15-859 Algorithms for Big Data

David Woodruff

Massive data sets

Examples

- Internet traffic logs
- Financial data
- etc.

Algorithms

- Want nearly linear time or less
- Usually at the cost of a randomized approximation

Regression analysis

Regression

- Statistical method to study dependencies between variables in the presence of noise.

Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

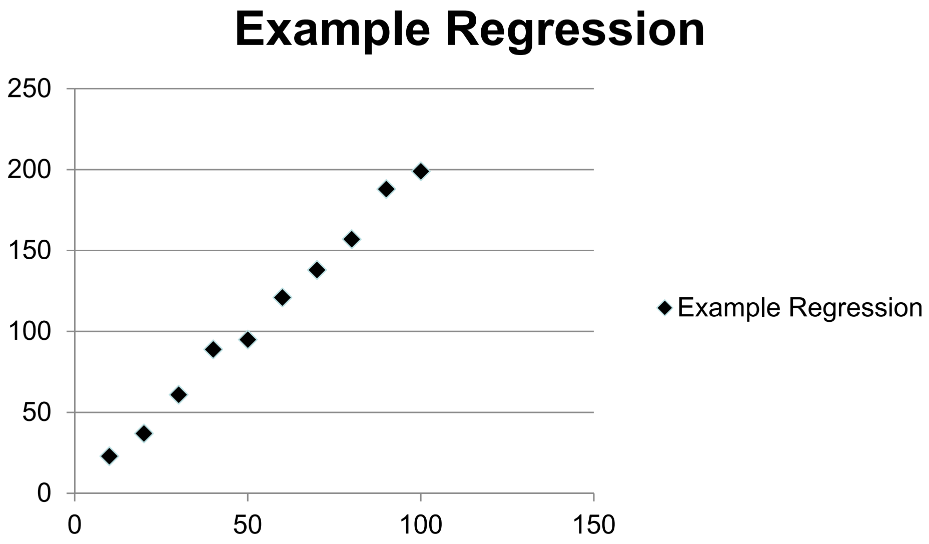
Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

Example

- Ohm's law V = R · I



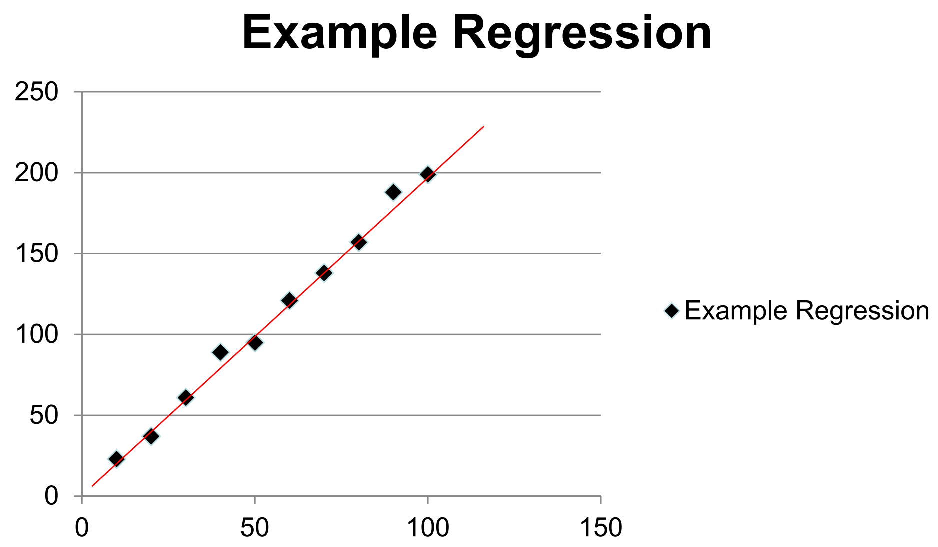
Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

Example

- Ohm's law V = R · I
- Find linear function that best fits the data



Regression analysis

Linear Regression

- Statistical method to study linear dependencies between variables in the presence of noise.

