JL Lemma, Dimensionality Reduction, and Subspace Embeddings

Lecture 11

September 29, 2020

F_2 estimation in turnstile setting

```
AMS-ℓ2-Estimate:
    Let Y_1, Y_2, ..., Y_n be {−1, +1} random variables that are
        4-wise independent
    z ← 0
    While (stream is not empty) do
        a_j = (i_j, Δ_j) is current update
        z ← z + Δ_j Y_{i_j}
    endWhile
    Output z^2
```

Claim: Output estimates $||x||_2^2$ where x is the vector at end of stream of updates.

Analysis

$Z = \sum_{i=1}^{n} x_i Y_i$ and output is Z^2

$$Z^2 = \sum_i x_i^2 Y_i^2 + 2 \sum_{i \neq j} x_i x_j Y_i Y_j$$

and hence

$$\mathbf{E}\left[Z^2\right] = \sum_i x_i^2 = ||x||_2^2.$$

One can show that $Var(Z^2) \leq 2(\mathbf{E}\left[Z^2\right])^2$.

Linear Sketching View

Recall that we take average of independent estimators and take median to reduce error. Can we view all this as a sketch?

AMS-ℓ_2-Sketch:
    $k = c \log(1/\delta)/\epsilon^2$
    `Let` M `be a` $\ell \times n$ `matrix with entries in` $\{-1, 1\}$ `s.t`
        `(i) rows are independent and`
        `(ii) in each row entries are` 4`-wise independent`
    z `is a` $\ell \times 1$ `vector initialized to` 0
    `While (stream is not empty) do`
        $a_j = (i_j, \Delta_j)$ `is current update`
        $z \leftarrow z + \Delta_j M e_{i_j}$
    `endWhile`
    `Output vector` z `as sketch.`

M is compactly represented via k hash functions, one per row, independently chosen from 4-wise independent hash family.

Geometric Interpretation

Given vector $x \in \mathbb{R}^n$ let M the random map $z = Mx$ has the following features

- $\mathbf{E}[z_i] = 0$ and $\mathbf{E}\left[z_i^2\right] = \|x\|_2^2$ for each $1 \leq i \leq k$ where k is number of rows of M
- Thus each z_i^2 is an estimate of length of x in Euclidean norm
- When $k = \Theta(\frac{1}{\epsilon^2} \log(1/\delta))$ one can obtain an $(1 \pm \epsilon)$ estimate of $\|x\|_2$ by averaging and median ideas

Thus we are able to compress x into k-dimensional vector z such that z contains information to estimate $\|x\|_2$ accurately

Geometric Interpretation

Given vector $x \in \mathbb{R}^n$ let M the random map $z = Mx$ has the following features

- $\mathbf{E}[z_i] = 0$ and $\mathbf{E}\left[z_i^2\right] = \|x\|_2^2$ for each $1 \leq i \leq k$ where k is number of rows of M
- Thus each z_i^2 is an estimate of length of x in Euclidean norm
- When $k = \Theta(\frac{1}{\epsilon^2} \log(1/\delta))$ one can obtain an $(1 \pm \epsilon)$ estimate of $\|x\|_2$ by averaging and median ideas

Thus we are able to compress x into k-dimensional vector z such that z contains information to estimate $\|x\|_2$ accurately

Question: Do we need median trick? Will averaging do?

Distributional JL Lemma

Lemma (Distributional JL Lemma)

Fix vector $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ *and let* $\mathbf{\Pi} \in \mathbb{R}^{\mathbf{k \times d}}$ *matrix where each entry* $\mathbf{\Pi_{ij}}$ *is chosen independently according to standard normal distribution* $\mathcal{N}(\mathbf{0}, \mathbf{1})$ *distribution. If* $\mathbf{k} = \mathbf{\Omega}(\frac{1}{\epsilon^2} \log(1/\delta))$*, then with probability* $(\mathbf{1} - \delta)$

$$\|\frac{1}{\sqrt{k}}\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2.$$

Can choose entries from $\{-1, 1\}$ as well.
Note: unlike ℓ_2 estimation, entries of $\mathbf{\Pi}$ are independent.

