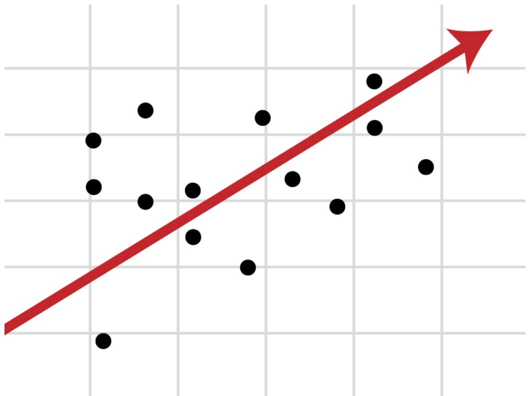
Subspace Embeddings and ℓ_p-Regression Using Exponential Random Variables

David P. Woodruff and Qin Zhang

IBM Research Almaden



COLT'13,
June 12, 2013

Subspace embeddings

- Subspace embeddings:

  A distribution over linear maps $\Pi : \mathbb{R}^n \to \mathbb{R}^m$, s.t., for any fixed d-dimensional subspace of $\mathbb{R}^n$ (denoted by M), w. pr. 0.99

  $$\|Mx\|_p \leq \|\Pi Mx\|_q \leq \kappa \|Mx\|_p$$

  simultaneously for all vectors $x \in \mathbb{R}^d$.

  Goal: to minimize

  1. m: the dimension of the subspace embedding.
  2. κ: the distortion of the embedding.
  3. t: the time to compute ΠM.

Subspace embeddings

- **Subspace embeddings:**

  A distribution over linear maps $\Pi : \mathbb{R}^n \to \mathbb{R}^m$, s.t., for any fixed d-dimensional subspace of $\mathbb{R}^n$ (denoted by M), w. pr. 0.99

  $$\|Mx\|_p \le \|\Pi M x\|_q \le \kappa \|Mx\|_p$$

  simultaneously for all vectors $x \in \mathbb{R}^d$.

  Goal: to minimize

  1. m: the dimension of the subspace embedding.
  2. κ: the distortion of the embedding.
  3. t: the time to compute ΠM.

- **Applications:**

  ℓ_p-regression (next slide), low-rank approximation, quantile regression, ...

All matter: embedding time, dimension and distortion

Using ℓ_p subspace embedding (SE) to solve ℓ_p regression:
$\min_{x \in \mathbb{R}^d} \left\| \bar{M}x - b \right\|_p$

For convenience, let $\bar{M} \in \mathbb{R}^{n \times (d-1)}$, and let $M = [\bar{M}, -b] \in \mathbb{R}^{n \times d}$. $n \gg d$.

Let Π be a SE with dimension m, distortion κ and embedding time t

All matter: embedding time, dimension and distortion

Using ℓ_p subspace embedding (SE) to solve ℓ_p regression:
$\min_{x \in \mathbb{R}^d} \left\| \bar{M}x - b \right\|_p$

For convenience, let $\bar{M} \in \mathbb{R}^{n \times (d-1)}$, and let $M = [\bar{M}, -b] \in \mathbb{R}^{n \times d}$. $n \gg d$.

Let Π be a SE with dimension m, distortion κ and embedding time t

1. Compute ΠM. (cost t)
2. Use ΠM to compute a matrix $R \in \mathbb{R}^{d \times d}$ (change-of-basis matrix) s.t. MR has some good properties. (cost $\uparrow$ if $m \uparrow$)
3. Given R, find a sampling matrix $\Pi^1 \in \mathbb{R}^{m' \times n}$. ($m' \uparrow$ if $\kappa \uparrow$)
4. Compute $\hat{x}$ of sub-sampled problem $\min_{x \in \mathbb{R}^d} \left\| \Pi^1 \bar{M}x - \Pi^1 b \right\|_p$.

   (cost $\uparrow$ if $m' \uparrow$, or $\kappa \uparrow$)

Total running time $\uparrow$ if $m \uparrow$ or $\kappa \uparrow$ or $t \uparrow$.

ℓ_1 regression

ℓ_1 regression: $\min_{x \in \mathbb{R}^d} \left\| \bar{M}x - b \right\|_1$ ($\bar{M} \in \mathbb{R}^{n \times (d-1)}$).

- Can be solved by linear programming, in time superlinear in n.
- Clarkson 2005 gave an $n \cdot \text{poly}(d)$ solution.
- $\ldots$

  Allow a $(1+\epsilon)$-approximation :

- Sohler & Woodruff 2011 used ℓ_1 subspace embedding (SE), gave $O(nd^{\omega-1}) + \text{poly}(d/\epsilon)$. ($\omega < 3$ is the exponent of matrix multiplication)
- Clarkson et al. 2012 used a more structured ℓ_1 SE, gave $O(nd \log n) + \text{poly}(d/\epsilon)$.
- Clarkson & Woodruff / Meng & Mahoney 2012 used other ℓ_1 SE's, gave $O(\text{nnz}(M) \log n) + \text{poly}(d/\epsilon)$, $\text{nnz}(M)$ is # non-zero entries of M.

ℓ_1 regression

ℓ_1 regression: $\min_{x \in \mathbb{R}^d} \left\| \bar{M}x - b \right\|_1$ ($\bar{M} \in \mathbb{R}^{n \times (d-1)}$).