Standard Setting

- One measured variable b
- A set of predictor variables $a_1 ,\dots, a_d$
- Assumption:

$$b = x_0 + a_1 x_1 + \dots + a_d x_d + \varepsilon$$

- ε is assumed to be noise and the x_i are model parameters we want to learn
- Can assume $x_0 = 0$
- Now consider n observations of b

Regression analysis

Matrix form

Input: n×d-matrix A and a vector $b=(b_1,\ldots, b_n)$
n is the number of observations; d is the number of predictor variables

Output: x^* so that Ax^* and b are close

- Consider the over-constrained case, when n ¨ d

Regression analysis

Least Squares Method

- Find x* that minimizes $|Ax-b|_2^2 = \Sigma\ (b_i – <A_{i*}, x>)^2$
- A_{i*} is i-th row of A
- Certain desirable statistical properties

Regression analysis

Geometry of regression

- We want to find an x that minimizes $|Ax-b|_2$
- The product Ax can be written as

$$A_{*1}x_1 + A_{*2}x_2 + \dots + A_{*d}x_d$$

where A_{*i} is the i-th column of A

- This is a linear d-dimensional subspace
- The problem is equivalent to computing the point of the column space of A nearest to b in l_2-norm

Regression analysis

Solving least squares regression via the normal equations

- How to find the solution x to $\min_x |Ax-b|_2$?

- Equivalent problem: $\min_x |Ax-b\ |_2^2$
  - Write b = Ax' + b', where b' orthogonal to columns of A
  - Cost is $|A(x-x')|_2^2 + |b'|_2^2$ by Pythagorean theorem
  - Optimal solution x if and only if $A^T(Ax-b) = A^T(Ax-Ax') = 0$
  - Normal Equation: $A^TAx = A^Tb$ for any optimal x
  - $x = (A^TA)^{-1} A^T b$

- If the columns of A are not linearly independent, the Moore-Penrose pseudoinverse gives a minimum norm solution x

Moore-Penrose Pseudoinverse

Singular Value Decomposition (SVD)

Any matrix A = U · Σ · V^{T}

- U has orthonormal columns
- Σ is diagonal with non-increasing non-negative entries down the diagonal
- V^{T} has orthonormal rows

- Pseudoinverse A^{-} = V $\Sigma^{-1}\mathrm{U}^{\mathrm{T}}$
  - Where Σ^{-1} is a diagonal matrix with i-th diagonal entry equal to 1/Σ_{ii} if Σ_{ii} > 0 and is 0 otherwise

- $\min_x |Ax\text{-}b|_2^2$ not unique when columns of A are linearly independent, but x = A^{-}b has minimum norm

Moore-Penrose Pseudoinverse

- Any optimal solution x has the form $\mathrm{A}^{-}\mathrm{b} + \left(\mathrm{I} - \mathrm{V}'\mathrm{V}'^{\mathrm{T}}\right)\mathrm{z}$, where $(\mathrm{V}')^{\mathrm{T}}$ corresponds to the rows i of V^{T} for which $\Sigma_{i,i} > 0$

- Why?

- Because A$\left(\mathrm{I} - \mathrm{V}'\mathrm{V}'^{\mathrm{T}}\right)\mathrm{z} = 0$, so $\mathrm{A}^{-}\mathrm{b} + \left(\mathrm{I} - \mathrm{V}'\mathrm{V}'^{\mathrm{T}}\right)\mathrm{z}$ is a solution. This is a (d-rank(A))-dimensional affine space so it spans all optimal solutions

- Since $\mathrm{A}^{-}\mathrm{b}$ is in column span of V', by the Pythagorean theorem, $|\mathrm{A}^{-}\mathrm{b} + \left(\mathrm{I} - \mathrm{V}'\mathrm{V}'^{\mathrm{T}}\right)\mathrm{z}|_2^2 = |\mathrm{A}^{-}\mathrm{b}|_2^2 + |(\mathrm{I} - \mathrm{V}'\mathrm{V}'^{\mathrm{T}})\mathrm{z}|_2^2 \geq |\mathrm{A}^{-}\mathrm{b}|_2^2$

