Input-sparsity Time Algorithms for Embeddings and Regression Problems

Michael Mahoney

Stanford Universtiy

September 2013

Joint work with Jiyan Yang and Xiangrui Meng.

1. Introduction and overview

2. Input-sparsity embeddings for ℓ_2 and ℓ_2 regression

3. Input-sparsity embeddings for ℓ_1 and ℓ_1 regression

4. Onto ℓ_p embeddings/regression

5. Onto quantile regression

6. Extensions and future directions

7. Conclusion

Succinct Data Representations

- PCA/SVD
- Sparsity and sparsification
- Hashing and sketching
- Streaming and sublinear algorithms
- Property testing
- Low-rank plus structured sparsity
- Feature/variable selection
- Random sampling and random projections
- Etc.

Quick History of Random Projections

- Johnson and Lindenstrauss (1982)
- Frankl and Meahara (1988)
- Indyk and Motwani (1998), Dasgupta and Gupta (1999)
- Achlioptas (2001)
- Charikar and Sahai (2002)
- Ailon and Chazelle (2006)
- Sohler and Woodruff (2011)
- Clarkson, Drineas, Magdon-Ismail, Mahoney, Meng, and Woodruff (2012)
- Clarkson and Woodruff (2012)

Quick History of Randomized Numerical Linear Algebra

(Mahoney, "Randomized Algorithms for Matrices and Data," FnTML, 2011.)

- Coarse sampling and additive-error algorithms
  - Frieze, Kannan, and Vempala (1998)
  - Achlipotas and McSherry (2001)
  - Drineas, Kannan, and Mahoney (2005)
- Leverage-score sampling/projection and relative-error algorithms
  - Drineas, Mahoney, and Muthukrishnan (2006)
  - Sarlos (2007); Drineas, Mahoney, Muthukrishnan, and Sarlos (2007)
  - Drineas, Magdon-Ismail, Mahoney, and Woodruff (2011)
- Preconditioning and iterating to high precision
  - Avron, Maymounkov, and Toledo (2009)
  - Rokhlin and Tygert (2008)
  - Meng, Saunders, and Mahoney (2011)
- Input-sparsity time regression algorithms
  - Clarkson and Woodruff (2012); Woodruff and Zhang (2013)
  - Meng and Mahoney (2012); Yang, Meng, and Mahoney (2013)
  - Nelson and Nguyen (2012)

ℓ_2 subspace embedding in input-sparsity time

Theorem ([Clarkson and Woodruff, STOC'13])

Given a matrix $A \in \mathbb{R}^{n \times d}$ with $n \gg d$, let $\Pi = SD$ where:

- *$D \in \mathbb{R}^{n \times n}$ is a diagonal matrix $\{\pm 1\}$ entries u.a.r.*
- *$S \in \mathbb{R}^{s \times n}$ has each column chosen independently and uniformly from the s standard basis vectors of $\mathbb{R}^s$.*

There is $s = \mathcal{O}((d/\epsilon)^4 \log^2(d/\epsilon))$ such that with a constant probability,

$$(1-\epsilon)\|Ax\|_2 \leq \|\Pi A x\|_2 \leq (1+\epsilon)\|Ax\|_2, \quad \forall x \in \mathbb{R}^d.$$

ΠA can be computed in $\mathcal{O}(\mathrm{nnz}(A))$ time.

$$\Pi = \begin{pmatrix} & 1 & & & & 1 \\ 1 & & & & & \\ & & \cdots & & & \\ & & & & 1 & \\ & & & 1 & & \end{pmatrix} \begin{pmatrix} \pm 1 & & & \\ & \pm 1 & & \\ & & \ddots & \\ & & & \pm 1 \end{pmatrix}.$$

CW proof of input-sparsity time ℓ_2 embedding result

Definition (ℓ_2 leverage scores)

Given any orthonormal basis U for the range(A) (where recall A is of size $n \times d$ with $n \gg d$), the ℓ_2 *leverage scores* of A are the squared ℓ_2 norms of U's rows: $\|U_{(i)}\|_2^2$, $i = 1, \ldots, n$.

Key idea: split rows of A into "heavy hitters" and "light hitters" based on their ℓ_2 leverage scores. Then show that:

- For high leverage score rows, the projection is an isometry.
- For low leverage score rows, use "a sparse Johnson-Lindenstrauss transform" of [DKS10].
- The cross-terms can be bounded separately.

Basic idea simple end elegant; actual proof long and detailed, using many ideas from TCS and streaming literature.

Simple proof of input-sparsity time ℓ_2 embedding result

Key idea: view it as approximating matrix multiplication.

