Concentration of Measure 2

Jarosław Błasiok

January 14, 2020

Concentration of Measure: Oblivious subspace embedding

Theorem

For any k, ε, δ, there is $m = \mathcal{O}(\frac{k+\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over matrices in $\mathbb{R}^{m \times d}$, satisfying for any $V \subset \mathbb{R}^d$:

$$\mathbb{P}_{T \sim \mathcal{D}}(\forall_{x \in V} \|Tx\|_2 = (1 \pm \varepsilon)\|x\|_2) > 1 - \delta.$$

Concentration of Measure: Oblivious subspace embedding

Theorem

For any k, ε, δ, there is $m = \mathcal{O}(\frac{k+\log \delta^{-1}}{\varepsilon^2})$, and a distribution $\mathcal{D}$ over matrices in $\mathbb{R}^{m \times d}$, satisfying for any $V \subset \mathbb{R}^d$:

$$\mathbb{P}_{T \sim \mathcal{D}}(\forall_{x \in V} \|Tx\|_2 = (1 \pm \varepsilon)\|x\|_2) > 1 - \delta.$$

Note

In fact we can chose distribution D to be any distribution with subgaussian independent entries, or more generally any distribution satisfying Distributional Johnson-Lindenstrauss.

Application: sketch and solve for overconstrained linear regression

Given $A \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^n$ we want to solve

$$\min \|Ax - b\|_2.$$

Application: sketch and solve for overconstrained linear regression

Given $A \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^n$ we want to solve

$$\min \|Ax - b\|_2.$$

If Π is subspace embedding for $span(a_1, \ldots a_k, b)$, where a_i are columns of i, we can instead solve

$$\min \|\Pi Ax - \Pi b\|_2.$$

Note that ΠA is only $\mathcal{O}(k) \times k$ size matrix.

Application: sketch and solve for overconstrained linear regression

Given $A \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^n$ we want to solve

$$\min \|Ax - b\|_2.$$

If Π is subspace embedding for $span(a_1, \ldots a_k, b)$, where a_i are columns of i, we can instead solve

$$\min \|\Pi Ax - \Pi b\|_2.$$

Note that ΠA is only $\mathcal{O}(k) \times k$ size matrix.

Unfortunately

If Π is unstructured, computing ΠA is as slow as solving the original regression problem. We can use more structured matrices Π that admit fast matrix-vector multiplication (sparse JL, Fast JL).

Oblivious subspace embedding

Fast Johnon-Lindenstrauss

Oblivious subspace embedding

Fast Johnon-Lindenstrauss

Plan of the proof

Definition (α-net, α-covering)

We say that a set $\mathcal{M}$ is an α-net for B, if for every $b \in B$, there is some $m_b \in \mathcal{M}$ such that $\|m - m_b\| \leq \alpha$.

Plan of the proof

Definition (α-net, α-covering)

We say that a set $\mathcal{M}$ is an α-net for B, if for every $b \in B$, there is some $m_b \in \mathcal{M}$ such that $\|m - m_b\| \leq \alpha$.

Strategy

- Consider $V \subset \mathbb{R}^d$, and $S_V := \{v \in V : \|v\|_2 = 1\}$. There is an α-net $\mathcal{M}_\alpha \subset S_V$ for S_V of size $(1 + \frac{1}{\alpha})^k$.

Plan of the proof

Definition (α-net, α-covering)

We say that a set $\mathcal{M}$ is an α-net for B, if for every $b \in B$, there is some $m_b \in \mathcal{M}$ such that $\|m - m_b\| \leq \alpha$.

