



Sampling algorithms for l_2 regression and applications

Michael W. Mahoney

Yahoo Research

<http://www.cs.yale.edu/homes/mmahoney>

(Joint work with P. Drineas and S. (Muthu) Muthukrishnan)

SODA 2006



Regression problems

$$\begin{aligned} Z_2 &= \min_{x \in \mathbb{R}^d} \|b - Ax\|_2 \\ &= \|b - A\hat{x}\|_2 \end{aligned}$$

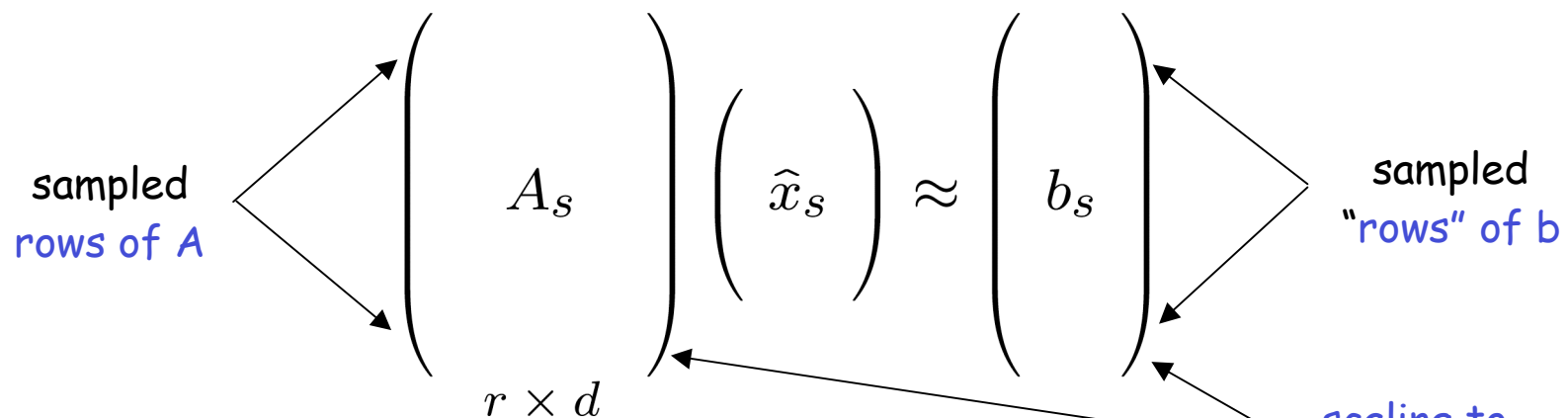


$$\begin{pmatrix} A \\ n \times d, n \gg d \end{pmatrix} \begin{pmatrix} \hat{x} \end{pmatrix} \approx \begin{pmatrix} b \end{pmatrix}$$

We seek **sampling-based algorithms** for solving l_2 regression.

We are interested in overconstrained problems, $n \gg d$. Typically, there is no x such that $Ax = b$.

"Induced" regression problems



$$\begin{aligned} \mathcal{Z}_{2,s} &= \min_{x \in \mathbb{R}^d} \|b_s - A_s x\|_2 \\ &= \|b_s - A_s \hat{x}_s\|_2 \end{aligned}$$

$$|\mathcal{Z}_2 - \mathcal{Z}_{2,s}| \leq ?$$

$$\|\hat{x} - \hat{x}_s\|_2 \leq ?$$

$$\|A \hat{x}_s - b\|_2 \leq ?$$



Regression problems, definition

$$\begin{aligned} Z_\xi &= \min_{x \in \mathbb{R}^d} \|b - Ax\|_\xi \\ &= \min_{x \in \mathbb{R}^d} \left(\sum_{i=1}^n |(b - Ax)_i|^\xi \right)^{1/\xi} \end{aligned} \quad \begin{cases} A \in \mathbb{R}^{n \times d} \\ b \in \mathbb{R}^n \end{cases}$$

There is work by [K. Clarkson](#) in SODA 2005 on sampling-based algorithms for l_1 regression ($\xi = 1$) for overconstrained problems.