- Can be solved by linear programming, in time superlinear in n.
- Clarkson 2005 gave an $n \cdot \text{poly}(d)$ solution.
- $\ldots$

  Allow a $(1+\epsilon)$-approximation :
- Sohler & Woodruff 2011 used ℓ_1 subspace embedding (SE), gave $O(nd^{\omega-1}) + \text{poly}(d/\epsilon)$. ($\omega < 3$ is the exponent of matrix multiplication)
- Clarkson et al. 2012 used a more structured ℓ_1 SE, gave $O(nd \log n) + \text{poly}(d/\epsilon)$.
- Clarkson & Woodruff / Meng & Mahoney 2012 used other ℓ_1 SE's, gave $O(\text{nnz}(M) \log n) + \text{poly}(d/\epsilon)$, $\text{nnz}(M)$ is # non-zero entries of M.

This paper: further improves the ℓ_1 SE, thus also ℓ_1 regression.

Our results

- ℓ_p subspace embeddings.
  Improved all previous results for $\forall p \in [1, \infty)\backslash 2$
  $p = 2$ has already been made optimal by Clarkson and Woodruff '12

Our results

- ℓ_p subspace embeddings.
  Improved all previous results for $\forall p \in [1, \infty) \backslash 2$
  $p = 2$ has already been made optimal by Clarkson and Woodruff '12

  In particular, $p = 1$

	Time	Distortion	Dimemsion
SW	$nd^{\omega-1}$	$\tilde{O}(d)$	$\tilde{O}(d)$
C$^+$	$nd \log d$	$\tilde{O}(d^{2+\gamma})$	$\tilde{O}(d^5)$
MM	$\text{nnz}(M)$	$\tilde{O}(d^3)$	$\tilde{O}(d^5)$
This paper	$\text{nnz}(M) + \tilde{O}(d^{2+\gamma})$ $\text{nnz}(M) + \tilde{O}(d^{2+\gamma})$	$\tilde{O}(d^2)$ $\tilde{O}(d^{3/2} \log^{1/2} n)$	$\tilde{O}(d)$ $\tilde{O}(d)$

SW: Sohler & Woodruff '11 ; C$^+$: Clarkson et al. '12; MM: Meng & Mahoney '12; $\omega < 3$ is the exponent of matrix multiplication. $\gamma = 0.0000001$.

Our results

- ℓ_p subspace embeddings.
  Improved all previous results for $\forall p \in [1, \infty) \backslash 2$
  $p = 2$ has already been made optimal by Clarkson and Woodruff '12

  In particular, $p = 1$

	Time	Distortion	Dimemsion
SW	$nd^{\omega-1}$	$\tilde{O}(d)$	$\tilde{O}(d)$
C+	$nd \log d$	$\tilde{O}(d^{2+\gamma})$	$\tilde{O}(d^5)$
MM	$\text{nnz}(M)$	$\tilde{O}(d^3)$	$\tilde{O}(d^5)$
This paper	$\text{nnz}(M) + \tilde{O}(d^{2+\gamma})$	$\tilde{O}(d^2)$	$\tilde{O}(d)$
	$\text{nnz}(M) + \tilde{O}(d^{2+\gamma})$	$\tilde{O}(d^{3/2} \log^{1/2} n)$	$\tilde{O}(d)$

  SW: Sohler & Woodruff '11 ; C+: Clarkson et al. '12; MM: Meng & Mahoney '12; $\omega < 3$ is the exponent of matrix multiplication. $\gamma = 0.0000001$.

- ℓ_p regression
  Improved all previous results for $\forall p \in [1, \infty) \backslash 2$
  Have efficient distributed implementations.

Our subspace embedding matrices

$(m, s) - \ell_2$-**SE** (oblivious subspace embedding for ℓ_2 norm)

A distribution over linear maps $S : \mathbb{R}^n \to \mathbb{R}^m$, s.t., for any fixed d-dimensional subspace of $\mathbb{R}^n$, w. pr. 0.99,

$$1/2 \cdot \|Mx\|_2 \leq \|SMx\|_2 \leq 3/2 \cdot \|Mx\|_2, \quad \forall x \in \mathbb{R}^d.$$

$s = O(1)$ is the the max of $\#$ non-zero entries of each colummn in S.

Our subspace embedding matrices

$(m, s) - \ell_2$-**SE** (oblivious subspace embedding for ℓ_2 norm)

A distribution over linear maps $S : \mathbb{R}^n \to \mathbb{R}^m$, s.t., for any fixed d-dimensional subspace of $\mathbb{R}^n$, w. pr. 0.99,

$$1/2 \cdot \|Mx\|_2 \leq \|SMx\|_2 \leq 3/2 \cdot \|Mx\|_2, \quad \forall x \in \mathbb{R}^d.$$

$s = O(1)$ is the the max of $\#$ non-zero entries of each colummn in S.

Our ℓ_p subspace embedding matrix

$$\left[\Pi \in \mathbb{R}^{m \times n} \right] = \underbrace{\left[S \in \mathbb{R}^{m \times n} \right]}_{\ell_2\text{-SE}} \times \overbrace{\begin{bmatrix} 1/u_1^{1/p} & & \\ & \ddots & \\ & & 1/u_n^{1/p} \end{bmatrix}}^{D \in \mathbb{R}^{n \times n}}$$

u_i are i.i.d. exponentials

Use different ℓ_2-SEs (from CW12, MM12, Nelson & Nguyen 12) for $1 \leq p < 2$ and $p > 2$.
Can compute ΠM in $O(\text{nnz}(M))$ time.