Letting $z = \frac{1}{\sqrt{k}}\Pi x$ we have projected x from d dimensions to $k = O(\frac{1}{\epsilon^2} \log(1/\delta))$ dimensions while preserving length to within $(1 \pm \epsilon)$-factor.

Dimensionality reduction

Theorem (Metric JL Lemma)

Let $v_1, v_2, \ldots, v_n$ be any n points/vectors in $\mathbb{R}^d$. For any $\epsilon \in (0, 1/2)$, there is linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ where $k \leq 8 \ln n/\epsilon^2$ such that for all $1 \leq i < j \leq n$,

$$(1 - \epsilon)||v_i - v_j||_2 \leq ||f(v_i) - f(v_j)||_2 \leq ||v_i - v_j||_2.$$

Moreover f can be obtained in randomized polynomial-time.

Linear map f is simply given by random matrix Π: $f(v) = \Pi v$.

Dimensionality reduction

Theorem (Metric JL Lemma)

Let $v_1, v_2, \ldots, v_n$ be any n points/vectors in $\mathbb{R}^d$. For any $\epsilon \in (0, 1/2)$, there is linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ where $k \leq 8 \ln n/\epsilon^2$ such that for all $1 \leq i < j \leq n$,

$$(1-\epsilon)||v_i - v_j||_2 \leq ||f(v_i) - f(v_j)||_2 \leq ||v_i - v_j||_2.$$

Moreover f can be obtained in randomized polynomial-time.

Linear map f is simply given by random matrix Π: $f(v) = \Pi v$.

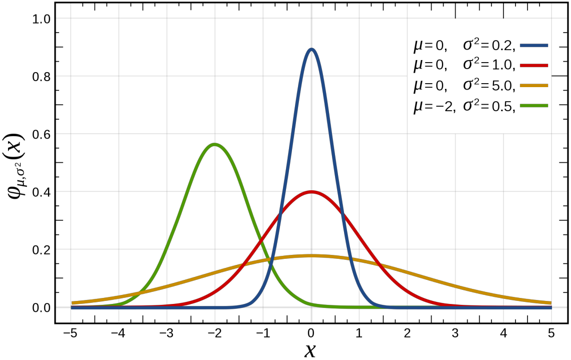
Proof.

Apply DJL with $\delta = 1/n^2$ and apply union bound to $\binom{n}{2}$ vectors $(v_i - v_j)$, $i \neq j$. □

Normal Distribution

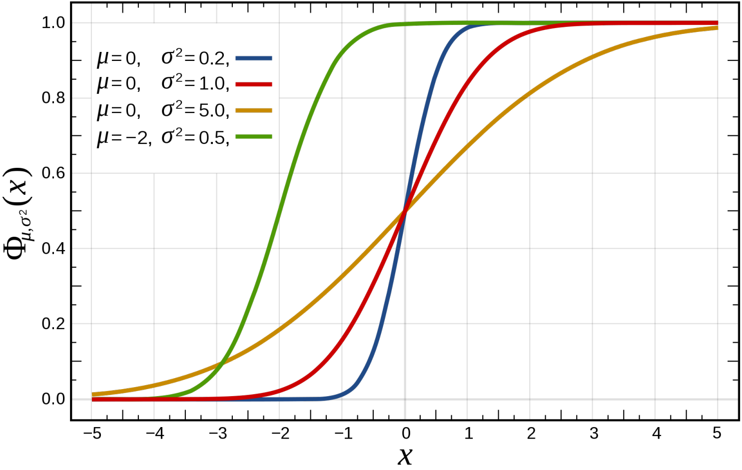
Density function: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Standard normal: $\mathcal{N}(0, 1)$ is when $\mu = 0, \sigma = 1$



Normal Distribution

Cumulative density function for standard normal:
$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{t} e^{-t^2/2}$ (no closed form)



Sum of independent Normally distributed variables

Lemma

Let X *and* Y *be independent random variables. Suppose* $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *and* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. *Let* $Z = X + Y$. *Then* $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Sum of independent Normally distributed variables

Lemma

Let X *and* Y *be independent random variables. Suppose* $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *and* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. *Let* $Z = X + Y$. *Then* $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Corollary

Let X *and* Y *be independent random variables. Suppose* $X \sim \mathcal{N}(0, 1)$ *and* $Y \sim \mathcal{N}(0, 1)$. *Let* $Z = aX + bY$. *Then* $Z \sim \mathcal{N}(0, a^2 + b^2)$.