Strategy

- Consider $V \subset \mathbb{R}^d$, and $S_V := \{v \in V : \|v\|_2 = 1\}$. There is an α-net $\mathcal{M}_\alpha \subset S_V$ for S_V of size $(1 + \frac{1}{\alpha})^k$.
- Consider $\alpha = 1/10$. From Distributional Johnson-Lindenstrauss, for target dimension $m = \mathcal{O}(\frac{k + \log \delta^{-1}}{\varepsilon^2})$, we can deduce

$$\mathbb{P}_{T \sim \mathcal{D}}(\forall s, v \in \mathcal{M}_{1/10} \langle Ts, Tv \rangle = \langle s, v \rangle \pm \varepsilon) > 1 - \delta. \quad (1)$$

Plan of the proof

Definition (α-net, α-covering)

We say that a set $\mathcal{M}$ is an α-net for B, if for every $b \in B$, there is some $m_b \in \mathcal{M}$ such that $\|m - m_b\| \leq \alpha$.

Strategy

- Consider $V \subset \mathbb{R}^d$, and $S_V := \{v \in V : \|v\|_2 = 1\}$. There is an α-net $\mathcal{M}_\alpha \subset S_V$ for S_V of size $(1 + \frac{1}{\alpha})^k$.
- Consider $\alpha = 1/10$. From Distributional Johnson-Lindenstrauss, for target dimension $m = \mathcal{O}(\frac{k + \log \delta^{-1}}{\varepsilon^2})$, we can deduce

$$\mathbb{P}_{T \sim \mathcal{D}}(\forall s, v \in \mathcal{M}_{1/10} \langle Ts, Tv \rangle = \langle s, v \rangle \pm \varepsilon) > 1 - \delta. \quad (1)$$

- If linear map T satisfies (1), then in fact $\forall v \in V, \|Tv\|_2 = (1 \pm 3\varepsilon)\|v\|_2$.

Construction of an α-net: packing/covering duality

Definition (Packing Numbers)

We say that a set $P \subset B$ is α-packing for B if and only if $\forall_{s \neq v \in P} \|s - v\| > \alpha$.
We write $\mathcal{P}(B, \alpha)$ for the size of the largest α-packing for B.
We write $\mathcal{N}(B, \alpha)$ for the size of smallest α-net for B.

Construction of an α-net: packing/covering duality

Definition (Packing Numbers)

We say that a set $P \subset B$ is α-packing for B if and only if $\forall_{s \neq v \in P} \|s - v\| > \alpha$.
We write $\mathcal{P}(B, \alpha)$ for the size of the largest α-packing for B.
We write $\mathcal{N}(B, \alpha)$ for the size of smallest α-net for B.

Fact: $\mathcal{P}(B, 2\alpha) \leq \mathcal{N}(B, \alpha)$

Consider covering N and packing P. If there are two points in P covered by the same ball in N, distance between them is at most 2α.

Construction of an α-net: packing/covering duality

Definition (Packing Numbers)

We say that a set $P \subset B$ is α-packing for B if and only if $\forall_{s \neq v \in P} \|s - v\| > \alpha$.
We write $\mathcal{P}(B, \alpha)$ for the size of the largest α-packing for B.
We write $\mathcal{N}(B, \alpha)$ for the size of smallest α-net for B.

Fact: $\mathcal{P}(B, 2\alpha) \leq \mathcal{N}(B, \alpha)$

Consider covering N and packing P. If there are two points in P covered by the same ball in N, distance between them is at most 2α.

Fact: $\mathcal{N}(B, \alpha) \leq \mathcal{P}(B, \alpha)$

Any maximal α-packing P is in fact an α-covering. Consider any $b \in B$: if there is no $p_b \in P$ with $\|p - b\| < \alpha$, then $P \cup \{b\}$ is also packing.

Construction of α-net: volume argument

Take $B_2^n = \{v \in \mathbb{R}^n : \|v\| \leq 1\|\}$. It is enough to show that $\mathcal{P}(B_2^n, \alpha) \leq (1 + \frac{1}{\alpha})^n$.