Exact solution

$$\mathcal{Z}_2 = \min_{x \in \mathbb{R}^d} \|b - Ax\|_2 = \|b - A\hat{x}\|_2$$

$$\begin{pmatrix} A \\ n \times d, \quad n \gg d \end{pmatrix} \begin{pmatrix} \hat{x} \end{pmatrix} \approx \begin{pmatrix} b \end{pmatrix}$$

Projection of b on
the subspace
spanned by the
columns of A

$$\mathcal{Z}_2 = \|b\|_2 - \|AA^+b\|_2$$

$$\hat{x} = A^+b$$

Pseudoinverse
of A



Singular Value Decomposition (SVD)

$$\begin{pmatrix} A \end{pmatrix}_{n \times d} = \begin{pmatrix} U \end{pmatrix}_{n \times \rho} \cdot \begin{pmatrix} \Sigma \end{pmatrix}_{\rho \times \rho} \cdot \begin{pmatrix} V \end{pmatrix}_{\rho \times d}^T$$

U (V): orthogonal matrix containing the left (right) singular vectors of A .

Σ : diagonal matrix containing the singular values of A .

ρ : **rank** of A .

Computing the SVD takes $O(d^2n)$ time. The pseudoinverse of A is

$$A^+ = V \Sigma^{-1} U^T \in \mathbb{R}^{d \times n}$$



Questions ...

$$Z_2 = \min_{x \in \mathbb{R}^d} \|b - Ax\|_2 = \|b - A\hat{x}\|_2$$

Can *sampling methods* provide accurate estimates for l_2 regression?

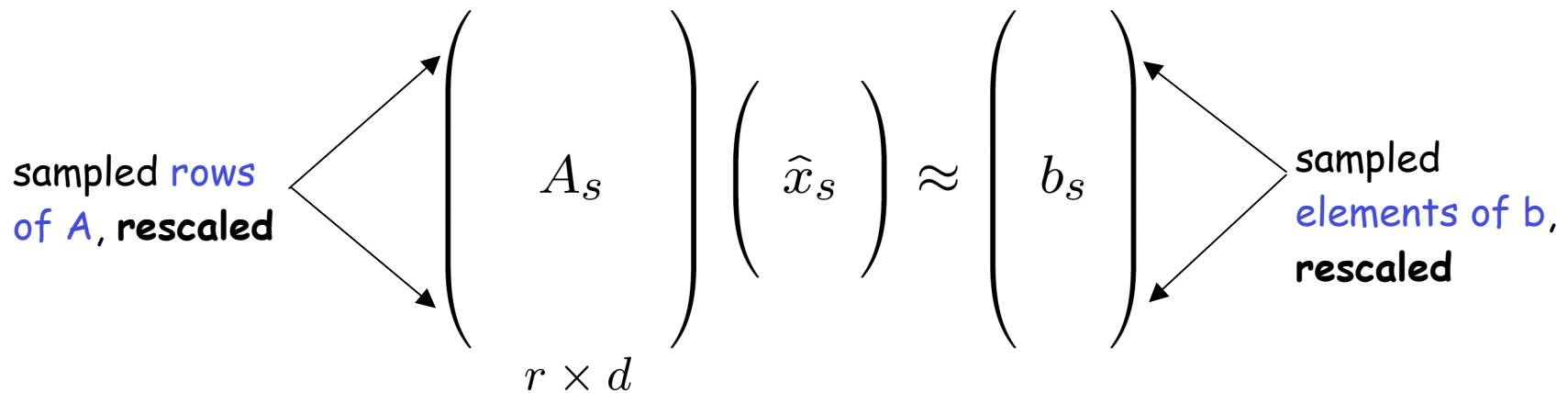
Is it possible to **approximate** the optimal vector and the optimal value Z_2 by only looking at a small **sample** from the input?

(Even if it takes some sophisticated oracles to actually perform the sampling ...)

Equivalently, is there an **induced subproblem** of the full regression problem, whose optimal solution and its value $Z_{2,s}$ approximates the optimal solution and its value Z_2 ?