- Let U be an orthonormal basis for range(A); thus $U^TU = I_d$.
  Define $X = (\Pi U)^T(\Pi U) = U^TD^TS^TSDU$.
  Show that $U^TU \approx U^TD^TS^TSDU = (SDU)^T(SDU)$.
- By computing a few moments, it is easy to obtain that

$$\mathbf{E}[\|X-I\|_F^2] = \frac{2}{s}\left(\sum_k (1-\|U_{*k}\|_4^4) + \sum_{k<l}(1-2\langle U_{*k}^2, U_{*l}^2\rangle)\right) \le \frac{d^2+d}{s}$$

- For any $\delta \in (0,1)$, set $s = (d^2+d)/(\epsilon^2\delta)$. By Markov's inequality,

$$\mathbf{Pr}[\|X-I\|_F \ge \epsilon] = \mathbf{Pr}[\|X-I\|_F^2 \ge \epsilon^2] \le \frac{d^2+d}{\epsilon^2 s} = \delta.$$

- Therefore, $\|X-I\|_2 \le \|X-I\|_F \le \epsilon$, w.p. $> 1-\delta$, which implies

$$(1-\epsilon)\|Uz\|_2 \le \|\Pi Uz\|_2 \le (1+\epsilon)\|Uz\|_2.$$

Important point to remember

The embedding matrix $\Pi = SD$ does *not* preserve the norm of an arbitrary set of e^d vectors.

- JL proofs: consider a fixed $x \in \mathbb{R}^d$; show that $||\Pi x||_2 = (1 \pm \epsilon)||x||_2$, w.p. $\geq 1 - 1/n^2$; and do a union bound to show all $\binom{n}{2}$ pairwise distances are preserved.
- Subspace embedding proofs: consider a fixed $x \in \mathbb{R}^d$; show that $||\Pi x||_2 = (1 \pm \epsilon)||x||_2$, w.p. $\geq 1 - e^{-d}$; put an ϵ-net on the d-dimensional space; and do a union bound to show all pairwise distances are preserved.

The CW proof—explicitly, and the MM proof implicitly—critically exploits that the e^d vectors come from a d-dimensional subspace of $\mathbb{R}^n$.

- have a very special structure—characterized by the leverage scores
- there can only be a small number of high-leverage components

Conditioning (for ℓ_1 and ℓ_p regression)

Definition ((α, β, p)-conditioning (from DDHKM09))

Given an $n \times d$ matrix A and $p \in [1, \infty]$, let q be the dual norm of p, and let $|A|_p^p = \sum_{ij} A_{ij}^p$. (Think $n \gg d$.) Then A is (α, β, p)-*conditioned* if:

- $|A|_p \leq \alpha$; and
- $\|z\|_q \leq \beta \|Az\|_p, \quad \forall z \in \mathbb{R}^d$.

Let $\bar{\kappa}_p(A)$ be the minimum value of $\alpha\beta$ such that A is (α, β, p)-conditioned. A basis U of range(A) is a *well-conditioned basis* if $\kappa = \bar{\kappa}_p(U)$ is a low-degree polynomial in d, independent of n.

Special cases:

- $p = 2$: Orthonormal basis ($\alpha = d^{1/2}$ & $\beta = 1$, and can find "quickly").
- $p = 1$: Auerbach basis ($\alpha = d$ & $\beta = 1$, "exists," but can construct approximate bases with $\alpha = d^{3/2}$ & $\beta = 1$ "quickly.")

Subspace-preserving sampling & approximate ℓ_p regression

Given a well-conditioned basis, we can do subspace-preserving sampling:

Lemma (Fast Subspace-preserving Sampling [DDHKM09,CDMMMW13])

Given a matrix $A \in \mathbb{R}^{n\times d}$, $p \in [1,\infty)$, $\epsilon > 0$, and a matrix $R \in \mathbb{R}^{d\times d}$ such that AR^{-1} is well-conditioned. It takes $\mathcal{O}(\mathrm{nnz}(A)\cdot \log n)$ time to compute a sampling matrix $S \in \mathbb{R}^{s\times n}$ with $s = \mathcal{O}(\bar{\kappa}_p^p(AR^{-1})d^{|p/2-1|+1}\log(1/\epsilon)/\epsilon^2)$ such that with a constant probability,

$$(1-\epsilon)\|Ax\|_p \le \|SAx\|_p \le (1+\epsilon)\|Ax\|_p, \quad \forall x \in \mathbb{R}^d .$$

Given a subspace-preserving sampling algorithm, we can compute a $1 \pm \epsilon$ approximate solution to an ℓ_p regression problem:

Lemma ($(1\pm\epsilon)$-ℓ_p Regression via Sampling [DDHKM09, CDMMMW13])

Given an ℓ_p regression problem: $\min_{x\in\mathbb{R}^d} \|Ax-b\|_p$.
Let S be a $(1\pm\epsilon)$-distortion embedding matrix of $\mathrm{range}([\ A \mid b\ ])$, and let $\hat{x} = \mathrm{argmin}_{x\in\mathbb{R}^d}\|SAx - Sb\|_p$.
Then $\hat{x}$ is a $1\pm\epsilon$-approximate solution to the original ℓ_p regression problem.

Low-distortion embeddings and regression problems

Succinct low-distortion embeddings and regression problems:

- Low-distortion ℓ_p subspace embeddings are the key building blocks for computing $(1 \pm \epsilon)$-approximation to an ℓ_p regression problem.
- (As a special case, for $p = 2$, this means finding, e.g., an orthogonal matrix from QR or SVD, or establishing a Johnson-Lindenstrauss result.)
- Given this embedding, we can get the solution to the ℓ_p regression problem in $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ additional time.
- For $p = 2$, this is actually only $\mathcal{O}(\text{nnz}(A))$ additional time.

