Robust Regression

Method of least absolute deviation (l_1 -regression)

- Find x^* that minimizes $|Ax-b|_1 = \Sigma\ |b_i – <A_{i*}, x>|$
- Cost is less sensitive to outliers than least squares
- Can solve via linear programming

Solving l_1 -regression via Linear Programming

- Minimize $(1,\ldots,1) \cdot (\alpha^+ + \alpha^-)$
- Subject to:

$$A x + \alpha^+ - \alpha^- = b$$
$$\alpha^+, \alpha^- \geq 0$$

- Generic linear programming gives poly(nd) time
- Want much faster time using sketching!

Well-Conditioned Bases

- For an n x d matrix A, can choose an n x d matrix U with orthonormal columns for which A = UW, and $|\mathrm{Ux}|_2 = |\mathrm{x}|_2$ for all x

- Can we find a U for which A = UW and $|\mathrm{Ux}|_1 \approx |\mathrm{x}|_1$ for all x?

- Let A = QW where Q has full column rank, and define $|\mathrm{z}|_{\mathrm{Q},1} = |\mathrm{Qz}|_1$
  - $|\mathrm{z}|_{\mathrm{Q},1}$ is a norm

- Let C = $\{\mathrm{z} \in \mathrm{R}^{\mathrm{d}} : |\mathrm{z}|_{\mathrm{Q},1} \leq 1\}$ be the unit ball of $|.|_{\mathrm{Q},1}$

- C is a convex set which is symmetric about the origin
  - Lowner-John Theorem: can find an ellipsoid E such that: $\mathrm{E} \subseteq \mathrm{C} \subseteq \sqrt{\mathrm{d}}\mathrm{E}$, where E = $\{\mathrm{z} \in \mathrm{R}^{\mathrm{d}} : \mathrm{z}^{\mathrm{T}}\mathrm{Fz} \leq 1\}$
  - $\left(\mathrm{z}^{\mathrm{T}}\mathrm{Fz}\right)^{.5} \leq |\mathrm{z}|_{\mathrm{Q},1} \leq \sqrt{\mathrm{d}}\left(\mathrm{z}^{\mathrm{T}}\mathrm{Fz}\right)^{.5}$
  - $\mathrm{F} = \mathrm{GG}^{\mathrm{T}}$ since F defines an ellipsoid

- Define $\mathrm{U} = \mathrm{QG}^{-1}$

Well-Conditioned Bases

- Recall $\mathrm{U} = \mathrm{QG}^{-1}$ where

  $\left(\mathrm{z}^{\mathrm{T}}\mathrm{Fz}\right)^{.5} \leq |\mathrm{z}|_{\mathrm{Q},1} \leq \sqrt{\mathrm{d}}\left(\mathrm{z}^{\mathrm{T}}\mathrm{Fz}\right)^{.5}$ and $\mathrm{F} = \mathrm{GG}^{\mathrm{T}}$

- $|\mathrm{Ux}|_1 = |\mathrm{QG}^{-1}\mathrm{x}|_1 = |\mathrm{Qz}|_1 = |\mathrm{z}|_{\mathrm{Q},1}$ where $\mathrm{z} = \mathrm{G}^{-1}\mathrm{x}$

- $\mathrm{z}^{\mathrm{T}}\mathrm{Fz} \quad = \left(\mathrm{x}^{\mathrm{T}}\left(\mathrm{G}^{-1}\right)^{\mathrm{T}}\mathrm{G}^{\mathrm{T}}\mathrm{G}\,(\mathrm{G}^{-1})\mathrm{x}\right) = \mathrm{x}^{\mathrm{T}}\mathrm{x} = |\mathrm{x}|_2^2$

- So $|\mathrm{x}|_2 \leq |\mathrm{Ux}|_1 \leq \sqrt{\mathrm{d}}|\mathrm{x}|_2$

- So $\frac{|\mathrm{x}|_1}{\sqrt{\mathrm{d}}} \leq |\mathrm{x}|_2 \leq \left|\mathrm{Ux}\right|_1 \leq \sqrt{\mathrm{d}}\left|\mathrm{x}\right|_2 \leq \sqrt{\mathrm{d}}|\mathrm{x}|_1$