Time Complexity

Solving least squares regression via the normal equations

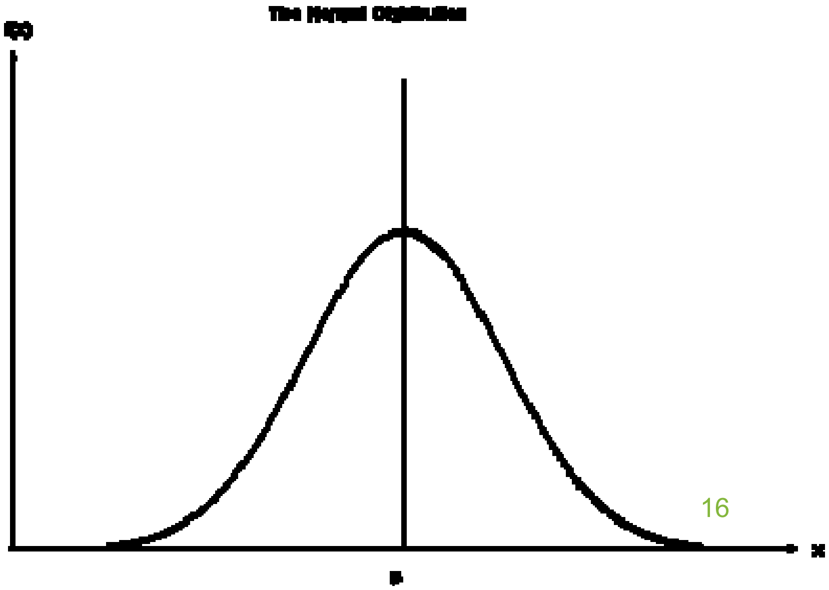
- Need to compute x = $A^{-}b$
- Naively this takes nd^2 time
- Can do $\mathrm{nd}^{1.376}$ using fast matrix multiplication
- But we want much better running time!

Sketching to solve least squares regression

- How to find an approximate solution x to $\min_x |Ax\text{-}b|_2$?
- Goal: output x‘ for which $|Ax‘\text{-}b|_2$ ‰ $(1+\varepsilon) \min_x |Ax\text{-}b|_2$ with high probability
- Draw S from a k x n random family of matrices, for a value k << n
- Compute S*A and S*b
- Output the solution x‘ to $\min_{x‘} |(SA)x\text{-}(Sb)|_2$
  - $x' = (SA)^{-}Sb$

How to choose the right sketching matrix S?

- Recall: output the solution x‘ to $\min_{x‘} |(SA)x-(Sb)|_2$
- Lots of matrices work
- S is d/ε^2 x n matrix of i.i.d. Normal random variables
- To see why this works, we introduce the notion of a subspace embedding



Subspace Embeddings

- Let $k = O(d/\varepsilon^2)$
- Let S be a k x n matrix of i.i.d. normal N(0,1/k) random variables
- For any fixed d-dimensional subspace, i.e., the column space of an n x d matrix A
  - W.h.p., for all x in R^d, $|SAx|_2 = (1\pm\varepsilon)|Ax|_2$
- Entire column space of A is preserved

Why is this true?

Subspace Embeddings – A Proof

- Want to show $|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$ for all x

- Can assume columns of A are orthonormal, since we prove this for all x

- Claim: SA is a k x d matrix of i.i.d. N(0,1/k) random variables

  - First property: for two independent random variables X and Y, with X drawn from $N(0,a^2)$ and Y drawn from $N(0,b^2)$, we have X+Y is drawn from $N(0, a^2 + b^2)$

X+Y is drawn from N(0, $a^2 + b^2$)

- Probability density function f_z of Z = X+Y is convolution of probability density functions f_X and f_Y

- $f_Z(z) = \int f_X(z-y) f_Y(y)\,dy$

- $f_X(x) = \frac{1}{a(2\pi)^{.5}} e^{-x^2/2a^2}$, $f_Y(y) = \frac{1}{b(2\pi)^{.5}} e^{-y^2/2b^2}$

- $f_Z(z) = \int \frac{1}{a(2\pi)^{.5}} e^{-(z-y)^2/2a^2} \frac{1}{b(2\pi)^{.5}} e^{-y^2/2b^2} dy$

$$= \frac{1}{(2\pi)^{.5}(a^2+b^2)^{.5}} e^{-z^2/2\left(a^2+b^2\right)} \int \frac{\left(a^2+b^2\right)^{.5}}{(2\pi)^{.5}ab} e^{-\frac{\left(y-\frac{b^2z}{a^2+b^2}\right)^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}} dy$$

