Statistical and Algorithmic Perspectives of Randomized Sketching Algorithms

Garvesh Raskutti (University of Wisconsin-Madison)
joint with Michael Mahoney

IHES, 2016
Paris, France

March 2016

Sketching

- Sketching one of most widely used data reduction techniques
- Based on properties of random sub-sampling/projections
- Ordinary least-squares (Drineas and Mahoney '11)
- CUR decomposition (Drineas and Mahoney '06)
- Spectral sparsification (Spielman and Teng '08)

Two Perspective on Sketching

Algorithmic:

- Achieve optimal worst-case error bounds up to constant.
- You gain both in terms of worst-case error and data reduction.

Statistical (intuitive):

- Effectively throwing away most of your data.
- How are we not losing something?

In this talk: We attempt to unify and explain these two perspectives on sketching in the context of ordinary least-squares.

Sketching for Large-scale Least-squares

Ordinary least-squares estimator, (X, Y), $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$ ($\text{rank}(X) = p$):

$$\beta_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

Often both n and p extremely large. Least-squares is computationally intensive.

One possible solution: Reduce data and perform least-squares. $S \in \mathbb{R}^{r \times n}$ where $r \ll n$, reduced data (SX, SY) and estimator:

$$\beta_S \in \arg\min_{\beta \in \mathbb{R}^p} \|SY - SX\beta\|_2^2.$$

Question: For what choices of S is β_S is a "good" approximation to β_{OLS}.

Prior work: Algorithmic Perspective

Assume $Y \in \mathbb{R}^n$ arbitrary and fixed (no distribution assumed) with known and fixed X.

Criterea: Minimize worst-case realtive approximation error:

$$C_{WC}(S) = \sup_{Y} \frac{\|Y - X\beta_S\|_2^2}{\|Y - X\beta_{OLS}\|_2^2}.$$

Drineas et al. 2011 prove that $C_{WC}(S) \leq 1 + \delta$ for some $\delta \in (0, 1)$ with high probability.

S involves random projection or leverage-score sampling.

Supremum taken over Y.

Statistical Perspective

Assume Y generated by the following linear model:

$$Y = X\beta + \epsilon,$$

where $\beta \in \mathbb{R}^p$, $\epsilon \in \mathbb{R}^n$ where $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon\epsilon^T] = I_{n\times n}$.

Performance metrics: Inverse residual efficiency and prediction efficiency:

$$\begin{aligned} C_{RE}(S) &= \frac{\mathbb{E}[\|Y - X\beta_S\|_2^2]}{\mathbb{E}[\|Y - X\beta_{OLS}\|_2^2]} \\ C_{PE}(S) &= \frac{\mathbb{E}[\|X\beta_S - X\beta\|_2^2]}{\mathbb{E}[\|X\beta_{OLS} - X\beta\|_2^2]}. \end{aligned}$$

Goals: (1) Determine good sketching schemes S in terms of above criterea; (2) Provide comparison between $C_{RE}(S)$, $C_{PE}(S)$ and $C_{WC}(S)$.

Randomized Sketching Schemes

- Leverage score-based sampling with re-scaling S_R and no re-scaling S_{NR}.
- Sub-Gaussian projection S_{SGP} and Hadamard projection S_{Had}.

Sampling-based Sketching

- Leverage score-based sampling with re-scaling S_R and no re-scaling S_{NR}.

Singular value decomposition for X:

$$X = \underbrace{U}_{n\times p}\underbrace{\Sigma}_{p\times p}\underbrace{V^T}_{p\times p},$$

where $U^T U = I_p$.
$\ell_i = \|U_{(i)}\|_2^2$ for $1 \leq i \leq n$.
Leverage-score sampling equivalent to row-norm sampling of U.
Note: $\sum_{i=1}^n \ell_i = p$, and $0 \leq \ell_i \leq 1$.
Leverage widely used in robust statistics. High leverage points are out-liers.
Here we want to include high-leverage points.
Fast computation of approximate leverage scores possible.

Comaprison of Metrics

$\Pi_U^S := U(SU)^\dagger S$ (oblique projection matrix). Then $X\beta_S = \Pi_U^S Y$.