Two distributions

- Exponential distribution PDF $f(x) = e^{-x}$, CDF $F(x) = 1 - e^{-x}$

  (Recently used by Andoni (2012) for approximating frequency moments).

  (**max stability**) If $u_1, \ldots, u_n$ are exponentially distributed, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{+n}$, then $\max\{\alpha_1/u_1, \ldots, \alpha_n/u_n\} \simeq \|\alpha\|_1 / u$, where u is exponential.

Two distributions

- Exponential distribution PDF $f(x) = e^{-x}$, CDF $F(x) = 1 - e^{-x}$

  (Recently used by Andoni (2012) for approximating frequency moments).

  (**max stability**) If $u_1, \ldots, u_n$ are exponentially distributed, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{+n}$, then $\max\{\alpha_1/u_1, \ldots, \alpha_n/u_n\} \simeq \|\alpha\|_1 / u$, where u is exponential.

- p-stable distribution: Previous pet for subspace embedding.

  $\mathcal{D}_p$ is p-**stable**, if for any vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ and $v_1, \ldots, v_n \stackrel{i.i.d.}{\sim} \mathcal{D}_p$, we have $\sum_{i \in [n]} \alpha_i v_i \simeq \|\alpha\|_p v$, where $v \sim \mathcal{D}_p$.

Two distributions

- Exponential distribution PDF $f(x) = e^{-x}$, CDF $F(x) = 1 - e^{-x}$

  (Recently used by Andoni (2012) for approximating frequency moments).

  (**max stability**) If $u_1, \ldots, u_n$ are exponentially distributed, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{+n}$, then $\max\{\alpha_1/u_1, \ldots, \alpha_n/u_n\} \simeq \|\alpha\|_1 / u$, where u is exponential.

- p-stable distribution: Previous pet for subspace embedding.

  $\mathcal{D}_p$ is p-**stable**, if for any vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ and $v_1, \ldots, v_n \overset{i.i.d.}{\sim} \mathcal{D}_p$, we have $\sum_{i \in [n]} \alpha_i v_i \simeq \|\alpha\|_p v$, where $v \sim \mathcal{D}_p$.

  E.g., for $p = 2$ it is the Gaussian distribution;
  for $p = 1$ it is the Cauchy distribution.

Two distributions

- Exponential distribution PDF $f(x) = e^{-x}$, CDF $F(x) = 1 - e^{-x}$
  (Recently used by Andoni (2012) for approximating frequency moments).

  (**max stability**) If $u_1, \ldots, u_n$ are exponentially distributed, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{+n}$, then $\max\{\alpha_1/u_1, \ldots, \alpha_n/u_n\} \simeq \|\alpha\|_1 / u$, where u is exponential.

- p-stable distribution: Previous pet for subspace embedding.

  $\mathcal{D}_p$ is p-**stable**, if for any vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ and $v_1, \ldots, v_n \overset{i.i.d.}{\sim} \mathcal{D}_p$, we have $\sum_{i \in [n]} \alpha_i v_i \simeq \|\alpha\|_p v$, where $v \sim \mathcal{D}_p$.

  E.g., for $p = 2$ it is the Gaussian distribution;
  for $p = 1$ it is the Cauchy distribution.

Similar embedding matrix

$$\left[\Pi \in \mathbb{R}^{m \times n}\right] = \underbrace{\left[S \in \mathbb{R}^{m \times n}\right]}_{\ell_2\text{-SE}} \times \begin{bmatrix} v_1 & & \\ & \ddots & \\ & & v_n \end{bmatrix}$$

$D' \in \mathbb{R}^{n \times n}$
v_i are i.i.d. p-stables

Exponential distribution is superior than p-stables

Why exponential distribution is better?

Exponential distribution is superior than p-stables

Why exponential distribution is better?

1. p-stables only exist for $p \in [1, 2]$; while exponential can be used for all ℓ_p-SE ($p \geq 1$).

Exponential distribution is superior than p-stables

Why exponential distribution is better?

1. p-stables only exist for $p \in [1, 2]$; while exponential can be used for all ℓ_p-SE ($p \geq 1$).