Many randomized matrix algorithms boil down to the $p = 2$ case.

But how long does it take to find a low-distortion embedding?

Oblivious linear subspace embeddings: previous work, our contribution, and more recent results

running time	ℓ_2	ℓ_1	$\ell_p, p \in (1,2)$
$\Omega(n \cdot d^2)$	QR/SVD	[Cla05]	[DDHKM09]
$\tilde{\mathcal{O}}(\mathrm{nnz}(A) \cdot d)$	JLT	CT [SW11]	✓
$\tilde{\mathcal{O}}(nd)$	FJLT	FCT [CDMMMW13]	
$\mathcal{O}(\mathrm{nnz}(A))$	[CW13]	✓	✓

More recent extensions and improvements:

- Clarkson and Woodruff 2013: other values of p
- Nelson and Nguyen 2013: ℓ_2 sparsity tradeoffs
- Miller and Peng 2013: optimizing various terms
- Yang, Meng, and Mahoney 2013: quantile regression
- Etc.

Input-sparsity time algorithms for RandNLA problems

Comments on “input-sparsity time” algorithms:

- Actual running time is $\mathcal{O}(\mathrm{nnz}(A) + \mathrm{poly}(d/\epsilon))$ time, where the second term is the time to solve the subproblem.
- So, running time is proportional to $\mathrm{nnz}(A)$ if $n \gg d$.
- Many tradeoffs, so can also get running times with leading terms of $\mathcal{O}(\mathrm{nnz}(A)\log(n))$ with better second-order terms.

More realistic models of data access:

- This is in the RAM model, idealized as $n \gg d$.
- Still open: some promising results in parallel/distributed environments, but not well-characterized theoretically.
- Still open: more realistic theoretical characterization of these ideas in more realistic data-access models, even on a single machine.

Onto ℓ_1: ℓ_1 subspace embedding in input-sparsity time

Replace $\{\pm 1\}$ random variables on diagonal of D by Cauchy variables.

Theorem

Given a matrix $A \in \mathbb{R}^{n \times d}$ with $n \gg d$, let $\Pi = SC$ where:

- *$C \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose entries are i.i.d. samples from the Cauchy distribution,*
- *$S \in \mathbb{R}^{s \times n}$ has each column chosen independently and uniformly from the s standard basis vectors of $\mathbb{R}^s$.*

There is $s = \mathcal{O}(d^5 \log^5 d)$ such that with a constant probability,

$$\|Ax\|_1 / \mathcal{O}(d^2 \log^2 d) \le \|\Pi A x\|_1 \le \mathcal{O}(d \log d) \|Ax\|_1, \quad \forall x \in \mathbb{R}^d.$$

In addition, ΠA can be computed in $\mathcal{O}(\mathrm{nnz}(A))$ time.

Cauchy variables used to approximate frequency moments (Indyk01), etc.

Sketch of proof: Auerbach basis matrix

Let the $n \times d$ matrix U be an Auerbach basis matrix of $\mathcal{A}_1$. By definition,

- U's columns are unit vectors in the ℓ_1 norm, thus $|U|_1 = d$,
- $\|x\|_\infty \leq \|Ux\|_1, \ \forall x \in \mathbb{R}^d$.

Things to note:

- This generalizes orthogonal matrices from $p = 2$.
- This generalizes to $p \in [1, \infty)$ [DDHKM09].
- This is a tool within the analysis—it must exist (it does!), but we do *not* need to compute it.

Definition (ℓ_1 leverage scores (from CDMMMW13))

Given a well-conditioned basis U for the range(A) (where recall A is of size $n \times d$ with $n \gg d$), the ℓ_1 *leverage scores* of A are the ℓ_1 norms of U's rows: $\|U_{(i)}\|_1$, $i = 1, \ldots, n$.

Sketch of proof: Auerbach basis matrix

Then, it is sufficient to prove

$$\|y\|_1/\mathcal{O}(d^2 \log^2 d) \leq \|\Pi y\|_1 \leq \mathcal{O}(d \log d)\|y\|_1, \quad \forall y \in Y,$$

where

$$Y = \{y \in \mathbb{R}^n \,|\, y = Ux,\ \|x\|_\infty = 1,\ x \in \mathbb{R}^d\}.$$

Notation:

- u_j, the j-th row of U, $j = 1, \ldots, n$,
- $v_j = \|u_j\|_1$, $j = 1, \ldots, n$, the ℓ_1 leverage scores of A.

Sketch of proof: stable distributions

- A distribution $\mathcal{D}$ over $\mathbb{R}$ is called p-stable, if for any m real numbers $a_1, \dots, a_m$, we have

$$\sum_{i=1}^{m} a_i X_i \simeq \left(\sum_{i=1}^{m} |a_i|^p \right)^{1/p} X,$$

  where $X_i \overset{\text{iid}}{\sim} \mathcal{D}$ and $X \sim \mathcal{D}$.