Lemma (R. and Mahoney ICML '15)

If $rank(SX) = p$, then

$$
\begin{aligned}
C_{WC}(S) &= 1 + \sup_{\epsilon \in \mathbb{R}^n, U^T\epsilon = 0} \frac{\|\Pi_U^S \epsilon\|_2^2}{\|\epsilon\|_2^2}, \\
C_{PE}(S) &= \frac{\|\Pi_S^U\|_F^2}{p}, \\
C_{RE}(S) &= 1 + \frac{C_{PE}(S) - 1}{n/p - 1}.
\end{aligned}
$$

From statistical perspective, need to control $\|\Pi_S^U\|_F^2$.
Note $C_{PE}(S) \asymp \frac{n}{p} C_{RE}(S)$.

Upper Bounds

Theorem (R. and Mahoney ICML '15)

For $S = S_R,\ S_{SGP},\ S_{Had}$,

$$\begin{aligned} C_{WC}(S) &\leq 1 + 12\frac{p}{r}, \\ C_{RE}(S) &\leq 1 + 44\frac{p}{r}, \\ C_{PE}(S) &\leq 44\frac{n}{r}, \end{aligned}$$

with high probability.

Note for $C_{WC}(S)$ and $C_{RE}(S)$, $r = \Omega(p)$ is sufficient whereas for $C_{PE}(S)$, we need $r = \Omega(n)$.

Non-uniform Leverage scores

Assume k leverage scores contain most of the mass.

Definition (k-heavy hitter distribution)

A sequence of leverage scores $\ell_1 \geq \ell_2 \geq ... \geq \ell_n$ is a *k-heavy hitter* distribution if there exist constants $c, C > 0$ such that for $1 \leq i \leq k$, $\frac{cp}{k} \leq \ell_i \leq \frac{Cp}{k}$ and $\sum_{i=k+1}^{n} \ell_i \leq \frac{3}{4}$.

Theorem (R. and Mahoney ICML '15)

$$\begin{aligned} C_{WC}(S_{NR}) &\leq 1 + \frac{44C^2}{c^2}\frac{p}{r}, \\ C_{RE}(S_{NR}) &\leq 1 + \frac{44C^4}{c^2}\frac{pk}{nr}, \\ C_{PE}(S_{NR}) &\leq \frac{44C^4}{c^2}\frac{k}{r}, \end{aligned}$$

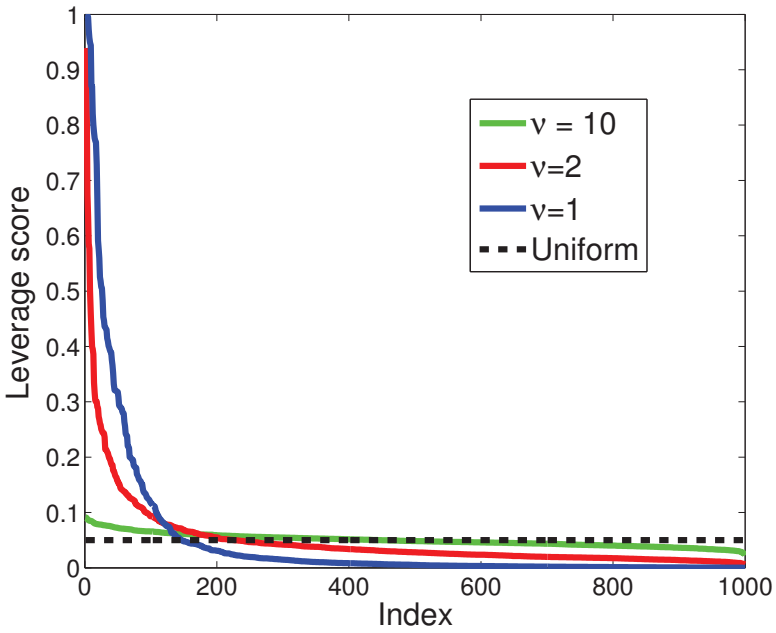
with high probability.