Construction of α-net: volume argument

Take $B_2^n = \{v \in \mathbb{R}^n : \|v\| \leq 1\|\}$. It is enough to show that $\mathcal{P}(B_2^n, \alpha) \leq (1 + \frac{1}{\alpha})^n$.
Take any α-packing $P \subset B_2^n$. Note for $p \in B_2^n$ balls $p + \frac{\alpha}{2} B_2$ are disjoint and all contained in $(1 + \alpha)B_2$. Therefore:

$$Vol(\cup_{p \in P} p + \alpha B_2) \leq Vol((1 + \alpha)B_2),$$

Construction of α-net: volume argument

Take $B_2^n = \{v \in \mathbb{R}^n : \|v\| \leq 1\|\}$. It is enough to show that $\mathcal{P}(B_2^n, \alpha) \leq (1 + \frac{1}{\alpha})^n$.
Take any α-packing $P \subset B_2^n$. Note for $p \in B_2^n$ balls $p + \frac{\alpha}{2} B_2$ are disjoint and all contained in $(1 + \alpha)B_2$. Therefore:

$$Vol(\cup_{p \in P} p + \alpha B_2) \leq Vol((1 + \alpha)B_2),$$

which implies

$$|P| \leq \frac{Vol((1 + \alpha)B_2)}{Vol(\alpha B_2)} = (1 + \frac{1}{\alpha})^n.$$

Johnson-Lindenstrauss implies preserving inner-products

Lemma

Consider $M \subset S_2^n$. Any distribution $\mathcal{D}$ over linear matrices satisfying DJL (preserving norms up to multiplicative error $1 + \varepsilon$), also satisfies with high probability over $T \sim \mathcal{D}$,

$$\forall u, v \in M, \langle Tu, Tv \rangle = \langle u, v \rangle \pm \varepsilon.$$

Johnson-Lindenstrauss implies preserving inner-products

Lemma

Consider $M \subset S_2^n$. Any distribution $\mathcal{D}$ over linear matrices satisfying DJL (preserving norms up to multiplicative error $1+\varepsilon$), also satisfies with high probability over $T \sim \mathcal{D}$,

$$\forall u, v \in M, \langle Tu, Tv \rangle = \langle u, v \rangle \pm \varepsilon.$$

Proof.

Note that

$$\langle u, v \rangle = \frac{\|u+v\|_2^2 - \|u-v\|_2^2}{4}.$$

If T preserves $\|u+v\|^2$ and $\|u-v\|^2$ for all u, v up to multiplicative error $(1+\varepsilon)$, we incur at most ε additive error in this expression. □

Extending from α-net to the entire subspace

Lemma

Assume that $\mathcal{M} \subset S_V$ is a 1/10-net for S_V. (remember: S_V is a unit sphere restricted to subspace V).
If $\forall s, t \in \mathcal{M} \, \langle Ts, Tv \rangle = \langle s, v \rangle \pm \varepsilon$, then

$$\forall v \in S_V, \, \|Tv\| = 1 \pm 2\varepsilon$$

Extending from α-net to the entire subspace

Lemma

Assume that $\mathcal{M} \subset S_V$ is a 1/10-net for S_V. (remember: S_V is a unit sphere restricted to subspace V).
If $\forall s, t \in \mathcal{M} \, \langle Ts, Tv \rangle = \langle s, v \rangle \pm \varepsilon$, then

$$\forall v \in S_V, \, \|Tv\| = 1 \pm 2\varepsilon$$

Proof

- Take $v_0 \in S_V$. We want to show that $\|Tv_0\| = 1 \pm \varepsilon$. We know that there is some $\pi_0 \in \mathcal{M}$, such that $\|v_0 - \pi_0\| < 1/10$.

Extending from α-net to the entire subspace

Lemma

Assume that $\mathcal{M} \subset S_V$ is a 1/10-net for S_V. (remember: S_V is a unit sphere restricted to subspace V).
If $\forall s, t \in \mathcal{M}\ \langle Ts, Tv \rangle = \langle s, v \rangle \pm \varepsilon$, then

$$\forall v \in S_V,\ \|Tv\| = 1 \pm 2\varepsilon$$

Proof

- Take $v_0 \in S_V$. We want to show that $\|Tv_0\| = 1 \pm \varepsilon$. We know that there is some $\pi_0 \in \mathcal{M}$, such that $\|v_0 - \pi_0\| < 1/10$.
- Take $v_1 := \frac{v_0 - \pi_0}{\|v_0 - \pi_0\|}$. We have $v_0 = \pi_0 + \gamma_1 v_1$. There is some $\pi_1 \in \mathcal{M}$, such that $\|v_1 - \pi_1\| < \frac{1}{10}$.