- Stable distributions exist for any $p \in (0, 2]$ ([Lévy, 1952]).
  - The standard Gaussian distribution $\mathcal{G}$ is 2-stable.
  - The standard Cauchy distribution $\mathcal{C}$ is 1-stable.
- Denote by $\mathcal{D}_p$ the "standard" p-stable distribution described by the characteristic function $\varphi(t) = e^{-|t|^p}$.

Sketch of proof: Cauchy tail inequalities

Lemma (Cauchy upper tail inequality [CDMMMW13])

For $i = 1, \ldots, m$, let C_i be m (not necessarily independent) standard Cauchy variables, and $\gamma_i > 0$ with $\gamma = \sum_i \gamma_i$. Let $X = \sum_i \gamma_i |C_i|$. For any $t > 1$,

$$\Pr[X > t\gamma] \leq \frac{1}{\pi t} \left(\frac{\log(1 + (2mt)^2)}{1 - 1/(\pi t)} + 1 \right).$$

Lemma (Cauchy lower tail inequality [CDMMMW13])

For $i = 1, \ldots, m$, let C_i be independent standard Cauchy random variables, and $\gamma_i \geq 0$ with $\gamma = \sum_i \gamma_i$. Let $X = \sum_i \gamma_i |C_i|$. Then, for any $t > 0$,

$$\log \Pr[X \leq (1 - t)\gamma] \leq \frac{-\gamma t^2}{3 \max_i \gamma_i}.$$

Sketch of proof: upper bound

With a constant probability,

$$|\Pi U|_1 = |SCU|_1 = \sum_{k=1}^{d} \sum_{i=1}^{s} \Big| \sum_{j=1}^{n} s_{ij} c_j u_{jk} \Big| \simeq \sum_{k=1}^{d} \sum_{i=1}^{s} \sum_{j=1}^{n} \left(|s_{ij} u_{jk}|\right) |\tilde{c}_{ik}|,$$

where $\{\tilde{c}_{ik}\}$ are *dependent* Cauchy random variables. Note that

$$\sum_{k=1}^{d} \sum_{i=1}^{s} \sum_{j=1}^{n} |s_{ij} u_{jk}| = \sum_{k=1}^{d} \sum_{j=1}^{n} |u_{jk}| = |U|_1 = d.$$

Applying the upper tail inequality, we get, with probability at least 0.9,

$$\|\Pi U x\|_1 \leq |\Pi U|_1 \|x\|_\infty \leq |\Pi U|_1 \|Ux\|_1 \leq \mathcal{O}(d \log d) \|Ux\|_1, \quad \forall x \in \mathbb{R}^d.$$

Sketch of proof: partition indexes

We partition indexes $[n]$ into two index sets: "heavy hitters" with large ℓ_1 leverage scores; and "light hitters" with small ℓ_1 leverage scores.

- $\tau = \omega^{1/4}/(d \log^2 d)$, where ω is a sufficiently large constant.
- Two index sets $H = \{j \mid v_j \geq \tau\}$ and $L = \{j \mid v_j < \tau\}$.
- When an index set appears as a superscript, we mean zeroing out elements or rows that do not belong to this index set.

It is easy to see that $|H| \leq \frac{d}{\tau}$ and $\|v^L\|_\infty \leq \tau$.

Sketch of proof: lower bound when $\|y^L\|_1 \geq \frac{1}{2}\|y\|_1$

Either $\|y^L\|_1$ or $\|y^H\|_1$ dominates $\|y\|_1$ because $\|y\|_1 = \|y^L\|_1 + \|y^H\|_1$. Define:

- $Y^H = \{y \in Y | \|y^H\|_1 \geq \frac{1}{2}\|y\|_1\}$,
- $Y^L = Y \backslash Y^H = \{y \in Y | \|y^L\|_1 > \frac{1}{2}\|y\|_1\}$.

For any *fixed* $y \in Y^L$, we can use the lower tail inequality to show that, with an exponentially small failure rate,

$$\|\Pi y\|_1 \geq \frac{1}{4}\|y\|_1.$$

Then by an ϵ-net argument, we prove that, with probability at least 0.9, the following union bound holds:

$$\|\Pi y\|_1 \geq \frac{1}{8}\|y\|_1, \quad \forall y \in Y^L.$$

Sketch of proof: lower bound when $\|y^H\|_1 \geq \frac{1}{2}\|y\|_1$

Given S, define a mapping $\phi : \{1, \ldots, n\} \to \{1, \ldots, s\}$ such that $s_{\phi(j),j} = 1$, $j = 1, \ldots, n$. Let $\hat{L} = \{j \in L \,|\, \phi(j) \in \phi(H)\}$. For any $y \in Y^H$,

$$\begin{aligned}
\|\Pi y\|_1 &\geq \|\Pi(y^H + y^{\hat{L}})\|_1 \geq \|\Pi y^H\|_1 - \|\Pi U^{\hat{L}} x\|_1 \\
&\geq \sum_{j \in H} |c_j||y_j| - |\Pi U^{\hat{L}}|_1 \|x\|_\infty \\
&\geq \left(\min_{j \in H} |c_j|\right) \|y^H\|_1 - |\Pi U^{\hat{L}}|_1.
\end{aligned}$$