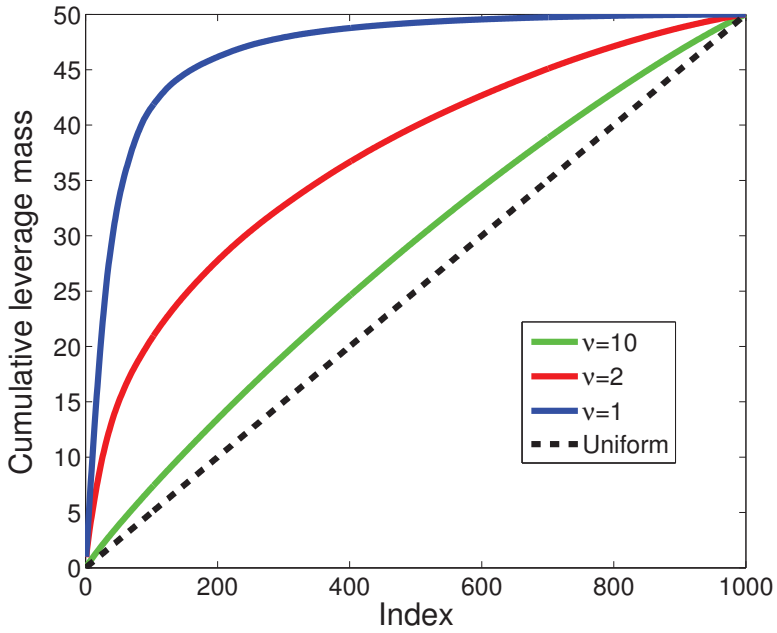
For non-uniform leverage scores, S_{NR} has better performance.

Empirical Setup

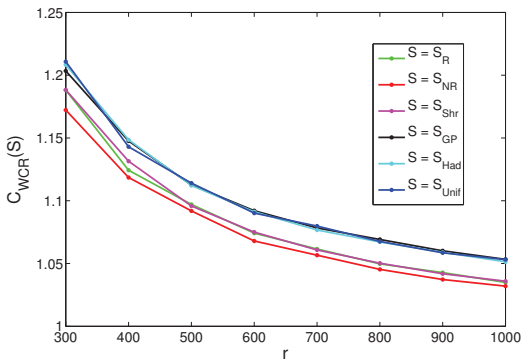
- Setup following Ma, Mahoney and Yu, 2015.
- $n = 1000$, $p = 50$, r varied Repated 100 times and averaged.
- Each row of X generated randomly using a multivariate t distribution, where $[\Sigma]_{ij} = 2 \times 0.5^{|i-j|}$, and ν degrees of freedom, $\nu = 1, 2, 10$.
- Different ν correspond to different uniformity of leverage score.
- Sketching schemes:
  - $S = S_U$ (uniform sampling)
  - $S = S_R$ (random leverage-score sampling)
  - $S = S_{NR}$ (random leverage-score sampling with re-scaling)
  - $S = S_{Shr}$ (Faster approximate leverage score-sampling)
  - $S = S_{GP}$ (Gaussian projection)
  - $S = S_{Had}$ (Hadamard projection)

Leverage Score Distributions

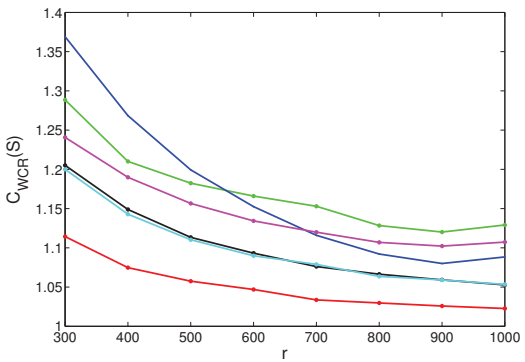




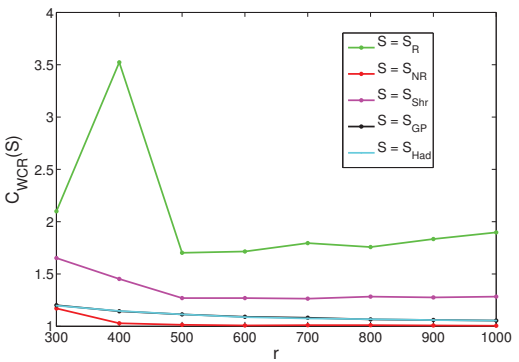
Empirical results: Algorithmic Perspective



