Subspace Embeddings and $\ell_p$-Regression Using Exponential Random Variables

David P. Woodruff and Qin Zhang
IBM Research Almaden

COLT’13,
June 12, 2013
Subspace embeddings:

A distribution over linear maps $\Pi : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 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. $\kappa$: the distortion of the embedding.
3. $t$: the time to compute $\Pi M$. 
Subspace embeddings:

A distribution over linear maps \( \Pi : \mathbb{R}^n \rightarrow \mathbb{R}^m \), 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. \( \kappa \): the distortion of the embedding.
3. \( t \): the time to compute \( \Pi M \).

Applications:

\( \ell_p \)-regression (next slide), low-rank approximation, quantile regression, ...
Using ℓ_p subspace embedding (SE) to solve ℓ_p regression:

$$\min_{x \in \mathbb{R}^d} \| \tilde{M} x - b \|_p$$

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

Let $\Pi$ be a SE with dimension $m$, distortion $\kappa$ and embedding time $t$. 
Using $\ell_p$ subspace embedding (SE) to solve $\ell_p$ regression:

$$\min_{x \in \mathbb{R}^d} \| \bar{M}x - b \|_p$$

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 $\Pi$ be a SE with dimension $m$, distortion $\kappa$ and embedding time $t$

1. Compute $\Pi M$. (cost $t$)

2. Use $\Pi M$ 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 $\min_{x \in \mathbb{R}^d} \| \Pi^1 \bar{M}x - \Pi^1 b \|_p$. (cost $\uparrow$ if $m' \uparrow$, or $\kappa \uparrow$)

Total running time $\uparrow$ if $m \uparrow$ or $\kappa \uparrow$ or $t \uparrow$. 
\textit{$\ell_1$ regression}

\textit{$\ell_1$ regression:} \( \min_{x \in \mathbb{R}^d} \| \tilde{M}x - b \|_1 \) \( (\tilde{M} \in \mathbb{R}^{n \times (d-1)}) \).

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

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

- Sohler & Woodruff 2011 used \( \ell_1 \) subspace embedding (SE), gave \( O(nd^{\omega-1}) + \text{poly}(d/\epsilon) \). \((\omega < 3 \text{ is the exponent of matrix multiplication})\)
- Clarkson et al. 2012 used a more structured \( \ell_1 \) SE, gave \( O(nd \log n) + \text{poly}(d/\epsilon) \).
- Clarkson & Woodruff / Meng & Mahoney 2012 used other \( \ell_1 \) SE’s, gave \( O(\text{nnz}(M) \log n) + \text{poly}(d/\epsilon) \), \text{nnz}(M) is \# non-zero entries of \( M \).
\( \ell_1 \) regression

\( \ell_1 \) regression: \( \min_{x \in \mathbb{R}^d} \| \tilde{M}x - b \|_1 \) \( (\tilde{M} \in \mathbb{R}^{n \times (d-1)}) \).

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

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

- Sohler & Woodruff 2011 used \( \ell_1 \) subspace embedding (SE), gave \( O(nd^{\omega-1}) + \text{poly}(d/\epsilon) \). \( (\omega < 3 \text{ is the exponent of matrix multiplication}) \)
- Clarkson et al. 2012 used a more structured \( \ell_1 \) SE, gave \( O(nd \log n) + \text{poly}(d/\epsilon) \).
- Clarkson & Woodruff / Meng & Mahoney 2012 used other \( \ell_1 \) 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 \( \ell_1 \) SE, thus also \( \ell_1 \) regression.
Our results

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

- \( \ell_p \) subspace embeddings.
  Improved all previous results for \( \forall p \in [1, \infty) \setminus 2 \)
  \( p = 2 \) has already been made optimal by Clarkson and Woodruff ’12

In particular, \( p = 1 \)

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Distortion</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW</td>
<td>( nd^{\omega - 1} )</td>
<td>( \tilde{O}(d) )</td>
<td>( \tilde{O}(d) )</td>
</tr>
<tr>
<td>( C^+ )</td>
<td>( nd \log d )</td>
<td>( \tilde{O}(d^{2+\gamma}) )</td>
<td>( \tilde{O}(d^5) )</td>
</tr>
<tr>
<td>MM</td>
<td>( \text{nnz}(M) )</td>
<td>( \tilde{O}(d^3) )</td>
<td>( \tilde{O}(d^5) )</td>
</tr>
<tr>
<td>This paper</td>
<td>( \text{nnz}(M) + \tilde{O}(d^{2+\gamma}) )</td>
<td>( \tilde{O}(d^2) )</td>
<td>( \tilde{O}(d) )</td>
</tr>
<tr>
<td></td>
<td>( \text{nnz}(M) + \tilde{O}(d^{2+\gamma}) )</td>
<td>( \tilde{O}(d^{3/2} \log^{1/2} n) )</td>
<td>( \tilde{O}(d) )</td>
</tr>
</tbody>
</table>