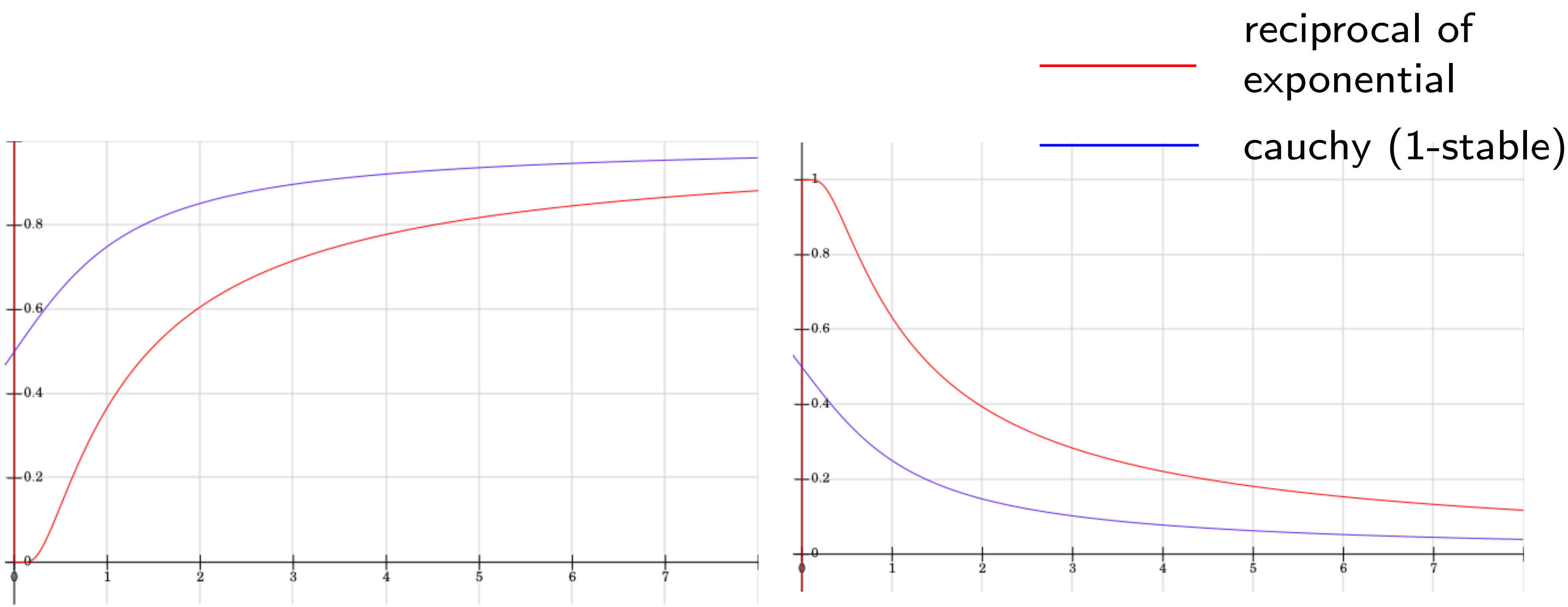
2. The lower tail of the reciprocal of exponential decreases faster than p-stable, while its the upper tail is similar to p-stables.

Exponential distribution is superior than p-stables

Why exponential distribution is better?

1. p-stables only exist for $p \in [1, 2]$; while exponential can be used for all ℓ_p-SE ($p \geq 1$).

2. The lower tail of the reciprocal of exponential decreases faster than p-stable, while its the upper tail is similar to p-stables.



lower tails upper tails

Analysis of distortions

Analysis for ℓ_1 subspace embedding

- Recall $\Pi = SD$:

u_i: exponential

$$\underbrace{\begin{bmatrix} \Pi \in \mathbb{R}^{m \times n} \end{bmatrix}}_{} = \underbrace{\begin{bmatrix} S \in \mathbb{R}^{m \times n} \end{bmatrix}}_{(O(d^{1.001}), O(1)) - \ell_2\text{-SE}} \times \underbrace{\begin{bmatrix} 1/u_1 & & \\ & \ddots & \\ & & 1/u_n \end{bmatrix}}_{D \in \mathbb{R}^{n \times n}}$$

Analysis for ℓ_1 subspace embedding

- Recall $\Pi = SD$:

$$\underbrace{\left[\Pi \in \mathbb{R}^{m \times n}\right]}_{} = \underbrace{\left[S \in \mathbb{R}^{m \times n}\right]}_{(O(d^{1.001}), O(1)) - \ell_2\text{-SE}} \times \overbrace{\begin{bmatrix} 1/u_1 & & & \\ & \bullet & & \\ & & \bullet & \\ & & & \bullet \\ & & & & 1/u_n \end{bmatrix}}^{u_i:\ \text{exponential}} D \in \mathbb{R}^{n \times n}$$

- **No underestimation**. For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.

$$\begin{aligned} & \|\Pi y\|_1 = \|SDy\|_1 \\ \geq\ & \|SDy\|_2 \geq 1/2 \cdot \|Dy\|_2 \quad \text{(property of } \ell_2\text{-SE)} \\ \geq\ & 1/2 \cdot \|Dy\|_\infty \sim 1/2 \cdot \|y\|_1 / u \quad (u \text{ is exponential, max stability}) \\ \geq\ & \Omega(d \log d) \cdot \|y\|_1 . \quad \text{(holds w.pr. } 1 - e^{-d \log d}, \\ & \qquad \text{lower tail of reciprocal of an exponential)} \end{aligned}$$

Analysis for ℓ_1 subspace embedding

- Recall $\Pi = SD$:

$$\underbrace{\begin{bmatrix} \Pi \in \mathbb{R}^{m\times n} \end{bmatrix} = \begin{bmatrix} S \in \mathbb{R}^{m\times n} \end{bmatrix}}_{(O(d^{1.001}), O(1)) - \ell_2\text{-SE}} \times \overset{u_i:\ \text{exponential}}{\begin{bmatrix} 1/u_1 & & \\ & \ddots & \\ & & 1/u_n \end{bmatrix}} \; D \in \mathbb{R}^{n\times n}$$

- **No underestimation**. For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.

$$\begin{aligned} & \|\Pi y\|_1 = \|SDy\|_1 \\ \geq\ & \|SDy\|_2 \geq 1/2 \cdot \|Dy\|_2 \quad \text{(property of } \ell_2\text{-SE)} \\ \geq\ & 1/2 \cdot \|Dy\|_\infty \sim 1/2 \cdot \|y\|_1 / u \quad (u \text{ is exponential, max stability}) \\ \geq\ & \Omega(d \log d) \cdot \|y\|_1 . \quad \text{(holds w.pr. } 1 - e^{-d \log d}, \\ & \qquad\qquad \text{lower tail of reciprocal of an exponential)} \end{aligned}$$

- This proves "for each y in the subspace" w.h.p.. To show this for all, we employ a standard net argument + a union bound.