(a) $\nu = 10$



(b) $\nu = 2$

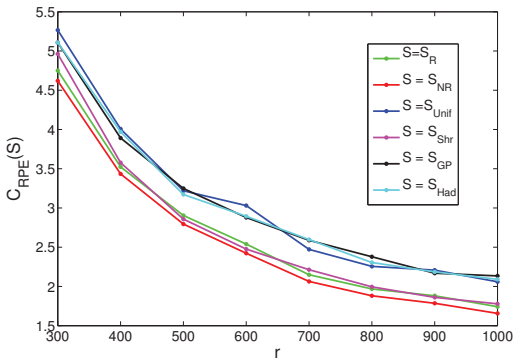


(c) $\nu = 1$

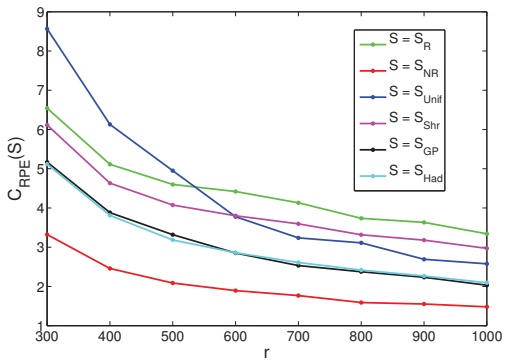
Figure : Comparison of $C_{WC}(S)$ (1 for OLS estimator) for determistic and random sampling.

Note: $C_{WC}(S)$ close to 1.

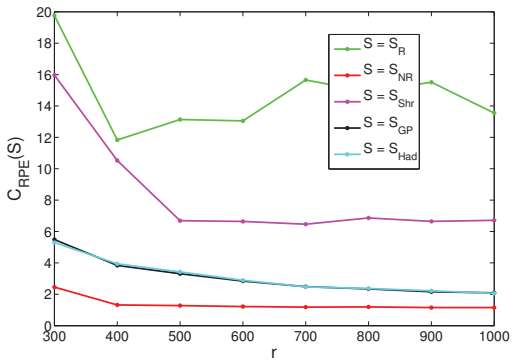
Empirical results: Statistical Perspective



(a) $\nu = 10$



(b) $\nu = 2$



(c) $\nu = 1$

Figure : Comparison of $C_{PE}(S)$ (1 for OLS estimator) for determistic and random sampling.

Note: $C_{PE}(S)$ much greater than 1.
$S = S_R$ performs poorly for non-uniform leverage scores whereas $S = S_{NR}$ performs well.

Lower Bound

- Subsequent work due to Pilanci and Wainwright, 2014 provides a general lower bound on $C_{PE}(S)$.

Theorem (Pilanci and Wainwright, 2014)

For any S where $\mathbb{E}[\|S^T(SS^T)^{-1}S\|_{op}] \leq \eta\frac{r}{n}$,

$$C_{PE}(S) \geq \frac{n}{128\eta r},$$

with probability greater than $1/2$.

Hence our upper bounds are sharp.
$C_{PE}(S)$ is an intrinsically harder metric that $C_{RE}(S)$ and $C_{WC}(S)$.
$S = S_{NR}$ does not satisfy assumption or lower bound.

Conclusions and Future Directions

- Conclusions:
  - Analysis shows statistical criteria $C_{PE}(S)$ is more challenging than standard algorithmic criteria $C_{WC}(S)$.
  - Leverage-score sampling without re-scaling ($S = S_{NR}$) has potential benefits, especially when leverage scores are non-uniform.
- Future directions:
  - Statistical persepctive of CUR decomposition.
  - Implicit regularization perspective on sketching features.
  - Sketching for low-rank tensors.