Extending from α-net to the entire subspace

Lemma

Assume that $\mathcal{M} \subset S_V$ is a 1/10-net for S_V. (remember: S_V is a unit sphere restricted to subspace V).
If $\forall s, t \in \mathcal{M}\ \langle Ts, Tv \rangle = \langle s, v \rangle \pm \varepsilon$, then

$$\forall v \in S_V,\ \|Tv\| = 1 \pm 2\varepsilon$$

Proof

- Take $v_0 \in S_V$. We want to show that $\|Tv_0\| = 1 \pm \varepsilon$. We know that there is some $\pi_0 \in \mathcal{M}$, such that $\|v_0 - \pi_0\| < 1/10$.
- Take $v_1 := \frac{v_0 - \pi_0}{\|v_0 - \pi_0\|}$. We have $v_0 = \pi_0 + \gamma_1 v_1$. There is some $\pi_1 \in \mathcal{M}$, such that $\|v_1 - \pi_1\| < \frac{1}{10}$.
- Take $v_2 = \frac{v_1 - \pi_1}{\|v_1 - \pi_1\|}$, we have $v_0 = \pi_0 + \gamma_1 \pi_1 + \gamma_2 v_2, \ldots$

Extending from α-net to the entire subspace

Lemma

Assume that $\mathcal{M} \subset S_V$ is a 1/10-net for S_V. (remember: S_V is a unit sphere restricted to subspace V).
If $\forall s, t \in \mathcal{M}\ \langle Ts, Tv \rangle = \langle s, v \rangle \pm \varepsilon$, then

$$\forall v \in S_V,\ \|Tv\| = 1 \pm 2\varepsilon$$

Proof

- Take $v_0 \in S_V$. We want to show that $\|Tv_0\| = 1 \pm \varepsilon$. We know that there is some $\pi_0 \in \mathcal{M}$, such that $\|v_0 - \pi_0\| < 1/10$.
- Take $v_1 := \frac{v_0 - \pi_0}{\|v_0 - \pi_0\|}$. We have $v_0 = \pi_0 + \gamma_1 v_1$. There is some $\pi_1 \in \mathcal{M}$, such that $\|v_1 - \pi_1\| < \frac{1}{10}$.
- Take $v_2 = \frac{v_1 - \pi_1}{\|v_1 - \pi_1\|}$, we have $v_0 = \pi_0 + \gamma_1 \pi_1 + \gamma_2 v_2, \ldots$
- We can decompose $v_0 := \sum_{k \le K} \gamma_k \pi_k + \gamma_{K+1} v_K$, where $\gamma_k \le (\frac{1}{10})^k$, $\pi_k \in \mathcal{M}$, and $\|v_K\| = 1$.

Extending from α-net to the entire subspace.

Where we are.

We can decompose $v_0 := \sum_{k \le K} \gamma_k \pi_k + \gamma_{K+1} v_K$, where $\gamma_k \le (\frac{1}{10})^k$, $\pi_k \in \mathcal{M}$, and $\|v_K\| = 1$.

We have $\langle T\pi_i, T\pi_j \rangle = \langle \pi_i, \pi_j \rangle \pm \varepsilon$ for all π_i.

Want: $\|Tv_0\|_2^2 = 1 \pm 3\varepsilon$.

Extending from α-net to the entire subspace.

Where we are.

We can decompose $v_0 := \sum_{k \le K} \gamma_k \pi_k + \gamma_{K+1} v_K$, where $\gamma_k \le (\frac{1}{10})^k$, $\pi_k \in \mathcal{M}$, and $\|v_K\| = 1$.
We have $\langle T\pi_i, T\pi_j \rangle = \langle \pi_i, \pi_j \rangle \pm \varepsilon$ for all π_i.
Want: $\|Tv_0\|_2^2 = 1 \pm 3\varepsilon$.

