Asymptotics for Sketching in Least Squares

Edgar Dobriban[1] Sifan (Sophia) Liu[2] [1]University of Pennsylvania, Wharton Statistics [2]Stanford Statistics

Introduction

- **Least squares regression problem.** Data generated from a linear model. Goal: learn the regression parameters.
- **"Sketch-and-solve" methods:** randomly project the data first, then do regression.
- Our work: find the accuracy loss (for estimation and test error) of popular sketching methods.

Problem setup

- Data generating model: $Y = X\beta + \varepsilon$, $X \in \mathbb{R}^{n\times p}$, $n > p$.
- Sketching matrix $S \in \mathbb{R}^{r\times n}$.
- Sketched data $(\tilde{X}, \tilde{Y}) = (SX, SY)$.
- Least squares estimators before and after sketching

$$\hat\beta = (X^\top X)^{-1}X^\top Y,$$
$$\hat\beta_s = (\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top \tilde{Y}.$$

Sketching methods

- **Uniform sampling**: sample each row with the same probability
- **Leverage score sampling**: sample the rows according to leverage scores $h_{ii} = x_i^\top (X^\top X)^{-1} x_i$
- **Random projection**: the entries of S are i.i.d.
- **Haar projection:** S are uniformly distributed on the manifold of $r \times n$ ortho-mxes S, $SS^\top = I_r$.
- **Subsampled randomized Hadamard transform (SRHT)**: $S = BHDP$, where $B \in \mathbb{R}^{n\times n}$ is a diagonal sampling matrix, $H \in \mathbb{R}^{n\times n}$ is a Hadamard (or Fourier) matrix, $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix of iid random variables equal to ± 1 with probability one half, and $P \in \mathbb{R}^{n\times n}$ is a uniformly distributed permutation matrix.

Error criteria

Mean Squared Error (MSE) of an estimator is $\mathbb{E}\|\hat\beta - \beta\|^2$. How much does the MSE increase after sketching? Given by the following efficiencies:

$$VE = \frac{\mathbb{E}\|\hat\beta_s - \beta\|^2}{\mathbb{E}\|\hat\beta - \beta\|^2}, \quad PE = \frac{\mathbb{E}\|X\hat\beta_s - X\beta\|^2}{\mathbb{E}\|X\hat\beta - X\beta\|^2}, \quad RE = \frac{\mathbb{E}\|Y - X\hat\beta_s\|^2}{\mathbb{E}\|Y - X\hat\beta\|^2}, \quad OE = \frac{\mathbb{E}(x_t^\top\hat\beta_s - y_t)^2}{\mathbb{E}(x_t^\top\hat\beta - y_t)^2}.$$

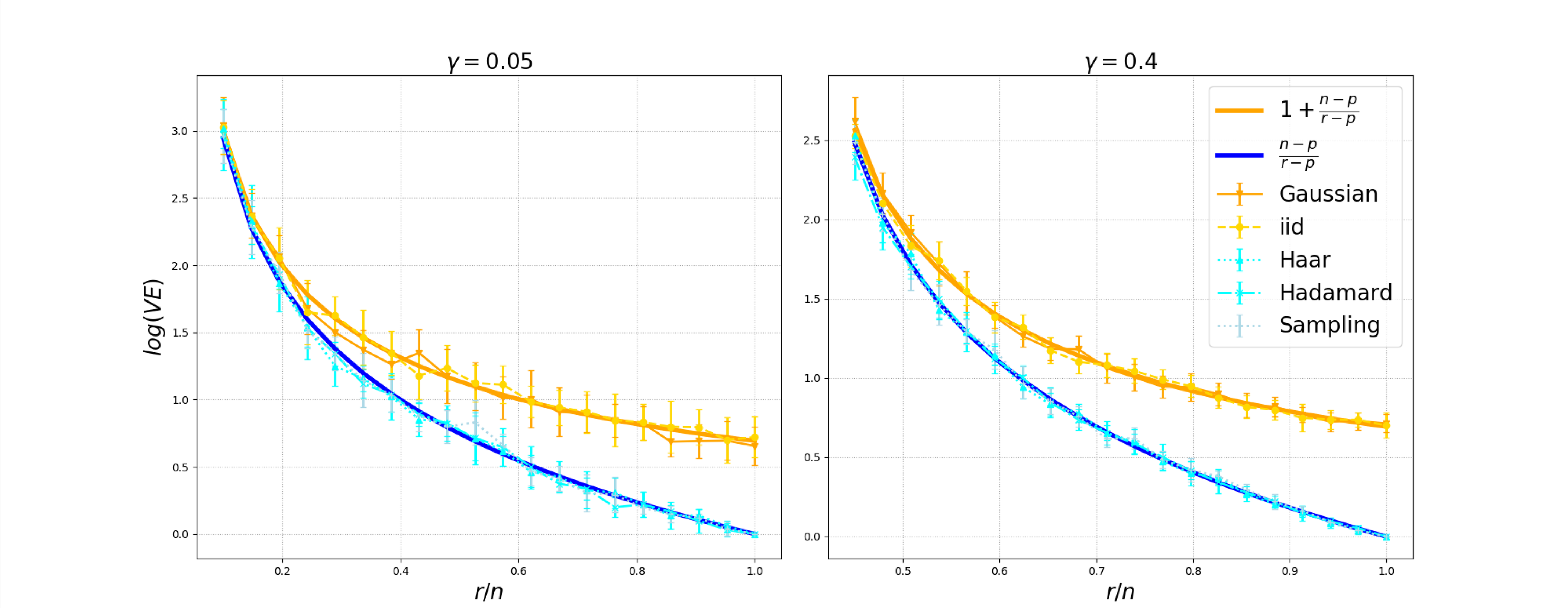
(Variance efficiency) *(Prediction efficiency)* *(Residual efficiency)* *(Out-of-sample efficiency)*

