Least Squares Estimation Using Sketched Data with Heteroskedastic Errors

ICML 2022

Sokbae (Simon) Lee, Columbia University

Serena Ng, Columbia University and NBER

Sketching

- Want a sketch $\widetilde{A} = \Pi A \in \mathbb{R}^{m \times d}$ that preserves features of $A \in \mathbb{R}^{n \times d}$
- Aim to have a much smaller m than n
- Random Sampling:
  - rows in $\widetilde{A}$ are rows of A.
  - e.g., Bernoulli sampling; uniform sampling w or w/o replacement; leverage score sampling
- Random Projections:
  - rows in $\widetilde{A}$ are linear combinations of rows of A.
  - e.g., Gaussian; SRHT; Countsketch

Literature: Algorithmic Perspective

- The early works of Sarlos (2006), Drineas, Mahoney, and Muthukrishnan (2006) and Drineas, Mahoney, Muthukrishnan, and Sarlos (2011) consider sketching of the least squares estimator from an algorithmic perspective (worst case analysis with the fixed data).
- See, e.g., Woodruff (2014), Drineas and Mahoney (2018) and Martinsson and Tropp (2020) for a review.

Literature: Statistical Perspective

- However, recent works due to Ma, Mahoney, and Yu (2015), Raskutti and Mahoney (2016), and Dobriban and Liu (2019) show that an optimal worse-case error may not yield an optimal mean-squared error.
- This led to interest in better understanding the statistical implications of sketching. For example, Geppert, Ickstadt, Munteanu, Qudedenfeld, and Sohler (2017) considers Bayesian estimation while Ahfock, Astle, and Richardson (2020) and Ma, Zhang, Xing, Ma, and Mahoney (2020) provide asymptotic distribution theory for the sketched least squares estimators under homoskedasticity.

Regression Model

- Given i.i.d. observations $\{(y_i, X_i) : i = 1, \ldots, n\}$, we consider a linear regression model:

$$y = X\beta_0 + e, \ \mathbb{E}(Xe) = 0$$

- $\sqrt{n}(\widehat{\beta}_{OLS} - \beta_0) \to_d N(0, V_1)$ as $n \to \infty$, where V_1 is the sandwich variance defined as

$$V_1 := [\mathbb{E}(X_i X_i^T)]^{-1} \mathbb{E}(e_i^2 X_i X_i^T)[\mathbb{E}(X_i X_i^T)]^{-1}.$$

- Under homoskedasticity, V_1 becomes

$$V_0 := \mathbb{E}(e_i^2)[\mathbb{E}(X_i X_i^T)]^{-1}.$$

Sketched OLS

- A sketch of the data (y, X) is $(\widetilde{y}, \widetilde{X})$, where $\widetilde{y} = \Pi y$, $\widetilde{X} = \Pi X$, and Π is usually an $m \times n$ random matrix.
- The sketched least squares estimator is $\widetilde{\beta}_{OLS} := (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \widetilde{y}$.

Assumption

(i) *The data* $\mathcal{D}_n := \{(y_i, X_i) \in \mathbb{R}^{1+p} : i = 1, \ldots, n\}$ *are independent and identically distributed (i.i.d.). Furthermore,* X *has singular value decomposition* $X = U_X \Sigma_X V_X^T$.

(ii) $\mathbb{E}(y_i^4) < \infty$, $\mathbb{E}(\|X_i\|^4) < \infty$, *and* $\mathbb{E}(X_i X_i^T)$ *has full rank* p.

(iii) *The random matrix* Π *is independent of* $\mathcal{D}_n$.

(iv) $m = m_n \to \infty$ *but* $m/n \to 0$ *as* $n \to \infty$, *while* p *is fixed.*

Two Leading Examples

For simplicity, we focus on Bernoulli sampling (BS) from Random Sampling and Countsketch (CS) from Random Projections.

- Bernoulli sampling (BS): $\Pi = \sqrt{\frac{n}{m}} B$, where B is a diagonal sampling matrix of i.i.d. Bernoulli random variables with success probability m/n.
- Countsketch (CS) : only one non-zero entry in each column of Π. The non-zero entry takes on value $\{+1, -1\}$ randomly drawn with equal probability, and is located uniformly at random for each column.

Theorem (OLS)

Let Assumption 1 hold and $\mathbb{E}(e_i X_i) = 0$.

(i) *Under BS,* $m^{1/2}(\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS}) \to_d N(0, V_1)$.