X+Y is drawn from N(0, $a^2 + b^2$)

Calculation: $e^{-\frac{(z-y)^2}{2a^2}-\frac{y^2}{2b^2}} \;=\; e^{-\frac{z^2}{2(a^2+b^2)}-\frac{\left(y-\frac{b^2z}{a^2+b^2}\right)^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}}$

Density of Gaussian distribution: $\int \frac{(a^2+b^2)^{.5}}{(2\pi)^{.5}ab} e^{-\frac{\left(y-\frac{b^2z}{a^2+b^2}\right)^2}{2\left(\frac{(ab)^2}{a^2+b^2}\right)}}$ dy = 1

Rotational Invariance

- Second property: if u, v are vectors with <u, v> = 0, then <g,u> and <g,v> are independent, where g is a vector of i.i.d. N(0,1/k) random variables
- Why?
- If g is an n-dimensional vector of i.i.d. N(0,1) random variables, and R is a fixed matrix, then the probability density function of Rg is

$$f \quad (x) = \frac{1}{\det(\mathrm{R} \quad \mathrm{R}^{\mathrm{T}})(2\pi)^{n/2}} e^{-\frac{x^T\left(\mathrm{R} \quad \mathrm{R}^{\mathrm{T}}\right)^{-1} x}{2}}$$

  - RR^{T} is the covariance matrix
  - For a rotation matrix R, the distribution of Rg and of g are the same

Orthogonal Implies Independent

- Want to show: if u, v are vectors with <u, v> = 0, then <g,u> and <g,v> are independent, where g is a vector of i.i.d. N(0,1/k) random variables

- Choose a rotation R which sends u to αe_1, and sends v to βe_2

- $< g, u > = < Rg, Ru > = < h, \alpha e_1 > = \alpha h_1$
- $< g, v > = < Rg, Rv > = < h, \beta e_2 > = \beta h_2$

  where h is a vector of i.i.d. N(0, 1/k) random variables

- Then h_1 and h_2 are independent by definition

Where were we?

- Claim: SA is a k x d matrix of i.i.d. N(0,1/k) random variables

- Proof: The rows of SA are independent

  - Each row is: $< \mathrm{g}, \mathrm{A}_1 >, < \mathrm{g}, \mathrm{A}_2 >, \ldots, < \mathrm{g}, \mathrm{A}_\mathrm{d} >$

  - First property implies the entries in each row are N(0,1/k) since the columns A_i have unit norm

  - Since the columns A_i are orthonormal, the entries in a row are independent by our second property

Back to Subspace Embeddings

- Want to show $|SAx|_2 = (1\pm\varepsilon)|Ax|_2$ for all x
- Can assume columns of A are orthonormal
- Can also assume x is a unit vector
- SA is a k x d matrix of i.i.d. N(0,1/k) random variables

- Consider any fixed unit vector $x \in R^d$
- $|SAx|_2^2 = \sum_{i\in[k]} < g_i, x >^2$, where g_i is i-th row of SA
- Each $< g_i, x >^2$ is distributed as $N\left(0, \frac{1}{k}\right)^2$
- E[$< g_i, x >^2$] = 1/k, and so E[$|SAx|_2^2$] = 1

How concentrated is $|SAx|_2^2$ about its expectation?

Johnson-Lindenstrauss Theorem

- Suppose $\mathrm{h}_1, \ldots, \mathrm{h}_\mathrm{k}$ are i.i.d. N(0,1) random variables
- Then G = $\sum_\mathrm{i} \mathrm{h}_\mathrm{i}^2$ is a χ^2-random variable
- Apply known tail bounds to G:
  - (Upper) $\mathrm{Pr}[\mathrm{G} \geq \mathrm{k} + 2(\mathrm{kx})^{.5} + 2\mathrm{x}] \leq \mathrm{e}^{-\mathrm{x}}$
  - (Lower) $\mathrm{Pr}[\mathrm{G} \leq \mathrm{k} - 2(\mathrm{kx})^{.5}] \leq \mathrm{e}^{-\mathrm{x}}$
- If $\mathrm{x} = \frac{\epsilon^2 \mathrm{k}}{16}$, then $\mathrm{Pr}[\mathrm{G} \in \mathrm{k}(1 \pm \epsilon)] \geq 1 - 2\mathrm{e}^{-\epsilon^2 \mathrm{k}/16}$
- If $\mathrm{k} = \Theta(\epsilon^{-2} \log(\frac{1}{\delta}))$, this probability is 1-δ