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

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

In particular, $p = 1$

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Distortion</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW</td>
<td>$nd^{\omega^{-1}}$</td>
<td>$\tilde{O}(d)$</td>
<td>$\tilde{O}(d)$</td>
</tr>
<tr>
<td>$C^+$</td>
<td>$nd \log d$</td>
<td>$\tilde{O}(d^{2+\gamma})$</td>
<td>$\tilde{O}(d^5)$</td>
</tr>
<tr>
<td>MM</td>
<td>$\text{nnz}(M)$</td>
<td>$\tilde{O}(d^3)$</td>
<td>$\tilde{O}(d^5)$</td>
</tr>
<tr>
<td>This paper</td>
<td>$\text{nnz}(M) + \tilde{O}(d^{2+\gamma})$</td>
<td>$\tilde{O}(d^2)$</td>
<td>$\tilde{O}(d)$</td>
</tr>
<tr>
<td></td>
<td>$\text{nnz}(M) + \tilde{O}(d^{2+\gamma})$</td>
<td>$\tilde{O}(d^{3/2} \log^{1/2} n)$</td>
<td>$\tilde{O}(d)$</td>
</tr>
</tbody>
</table>

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

- $\ell_p$ regression
  Improved all previous results for $\forall p \in [1, \infty) \setminus 2$
  Have efficient distributed implementations.
Our subspace embedding matrices

\((m, s) - \ell_2\text{-SE}\) (oblivious subspace embedding for \(\ell_2\) 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 columnn in \(S\).
Our subspace embedding matrices

\((m, s) - \ell_2\text{-SE}\) (oblivious subspace embedding for \(\ell_2\) 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,

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

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

Our \(\ell_p\) subspace embedding matrix

\[
\begin{bmatrix}
\Pi & \in & \mathbb{R}^{m \times n} \\
S & \in & \mathbb{R}^{m \times n}
\end{bmatrix}
\text{ \(\ell_2\)-SE}
\times
\begin{bmatrix}
1/u_1^{1/p} \\
\cdot \\
\cdot \\
1/u_n^{1/p}
\end{bmatrix}
D \in \mathbb{R}^{n \times n}

u_i \text{ are i.i.d. exponentials}

Use different \(\ell_2\)-SEs (from CW12, MM12, Nelson & Nguyen 12) for \(1 \leq p < 2\) and \(p > 2\).
Can compute \(\Pi M\) 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 $\max\{\alpha_1 / u_1, \ldots, \alpha_n / u_n\} \approx \|\alpha\|_1 / u$, 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 \( \max\{\alpha_1/u_1, \ldots, \alpha_n/u_n\} \sim \|\alpha\|_1 / u \), 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 $\nu_1, \ldots, \nu_n \overset{i.i.d.}{\sim} \mathcal{D}_p$, we have $\sum_{i \in [n]} \alpha_i \nu_i \sim \|\alpha\|_p \nu$, where $\nu \sim \mathcal{D}_p$. 
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 
  $\max\{\alpha_1/u_1, \ldots, \alpha_n/u_n\} \approx \|\alpha\|_1 / u$, 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 
  $\sum_{i \in [n]} \alpha_i v_i \approx \|\alpha\|_p v$, 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 $\max\{\alpha_1/u_1, \ldots, \alpha_n/u_n\} \approx \|\alpha\|_1 / u$, 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 $v_1, \ldots, v_n \sim i.i.d. \mathcal{D}_p$, we have $\sum_{i \in [n]} \alpha_i v_i \approx \|\alpha\|_p v$, 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

$$ \begin{bmatrix} \Pi \in \mathbb{R}^{m \times n} \\ \ell_2\text{-SE} \end{bmatrix} = \begin{bmatrix} S \in \mathbb{R}^{m \times n} \\ v_1 \end{bmatrix} \times \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \quad D' \in \mathbb{R}^{n \times n} \quad v_i \text{ are i.i.d. } p\text{-stables} $$
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 $\ell_p$-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 $\ell_p$-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 $\ell_p$-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.

