at least $1 - 1/\text{poly}(n)$. The quantity $\|HDx\|_4^4 = \|y\|_4^4 = \sum_i z_i^2$ is a convex function of the z variables which is defined over a polytop $z_i \in [0, 1]$ and $\sum_i z_i = 1$ (this is because $\|y\|_2^2 = 1$). This means that its maximal value is obtained on an extreme point of this polytope. In other words, the point $z_1, \ldots, z_{1/\eta} = \eta$ and $z_{1/\eta+1}, \ldots, z_d = 0$ or $z = [\eta, \eta, \ldots, \eta, \eta, 0, 0, 0, \ldots, 0, 0, 0]$. Computing the value of the function in this point gives $\sum_i z_i^2 \leq (1/\eta) \cdot (\eta^2) = \eta$. Recalling the $\eta = \frac{c\log(n)}{d}$ completes the proof. □

0.2 Sampling from vectors with low ℓ_4 norms

Here we prove a very simple fact. For vectors whose ℓ_4 is low, dimensionality reduction can be obtained by sampling.

Let y be a vector such that $\|y\|_2 = 1$. Let z be a sampled version of y such that $z_i = y_i/\sqrt{p}$ with probability p and 0 else. This is akin to sampling, in expectation, $d \cdot p$ coordinates from y (and scaling them up by $1/\sqrt{p}$). Note the $\mathbb{E}[\|z\|^2] = \mathbb{E}[\|y\|^2] = 1$ moreover:

$$\Pr[|\|z\|^2 - 1| > \varepsilon] = \Pr[|\sum z_i^2 - 1| > \varepsilon] = \Pr[|\sum b_i y_i^2/p - 1| > \varepsilon]$$

Where b_i are independent random indicator variables taking the $b_i = 1$ with probability p and $b_i = 0$ else. To apply Chernoff's bound we must assert that $y_i^2/p \leq 1$. Let's assume this for now and return to it later. Applying Chernoff's bound we get

$$\Pr[|\sum b_i y_i^2/p - 1| > \varepsilon] \leq e^{-\frac{c\varepsilon^2}{\sigma^2}}$$

where $\sigma^2 = \sum_i \mathbb{E}[(b_i y_i^2/p)^2] = \|y\|_4^4/p$. Concluding that

$$\Pr[|\|z\|^2 - 1| > \varepsilon] \leq e^{-\frac{cp\varepsilon^2}{\|y\|_4^4}}$$

This shows that the concentration of the sampling procedure really depends directly on the ℓ_4 norm of the sampled vector. If we plug in the bound on $\|y\|_4^4 = \|HDx\|_4^4$ from the previous section we get

$$\Pr[|\|z\|^2 - 1| > \varepsilon] \leq e^{-\frac{cp\varepsilon d}{\log(n)}} \leq \frac{1}{\text{poly}(n)}$$

For some $p \in O(\log^2(n)/d\varepsilon^2)$.

0.3 Random Projection by Sampling

Putting it all together we obtain the following.

Lemma 0.2. *Define the following matrices*

- *D: A diagonal matrix such that $D_{i,i} \in \{+1, -1\}$ uniformly.*
- *H: The $d \times d$ Walsh Hadamard Transform matrix.*
- *P: A 'sampling matrix' which contains each row of matrix $I_d \cdot \sqrt{p}$ with probability $p = c\log^2(n)/d\varepsilon^2$.*

Then, with at least constant probability the following holds.

1. *The target dimension of the mapping is* $k = c\log^2(n)/\varepsilon^2$ *(a factor log*(n) *worse than optimal).*

2. *The mapping* $x \mapsto PHDx$ *is a* $(1 \pm \varepsilon)$*-distortion mapping for any set of* n *points. That is, for any set* $x_1, \ldots, x_n \in \mathbb{R}^d$ *we have*

$$\|x_i - x_j\|(1-\varepsilon) \le \|PHDx_i - PHDx_j\| \le \|x_i - x_j\|(1+\varepsilon)$$

3. *Storing* PHD *requires at most* $O(d + k\log(d))$ *space.*

4. *Applying the mapping* $x \mapsto PHDx$ *requires at most* $d\log(d)$ *floating point operations.*

References

[1] Jelani Nelson and Huy L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *arXiv:1211.0995v1*, 2012.

[2] Daniel M. Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. In *SODA*, pages 1195–1206, 2012.

[3] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38st Annual Symposium on the Theory of Compututing (STOC)*, pages 557–563, Seattle, WA, 2006.

[4] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast johnson-lindenstrauss transform. In *SODA*, pages 185–191, 2011.