Sum of independent Normally distributed variables

Lemma

Let X *and* Y *be independent random variables. Suppose* $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ *and* $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$*. Let* $Z = X + Y$*. Then* $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$*.*

Corollary

Let X *and* Y *be independent random variables. Suppose* $X \sim \mathcal{N}(0, 1)$ *and* $Y \sim \mathcal{N}(0, 1)$*. Let* $Z = aX + bY$*. Then* $Z \sim \mathcal{N}(0, a^2 + b^2)$*.*

Normal distribution is a *stable* distributions: adding two independent random variables within the same class gives a distribution inside the class. Others exist and useful in F_p estimation for $p \in (0, 2)$.

Concentration of sum of squares of normally distributed variables

$\chi^2(k)$ distribution: distribution of sum of k independent standard normally distributed variables
$Y = \sum_{i=1}^{k} Z_i$ where each $Z_i \simeq \mathcal{N}(0, 1)$.

Concentration of sum of squares of normally distributed variables

$\chi^2(k)$ distribution: distribution of sum of k independent standard normally distributed variables

$Y = \sum_{i=1}^{k} Z_i$ where each $Z_i \simeq \mathcal{N}(0,1)$.

$\mathbf{E}\left[Z_i^2\right] = 1$ hence $\mathbf{E}[Y] = k$.

Concentration of sum of squares of normally distributed variables

$\chi^2(k)$ distribution: distribution of sum of k independent standard normally distributed variables
$Y = \sum_{i=1}^{k} Z_i$ where each $Z_i \simeq \mathcal{N}(0,1)$.
$\mathbf{E}\left[Z_i^2\right] = 1$ hence $\mathbf{E}[Y] = k$.

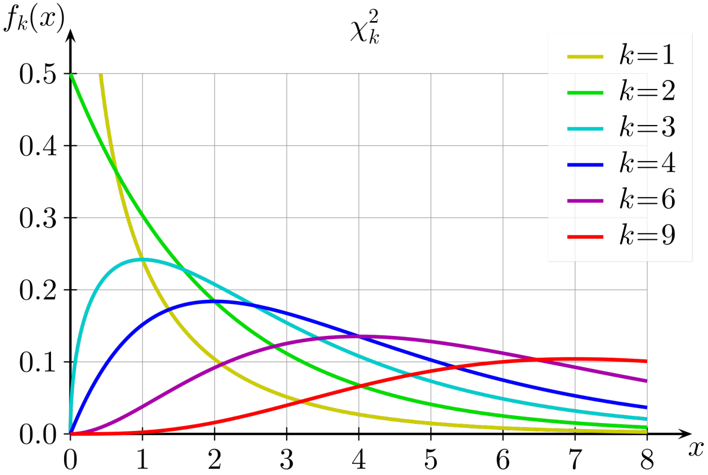
Lemma

Let $Z_1, Z_2, \ldots, Z_k$ be independent $\mathcal{N}(0,1)$ random variables and let $Y = \sum_i Z_i^2$. Then, for $\epsilon \in (0, 1/2)$, there is a constant c such that,

$$\Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - 2e^{c\epsilon^2 k}.$$

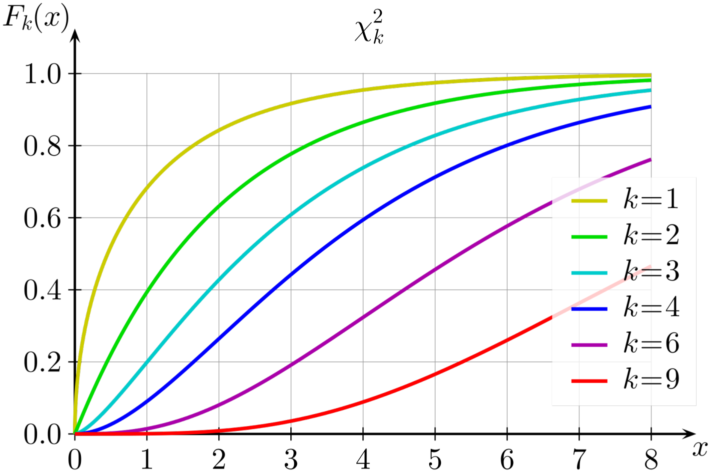
χ^2 distribution

Density function



χ^2 distribution

Cumulative density function



Concentration of sum of squares of normally distributed variables

$\chi^2(k)$ distribution: distribution of sum of k independent standard normally distributed variables