Net for ℓ_1 – Ball

- Consider the unit ℓ_1-ball B = $\{\mathrm{x} \ \in \mathrm{R}^{\mathrm{d}} : |\mathrm{x}|_1 = 1\}$
- Subset N is a γ-net if for all $\mathrm{x} \in \mathrm{B}$, there is a $\mathrm{y} \in \mathrm{N}$, such that $|\mathrm{x} - \mathrm{y}|_1 \leq \gamma$
- Greedy construction of N
  - While there is a point $\mathrm{x} \in \mathrm{B}$ of distance larger than γ from every point in N, include x in N
- The ℓ_1-ball of radius γ/2 around every point in N is contained in the ℓ_1-ball of radius 1+ γ/2 around 0^{d}
- Further, all such ball are disjoint
- Ratio of volume of d-dimensional similar polytopes of radius 1+ γ/2 to radius $\gamma/2$ 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 $\ell_1 - \mathrm{Subspace}$

- Let A = UW for a well-conditioned basis U
  - $|x|_1 \leq |Ux|_1 \leq d|x|_1$ for all x

- Let N be a (γ/d) −net for the unit ℓ_1-ball B

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

- Claim: For every x in B, there is a y in M for which $|Ux - y|_1 \leq \gamma$

- Proof: Let x’ in B be such that $|x - x'|_1 \leq \gamma/d$

  Then $|Ux - Ux'|_1 \leq d|x - x'|_1 \leq \gamma$, using the well-conditioned basis property. Set y = Ux’

- $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$

Rough Algorithm Overview

$\min_{x \text{ in } R^d} |Ax-b|_1 = \min_{x \text{ in } R^d} |Ux - b'|_1$

Sample poly(d/ε) rows of U∘b' proportional to their l_1-norm.

Compute poly(d)-approximation

Compute well-conditioned basis

Find x' such that $|Ax'-b|_1 \leq \text{poly}(d) \min_{x \text{ in } R^d}$ Let b' = b-Ax' be the residual

Find a basis A=UW so that for all x in R^d, $|x|_1/\text{poly}(d) \leq |Ux|_1 \leq \text{poly}(d)\, |x|_1$

Takes nnz(A) t

Now generic linear programming is efficient

Solve l_1-regression on the sample, obtaining vector x, and output x

Will focus on showing how to quickly compute

1. A poly(d)-approximation

2. A well-conditioned basis

Sketching Theorem

Theorem

- There is a probability space over (d log d) × n matrices R such that for any n×d matrix A, with probability at least 99/100 we have for all x:

$$|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1$$

Embedding

- is linear
- is independent of A
- preserves lengths of an infinite number of vectors

Application of Sketching Theorem

Computing a d(log d)-approximation

- Compute RA and Rb
- Solve $x' = \text{argmin}_x |RAx-Rb|_1$
- Main theorem applied to A◦b implies x' is a d log d – approximation
- RA, Rb have d log d rows, so can solve l_1-regression efficiently

Application of Sketching Theorem

Computing a well-conditioned basis

1. Compute RA
2. Compute W so that RAW is orthonormal (in the l_2-sense)
3. Output U = AW

U = AW is well-conditioned because

$|AWx|_1 \leq |RAWx|_1 \leq (d \log d)^{1/2} |RAWx|_2 = (d \log d)^{1/2} |x|_2 \leq (d \log d)^{1/2} |x|_1$

and

$|AWx|_1 \geq |RAWx|_1/(d \log d) \gtrsim |RAWx|_2/(d \log d) = |x|_2/(d \log d) \geq |x|_1 /(d^{3/2} \log d)$

Sketching Theorem

Theorem:

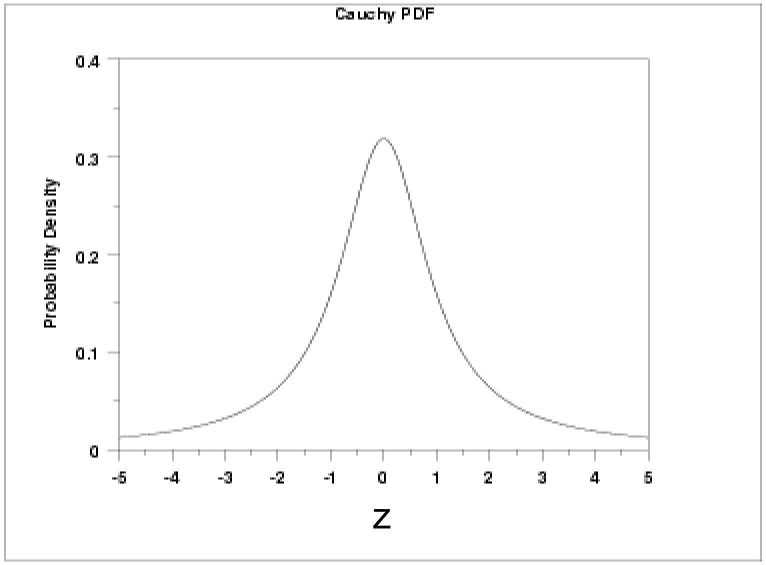
- There is a probability space over (d log d) × n matrices R such that for any n×d matrix A, with probability at least 99/100 we have for all x:

$$|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1$$

A dense R that works:

The entries of R are i.i.d. Cauchy random variables, scaled by 1/(d log d)

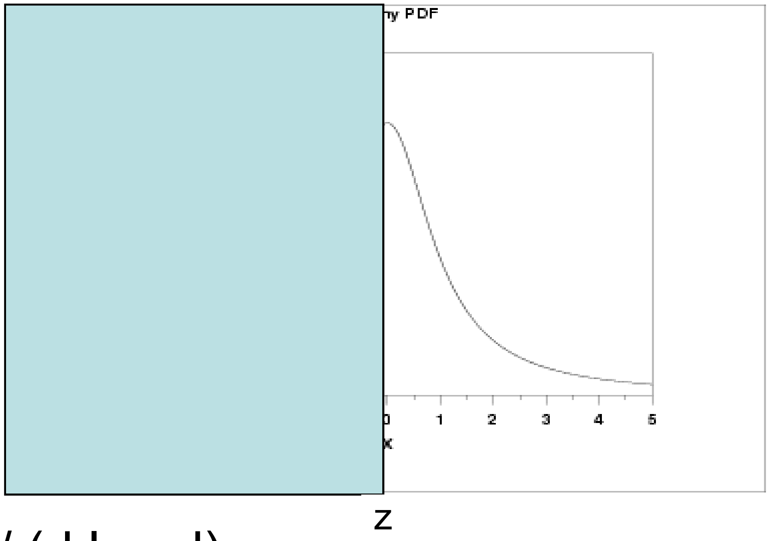
Cauchy Random Variables



- pdf(z) = $1/(\pi(1+z^2))$ for z in (-4 , 4)
- Undefined expectation and infinite variance
- 1-stable:
  - If $z_1, z_2, \ldots, z_n$ are i.i.d. Cauchy, then for a 5 R^n,

    $a_1 \cdot z_1 + a_2 \cdot z_2 + \ldots + a_n \cdot z_n \;\hat{\sim}\; |a|_1 \cdot z$, where z is Cauchy
- Can generate as the ratio of two standard normal random variables

Proof of Sketching Theorem



- By 1-stability,
  - For all rows r of R,
    - $\langle r, Ax \rangle = |Ax|_1 \cdot Z / (d \log d)$,
      where Z is a Cauchy
- $RAx =$ϒ$(|Ax|_1 \cdot Z_1, \ldots, |Ax|_1 \cdot Z_{d \log d}) / (d \log d)$,
  where $Z_1, \ldots, Z_{d \log d}$ are i.i.d. Cauchy
- $|RAx|_1 = |Ax|_1 \sum_j |Z_j| / (d \log d)$
  - The $|Z_j|$ are half-Cauchy
- $\sum_j |Z_j| = \Omega(d \log d)$ with probability 1-exp(-d log d) by Chernoff
- But the $|Z_j|$ are heavy-tailed…

Proof of Sketching Theorem

- $\sum_j |Z_j|$ is heavy-tailed, so $|RAx|_1 = |Ax|_1 \sum_j |Z_j| / (d \log d)$ may be large

- Each $|Z_j|$ has c.d.f. asymptotic to 1-Θ(1/z) for z in [0, 4)