- $\mathrm{Pr}[|\mathrm{SAx}|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(\mathrm{d})}$

This only holds for a fixed x, how to argue for all x?

Net for Sphere

- Consider the sphere $S^{\mathrm{d}-1}$
- Subset N is a γ-net if for all $\mathrm{x} \in S^{\mathrm{d}-1}$, there is a $\mathrm{y} \in \mathrm{N}$, such that $|\mathrm{x} - \mathrm{y}|_2 \leq \gamma$
- Greedy construction of N
  - While there is a point $\mathrm{x} \in S^{\mathrm{d}-1}$ of distance larger than γ from every point in N, include x in N
- The ball of radius γ/2 around every point in N is contained in the ball of radius 1+ γ/2 around 0^{d}
- Further, all such balls are disjoint
- Ratio of volume of d-dimensional ball of radius 1+ $\gamma/2$ to d-dimensional sphere of radius γ is $(1+\gamma/2)^{\mathrm{d}}/(\gamma/2)^{\mathrm{d}}$, so $|\mathrm{N}| \leq (1+\gamma/2)^{\mathrm{d}}/(\gamma/2)^{\mathrm{d}}$

Net for Subspace

- Let M = {Ax | x in N}, so $|\mathrm{M}| \leq (1+\gamma/2)^{\mathrm{d}}/(\gamma/2)^{\mathrm{d}}$

- Claim: For every x in $S^{\mathrm{d}-1}$, there is a y in M for which $|\mathrm{Ax}-\mathrm{y}|_2 \leq \gamma$

- Proof: Let x’ in $S^{\mathrm{d}-1}$ be such that $|\mathrm{x}-\mathrm{x}'|_2 \leq \gamma$

  Then $|\mathrm{Ax}-\mathrm{Ax}'|_2 = |\mathrm{x}-\mathrm{x}'|_2 \leq \gamma$, using that the columns of A are orthonormal. Set y = Ax’

Net Argument

- For a fixed unit x, $\Pr[|SAx|_2^2 \in (1 \pm \epsilon)] \geq 1 - 2^{-\Theta(d)}$
- For a fixed pair of unit x, x', $|SAx|_2^2$, $|SAx'|_2^2, |SA(x - x')|_2^2$ are preserved up to a $1 \pm \epsilon$ factor with prob. $1 - 2^{-\Theta(d)}$
- $|SA(x - x')|_2^2 = |SAx|_2^2 + |SAx'|_2^2 - 2 < SAx, SAx' >$
- $|A(x - x')|_2^2 = |Ax|_2^2 + |Ax'|_2^2 - 2 < Ax, Ax' >$
  - So $\Pr[< Ax, Ax' > = < SAx, SAx' > \ \pm O(\epsilon)] = 1 \ - 2^{-\Theta(d)}$
- Choose a ½-net M = {Ax | x in N} of size 5^d
- By a union bound, for all pairs y, y' in M,

$$< y, y' > = < Sy, Sy' > \ \pm O(\epsilon)$$

- Condition on this event
- By linearity, if this holds for y, y' in M, for $\alpha y, \beta y'$ we have

$$< \alpha y, \beta y' > = \alpha\beta < Sy, Sy' > \ \pm O(\epsilon \, \alpha\beta)$$