Analysis for ℓ_1 subspace embedding

- Recall $\Pi = SD$:

$$\underbrace{\left[\Pi \in \mathbb{R}^{m\times n}\right]}_{} = \underbrace{\left[S \in \mathbb{R}^{m\times n}\right]}_{(O(d^{1.001}),\, O(1)) - \ell_2\text{-SE}} \times \overset{u_i:\ \text{exponential}}{\begin{bmatrix} 1/u_1 & & \\ & \ddots & \\ & & 1/u_n \end{bmatrix}} \quad D \in \mathbb{R}^{n\times n}$$

- **No underestimation**. For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.

$$\begin{aligned} & \|\Pi y\|_1 = \|SDy\|_1 \\ \geq\ & \|SDy\|_2 \geq 1/2 \cdot \|Dy\|_2 \quad \text{(property of } \ell_2\text{-SE)} \\ \geq\ & 1/2 \cdot \|Dy\|_\infty \sim 1/2 \cdot \|y\|_1 / u \quad (u \text{ is exponential, max stability}) \\ \geq\ & \Omega(d \log d) \cdot \|y\|_1 . \quad \text{(holds w.pr. } 1 - e^{-d \log d}, \\ & \qquad\qquad \text{lower tail of reciprocal of an exponential)} \end{aligned}$$

- This proves "for each y in the subspace" w.h.p.. To show this for all, we employ a standard net argument + a union bound.

  For $d \geq \log n$, distortion can be improved to $\tilde{O}(\sqrt{d \log n})$.

Analysis for ℓ_1 subspace embedding

- Recall $\Pi = SD$:

$$\underbrace{\left[\Pi \in \mathbb{R}^{m\times n}\right]}_{} = \underbrace{\left[S \in \mathbb{R}^{m\times n}\right]}_{(O(d^{1.001}), O(1)) - \ell_2\text{-SE}} \times \overbrace{\begin{bmatrix} 1/u_1 & & \\ & \ddots & \\ & & 1/u_n \end{bmatrix}}^{u_i:\ \text{exponential}} \; D \in \mathbb{R}^{n\times n}$$

- **No underestimation**. For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.

$$\begin{aligned} & \|\Pi y\|_1 = \|SDy\|_1 \\ \geq\ & \|SDy\|_2 \geq 1/2 \cdot \|Dy\|_2 \quad \text{(property of } \ell_2\text{-SE)} \\ \geq\ & 1/2 \cdot \|Dy\|_\infty \sim 1/2 \cdot \|y\|_1 / u \quad (u \text{ is exponential, max stability}) \\ \geq\ & \Omega(d \log d) \cdot \|y\|_1 . \quad \text{(holds w.pr. } 1 - e^{-d \log d}, \\ & \qquad\qquad \text{lower tail of reciprocal of an exponential)} \end{aligned}$$

- This proves "for each y in the subspace" w.h.p.. To show this for all, we employ a standard net argument + a union bound.

  For $d \geq \log n$, distortion can be improved to $\tilde{O}(\sqrt{d \log n})$.

- Similar arguments work for general $1 \leq p < 2$.

Analysis for ℓ_1 subspace embedding (cont.)

- Recall $\Pi = SD$:

u_i: exponential

$$\begin{bmatrix}\Pi \in \mathbb{R}^{m\times n}\end{bmatrix} = \underbrace{\begin{bmatrix}S \in \mathbb{R}^{m\times n}\end{bmatrix}}_{(O(d^{1.001}),\,O(1))-\ell_2\text{-SE}} \times \underbrace{\begin{bmatrix}1/u_1 & & \\ & \ddots & \\ & & 1/u_n\end{bmatrix}}_{D \in \mathbb{R}^{n\times n}}$$

- **No overestimation**. For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.

$$\begin{aligned} & \|\Pi y\|_1 = \|SDy\|_1 \\ \le\ & O(1)\cdot\|Dy\|_1 \quad (\ell_2\text{-SE only contracts } \ell_1\text{-norm}) \\ \preceq\ & O(1)\cdot\gamma\,\|D'y\|_1 \quad (\text{for a constant } \gamma, \\ & \qquad \text{upper tails of reciprocal of exponential and Cauchy are similar}) \\ \le\ & O(d\log d\cdot\|y\|_1) \quad (\text{holds for all } y = Mx \text{ w.pr. } 0.99\text{, previously known}) \end{aligned}$$

Analysis for ℓ_1 subspace embedding (cont.)