- There *exists* a well-conditioned basis of A
  - Suppose w.l.o.g. the basis vectors are $A_{*1}, \ldots, A_{*d}$

- $|RA_{*i}|_1 \hat{=} |A_{*i}|_1 \ f\sum_j |Z_{i,j}| / (d \log d)$

- Let $E_{i,j}$ be the event that $|Z_{i,j}| \le d^3$
  - Define $Z'_{i,j} = |Z_{i,j}|$ if $|Z_{i,j}| \le d^3$, and $Z'_{i,j} = d^3$ otherwise
  - $E\left[Z_{i,j} \mid E_{i,j}\right] = E\left[Z'_{i,j} \mid E_{i,j}\right] = O(\log d)$

- Let E be the event that for all i,j, $E_{i,j}$ occurs
  - $\Pr[E] \ge 1 - \frac{\log d}{d}$
- What is $E\left[Z'_{i,j} \mid E\right]$?

Proof of Sketching Theorem

- What is $\mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \mid \mathrm{E}\right]$?

- $$\begin{aligned}\mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \middle| \mathrm{E}_{\mathrm{i,j}}\right] &= \mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \middle| \mathrm{E}_{\mathrm{i,j}}, \mathrm{E}\right] \Pr[\mathrm{E} \mid \mathrm{E}_{\mathrm{i,j}}] + \mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \middle| \mathrm{E}_{\mathrm{i,j}}, \neg\mathrm{E}\right] \Pr[\neg\mathrm{E} \mid \mathrm{E}_{\mathrm{i,j}}] \\ &\geq \mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \middle| \mathrm{E}_{\mathrm{i,j}}, \mathrm{E}\right] \Pr[\mathrm{E} \mid \mathrm{E}_{\mathrm{i,j}}] \\ &= \mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \middle| \mathrm{E}\right] \cdot \left(\frac{\Pr\left[\mathrm{E}_{\mathrm{i,j}} \middle| \mathrm{E}\right] \Pr[\mathrm{E}]}{\Pr\left[\mathrm{E}_{\mathrm{i,j}}\right]}\right) \\ &\geq \mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \middle| \mathrm{E}\right] \cdot \left(1 - \frac{\log \mathrm{d}}{\mathrm{d}}\right)\end{aligned}$$

- So, $\mathrm{E}\left[\mathrm{Z}'_{\mathrm{i,j}} \middle| \mathrm{E}\right] = \mathrm{O}(\log \mathrm{d})$
- $|RA_{*i}|_1 \hat{=} |A_{*i}|_1 \cdot \sum_{i,j} |Z_{i,j}| / (d \log d)$
- With constant probability, $\sum_i |RA_{*i}|_1 = O(\log d) \sum_i |A_{*i}|_1$

Proof of Sketching Theorem

- With constant probability, $\sum_i |RA_{*i}|_1 = O(\log d) \sum_i |A_{*i}|_1$

- Recall $A_{*1}, \ldots, A_{*d}$ is a well-conditioned basis, and we showed the existence of such a basis earlier

- We will use the Auerbach basis which always exists:
  - For all x, $|x|_4 \leq |Ax|_1$
  - $\sum_i |A_{*i}|_1 = d$

- $\sum_i |RA_{*i}|_1 = O(d \log d)$

- For all x, $|RAx|_1 \leq \sum_i |RA_{*i} x_i| \leq |x|_4 \sum_i |RA_{*i}|_1$

  $= |x|_4 \, O(d \log d)$

  $= O(d \log d) \, |Ax|_1$

Where are we?

- Suffices to show for all x with $|x|_1 = 1$, that $|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1$
- We know
  - (1) there is a γ-net M, with $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$, of the set {Ax such that $|x|_1 = 1$}
  - (2) for any fixed x, $|RAx|_1 \geq |Ax|_1$ with probability $1 - \exp(-d \log d)$
  - (3) for all x, $|RAx|_1 = O(d \log d)|Ax|_1$

- Set $\gamma = 1/(d^3 \log d)$ so $|M| \leq d^{O(d)}$
  - By a union bound, for all y in M, $|Ry|_1 \geq |y|_1$

- Let x with $|x|_1 = 1$ be arbitrary. Let y in M satisfy $|Ax - y|_1 \leq \gamma = 1/(d^3 \log d)$