Finishing the Net Argument

- Let y = Ax for an arbitrary $x \in S^{d-1}$
- Let $y_1 \in M$ be such that $|y - y_1|_2 \leq \gamma$
- Let α be such that $|\alpha(y - y_1)|_2 = 1$
  - $\alpha \geq 1/\gamma$ (could be infinite)
- Let $y_2' \in M$ be such that $|\alpha(y - y_1) - y_2{}'|_2 \leq \gamma$
- Then $\left|y - y_1 - \frac{y_2{}'}{\alpha}\right|_2 \leq \frac{\gamma}{\alpha} \leq \gamma^2$
- Set $y_2 = \frac{y_2'}{\alpha}$. Repeat, obtaining $y_1, y_2, y_3, \ldots$ such that for all integers i,

  $$|y - y_1 - y_2 - \ \ldots - y_i|_2 \leq \gamma^i$$
- Implies $|y_i|_2 \leq \gamma^{i-1} + \gamma^i \leq 2\gamma^{i-1}$

Finishing the Net Argument

- Have $y_1, y_2, y_3, \ldots$ such that y $= \sum_i y_i$ and $|y_i|_2 \le 2\gamma^{i-1}$

- $|Sy|_2^2 = |S \sum_i y_i|_2^2$

  = $\sum_i |Sy_i|_2^2 + 2 \sum_{i,j} < Sy_i, Sy_j >$

  = $\sum_i |y_i|_2^2 + 2 \ \sum_{i,j} \ < y_i, y_j > \pm O(\epsilon) \sum_{i,j} |y_i|_2 |y_j|_2$

  = $| \sum_i y_i |_2^2 \ \pm O(\epsilon)$

  = $|y|_2^2 \pm O(\epsilon)$

  = $1 \pm O(\epsilon)$

- Since this held for an arbitrary y = Ax for unit x, by linearity it follows that for all x, $|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$

Back to Regression

- We showed that S is a subspace embedding, that is, simultaneously for all x,

$$|SAx|_2 = (1 \pm \varepsilon)|Ax|_2$$

What does this have to do with regression?

Subspace Embeddings for Regression

- Want x so that $|Ax-b|_2$ ‰ $(1+\varepsilon)\ \mathrm{min}_y\ |Ay-b|_2$
- Consider subspace L spanned by columns of A together with b
- Then for all y in L, $|Sy|_2 = (1\pm \varepsilon)\ |y|_2$
- Hence, $|S(Ax-b)|_2 = (1\pm \varepsilon)\ |Ax-b|_2$ for all x
- Solve $\mathrm{argmin}_y\ |(SA)y - (Sb)|_2$
- Given SA, Sb, can solve in poly(d/ε) time

Only problem is computing SA takes $O(nd^2)$ time

How to choose the right sketching matrix S? [S]

- S is a Subsampled Randomized Hadamard Transform
  - S = P*H*D
  - D is a diagonal matrix with +1, -1 on diagonals
  - H is the Hadamard matrix: $\mathrm{H}_{\mathrm{i,j}} = (-1/\mathrm{n}^{.5})^{<\mathrm{i,j}>}$
  - P just chooses a random (small) subset of rows of H*D
  - S*A can be computed in O(nd log n) time

Why does it work?

Why does this work?

- We can again assume columns of A are orthonormal
- It suffices to show $|\mathrm{SAx}|_2^2 = |\mathrm{PHDAx}|_2^2 = 1 \pm \epsilon$ for all x
- HD is a rotation matrix, so $|\mathrm{HDAx}|_2^2 = |\mathrm{Ax}|_2^2 = 1$ for any x
  - Notation: let y = Ax
- Flattening Lemma: For any fixed y,

$$\Pr\left[|\mathrm{HDy}|_\infty \geq \mathrm{C} \quad \frac{\log^{.5}(\frac{\mathrm{nd}}{\delta})}{\mathrm{n}^{.5}}\right] \leq \frac{\delta}{2\mathrm{d}}$$

Proving the Flattening Lemma

- Flattening Lemma: Pr $[|\mathrm{HDy}|_\infty \geq \mathrm{C} \; \frac{\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}] \leq \frac{\delta}{2\mathrm{d}}$
- Let C > 0 be a constant. We will show for a fixed i in [n],

$$\Pr\,[|(\mathrm{HDy})_{\mathrm{i}}| \quad \geq \mathrm{C} \; \frac{\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}] \leq \frac{\delta}{2\mathrm{nd}}$$