Proof

$$\|Tv_0\|^2 = \sum_{i,j \le K} \gamma_i \gamma_j \langle T\pi_i, T\pi_j \rangle + \gamma_{K+1} \langle Tv_K, Tv_0 \rangle ,$$

Extending from α-net to the entire subspace.

Where we are.

We can decompose $v_0 := \sum_{k \le K} \gamma_k \pi_k + \gamma_{K+1} v_K$, where $\gamma_k \le (\frac{1}{10})^k$, $\pi_k \in \mathcal{M}$, and $\|v_K\| = 1$.
We have $\langle T\pi_i, T\pi_j \rangle = \langle \pi_i, \pi_j \rangle \pm \varepsilon$ for all π_i.
Want: $\|Tv_0\|_2^2 = 1 \pm 3\varepsilon$.

Proof

$$\|Tv_0\|^2 = \sum_{i,j \le K} \gamma_i \gamma_j \langle T\pi_i, T\pi_j \rangle + \gamma_{K+1} \langle Tv_K, Tv_0 \rangle ,$$

hence

$$\|Tv_0\|^2 - \|v_0\|^2 \le \varepsilon \sum_{i,j} \gamma_i \gamma_j + \gamma_{K+1}(\|T\|^2 + 1).$$

Extending from α-net to the entire subspace.

Where we are.

We can decompose $v_0 := \sum_{k \le K} \gamma_k \pi_k + \gamma_{K+1} v_K$, where $\gamma_k \le (\frac{1}{10})^k$, $\pi_k \in \mathcal{M}$, and $\|v_K\| = 1$.
We have $\langle T\pi_i, T\pi_j \rangle = \langle \pi_i, \pi_j \rangle \pm \varepsilon$ for all π_i.
Want: $\|Tv_0\|_2^2 = 1 \pm 3\varepsilon$.

Proof

$$\|Tv_0\|^2 = \sum_{i,j \le K} \gamma_i \gamma_j \langle T\pi_i, T\pi_j \rangle + \gamma_{K+1} \langle Tv_K, Tv_0 \rangle ,$$

hence

$$\|Tv_0\|^2 - \|v_0\|^2 \le \varepsilon \sum_{i,j} \gamma_i \gamma_j + \gamma_{K+1}(\|T\|^2 + 1).$$

We have $\sum_{i,j} \gamma_i \gamma_j \le (\sum_i \frac{1}{10^i})^2 \le 2$.
We can chose $\gamma_{K+1} \ll \frac{\varepsilon}{\|T\|^2}$.

□

Oblivious subspace embedding

Fast Johnon-Lindenstrauss

Fast Johnson-Lindenstrauss

Theorem

Given δ, m, d, there is a distribution $\mathcal{D}$ over matrices in $\mathbb{R}^{m \times d}$, where $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, such that for any vector $v \in \mathbb{R}^d$, we have

$$\mathbb{P}_{T \sim \mathcal{D}}(\|Tv\|_2 \neq (1 \pm \varepsilon)\|v\|_2) < \delta,$$

and moreover matrix-vector multiplication Tv can be computed in $\mathcal{O}(d \log d + \frac{\log^2 \frac{1}{\delta}}{\varepsilon^2} \log \frac{d}{\delta})$ time.

Fast Johnson-Lindenstrauss

Theorem

Given δ, m, d, there is a distribution $\mathcal{D}$ over matrices in $\mathbb{R}^{m \times d}$, where $m = \mathcal{O}(\frac{\log \delta^{-1}}{\varepsilon^2})$, such that for any vector $v \in \mathbb{R}^d$, we have

$$\mathbb{P}_{T \sim \mathcal{D}}(\|Tv\|_2 \neq (1 \pm \varepsilon)\|v\|_2) < \delta,$$

and moreover matrix-vector multiplication Tv can be computed in $\mathcal{O}(d \log d + \frac{\log^2 \frac{1}{\delta}}{\varepsilon^2} \log \frac{d}{\delta})$ time.