The induced subproblem

$$\begin{aligned} \mathcal{Z}_{2,s} &= \min_{x \in \mathbb{R}^d} \|b_s - A_s x\|_2 \\ &= \|b_s - A_s \hat{x}_s\|_2 \end{aligned}$$



$$|\mathcal{Z}_2 - \mathcal{Z}_{2,s}| \leq ?$$

$$\|\hat{x} - \hat{x}_s\|_2 \leq ?$$

$$\|A\hat{x}_s - b\|_2 \leq ?$$



Our results

If the p_i satisfy certain **conditions**, then with probability at least $1-\delta$,

$$\mathcal{Z}_{2,s} \leq (1 + \epsilon) \mathcal{Z}_2$$

$$\mathcal{Z}_2 \leq \|A\hat{x}_s - b\|_2 \leq (1 + \epsilon) \mathcal{Z}_2$$

$$\|\hat{x} - \hat{x}_s\|_2 \leq \frac{\epsilon}{\sigma_{\min}(A)} \mathcal{Z}_2$$

The sampling complexity is

$$r = O\left(d^2 \log(1/\delta)/\epsilon^2\right)$$



Our results, cont'd

If the p_i satisfy certain **conditions**, then with probability at least $1-\delta$,

$$\mathcal{Z}_{2,s} \leq (1 + \epsilon) \mathcal{Z}_2$$

$$\mathcal{Z}_2 \leq \|A\hat{x}_s - b\|_2 \leq (1 + \epsilon) \mathcal{Z}_2$$

$$\|\hat{x} - \hat{x}_s\|_2 \leq \epsilon \left(\frac{\kappa(A)}{\gamma} \right) \|\hat{x}\|_2$$

$\kappa(A)$: condition
number of A

$$\gamma = \|AA^+b\|_2 / \|b\|_2$$

The sampling complexity is

$$r = O\left(d^2 \log(1/\delta) / \epsilon^2\right)$$

Back to induced subproblems ...

$$Z_{2,s} = \min_{x \in \mathbb{R}^d} \|b_s - A_s x\|_2 = \|b_s - A_s \hat{x}_s\|_2$$

sampled rows of A , rescaled

$$\begin{pmatrix} A_s \\ O(d^2) \times d \end{pmatrix} \begin{pmatrix} \hat{x}_s \end{pmatrix} \approx \begin{pmatrix} b_s \\ O(d^2) \times 1 \end{pmatrix}$$

sampled elements of b , rescaled

The relevant information for l_2 regression if $n \gg d$ is contained in an induced subproblem of size $O(d^2)$ -by- d .

(upcoming writeup: we can reduce the sampling complexity to $r = O(d)$.)



Conditions on the probabilities, SVD

$$\begin{pmatrix} A \end{pmatrix}_{n \times d} = \begin{pmatrix} U \end{pmatrix}_{n \times \rho} \cdot \begin{pmatrix} \Sigma \end{pmatrix}_{\rho \times \rho} \cdot \begin{pmatrix} V \end{pmatrix}_{\rho \times d}^T$$

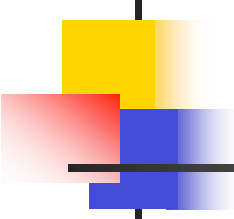
U (V): orthogonal matrix containing the left (right) singular vectors of A .

Σ : diagonal matrix containing the singular values of A .

ρ : **rank** of A .

Let $U_{(i)}$ denote the i -th row of U .

Let $U^\perp \in \mathbb{R}^{n \times (n-\rho)}$ denote the **orthogonal complement** of U .



Conditions on the probabilities, interpretation

What do the lengths of the rows of the $n \times d$ matrix $U = U_A$ "mean"?