![Graph showing lower tails and upper tails comparison between reciprocal of exponential and Cauchy (1-stable) distributions.]
Analysis of distortions
Analysis for $\ell_1$ subspace embedding

Recall $\Pi = SD$:

$$\left[ \begin{array}{c} \Pi \in \mathbb{R}^{m \times n} \\ \end{array} \right] = \left[ \begin{array}{c} S \in \mathbb{R}^{m \times n} \\ \end{array} \right] \times \left( O(d^{1.001}), O(1) \right) - \ell_2$-SE

$$\left[ \begin{array}{c} 1/ u_1 \\ \vdots \\ 1/ u_n \\ \end{array} \right] \quad D \in \mathbb{R}^{n \times n}$$

$u_i$: exponential
Recall $\Pi = SD$:

$$\left[ \begin{array}{c} \Pi \in \mathbb{R}^{m \times n} \\ \end{array} \right] = \left[ \begin{array}{c} S \in \mathbb{R}^{m \times n} \\ \end{array} \right] \times \left[ \begin{array}{c} 1/u_1 \\ \vdots \\ 1/u_n \end{array} \right] \quad \text{for} \quad D \in \mathbb{R}^{n \times n}.$$

$$(O(d^{1.001}), O(1)) - \ell_2\text{-SE}$$

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

$$\|\Pi y\|_1 = \|SDy\|_1$$

$$\geq \|SDy\|_2 \geq 1/2 \cdot \|Dy\|_2 \quad \text{(property of } \ell_2\text{-SE)}$$

$$\geq 1/2 \cdot \|Dy\|_\infty \sim 1/2 \cdot \|y\|_1/u \quad \text{(} u \text{ is exponential, max stability)}$$

$$\geq \Omega(d \log d) \cdot \|y\|_1. \quad \text{(holds w.pr. } 1 - e^{-d \log d},$$

lower tail of reciprocal of an exponential)
Analysis for $\ell_1$ subspace embedding

- Recall $\Pi = SD$:

\[
\begin{bmatrix}
\Pi 
\in \mathbb{R}^{m \times n}
\end{bmatrix}
= \begin{bmatrix}
S 
\in \mathbb{R}^{m \times n}
\end{bmatrix} \times
\begin{bmatrix}
\frac{1}{u_1} \\
\vdots \\
\frac{1}{u_n}
\end{bmatrix}
\quad D \in \mathbb{R}^{n \times n}
\]

\[(O(d^{1.001}), O(1)) - \ell_2\text{-SE}\]

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

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

- 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 $\ell_1$ subspace embedding

- Recall $\Pi = SD$:
  \[
  \begin{bmatrix}
  \Pi \\
  \in \mathbb{R}^{m \times n}
  \end{bmatrix}
  =
  \begin{bmatrix}
  S \\
  \in \mathbb{R}^{m \times n}
  \end{bmatrix}
  \times
  \begin{bmatrix}
  1/\mu_1 \\
  \bullet \\
  \bullet \\
  \bullet \\
  1/\mu_n
  \end{bmatrix}
  \]

  $(O(d^{1.001}), O(1)) - \ell_2$-SE

- **No underestimation.** For each $x \in \mathbb{R}^d$, let $y = Mx \in \mathbb{R}^n$.
  \[
  \|\Pi y\|_1 = \|SDy\|_1
  \geq \|SDy\|_2 \geq 1/2 \cdot \|Dy\|_2 \quad \text{(property of $\ell_2$-SE)}
  \geq 1/2 \cdot \|Dy\|_\infty \sim 1/2 \cdot \|y\|_1 / u \quad \text{($u$ is exponential, max stability)}
  \geq \Omega(d \log d) \cdot \|y\|_1 . \quad \text{(holds w.pr. $1 - e^{-d \log d}$, lower tail of reciprocal of an exponential)}
  
- 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 $\ell_1$ subspace embedding

- Recall $\Pi = SD$:
  \[
  \begin{bmatrix}
  \Pi 
  \end{bmatrix} 
  \in \mathbb{R}^{m \times n} = \begin{bmatrix}
  S 
  \end{bmatrix} \in \mathbb{R}^{m \times n} \times \begin{bmatrix}
  1 / u_1 \\
  \vdots \\
  1 / u_n
  \end{bmatrix} 
  
  \text{(property of $\ell_2$-SE)}
  \]