The proof is done by showing that there exist constants ω_3 and ω_4 with $\omega_3 > 4\omega_4$ such that

- $\min_{j \in H} |c_j| > \omega_3/(d^2 \log^2 d)$ with probability at least 0.9,
- $|\Pi U^{\hat{L}}|_1 \leq \omega_4/(d^2 \log^2 d)$ with probability at least 0.9,

Onto ℓ_p: Generalization to ℓ_p subspace embeddings

To generalize our ℓ_1 result to ℓ_p, we need to:

- prove tail inequalities for p-stable distributions by establishing an order among p-stable variables
- prove upper and lower tail inequalities for p-stable distributions
- easily generalize the rest of the analysis for $p = 1$

We can prove:

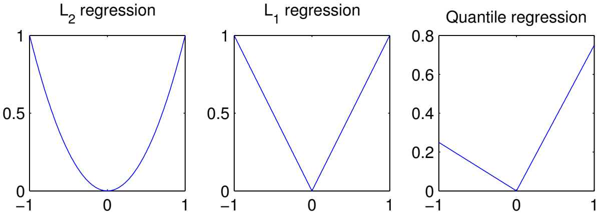
- Input-sparsity time low-distortion embedding for ℓ_p
- Nearly-input-sparsity time solution to ℓ_p regression problems

Differences between our approach and CW's:

- Our approach works for $p \in [1, 2)$ while CW's works for all $p \in [1, \infty)$.
- CW doesn't get the embedding in input-sparsity time.
- CW solves a rounding problem of size $n/\operatorname{poly}(d) \times d$, while ours does not have this intermediate step.

What is quantile regression?

	ℓ_2 regression	ℓ_1 regression	quantile regression
estimation	mean	median	quantile τ
loss function	x^2	$\lvert x \rvert$	$\rho_\tau(x)$
formulation	$\lVert Ax - b \rVert_2^2$	$\lVert Ax - b \rVert_1$	$\rho_\tau(Ax - b)$
is a norm?	yes	yes	no



(Note, ℓ_1 regression is a special case of quantile regression with $\tau = 0.5$.)

What is quantile regression?

- Quantile regression is a method to estimate the quantiles of the conditional distribution of response; it involves minimizing asymmetrically weighted absolute residuals:

$$\rho_\tau(z) = \begin{cases} \tau z, & z \geq 0; \\ (\tau - 1)z, & z < 0. \end{cases}$$

- Given $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and a parameter $\tau \in (0,1)$, quantile regression problem can be solved via the optimization problem:

$$\text{minimize}_{x \in \mathbb{R}^d} \quad \rho_\tau(Ax - b), \tag{1}$$

  where $\rho_\tau(y) = \sum_{i=1}^n \rho_\tau(y_i)$, for $y \in \mathbb{R}^n$.

- We use A to denote $\begin{bmatrix} A & -b \end{bmatrix}$, the quantile regression problem (1) can equivalently be expressed as the following,

$$\text{minimize}_{x \in \mathcal{C}} \quad \rho_\tau(Ax), \tag{2}$$

  where $\mathcal{C} = \{x \in \mathbb{R}^d \mid c^T x = 1\}$ and c is a unit vector with the last coordinate 1.

Alorithms for quantile regression

Previous algorithms:

- Standard solver for quantile regression problem: an interior-point method `ipm` [Portnoy and Koenker, 1997], which might be applicable for medium-large scale problem with size $1e6$ by 50.
- Best previous sampling algorithm for quantile regression problems, namely `prqfn`, is using an interior-point method on a smaller problem that has been preprocessed by randomly sampling a subset of the data; see [Portnoy and Koenker, 1997].

Our approach:

- Conditioning: find an ℓ_1-well-conditioned basis U for the span of A.
- Leveraging: compute/estimate the ℓ_1 leverage scores from the rows of U, and construct sampling matrix S to draw a random sample.
- Solving the subproblem: $\text{minimize}_{x \in \mathcal{C}}\ \rho_\tau(SAx)$.

Conditioning: finding an ℓ_1 well-conditioned basis

Recall, given an $n \times d$ matrix A and $p \in [1, \infty]$, we want to find a low-distortion embedding $\Pi \in \mathbb{R}^{s \times n}$ s.t. $s = \mathcal{O}(\text{poly}(d))$ and

$$1/\,\mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p \le \|\Pi A x\|_p \le \mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p, \quad \forall x \in \mathbb{R}^d .$$

There are two main ways:

Lemma (Conditioning via QR on low-distortion embedding)

Given a low-distortion embedding matrix Π of $\mathcal{A}_p$, let R be the "R" matrix from the QR decomposition of ΠA. Then, AR^{-1} is ℓ_p-well-conditioned.