Lemma

Let $Z_1, Z_2, \ldots, Z_k$ be independent $\mathcal{N}(0,1)$ random variables and let $Y = \sum_i Z_i^2$. Then, for $\epsilon \in (0, 1/2)$, there is a constant c such that, $\Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - 2e^{c\epsilon^2 k}$.

Recall Chernoff-Hoeffding bound for *bounded* independent non-negative random variables. Z_i^2 is not bounded, however Chernoff-Hoeffding bounds extend to sums of random variables with exponentially decaying tails.

Proof of DJL Lemma

Without loss of generality assume $\|x\|_2 = 1$ (unit vector)

$Z_i = \sum_{j=1}^{n} \Pi_{ij} x_i$

- $Z_i \sim \mathcal{N}(0, 1)$

Proof of DJL Lemma

Without loss of generality assume $\|x\|_2 = 1$ (unit vector)

$Z_i = \sum_{j=1}^{n} \Pi_{ij} x_i$

- $Z_i \sim \mathcal{N}(0, 1)$
- Let $Y = \sum_{i=1}^{k} Z_i^2$. Y's distribution is χ^2 since $Z_1, \ldots, Z_k$ are iid.

Proof of DJL Lemma

Without loss of generality assume $\|x\|_2 = 1$ (unit vector)

$Z_i = \sum_{j=1}^{n} \Pi_{ij} x_i$

- $Z_i \sim \mathcal{N}(0, 1)$
- Let $Y = \sum_{i=1}^{k} Z_i^2$. Y's distribution is χ^2 since $Z_1, \ldots, Z_k$ are iid.
- Hence $\Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - 2e^{c\epsilon^2 k}$

Proof of DJL Lemma

Without loss of generality assume $\|x\|_2 = 1$ (unit vector)

$Z_i = \sum_{j=1}^{n} \Pi_{ij} x_i$

- $Z_i \sim \mathcal{N}(0, 1)$
- Let $Y = \sum_{i=1}^{k} Z_i^2$. Y's distribution is χ^2 since $Z_1, \ldots, Z_k$ are iid.
- Hence $\Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - 2e^{c\epsilon^2 k}$
- Since $k = \Omega(\frac{1}{\epsilon^2} \log(1/\delta))$ we have $\Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - \delta$

Proof of DJL Lemma

Without loss of generality assume $\|x\|_2 = 1$ (unit vector)

$Z_i = \sum_{j=1}^{n} \Pi_{ij} x_i$

- $Z_i \sim \mathcal{N}(0, 1)$
- Let $Y = \sum_{i=1}^{k} Z_i^2$. Y's distribution is χ^2 since $Z_1, \ldots, Z_k$ are iid.
- Hence $\Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - 2e^{c\epsilon^2 k}$
- Since $k = \Omega(\frac{1}{\epsilon^2} \log(1/\delta))$ we have $\Pr[(1-\epsilon)^2 k \leq Y \leq (1+\epsilon)^2 k] \geq 1 - \delta$
- Therefore $\|z\|_2 = \sqrt{Y/k}$ has the property that with probability $(1-\delta)$, $\|z\|_2 = (1 \pm \epsilon)\|x\|_2$.

JL lower bounds

Question: Are the bounds achieved by the lemmas tight or can we do better? How about non-linear maps?

Essentially optimal modulo constant factors for worst-case point sets.

Fast JL and Sparse JL

Projection matrix Π is dense and hence Πx takes $\Theta(kd)$ time.

Question: Can we find Π to improve time bound?

Two scenarios: x is dense and x is sparse

Fast JL and Sparse JL

Projection matrix Π is dense and hence Πx takes $\Theta(kd)$ time.

Question: Can we find Π to improve time bound?