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

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

- 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 $1 \leq p < 2$. 
Recall $\Pi = SD$:

$$\Pi \in \mathbb{R}^{m \times n} = \begin{bmatrix} S \in \mathbb{R}^{m \times n} \end{bmatrix} \times$$

$$\begin{bmatrix} 1/u_1 \\ \vdots \\ 1/u_n \end{bmatrix}$$

$(O(d^{1.001}), O(1)) - \ell_2$-SE

$D \in \mathbb{R}^{n \times n}$

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

$$\|\Pi y\|_1 = \|SDy\|_1$$

$$\leq O(1) \cdot \|Dy\|_1 \quad (\ell_2$-SE only contracts $\ell_1$-norm)$$

$$\leq O(1) \cdot \gamma \|D'y\|_1 \quad \text{(for a constant } \gamma, \text{ upper tails of reciprocal of exponential and Cauchy are similar)}$$

$$\leq O(d \log d \cdot \|y\|_1) \quad \text{(holds for all } y = Mx \text{ w.pr. 0.99, previously known)}$$
Analysis for $\ell_1$ subspace embedding (cont.)

- Recall $\Pi = SD$:

\[
\begin{bmatrix}
\Pi \\
\in \mathbb{R}^{m \times n}
\end{bmatrix} = \begin{bmatrix}
S \\
\in \mathbb{R}^{m \times n}
\end{bmatrix} \times \begin{bmatrix}
1/u_1 \\
1/u_n
\end{bmatrix}
\]

\[(O(d^{1.001}), O(1)) - \ell_2$-SE

\[D \in \mathbb{R}^{n \times n} \quad D' \in \mathbb{R}^{n \times n}
\]

\[u_i: \text{exponential} \quad v_i: \text{Cauchy}
\]

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

\[
\|\Pi y\|_1 = \|SDy\|_1
\]

\[\leq O(1) \cdot \|Dy\|_1 \quad (\ell_2$-SE only contracts $\ell_1$-norm)

\[\leq O(1) \cdot \gamma \|D'y\|_1 \quad (\text{for a constant } \gamma,
\text{upper tails of reciprocal of exponential and Cauchy are similar})
\]

\[\leq O(d \log d \cdot \|y\|_1) \quad (\text{holds for all } y = Mx \text{ w.pr. } 0.99, \text{previously known})
\]
Analysis for $\ell_1$ subspace embedding (cont.)

- Recall $\Pi = SD$:

\[
P \in \mathbb{R}^{m \times n} = S \in \mathbb{R}^{m \times n} \times \begin{bmatrix} 1/u_1 \\ 1/u_n \end{bmatrix} \preceq \gamma \begin{bmatrix} v_1 \\ v_n \end{bmatrix}
\]

$(O(d^{1.001}), O(1)) - \ell_2$-SE

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

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

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

11-3
High level ideas for $\ell_p$ ($p > 2$)

- Recall $\Pi = SD$:
  \[
  \begin{bmatrix}
  \Pi 
  \end{bmatrix} \in \mathbb{R}^{m \times n} = \begin{bmatrix}
  S 
  \end{bmatrix} \in \mathbb{R}^{m \times n} \times \begin{bmatrix}
  D 
  \end{bmatrix} \in \mathbb{R}^{n \times n}
  \]
  \[
  \left(\tilde{O}(n^{1-2/p}d^{1+2/p}) + \text{poly}(d), 1\right) - \ell_2\text{-SE}
  \]

- We actually can embed the subspace into $\ell_\infty$.
  \[
  \Omega(1/(d \log d)^{1/p}) \cdot ||Mx||_p \leq ||\Pi Mx||_\infty \leq O((d \log d)^{1/p}) \cdot ||Mx||_p.
  \]
High level ideas for $\ell_p$ ($p > 2$)

- Recall $\Pi = SD$:
  $$
  \Pi \in \mathbb{R}^{m \times n} = S \in \mathbb{R}^{m \times n} \times D \in \mathbb{R}^{n \times n} \\
  \left( \tilde{O}(n^{1-2/p}d^{1+2/p}) + \text{poly}(d), 1 \right) - \ell_2\text{-SE}
  $$

- We actually can embed the subspace into $\ell_\infty$.
  $$
  \Omega(1/(d \log d)^{1/p}) \cdot \|Mx\|_p \leq \|\Pi Mx\|_\infty \leq O((d \log d)^{1/p}) \cdot \|Mx\|_p.
  $$
  Good news: $\ell_\infty$-regression can be solved efficiently by LP.
High level ideas for $\ell_p$ ($p > 2$)