- Recall $\Pi = SD$:

u_i: exponential $\qquad$ v_i: Cauchy

$$\underbrace{\begin{bmatrix} \Pi \in \mathbb{R}^{m \times n} \end{bmatrix}}_{} = \underbrace{\begin{bmatrix} S \in \mathbb{R}^{m \times n} \end{bmatrix}}_{(O(d^{1.001}), O(1)) - \ell_2\text{-SE}} \times \underbrace{\begin{bmatrix} 1/u_1 & & \\ & \ddots & \\ & & 1/u_n \end{bmatrix}}_{D \in \mathbb{R}^{n \times n}} \preceq \gamma \underbrace{\begin{bmatrix} v_1 & & \\ & \ddots & \\ & & v_n \end{bmatrix}}_{D' \in \mathbb{R}^{n \times n}}$$

- **No overestimation**. For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.

$$\begin{aligned} & \|\Pi y\|_1 = \|SDy\|_1 \\ \leq \quad & O(1) \cdot \|Dy\|_1 \quad (\ell_2\text{-SE only contracts } \ell_1\text{-norm}) \\ \preceq \quad & O(1) \cdot \gamma \left\|D' y\right\|_1 \quad (\text{for a constant } \gamma, \\ & \qquad \text{upper tails of reciprocal of exponential and Cauchy are similar}) \\ \leq \quad & O(d \log d \cdot \|y\|_1) \quad (\text{holds for all } y = Mx \text{ w.pr. } 0.99, \text{ previously known}) \end{aligned}$$

Analysis for ℓ_1 subspace embedding (cont.)

- Recall $\Pi = SD$:

$$\begin{array}{ccccccc} & & & & u_i\text{: exponential} & & v_i\text{: Cauchy} \\ \left[\Pi \in \mathbb{R}^{m\times n}\right] & = & \underset{(O(d^{1.001}), O(1)) - \ell_2\text{-SE}}{\left[S \in \mathbb{R}^{m\times n}\right]} & \times & \underset{D \in \mathbb{R}^{n\times n}}{\begin{bmatrix} 1/u_1 & & \\ & \ddots & \\ & & 1/u_n \end{bmatrix}} & \preceq \gamma & \underset{D' \in \mathbb{R}^{n\times n}}{\begin{bmatrix} v_1 & & \\ & \ddots & \\ & & v_n \end{bmatrix}} \end{array}$$

- **No overestimation.** For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.

$$\begin{aligned} & \|\Pi y\|_1 = \|SDy\|_1 \\ \leq \quad & O(1) \cdot \|Dy\|_1 \quad (\ell_2\text{-SE only contracts } \ell_1\text{-norm}) \\ \preceq \quad & O(1) \cdot \gamma \left\|D' y\right\|_1 \quad (\text{for a constant } \gamma, \\ & \qquad \text{upper tails of reciprocal of exponential and Cauchy are similar}) \\ \leq \quad & O(d \log d \cdot \|y\|_1) \quad (\text{holds for all } y = Mx \text{ w.pr. } 0.99\text{, previously known}) \end{aligned}$$

- Similar arguments work for general $1 \leq p < 2$.

High level ideas for ℓ_p $(p > 2)$

- Recall $\Pi = SD$:

u_i: exponential

$$\left[\Pi \in \mathbb{R}^{m \times n}\right] = \left[S \in \mathbb{R}^{m \times n}\right] \times \begin{bmatrix} 1/u_1^{1/p} & & & \\ & \bullet & & \\ & & \bullet & \\ & & & \bullet \\ & & & & 1/u_n^{1/p} \end{bmatrix} D \in \mathbb{R}^{n \times n}$$

$\left(\tilde{O}(n^{1-2/p} d^{1+2/p}) + \text{poly}(d), 1\right) - \ell_2\text{-SE}$

- We actually can embed the subspace into ℓ_∞.

$$\Omega(1/(d \log d)^{1/p}) \cdot \|Mx\|_p \le \|\Pi M x\|_\infty \le O((d \log d)^{1/p}) \cdot \|Mx\|_p.$$

High level ideas for ℓ_p $(p > 2)$

- Recall $\Pi = SD$:

$$\left[\Pi \in \mathbb{R}^{m\times n}\right] = \left[S \in \mathbb{R}^{m\times n}\right] \times \begin{bmatrix} 1/u_1^{1/p} & & & \\ & \bullet & & \\ & & \bullet & \\ & & & \bullet \\ & & & & 1/u_n^{1/p} \end{bmatrix} D \in \mathbb{R}^{n\times n}$$

u_i: exponential

$\left(\tilde{O}(n^{1-2/p}d^{1+2/p}) + \mathrm{poly}(d), 1\right) - \ell_2$-SE

- We actually can embed the subspace into ℓ_∞.

$$\Omega(1/(d\log d)^{1/p}) \cdot \|Mx\|_p \le \|\Pi M x\|_\infty \le O((d \log d)^{1/p}) \cdot \|Mx\|_p.$$

Good news: ℓ_∞-regression can be solved efficiently by LP.

High level ideas for ℓ_p $(p > 2)$

- Recall $\Pi = SD$:

$$\left[\Pi \in \mathbb{R}^{m\times n}\right] = \left[S \in \mathbb{R}^{m\times n}\right] \times \overbrace{\begin{bmatrix} 1/u_1^{1/p} & & \\ & \ddots & \\ & & 1/u_n^{1/p} \end{bmatrix}}^{u_i\text{: exponential}} D \in \mathbb{R}^{n\times n}$$

$$\left(\tilde{O}(n^{1-2/p}d^{1+2/p}) + \mathrm{poly}(d), 1\right) - \ell_2\text{-SE}$$

- We actually can embed the subspace into ℓ_∞.

$$\Omega(1/(d\log d)^{1/p}) \cdot \|Mx\|_p \le \|\Pi Mx\|_\infty \le O((d\log d)^{1/p}) \cdot \|Mx\|_p.$$

Good news: ℓ_∞-regression can be solved efficiently by LP.