Two scenarios: x is dense and x is sparse

Known results:

- Choose Π_{ij} to be $\{-1, 0, 1\}$ with probability $1/6, 1/3, 1/6$. Also works. Roughly $1/3$ entries are 0
- Fast JL: Choose Π in a dependent way to ensure Πx can be computed in $O(d \log d + k^2)$ time. For dense x.
- Sparse JL: Choose Π such that each column is s-sparse. The best known is $s = O(\frac{1}{\epsilon} \log(1/\delta))$. Helps in sparse x.

Part I

(Oblivious) Subspace Embeddings

Subspace Embedding

Question: Suppose we have linear subspace E of $\mathbb{R}^d$ of dimension ℓ. Can we find a projection $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ such that for *every* $x \in E$, $\|\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$?

Subspace Embedding

Question: Suppose we have linear subspace E of $\mathbb{R}^d$ of dimension ℓ. Can we find a projection $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ such that for *every* $x \in E$, $\|\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$?

- Not possible if $k < \ell$. Why?

Subspace Embedding

Question: Suppose we have linear subspace E of $\mathbb{R}^d$ of dimension ℓ. Can we find a projection $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ such that for *every* $x \in E$, $\|\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$?

- Not possible if $k < \ell$. Why? Π maps E to a lower dimension. Implies some non-zero vector $x \in E$ mapped to $\mathbf{0}$

Subspace Embedding

Question: Suppose we have linear subspace E of $\mathbb{R}^d$ of dimension ℓ. Can we find a projection $\Pi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ such that for *every* $x \in E$, $\|\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$?

- Not possible if $k < \ell$. Why? Π maps E to a lower dimension. Implies some non-zero vector $x \in E$ mapped to $\mathbf{0}$
- Possible if $k = \ell$. Why?

Subspace Embedding

Question: Suppose we have linear subspace $\boldsymbol{E}$ of $\mathbb{R}^{\boldsymbol{d}}$ of dimension ℓ. Can we find a projection $\boldsymbol{\Pi} : \mathbb{R}^{\boldsymbol{d}} \rightarrow \mathbb{R}^{\boldsymbol{k}}$ such that for *every* $\boldsymbol{x} \in \boldsymbol{E}$, $\|\boldsymbol{\Pi x}\|_2 = (1 \pm \epsilon)\|x\|_2$?

- Not possible if $\boldsymbol{k} < \ell$. Why? $\boldsymbol{\Pi}$ maps $\boldsymbol{E}$ to a lower dimension. Implies some non-zero vector $\boldsymbol{x} \in \boldsymbol{E}$ mapped to $\mathbf{0}$
- Possible if $\boldsymbol{k} = \ell$. Why? Pick $\boldsymbol{\Pi}$ to be an orthonormal basis for $\boldsymbol{E}$.

Subspace Embedding

Question: Suppose we have linear subspace E of $\mathbb{R}^d$ of dimension ℓ. Can we find a projection $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ such that for *every* $x \in E$, $\|\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$?

- Not possible if $k < \ell$. Why? Π maps E to a lower dimension. Implies some non-zero vector $x \in E$ mapped to $\mathbf{0}$
- Possible if $k = \ell$. Why? Pick Π to be an orthonormal basis for E. **Disadvantage:** This requires knowing E and computing orthonormal basis which is slow.

Subspace Embedding

Question: Suppose we have linear subspace E of $\mathbb{R}^d$ of dimension ℓ. Can we find a projection $\Pi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ such that for *every* $x \in E$, $\|\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$?

- Not possible if $k < \ell$. Why? Π maps E to a lower dimension. Implies some non-zero vector $x \in E$ mapped to $\mathbf{0}$
- Possible if $k = \ell$. Why? Pick Π to be an orthonormal basis for E. **Disadvantage:** This requires knowing E and computing orthonormal basis which is slow.

What we really want: *Oblivious* subspace embedding ala JL based on random projections

Oblivious Supspace Embedding

Theorem

Suppose E is a linear subspace of $\mathbb{R}^n$ of dimension d. Let Π be a DJL matrix $\Pi \in \mathbb{R}^{k \times n}$ with $k = O(\frac{d}{\epsilon^2} \log(1/\delta))$ rows. Then with probability $(1 - \delta)$ for every $x \in E$,

$$\|\frac{1}{\sqrt{k}}\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2.$$

In other words JL Lemma extends from one dimension to arbitrary number of dimensions in a graceful way.