For sketch-and-solve linear regression

The time of computing a sketch ΠA where $A \in \mathbb{R}^{d \times k}$ is $\mathcal{O}(dk \log d + \frac{k^3}{\varepsilon^2} + \frac{k^2 \log d}{\varepsilon^2})$.
After sketching we can solve linear regression $argmin_x \|\Pi Ax - \Pi b\|_2^2$ in time $\mathcal{O}(\frac{k^3}{\varepsilon^2})$.

Naive way for Fast Johnson-Lindenstrauss: coordinate subsampling

Fix a vector $v \in \mathbb{R}^d$. Pick a random coordinate $i \sim [d]$. Then $\mathbb{E}_{i\sim[d]} v_i^2 = \frac{\|v\|_2^2}{d}$.

Let $T \in \mathbb{R}^{m\times d}$ be a matrix where each row has exactly one $\frac{\sqrt{d}}{\sqrt{m}}$ in a uniformly random location (other entries in the row are 0). Then

$$\mathbb{E}_T \|Tv\|_2^2 = \|v\|_2^2.$$

Naive way for Fast Johnson-Lindenstrauss: coordinate subsampling

Fix a vector $v \in \mathbb{R}^d$. Pick a random coordinate $i \sim [d]$. Then $\mathbb{E}_{i\sim[d]} v_i^2 = \frac{\|v\|_2^2}{d}$.
Let $T \in \mathbb{R}^{m\times d}$ be a matrix where each row has exactly one $\frac{\sqrt{d}}{\sqrt{m}}$ in a uniformly random location (other entries in the row are 0). Then

$$\mathbb{E}_T \|Tv\|_2^2 = \|v\|_2^2.$$

In fact we can write

$$\|Tv\|_2^2 = \frac{1}{m}\sum_{i\le m} Y_i,$$

where all Y_i are i.i.d. random variables with $\mathbb{E}Y_i = \|v\|_2^2$.
Concretely $Y_i = v_{t_i}^2$, where $t_i \in [d]$ are i.i.d. uniform.

Naive way for Fast Johnson-Lindenstrauss: coordinate subsampling

Fix a vector $v \in \mathbb{R}^d$. Pick a random coordinate $i \sim [d]$. Then $\mathbb{E}_{i\sim[d]} v_i^2 = \frac{\|v\|_2^2}{d}$.
Let $T \in \mathbb{R}^{m\times d}$ be a matrix where each row has exactly one $\frac{\sqrt{d}}{\sqrt{m}}$ in a uniformly random location (other entries in the row are 0). Then

$$\mathbb{E}_T \|Tv\|_2^2 = \|v\|_2^2.$$

In fact we can write

$$\|Tv\|_2^2 = \frac{1}{m}\sum_{i\le m} Y_i,$$

where all Y_i are i.i.d. random variables with $\mathbb{E}Y_i = \|v\|_2^2$.
Concretely $Y_i = v_{t_i}^2$, where $t_i \in [d]$ are i.i.d. uniform.
Is it concentrated?

Coordinate subsampling: is it concentrated?

Without loss of generality, assume that $\|v\|_2^2 = 1$. Take $Y_i = v_{t_i}^2$, where $t_i \in [d]$ are i.i.d. uniform. When

$$\frac{1}{m} \sum_{i \leq m} Y_i$$

is concentrated around 1?

Coordinate subsampling: is it concentrated?

Without loss of generality, assume that $\|v\|_2^2 = 1$. Take $Y_i = v_{t_i}^2$, where $t_i \in [d]$ are i.i.d. uniform. When

$$\frac{1}{m} \sum_{i \leq m} Y_i$$

is concentrated around 1?
If $v = e_1$, we need $m \gg d$, if we hope to hit the non-zero coordiante at all.

Coordinate subsampling: is it concentrated?

Without loss of generality, assume that $\|v\|_2^2 = 1$. Take $Y_i = v_{t_i}^2$, where $t_i \in [d]$ are i.i.d. uniform. When

$$\frac{1}{m} \sum_{i \leq m} Y_i$$

is concentrated around 1?

If $v = e_1$, we need $m \gg d$, if we hope to hit the non-zero coordiante at all.