Lemma (Conditioning via ellipsoidal rounding)

Given an $n \times d$ matrix A and $p \in [1, \infty]$, it takes at most $\mathcal{O}(nd^3 \log n)$ time to find a matrix $R \in \mathbb{R}^{d \times d}$ such that $\kappa_p(AR^{-1}) \le 2d$.

Comparison of conditioning methods

name	running time	κ	type
SC[SW11]	$\mathcal{O}(nd^2 \log d)$	$\mathcal{O}(d^{5/2} \log^{3/2} n)$	QR
FC [CDMMMW13]	$\mathcal{O}(nd \log d)$	$\mathcal{O}(d^{7/2} \log^{5/2} n)$	QR
Ellipsoid rounding [Cla05]	$\mathcal{O}(nd^5 \log n)$	$d^{3/2}(d+1)^{1/2}$	ER
Fast ER [CDMMMW13]	$\mathcal{O}(nd^3 \log n)$	$2d^2$	ER
SPC1 [MM13]	$\mathcal{O}(\mathrm{nnz}(A))$	$\mathcal{O}(d^{\frac{13}{2}} \log^{\frac{11}{2}} d)$	QR
SPC2 [MM13]	$\mathcal{O}(\mathrm{nnz}(A) \cdot \log(n)) + \texttt{ER\_small}$	$6d^2$	QR+ER
SPC3 (YMM13)	$\mathcal{O}(\mathrm{nnz}(A) \cdot \log(n)) + \texttt{QR\_small}$	$\mathcal{O}(d^{\frac{19}{4}} \log^{\frac{11}{4}} d)$	QR+QR

Table: Summary of running time, condition number, and type of conditioning methods proposed recently. QR and ER refer, respectively, to methods based on the QR factorization and methods based on Ellipsoid Rounding.

SC := Slow Cauchy Transform
FC := Fast Cauchy Transform
SPC := Sparse Cauchy Transform

Fast Randomized Algorithm for Quantile Regression

```
Input: $A \in \mathbb{R}^{n \times d}$ with full column rank, $\epsilon \in (0, 1/2)$, $\tau \in [1/2, 1)$.
Output: An approximate solution $\hat{x} \in \mathbb{R}^d$ to problem $\text{minimize}_{x \in \mathcal{C}}\ \rho_\tau(Ax)$.
Compute $R \in \mathbb{R}^{d \times d}$ such that $AR^{-1}$ is a well-conditioned basis for range($A$).
Compute a $(1 \pm \epsilon)$-distortion subspace-preserving embedding $S \in \mathbb{R}^{s \times n}$.
Return $\hat{x} \in \mathbb{R}^d$ that minimizes $\rho_\tau(SAx)$ with respect to $x \in \mathcal{C}$.
```

Theorem (Fast Quantile Regression)

Given $A \in \mathbb{R}^{n \times d}$ and $\varepsilon \in (0, 1/2)$, the above algorithm returns a vector $\hat{x}$ that, with probability at least 0.8, satisfies

$$\rho_\tau(A\hat{x}) \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \rho_\tau(Ax^*),$$

where x^* is an optimal solution to the original problem. In addition, the algorithm to construct $\hat{x}$ runs in time

$$\mathcal{O}(\text{nnz}(A) \cdot \log n) + \phi\left(\mathcal{O}(\mu d^3 \log(\mu/\epsilon)/\epsilon^2), d\right),$$

where $\mu = \frac{\tau}{1-\tau}$ and $\phi(s, d)$ is the time to solve an $s \times d$ quantile regression problem.

Types of data considered

Synthetic data

Following the construction of CDMMMW13:

- Each row of the design matrix A (size $\sim 10^6 \times 10^2$) is a canonical vector. Suppose the number of measurements on the j-th column are c_j, where $c_j = qc_{j-1}$, for $j = 2, \ldots, d$. Here $1 < q \leq 2$. A is a $n \times d$ matrix.
- The true vector x^* with length d is a vector with independent Gaussian entries. Let $b^* = Ax^*$.
- The response vector b is obtained by adding noise to b^*.

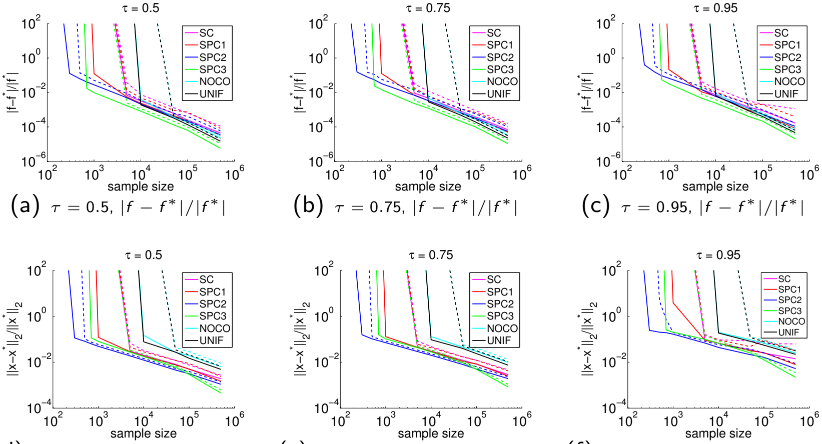
Real data

A data set consisting of a 5% sample of the U.S. 2000 Census data consisting of annual salary and related features. The size of the design matrix is 5×10^6 by 11.