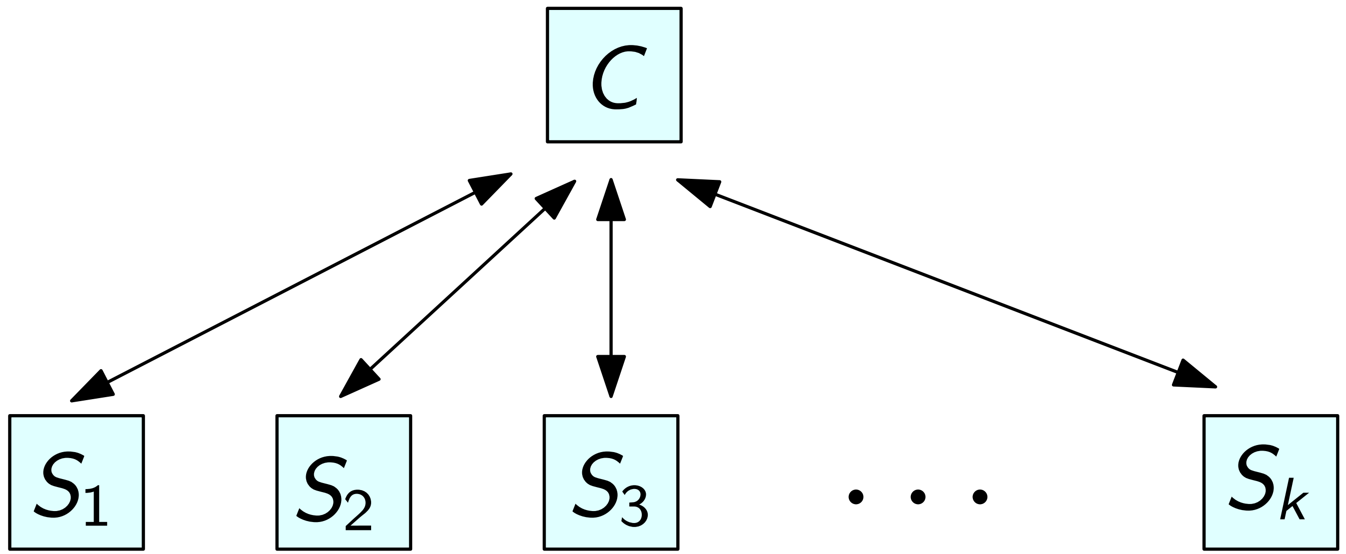
- Main technical ingredients

1. **No underestimation**: max stability of exponentials, like before
2. **No overestimation**: More complicated. Use leverage scores to upper bound coordinates of the vectors in a subspace (an idea previously used in CW12).

Distributed implementation ℓ_p-regression

- **The distributed model**: We have k machines and one central server.

  – Each machine has a 2-way communication channel with the server.
  – Each machine has a subset of rows of $\bar{M} \in \mathbb{R}^{n\times(d-1)}$ and $b \in \mathbb{R}^d$.
  – Goal is to solve ℓ_p-regression: $\min_{x\in\mathbb{R}^d} \left\|\bar{M}x - b\right\|_p$



Distributed implementation ℓ_p-regression (cont.)

- Recall the ℓ_p regression framework

1. Compute ΠM. (by machines, each computes ΠM_i where M_i is its local submatrix)
2. Use ΠM to compute a matrix $R \in \mathbb{R}^{d \times d}$ s.t. MR has good propertities. (by server)
3. Given R, find a sampling matrix $\Pi^1 \in \mathbb{R}^{m' \times n}$. (by machines, actually Π^1_i)
4. Solve sub-sampled problem $\min_{x \in \mathbb{R}^d} \left\| \Pi^1 \bar{M} x - \Pi^1 b \right\|_p$. (by server)

Distributed implementation ℓ_p-regression (cont.)

- Recall the ℓ_p regression framework

  1. Compute ΠM. (by machines, each computes ΠM_i where M_i is its local submatrix)
  2. Use ΠM to compute a matrix $R \in \mathbb{R}^{d \times d}$ s.t. MR has good propertities. (by server)
  3. Given R, find a sampling matrix $\Pi^1 \in \mathbb{R}^{m' \times n}$. (by machines, actually Π^1_i)
  4. Solve sub-sampled problem $\min_{x \in \mathbb{R}^d} \left\| \Pi^1 \bar{M} x - \Pi^1 b \right\|_p$. (by server)

- Total running time of the system

  - Running time of the centralized version + communication cost,
  - Most work is distributed on the k machines.

    Running time on the server + total communication = sublinear in n: poly(d) for $1 \leq p < 2$, and $n^{1-2/p}$poly(d) for $p > 2$.

    – Previous results either have $n/$poly(d) communication or only work for $1 \leq p \leq 2$.