(ii) *Under CS,* $m^{1/2}(\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS}) \to_d N(0, V_0)$.

Theorem 1 indicates that both sampling schemes yield asymptotically normal estimates, but for different reasons have different asymptotic variances.

Practical Inference

- In applications, researchers would like to test a hypothesis about β_0 using a sketched estimate, and our results provide all the quantities required for inference.
- Since $m/n \to 0$,

$$
\begin{aligned}
m^{1/2}(\widetilde{\beta} - \widehat{\beta}\,) &= m^{1/2}(\widetilde{\beta} - \beta_0) - (m/n)^{1/2}\, n^{1/2}(\widehat{\beta} - \beta_0) \\
&= m^{1/2}(\widetilde{\beta} - \beta_0) + o_p(1).
\end{aligned}
$$

- Then, asymptotic normality of $m^{1/2}(\widetilde{\beta} - \widehat{\beta})$ provides a guide to conduct inference for β_0:

$$
\widetilde{V}_m^{-1/2}(\widetilde{\beta}_{OLS} - \beta_0) \approx N(0, I_p),
$$

where the form of $\widetilde{V}_m$ will be given below.

Monte Carlo Experiments

	(1)	(2)	(3)	(4)
	Size		Power	
	s.e.0	s.e.1	s.e.0	s.e.1
(i) Homoskedastic Design				
bernoulli	0.046	0.050	0.490	0.496
uniform	0.047	0.052	0.489	0.490
leverage	0.045	0.053	0.483	0.513
countsketch	0.049	0.051	0.479	0.489
srht	0.056	0.061	0.492	0.498
srft	0.055	0.057	0.484	0.489
(ii) Heteroskedastic Design				
bernoulli	0.310	0.047	0.713	0.436
uniform	0.301	0.053	0.719	0.435
leverage	0.183	0.051	0.727	0.529
countsketch	0.054	0.057	0.813	0.812
srht	0.054	0.056	0.804	0.809
srft	0.050	0.052	0.799	0.806

References i

Ahfock, D. C., W. J. Astle, and S. Richardson (2020): "Statistical properties of sketching algorithms," *Biometrika*, 108(2), 283–297.

Dobriban, E., and S. Liu (2019): "Asymptotics for sketching in least squares," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 3675–3685.

Drineas, P., M. Mahoney, and S. Muthukrishnan (2006): "Sampling Algorithms for L2 Regression and Applications," *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136.

Drineas, P., M. Mahoney, S. Muthukrishnan, and T. Sarlos (2011): "Faster Least Squares Approximation," *Numerical Mathematics*, 117, 219–249.

DRINEAS, P., AND M. W. MAHONEY (2018): "Lectures on randomized numerical linear algebra," in *The Mathematics of Data*, ed. by M. W. Mahoney, J. C. Duchi, and A. C. Gilbert, pp. 1–48. AMS/IAS/SIAM.

GEPPERT, L., K. ICKSTADT, A. MUNTEANU, J. QUDEDENFELD, AND C. SOHLER (2017): "Random Projections for Bayesian Regressions," *Statistical Computing*, 27:, 79–101.

MA, P., M. W. MAHONEY, AND B. YU (2015): "A statistical perspective on algorithmic leveraging," *Journal of Machine Learning Research*, 16(1), 861–911.

MA, P., X. ZHANG, X. XING, J. MA, AND M. MAHONEY (2020): "Asymptotic Analysis of Sampling Estimators for Randomized Numerical Linear Algebra Algorithms," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ed. by S. Chiappa, and R. Calandra, vol. 108 of *Proceedings of Machine Learning Research*, pp. 1026–1035. PMLR.

MARTINSSON, P.-G., AND J. A. TROPP (2020): "Randomized numerical linear algebra: Foundations and algorithms," *Acta Numerica*, 29, 403–572.

RASKUTTI, G., AND M. W. MAHONEY (2016): "A statistical perspective on randomized sketching for ordinary least-squares," *Journal of Machine Learning Research*, 17(1), 7508–7538.

SARLOS, T. (2006): “Improved Approximation Algorithms for Large Matrices via Random Projections,” *Proceedings of the 47 IEEE Symposium on Foundations of Computer Science*.

WOODRUFF, D. P. (2014): “Sketching as a tool for numerical linear algebra,” *Foundations and Trends in Theoretical Computer Science*, 10(1–2), 1–157.