"large" vs. "LARGE" data:

(These data are only "large," but we also have implementations on "LARGE" data—embarrassingly parallel, so only 3 passes through the data in Hadoop.)

Relative error when the sampling size *s* changes



(a) $\tau = 0.5$, $|f - f^*|/|f^*|$

(b) $\tau = 0.75$, $|f - f^*|/|f^*|$

(c) $\tau = 0.95$, $|f - f^*|/|f^*|$

(d) $\tau = 0.5$, $\|x - x^*\|_2/\|x^*\|_2$

(e) $\tau = 0.75$, $\|x - x^*\|_2/\|x^*\|_2$

(f) $\tau = 0.95$, $\|x - x^*\|_2/\|x^*\|_2$

Figure: First (solid lines) and third (dashed lines) quartiles of the relative errors of the objective value and solution vector. The test is on synthetic data with size $1e6$ by 50.

Comparison of the running time of each conditioning method

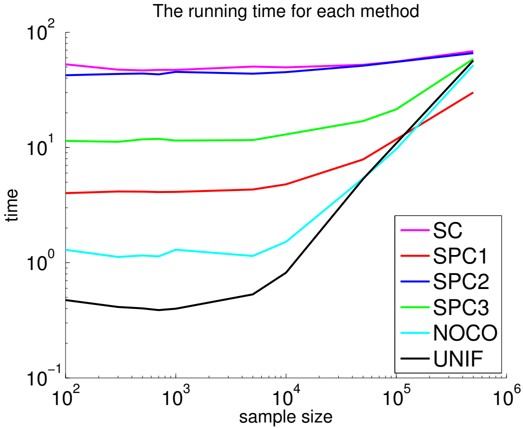
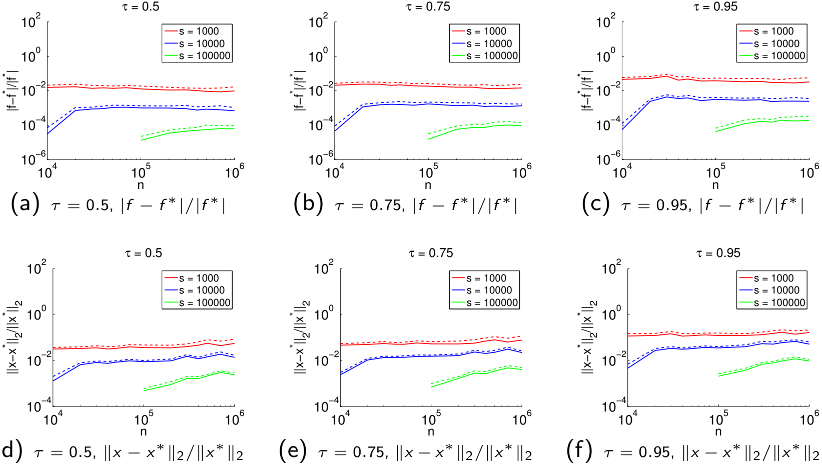


Figure: The running time for solving the problems associated with three different τ values when the sampling size s changes.

Relative error when the higher dimension n changes



(a) $\tau = 0.5$, $|f - f^*|/|f^*|$

(b) $\tau = 0.75$, $|f - f^*|/|f^*|$

(c) $\tau = 0.95$, $|f - f^*|/|f^*|$

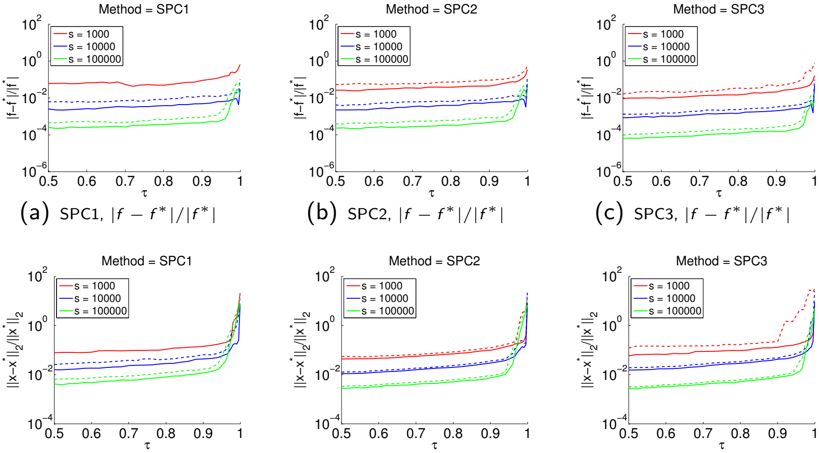
(d) $\tau = 0.5$, $\|x - x^*\|_2/\|x^*\|_2$

(e) $\tau = 0.75$, $\|x - x^*\|_2/\|x^*\|_2$

(f) $\tau = 0.95$, $\|x - x^*\|_2/\|x^*\|_2$

Figure: The first (solid lines) and the third (dashed lines) quartiles of the relative errors of the objective value and solution vector, when n varying from $1e4$ to $1e6$ and $d = 50$ by using SPC3.