Conclusions and open problems

- 1. We have proposed algorithms for ℓ_p ($p \in [1, \infty]\backslash 2$) subspace embeddings using exponential random variables, which improve all previous work on embedding distortions and dimensions, given the optimal running time.

  2. Improved subspace embeddings also lead to improved ℓ_p regressions

  3. Our algorithms can be efficiently implemented in the distributed setting.

Conclusions and open problems

- 1. We have proposed algorithms for ℓ_p ($p \in [1, \infty] \backslash 2$) subspace embeddings using exponential random variables, which improve all previous work on embedding distortions and dimensions, given the optimal running time.

  2. Improved subspace embeddings also lead to improved ℓ_p regressions

  3. Our algorithms can be efficiently implemented in the distributed setting.

- What is the best distortion given $O(\mathrm{nnz}(M) + \mathrm{poly}(d))$ embedding time and $\tilde{O}(d)$ embedding dimension, for ℓ_1 subspace embedding? Currently it is $\min\{\tilde{O}(d^2, d^{3/2} \log^{1/2} n)\}$.
  Is it possible to make it $\tilde{O}(d^{3/2})$ or even $\tilde{O}(d)$?

- Can we prove any tradeoff lower bounds?

Thank you!
Questions?

High level ideas for ℓ_p $(p > 2)$ (cont.)

- High level idea for no overestimation is similar as before.

  No overestimation of $\Pi M x$ for each $x \in \mathbb{R}^d$, w. pr. $1 - e^{-d \log d}$.

  $+$ a standard net argument to extend this to all $x \in \mathbb{R}^d$.

High level ideas for ℓ_p ($p > 2$) (cont.)

- High level idea for no overestimation is similar as before.

  No overestimation of $\Pi M x$ for each $x \in \mathbb{R}^d$, w. pr. $1 - e^{-d \log d}$.

  $+$ a standard net argument to extend this to all $x \in \mathbb{R}^d$.

  Cannot show for arbitrary vectors!

  We should use the the property of a subspace.

High level ideas for ℓ_p ($p > 2$) (cont.)

- High level idea for no overestimation is similar as before.

  No overestimation of $\Pi M x$ for each $x \in \mathbb{R}^d$, w. pr. $1 - e^{-d \log d}$.

  $+$ a standard net argument to extend this to all $x \in \mathbb{R}^d$.

  Cannot show for arbitrary vectors!

  We should use the the property of a subspace.

- Use leverage scores of M (An idea in Clarkson & Woodruff, '13)

  $\ell_i^p = \|M_i\|_p^p$, where M_i is the i-th row of M.

High level ideas for ℓ_p ($p > 2$) (cont.)

- High level idea for no overestimation is similar as before.

  No overestimation of $\Pi M x$ for each $x \in \mathbb{R}^d$, w. pr. $1 - e^{-d \log d}$.

  $+$ a standard net argument to extend this to all $x \in \mathbb{R}^d$.

  Cannot show for arbitrary vectors!

  We should use the the property of a subspace.

- Use leverage scores of M (An idea in Clarkson & Woodruff, '13)

  $\ell_i^p = \|M_i\|_p^p$, where M_i is the i-th row of M.

  Can assume M is the Auerbach basis (since we prove for all $x \in \mathbb{R}^d$), which has the property $\sum_{i \in [n]} \ell_i^p \le d$. Thus not many big ℓ_i.

High level ideas for ℓ_p ($p > 2$) (cont.)

- High level idea for no overestimation is similar as before.

  No overestimation of $\Pi M x$ for each $x \in \mathbb{R}^d$, w. pr. $1 - e^{-d \log d}$.

  $+$ a standard net argument to extend this to all $x \in \mathbb{R}^d$.

  Cannot show for arbitrary vectors!

  We should use the the property of a subspace.

- Use leverage scores of M (An idea in Clarkson & Woodruff, '13)

  $\ell_i^p = \|M_i\|_p^p$, where M_i is the i-th row of M.

  Can assume M is the Auerbach basis (since we prove for all $x \in \mathbb{R}^d$), which has the property $\sum_{i \in [n]} \ell_i^p \le d$. Thus not many big ℓ_i.

  Also, $\forall x \in \mathbb{R}^d$ and $y = Mx$, for all $i \in [n]$, we have $y_i \le d^{1-1/p} \ell_i$.
  Thus only a few big y_i's.

High level ideas for ℓ_p $(p > 2)$ (cont.)

- High level idea for no overestimation is similar as before.

  No overestimation of $\Pi M x$ for each $x \in \mathbb{R}^d$, w. pr. $1 - e^{-d \log d}$.

  $+$ a standard net argument to extend this to all $x \in \mathbb{R}^d$.

  Cannot show for arbitrary vectors!

  We should use the the property of a subspace.

- Use leverage scores of M (An idea in Clarkson & Woodruff, '13)

  $\ell_i^p = \|M_i\|_p^p$, where M_i is the i-th row of M.

  Can assume M is the Auerbach basis (since we prove for all $x \in \mathbb{R}^d$), which has the property $\sum_{i \in [n]} \ell_i^p \le d$. Thus not many big ℓ_i.

  Also, $\forall x \in \mathbb{R}^d$ and $y = Mx$, for all $i \in [n]$, we have $y_i \le d^{1-1/p} \ell_i$. Thus only a few big y_i's.

  Use this property, together with max stability, can design an embedding matrix Π works for arbitrary vectors.