Theoretical results

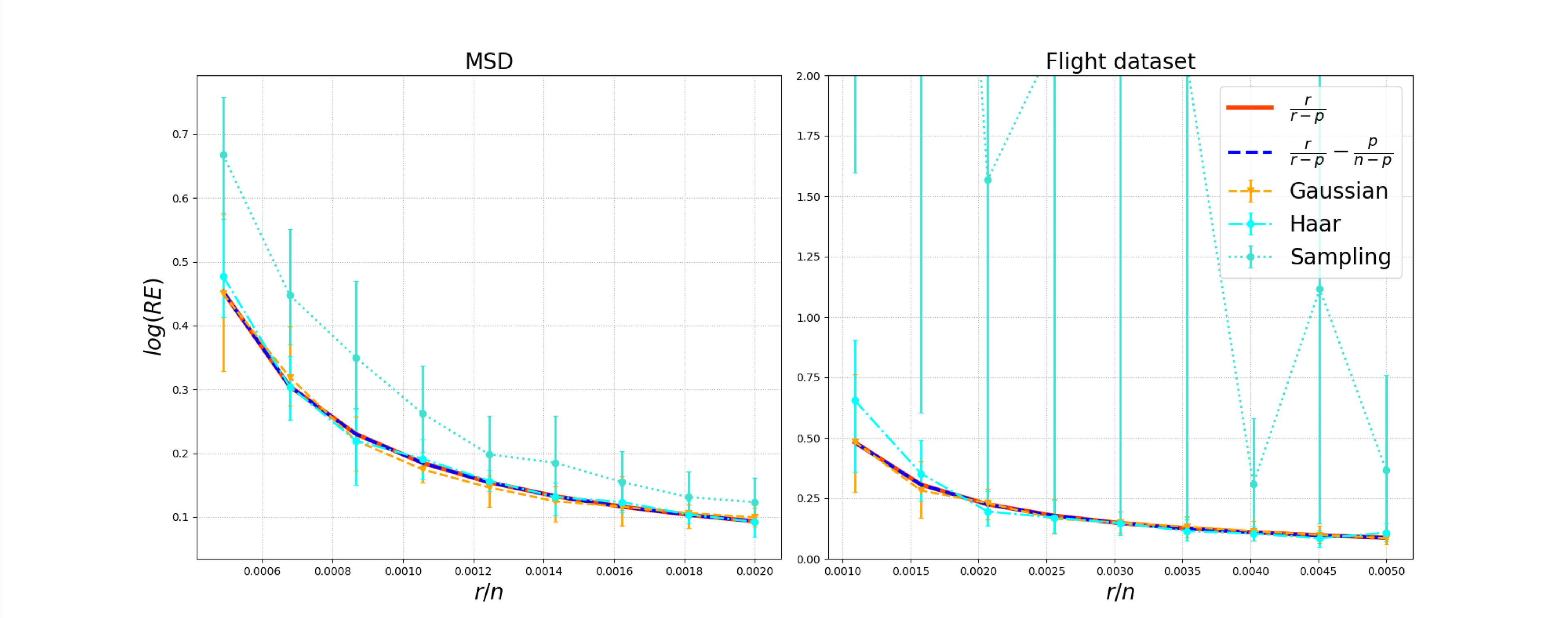
Assumption on X	Arbitrary	Arbitrary	Ortho-invariant	Elliptical: $WZ\Sigma^{1/2}$
Assumption on S	iid entries	Haar/Hadamard	Uniform sampling	Leverage sampling
VE	$1+\frac{n-p}{r-p}$	$\frac{n-p}{r-p}$	$\frac{n-p}{r-p}$	$\frac{\eta^{-1}_{sw^2}(1-p/n)}{\eta^{-1}_{w^2}(1-p/n)}$
PE	$1+\frac{n-p}{r-p}$	$\frac{n-p}{r-p}$	$\frac{n-p}{r-p}$	$1+\mathbb{E}[w^2(1-s)]\frac{\eta^{-1}_{sw^2}(1-p/n)}{p/n}$
OE	$\frac{nr-p^2}{n(r-p)}$	$\frac{r(n-p)}{n(r-p)}$	$\frac{r(n-p)}{n(r-p)}$	$\frac{1+\mathbb{E}w^2\eta^{-1}_{sw^2}(1-\gamma)}{1+\mathbb{E}w^2\eta^{-1}_{w^2}(1-\gamma)}$

For instance, when X is arbitrary and S is a matrix with iid entries, the estimation error increases by $1+(n-p)/(r-p)$ due to sketching. E.g., $n = 10^6$, $p = 10^5$, $r = 5\cdot 10^5$, then $RE = 3.25$

Numerical results



Above: simulation results for $\gamma = p/n = 0.05$ (left), and 0.4 (right). Showing SD over 10 trials of Gaussian, iid, Haar, Hadamard sketching, and sampling. Below: empirical data analysis for Million Song dataset (left) and Flight dataset (right).



Contributions

- **Accurate** and **easy to use** formulas
- **Separation** between sketching methods: Hadamard is better than Gaussian
- **Tradeoff** between computation and statistical accuracy

Acknowledgments

Thanks to K Clarkson, M Lopes, M Mahoney, M Pilanci, G Raskutti, D Woodruff. Support: NSF BIGDATA IIS 1837992, Tsinghua Summer Research award. arXiv: `arxiv.org/abs/1810.06089`.. Code `github.com/liusf15/Sketching-lr`

References

[1] Garvesh Raskutti and Michael W Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

[2] Petros Drineas and Michael W Mahoney. RandNLA: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

[3] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[4] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Contact Information

- dobriban@wharton.upenn.edu
- sfliu@stanford.edu