Observation:

Subsampling works if there are no large coordinates.

By Hoeffding bound, if all $v_i \leq \sqrt{\frac{B}{d}}$ it is enough to pick $m = \mathcal{O}(\frac{B^2}{\varepsilon^2} \log \frac{1}{\delta})$.

In this case we have $\mathbb{E} Y_i = 1$, and $Y_i \leq B$ with probability 1.

Making sure there are no large coordinates.

We want to find a random orthonormal matrix M, such that

- For any v, with probability at least $1 - \delta$, we have $\|Mv\|_\infty \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{n}}$.

Making sure there are no large coordinates.

We want to find a random orthonormal matrix M, such that

- For any v, with probability at least $1 - \delta$, we have $\|Mv\|_\infty \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{n}}$.
- We can compute Mv in time $\mathcal{O}(d \log d)$.

Making sure there are no large coordinates.

We want to find a random orthonormal matrix M, such that

- For any v, with probability at least $1 - \delta$, we have $\|Mv\|_\infty \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{n}}$.
- We can compute Mv in time $\mathcal{O}(d \log d)$.

Note

This $\frac{\sqrt{\log(n/\delta)}}{\sqrt{n}}$ makes sense. If the coordinates of Mv were gaussians with std. deviation $\frac{1}{\sqrt{n}}$ we would expect maximum to be that large.

Hadamard matrix

$$H = \frac{1}{\sqrt{d}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Hadamard matrix

$$H = \frac{1}{\sqrt{d}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

- We can multiply by H in time $\mathcal{O}(d \log d)$ using divide-and-conquer. (A bit like FFT).

Hadamard matrix

$$H = \frac{1}{\sqrt{d}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

- We can multiply by H in time $\mathcal{O}(d \log d)$ using divide-and-conquer. (A bit like FFT).
- All entries are $\pm\frac{1}{\sqrt{d}}$.

Hadamard matrix

$$H = \frac{1}{\sqrt{d}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

- We can multiply by H in time $\mathcal{O}(d \log d)$ using divide-and-conquer. (A bit like FFT).
- All entries are $\pm\frac{1}{\sqrt{d}}$.
- This matrix is orthonormal.

Hadamard matrix

$$H = \frac{1}{\sqrt{d}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

- We can multiply by H in time $\mathcal{O}(d \log d)$ using divide-and-conquer. (A bit like FFT).
- All entries are $\pm\frac{1}{\sqrt{d}}$.
- This matrix is orthonormal.

Flattening matrix M

Take $D \in \mathbb{R}^{d \times d}$ to be diagonal with random ± 1 entries. We pick $M = HD$. (Flip randomly sign for each coordinate, then apply H).

Flattening matrix: concentration

We want to show that with probability $1 - \delta$, we have $\|Mv\|_\infty \leq \frac{\sqrt{\log d/\delta}}{\sqrt{d}}$, where $M = HD$, Hadamard times ± 1 diagonal.

Flattening matrix: concentration

We want to show that with probability $1 - \delta$, we have $\|Mv\|_\infty \leq \frac{\sqrt{\log d/\delta}}{\sqrt{d}}$, where $M = HD$, Hadamard times ± 1 diagonal.

- Note that any coordinate $(Mv)_i = \sum_j H_{ij} D_{jj} v_j = \frac{1}{\sqrt{d}} \sum_j D_{jj} w_j$, where $w_j = v_j H_{ij}$.

Flattening matrix: concentration

We want to show that with probability $1 - \delta$, we have $\|Mv\|_\infty \leq \frac{\sqrt{\log d/\delta}}{\sqrt{d}}$, where $M = HD$, Hadamard times ± 1 diagonal.

- Note that any coordinate $(Mv)_i = \sum_j H_{ij} D_{jj} v_j = \frac{1}{\sqrt{d}} \sum_j D_{jj} w_j$, where $w_j = v_j H_{ij}$.
- We know that $D_{jj} w_j$ is $\frac{|v_j|}{\sqrt{d}}$-subgaussian, therefore every $(Mv)_i$ is $\frac{1}{\sqrt{d}}$-subgaussian (Hoeffding).