Relative error when the quantile τ changes



(a) SPC1, $|f - f^*|/|f^*|$

(b) SPC2, $|f - f^*|/|f^*|$

(c) SPC3, $|f - f^*|/|f^*|$

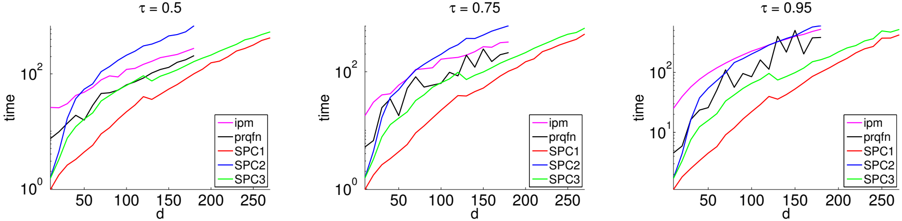
(d) SPC1, $\|x - x^*\|_2/\|x^*\|_2$

(e) SPC2, $\|x - x^*\|_2/\|x^*\|_2$

(f) SPC3, $\|x - x^*\|_2/\|x^*\|_2$

Figure: The first (solid lines) and the third (dashed lines) quartiles of the relative errors of the objective value, and solution vector. The test data size 1e6 by 50.

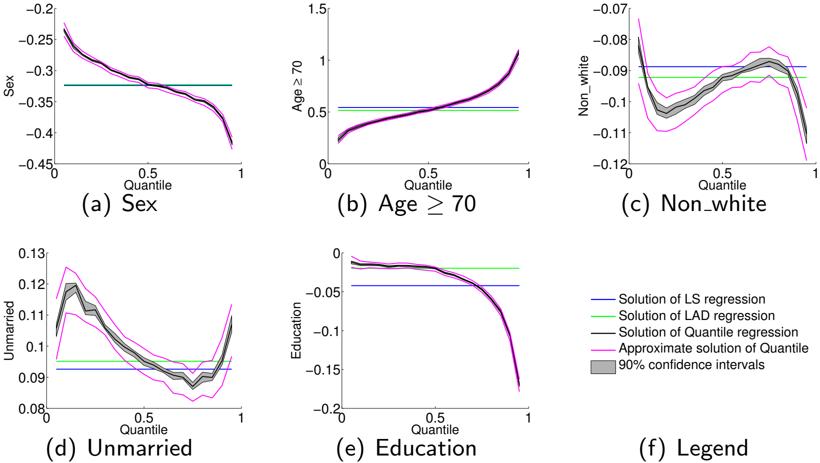
Running time when the lower dimension *d* changes



(a) $\tau = 0.5$ (b) $\tau = 0.75$ (c) $\tau = 0.95$

Figure: The running time for five methods for solving simulated problem, with $n = 1e6$, when d varies.

Plots for real data



(a) Sex

(b) Age $\geq$ 70

(c) Non_white

(d) Unmarried

(e) Education

(f) Legend

Figure: Each subfigure is associated with a coefficient in the census data. The two magenta curves show the first and third quartiles of solutions obtained

Obvious/non-obvious extensions/improvements

- Extend input-sparsity ideas to get improved low-rank approximation of arbitrary $n \times d$ matrices A, with $n \approx d$ and rank parameter k.
    - The $1 \pm \epsilon$ input-sparsity time algorithm for ℓ_2 regression implies $1 \pm \epsilon$ low-rank approximation for a weak (i.e., on the error w.r.t. only the top part of the spectrum) notion of approximation [DMM08,CW12].
    - Getting finer bounds (e.g., exactly k columns, projecting onto $k + 10$ columns, etc.) requires finer control on how the top and bottom part of the spectrum interact (in the sketched space).
    - Genetics, astronomy, numerical, etc. applications/implementations of these ideas typically use these finer notions of low-rank approximation.
- Extend input-sparsity ideas to get good statistical properties, e.g., by exploiting sparsity tradeoffs of [NN12].
- Go beyond linear regression and extend input-sparsity ideas to, e.g., logistic regression and other convex optimization problems.
- Understand connections of input-sparsity ideas with SGD, coordinate descent, etc., at both "large" scale and "LARGE" scale.

Conclusion

- Extremely sparse random projection algorithms for least-squares, least absolute deviations, quantile, etc. regression problems that underlie many common matrix algorithms
- Key step is constructing a succinct data representation that provides a low-distortion embedding
- Run in input-sparsity or nearly-input-sparsity time (plus the time for solving a subproblem whose size depends only on the lower dimension of the input matrix) in an idealized theoretical model
- Implementations in RAM (and parallel/distributed environments) perform well and illustrate tradeoffs, e.g., between running time to construct embedding and distortion quality of that embedding
- Many interesting algorithmic/statistical questions (about both "large" and "LARGE" data) raised by these results . . .