- Recall $\Pi = SD$:
  \[
  \Pi \in \mathbb{R}^{m \times n} = S \in \mathbb{R}^{m \times n} \times D \in \mathbb{R}^{n \times n}
  \]
  \[
  (\tilde{O}(n^{1-2/p}d^{1+2/p}) + \text{poly}(d), 1) - \ell_2\text{-SE}
  \]

- We actually can embed the subspace into $\ell_\infty$.
  \[
  \Omega(1/(d \log d)^{1/p}) \cdot \|Mx\|_p \leq \|\Pi Mx\|_\infty \leq O((d \log d)^{1/p}) \cdot \|Mx\|_p.
  \]
  Good news: $\ell_\infty$-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).
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 \( \ell_p \)-regression: \( \min_{x \in \mathbb{R}^d} \| \bar{M} x - b \|_p \)
Recall the $\ell_p$ regression framework

1. Compute $\Pi M$. (by machines, each computes $\Pi M_i$ where $M_i$ is its local submatrix)

2. Use $\Pi M$ to compute a matrix $R \in \mathbb{R}^{d \times d}$ s.t. $MR$ has good properties. (by server)

3. Given $R$, find a sampling matrix $\Pi^1 \in \mathbb{R}^{m' \times n}$. (by machines, actually $\Pi^1_i$)

4. Solve sub-sampled problem $\min_{x \in \mathbb{R}^d} \| \Pi^1 \bar{M}x - \Pi^1 b \|_p$. (by server)
Distributed implementation $\ell_p$-regression (cont.)

- Recall the $\ell_p$ regression framework
  1. Compute $\Pi M$. (by machines, each computes $\Pi M_i$ where $M_i$ is its local submatrix)
  2. Use $\Pi M$ to compute a matrix $R \in \mathbb{R}^{d \times d}$ s.t. $MR$ has good properties. (by server)
  3. Given $R$, find a sampling matrix $\Pi^1 \in \mathbb{R}^{m' \times n}$. (by machines, actually $\Pi^1_i$)
  4. Solve sub-sampled problem $\min_{x \in \mathbb{R}^d} \| \hat{\Pi}^1 \tilde{M}x - \Pi^1 b \|_p$. (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 $n$
- $\text{poly}(d)$ for $1 \leq p < 2$, and $n^{1-2/p}\text{poly}(d)$ for $p > 2$.
- Previous results either have $n/\text{poly}(d)$ communication or only work for $1 \leq p \leq 2$. 
Conclusions and open problems

1. We have proposed algorithms for $\ell_p$ ($p \in [1, \infty] \setminus 2$) 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 $\ell_p$ regressions.

3. Our algorithms can be efficiently implemented in the distributed setting.
Conclusions and open problems

1. We have proposed algorithms for $\ell_p$ ($p \in [1, \infty]\backslash 2$) 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 $\ell_p$ regressions.

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

What is the best distortion given $O(\text{nnz}(M) + \text{poly}(d))$ embedding time and $\tilde{O}(d)$ embedding dimension, for $\ell_1$ 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 idea for no overestimation is similar as before.

No overestimation of $\Pi Mx$ 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 $\ell_p$ ($p > 2$) (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 property of a subspace.
High level ideas for $\ell_p$ ($p > 2$) (cont.)

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

  No overestimation of $\Pi Mx$ 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 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 $\ell_p$ ($p > 2$) (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 property of a subspace.

- Use leverage scores of $M$ (An idea in Clarkson & Woodruff, ’13)
  
  $\ell^p_i = \|M_i\|^p$, 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 $\sum_{i \in [n]} \ell^p_i \leq d$. Thus not many big $\ell_i$. 

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

- High level idea for no overestimation is similar as before.
  
  No overestimation of $\prod Mx$ 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 property of a subspace.

- Use leverage scores of $M$ (An idea in Clarkson & Woodruff, ’13)
  
  $\ell_p^i = \|M_i\|_p^p$, 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 $\sum_{i \in [n]} \ell_i^p \leq d$. Thus not many big $\ell_i$.

  Also, $\forall x \in \mathbb{R}^d$ and $y = Mx$, for all $i \in [n]$, we have $y_i \leq d^{1-1/p} \ell_i$.

  Thus only a few big $y_i$’s.
High level ideas for $\ell_p$ ($p > 2$) (cont.)

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

  No overestimation of $\Pi Mx$ 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 $\sum_{i \in [n]} \ell_i^p \leq d$. Thus not many big $\ell_i$.

  Also, $\forall x \in \mathbb{R}^d$ and $y = Mx$, for all $i \in [n]$, we have $y_i \leq d^{1-1/p} \ell_i$.

  Thus only a few big $y_i$’s.

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