- $$\begin{aligned} |RAx|_1 &\geq |Ry|_1 - |R(Ax - y)|_1 \\ &\geq |y|_1 - O(d \log d)|Ax - y|_1 \\ &\geq |y|_1 - O(d \log d)\gamma \\ &\geq |y|_1 - O\left(\frac{1}{d^2}\right) \\ &\geq |y|_1/2 \quad \text{(why?)} \end{aligned}$$

Sketching to solve l_1-regression [CW, MM]

- Most expensive operation is computing R*A where R is the matrix of i.i.d. Cauchy random variables
- All other operations are in the “smaller space”
- Can speed this up by choosing R as follows:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} C_1 & & & & \\ & C_2 & & & \\ & & C_3 & & \\ & & & \ldots & \\ & & & & C_n \end{bmatrix}$$

Further sketching improvements [WZ]

- Can show you need a fewer number of sampled rows in later steps if instead choose R as follows

- Instead of diagonal of Cauchy random variables, choose diagonal of reciprocals of exponential random variables

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1/E_1 & & & & \\ & 1/E_2 & & & \\ & & 1/E_3 & & \\ & & & \ldots & \\ & & & & 1/E_n \end{bmatrix}$$

Turnstile Streaming Model

- Underlying n-dimensional vector x initialized to 0^n

- Long stream of updates $x_i \leftarrow x_i + \Delta_i$ for Δ_i in $\{-1,1\}$

- At end of the stream, x is promised to be in the set $\{-M, -M+1, \ldots, M-1, M\}^n$ for some bound $M \leq \text{poly}(n)$

- Output an approximation to f(x) whp

- Goal: use as little space (in bits) as possible
  - Massive data: stock transactions, weather data, genomes

Example Problem: Norms

- Suppose you want $|x|_p^p = \Sigma_{i=1}^n |x_i|^p$

- Want Z for which $(1-\varepsilon) |x|_p^p \leq Z \leq (1+\varepsilon) |x|_p^p$ with probability > 9/10

Example Problem: Euclidean Norm

- Want Z for which (1-ε) $|x|_2^2 \leq Z \leq$ (1+ε) $|x|_2^2$
- Sample a random CountSketch matrix S with $1/\epsilon^2$ rows
- Can store S efficiently using limited independence
- If $x_i \leftarrow x_i + \Delta_i$ in the stream, then $\mathrm{Sx} \leftarrow \mathrm{Sx} + \Delta_\mathrm{i} S_{*,\mathrm{i}}$
- At end of stream, output $|\mathrm{Sx}|_2^2$
- With probability at least 9/10, $|\mathrm{Sx}|_2^2 = (1 \pm \epsilon)|\mathrm{x}|_2^2$
- Space complexity is $1/\epsilon^2$ words, each word is O(log n) bits

Example Problem: 1-Norm

- Want Z for which (1-Ɛ) $|x|_1 \leq Z \leq$ (1+Ɛ) $|x|_1$
- Sample a random Cauchy matrix S?
- Can store S with $\frac{1}{\epsilon}$ words of space [Kane, Nelson, W]
- If $x_i \leftarrow x_i + \Delta_i$ in the stream, then $Sx \leftarrow Sx + \Delta_i S_{*,i}$
- Space complexity is $1/\epsilon^2$ words, each word is O(log n) bits ?
- At end of stream, output $|Sx|_1$?
- *Cauchy random variables have no concentration…*

1-Norm Estimator

- Probability density function f(x) of |C| for a Cauchy random variable C is $\mathrm{f(x)} = \frac{2}{\pi(1+x^2)}$

- Cumulative distribution function F(z):

$$\mathrm{F(z)} = \int_0^z \mathrm{f(x)dx} = \frac{2}{\pi}\mathrm{arctan(z)}$$

- Since $\tan(\pi/4) = 1$, F(1) = ½, so median(|C|) = 1

- If you take $\mathrm{r} = \frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}$ independent samples $X_1, \ldots, X_r$ from F, and $X = \mathrm{median}_i X_i$, then F(X) in [1/2-ϵ, 1/2+ ϵ] with large probability

- $F^{-1}(X) = \tan\left(\frac{X\pi}{2}\right) \in [1 - 4\epsilon, 1 + 4\epsilon]$