Consider possible $n \times d$ matrices U of d left singular vectors:

$I_n|_k = k$ columns from the identity

row lengths = 0 or 1

$I_n|_k \times \rightarrow x$

$H_n|_k = k$ columns from the $n \times n$ Hadamard (real Fourier) matrix

row lengths all equal

$H_n|_k \times \rightarrow$ maximally dispersed

$U_k = k$ columns from any orthogonal matrix

row lengths between 0 and 1

The lengths of the rows of $U = U_A$ correspond to a notion of information dispersal

Conditions for the probabilities

The conditions that the p_i must satisfy, for some $\beta_1, \beta_2, \beta_3 \in (0,1]$:

$$p_i \geq \beta_1 \frac{\|U_{(i)}\|_2^2}{\sum_{j=1}^n \|U_{(j)}\|_2^2}$$

lengths of **rows** of
matrix of **left**
singular vectors of A

$$p_i \geq \beta_2 \frac{\|U_{(i)}\| \left((U^\perp U^{\perp T} b)_i \right)}{\sum_{j=1}^n \|U_{(j)}\| \left((U^\perp U^{\perp T} b)_j \right)}$$

Component of b
not in the span of
the **columns** of A

$$p_i \geq \beta_3 \frac{\left((U^\perp U^{\perp T} b)_i \right)^2}{\sum_{j=1}^n \left((U^\perp U^{\perp T} b)_j \right)^2}$$

Small β_i)
more sampling

The sampling complexity is: $r = O \left(d^2 \log(1/\delta) / \left(\epsilon^2 \min \{ \beta_1^2, \beta_2^2, \beta_3^2 \} \right) \right)$



Computing “good” probabilities

In $O(nd^2)$ time we can easily compute p_i 's that satisfy all three conditions, with $\beta_1 = \beta_2 = \beta_3 = 1/3$.

(Too expensive in practice for this problem!)

Open question: can we compute “good” probabilities faster, in a pass efficient manner?

Some assumptions might be acceptable (e.g., bounded condition number of A , etc.)



Critical observation

$$\mathcal{Z}_2 = \min_{x \in \mathbb{R}^d} \|b - Ax\|_2 = \|b - A\hat{x}\|_2$$

sample & rescale \rightarrow

$$\begin{pmatrix} A \\ n \times d, \quad n \gg d \end{pmatrix} \begin{pmatrix} \hat{x} \end{pmatrix} \approx \begin{pmatrix} b \end{pmatrix} \leftarrow \text{sample \& rescale}$$



Critical observation, cont'd

$$\mathcal{Z}_2 = \min_{x \in \mathbb{R}^d} \|b - Ax\|_2 = \|b - A\hat{x}\|_2$$

sample & rescale only U \rightarrow

$$\begin{pmatrix} U \end{pmatrix} \cdot \begin{pmatrix} \Sigma \end{pmatrix} \cdot \begin{pmatrix} V \end{pmatrix}^T \begin{pmatrix} \hat{x} \end{pmatrix} \approx \begin{pmatrix} b \end{pmatrix}$$

sample & rescale \leftarrow



Critical observation, cont'd

$$\mathcal{Z}_{2,s} = \min_{x \in \mathbb{R}^d} \|b_s - A_s x\|_2 = \|b_s - A_s \hat{x}_s\|_2$$

$$\begin{pmatrix} U_s \end{pmatrix} \cdot \begin{pmatrix} \Sigma \end{pmatrix} \cdot \begin{pmatrix} V \end{pmatrix}^T \begin{pmatrix} \hat{x}_s \end{pmatrix} \approx \begin{pmatrix} b_s \end{pmatrix}$$

Important observation: U_s is almost orthogonal, and we can bound the spectral and the Frobenius norm of

$$U_s^T U_s - I.$$

(FKV98, DK01, DKM04, RV04)

Application: CUR-type decompositions

Create an approximation to A , using rows and columns of A

$$\begin{pmatrix} A \end{pmatrix} \approx \begin{pmatrix} C \end{pmatrix} \cdot \begin{pmatrix} U \end{pmatrix} \cdot \begin{pmatrix} R \end{pmatrix}$$

Carefully chosen U

$O(1)$ columns

$O(1)$ rows

Goal: provide (good) bounds for some norm of the error matrix

$A - CUR$

1. How do we draw the rows and columns of A to include in C and R ?
2. How do we construct U ?