Proof Idea

How do we prove that Π works for *all* $x \in E$ which is an infinite set?

Several proofs but one useful argument that is often a starting hammer is the "net argument"

- Choose a large but finite set of vectors T carefully (the net)
- Prove that Π preserves lengths of vectors in T (via naive union bound)
- Argue that *any* vector $x \in E$ is sufficiently close to a vector in T and hence Π also preserves length of x

Net argument

Sufficient to focus on unit vectors in E. Why?

Net argument

Sufficient to focus on unit vectors in E. Why?

Also assume wlog and ease of notation that E is the subspace formed by the first d coordinates in standard basis.

Net argument

Sufficient to focus on unit vectors in E. Why?

Also assume wlog and ease of notation that E is the subspace formed by the first d coordinates in standard basis.

Claim: There is a net T of size $e^{O(d)}$ such that preserving lengths of vectors in T suffices.

Net argument

Sufficient to focus on unit vectors in E. Why?

Also assume wlog and ease of notation that E is the subspace formed by the first d coordinates in standard basis.

Claim: There is a net T of size $e^{O(d)}$ such that preserving lengths of vectors in T suffices.

Assuming claim: use DJL with $k = O(\frac{d}{\epsilon^2}\log(1/\delta))$ and union bound to show that all vectors in T are preserved in length up to $(1 \pm \epsilon)$ factor.

Net argument

Sufficient to focus on unit vectors in E.

Also assume wlog and ease of notation that E is the subspace formed by the first d coordinates in standard basis.

A weaker net:

- Consider the box $[-1, 1]^d$ and make a grid with side length ϵ/d
- Number of grid vertices is $(2d/\epsilon)^d$
- Sufficient to take T to be the grid vertices
- Gives a weaker bound of $O(\frac{1}{\epsilon^2} d \log(d/\epsilon))$ dimensions
- A more careful net argument gives tight bound

Net argument: analysis

Fix any $x \in E$ such that $\|x\|_2 = 1$ (unit vector)
There is grid point y such that $\|y\|_2 \leq 1$ and x is close to y
Let $z = x - y$. We have $|z_i| \leq \epsilon/d$ for $1 \leq i \leq i \leq d$ and $z_i = 0$ for $i > d$

Net argument: analysis

Fix any $x \in E$ such that $\|x\|_2 = 1$ (unit vector)
There is grid point y such that $\|y\|_2 \leq 1$ and x is close to y
Let $z = x - y$. We have $|z_i| \leq \epsilon/d$ for $1 \leq i \leq i \leq d$ and $z_i = 0$ for $i > d$

$$\begin{aligned} \|\Pi x\| = \|\Pi y + \Pi z\| &\leq \|\Pi y\| + \|\Pi z\| \\ &\leq (1+\epsilon) + (1+\epsilon)\sum_{i=1}^{d} |z_i| \\ &\leq (1+\epsilon) + \epsilon(1+\epsilon) \leq 1 + 3\epsilon \end{aligned}$$

Net argument: analysis

Fix any $x \in E$ such that $\|x\|_2 = 1$ (unit vector)
There is grid point y such that $\|y\|_2 \leq 1$ and x is close to y
Let $z = x - y$. We have $|z_i| \leq \epsilon/d$ for $1 \leq i \leq i \leq d$ and $z_i = 0$ for $i > d$

$$
\begin{aligned}
\|\Pi x\| = \|\Pi y + \Pi z\| &\leq \|\Pi y\| + \|\Pi z\| \\
&\leq (1+\epsilon) + (1+\epsilon)\sum_{i=1}^{d} |z_i| \\
&\leq (1+\epsilon) + \epsilon(1+\epsilon) \leq 1 + 3\epsilon
\end{aligned}
$$

Similarly $\|\Pi x\| \geq 1 - O(\epsilon)$.

Application of Subspace Embeddings

Faster algorithms for approximate

- matrix multiplication
- regression
- SVD

Basic idea: Want to perform operations on matrix A with n data columns (say in large dimension $\mathbb{R}^h$) with small effective rank d. Want to reduce to a matrix of size roughly $\mathbb{R}^{d \times d}$ by spending time proportional to $\text{nnz}(A)$.

Later in course.