- If we show this, we can apply a union bound over all i
- $|(\mathrm{HDy})_{\mathrm{i}}| = \sum_{\mathrm{j}} \mathrm{H}_{\mathrm{i,j}}\mathrm{D}_{\mathrm{j,j}}\mathrm{y}_{\mathrm{j}}$
- (Azuma-Hoeffding) For independent zero-mean random variables Z_{j}:

$\Pr\left[|\sum_{\mathrm{j}} \mathrm{Z}_{\mathrm{j}}| > \mathrm{t}\right] \leq 2\mathrm{e}^{-(\frac{\mathrm{t}^2}{2\sum_{\mathrm{j}} \beta_{\mathrm{j}}^2})}$, where $|\mathrm{Z}_{\mathrm{j}}| \leq \beta_{\mathrm{j}}$ with probability 1

  - $\mathrm{Z}_{\mathrm{j}} = \mathrm{H}_{\mathrm{i,j}}\mathrm{D}_{\mathrm{j,j}}\mathrm{y}_{\mathrm{j}}$ has 0 mean
  - $|\mathrm{Z}_{\mathrm{j}}| \leq \frac{|\mathrm{y}_{\mathrm{j}}|}{\mathrm{n}^{.5}} = \beta_{\mathrm{j}}$ with probability 1
  - $\sum_{\mathrm{j}} \beta_{\mathrm{j}}^2 = \frac{1}{\mathrm{n}}$

- $\Pr\left[|\sum_{\mathrm{j}} \mathrm{Z}_{\mathrm{j}}| > \frac{\mathrm{C}\log^{.5}\left(\frac{\mathrm{nd}}{\delta}\right)}{\mathrm{n}^{.5}}\right] \leq 2\mathrm{e}^{-\frac{\mathrm{C}^2\log\left(\frac{\mathrm{nd}}{\delta}\right)}{2}} \leq \frac{\delta}{2\mathrm{nd}}$

Consequence of the Flattening Lemma

- Recall columns of A are orthonormal
- HDA has orthonormal columns
- Flattening Lemma implies $|\mathrm{HDAe_i}|_\infty \leq \mathrm{C} \quad \frac{\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}$ with probability $1 - \frac{\delta}{2\mathrm{d}}$ for a fixed $\mathrm{i} \in [\mathrm{d}]$
- With probability $1 - \frac{\delta}{2}$, $\left|\mathrm{e_j HDAe_i}\right| \quad \leq \mathrm{C} \quad \frac{\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}$ for all i,j
- Given this, $\left|\mathrm{e_j HDA}\right|_2 \leq \mathrm{C} \quad \frac{\mathrm{d}^{.5}\log^{.5} \mathrm{nd}/\delta}{\mathrm{n}^{.5}}$ for all j

(Can be optimized further)

Matrix Chernoff Bound

- Let $X_1, \ldots, X_s$ be independent copies of a symmetric random matrix $X \in R^{dxd}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $\left|E[X^T X]\right|_2 \leq \sigma^2$. Let W = $\frac{1}{s}\sum_{i \in [s]} X_i$. For any $\epsilon > 0$,

$$\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-s\epsilon^2/(\sigma^2 + \frac{\gamma\epsilon}{3})}$$

(here $|W|_2 = \sup |Wx|_2/|x|_2$)

- Let V = HDA, and recall V has orthonormal columns

- Suppose P in the S = PHD definition samples s rows uniformly with replacement. If row i is sampled in the j-th sample, $P_{j,i} = \frac{\sqrt{n}}{\sqrt{s}}$, and is 0 otherwise

- Let Y_i be the i-th sampled row of V = HDA

- Let $X_i = I_d - n \cdot Y_i^T Y_i$
  - $E[X_i] = I_d - n \cdot \sum_j \left(\frac{1}{n}\right) V_j^T V_j = I_d - V^T V = 0^{d \times d}$
  - $|X_i|_2 \leq |I_d|_2 + n \cdot \max \left|e_j HDA\right|_2^2 = 1 + n \cdot C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = \Theta\left(d \log\left(\frac{nd}{\delta}\right)\right)$