Flattening matrix: concentration

We want to show that with probability $1-\delta$, we have $\|Mv\|_\infty \le \frac{\sqrt{\log d/\delta}}{\sqrt{d}}$, where $M = HD$, Hadamard times ± 1 diagonal.

- Note that any coordinate $(Mv)_i = \sum_j H_{ij} D_{jj} v_j = \frac{1}{\sqrt{d}} \sum_j D_{jj} w_j$, where $w_j = v_j H_{ij}$.
- We know that $D_{jj} w_j$ is $\frac{|v_j|}{\sqrt{d}}$-subgaussian, therefore every $(Mv)_i$ is $\frac{1}{\sqrt{d}}$-subgaussian (Hoeffding).
- For given i, $\mathbb{P}((Mv)_i > \frac{\sqrt{\log(d/\delta)}}{\sqrt{d}}) < \frac{\delta}{d}$. By union bound $\|Mv\|_\infty \le \frac{\sqrt{\log(d/\delta)}}{\sqrt{d}}$ with probability $1-\delta$.

Flattening matrix: concentration

We want to show that with probability $1-\delta$, we have $\|Mv\|_\infty \le \frac{\sqrt{\log d/\delta}}{\sqrt{d}}$, where $M = HD$, Hadamard times ± 1 diagonal.

- Note that any coordinate $(Mv)_i = \sum_j H_{ij} D_{jj} v_j = \frac{1}{\sqrt{d}} \sum_j D_{jj} w_j$, where $w_j = v_j H_{ij}$.
- We know that $D_{jj} w_j$ is $\frac{|v_j|}{\sqrt{d}}$-subgaussian, therefore every $(Mv)_i$ is $\frac{1}{\sqrt{d}}$-subgaussian (Hoeffding).
- For given i, $\mathbb{P}((Mv)_i > \frac{\sqrt{\log(d/\delta)}}{\sqrt{d}}) < \frac{\delta}{d}$. By union bound $\|Mv\|_\infty \le \frac{\sqrt{\log(d/\delta)}}{\sqrt{d}}$ with probability $1-\delta$.
- Note that $(Mv)_i$ and $(Mv)_j$ are <u>not</u> independent. This is not a problem.

Putting it all together

Final construction of dimensionality reduction

- Apply D (flip randomly sign of each coordinate).

Putting it all together

Final construction of dimensionality reduction

- Apply D (flip randomly sign of each coordinate).
- Apply H (Hadamard matrix). Now each coordinate is at most $\frac{\sqrt{\log d/\delta}}{\sqrt{d}}$.

Putting it all together

Final construction of dimensionality reduction

- Apply D (flip randomly sign of each coordinate).
- Apply H (Hadamard matrix). Now each coordinate is at most $\frac{\sqrt{\log d/\delta}}{\sqrt{d}}$.
- Subsample coordinates. We reduce dimension to $\frac{\log \delta^{-1}}{\varepsilon^2} \log \frac{d}{\delta}$.

Putting it all together

Final construction of dimensionality reduction

- Apply D (flip randomly sign of each coordinate).
- Apply H (Hadamard matrix). Now each coordinate is at most $\frac{\sqrt{\log d/\delta}}{\sqrt{d}}$.
- Subsample coordinates. We reduce dimension to $\frac{\log \delta^{-1}}{\varepsilon^2} \log \frac{d}{\delta}$.
- Apply slow JL (say, with independent random signs) to further reduce dimension to the optimal one.

Putting it all together

Final construction of dimensionality reduction

- Apply D (flip randomly sign of each coordinate).
- Apply H (Hadamard matrix). Now each coordinate is at most $\frac{\sqrt{\log d/\delta}}{\sqrt{d}}$.
- Subsample coordinates. We reduce dimension to $\frac{\log \delta^{-1}}{\varepsilon^2} \log \frac{d}{\delta}$.
- Apply slow JL (say, with independent random signs) to further reduce dimension to the optimal one.

Composition of linear maps is a linear map. □