Least Squares Estimation using Sketched Data with Heteroskedastic Errors

Sokbae Lee \(^1\) Serena Ng \(^1\)

Abstract

Researchers may perform regressions using a sketch of data of size \(m\) instead of the full sample of size \(n\) for a variety of reasons. This paper considers the case when the regression errors do not have constant variance and heteroskedasticity robust standard errors would normally be needed for test statistics to provide accurate inference. We show that estimates using data sketched by random projections will behave ‘as if’ the errors were homoskedastic. Estimation by random sampling would not have this property. The result arises because the sketched estimates in the case of random projections can be expressed as degenerate \(U\)-statistics, and under certain conditions, these statistics are asymptotically normal with homoskedastic variance. We verify that the conditions hold not only in the case of least squares regression when the covariates are exogenous, but also in instrumental variables estimation when the covariates are endogenous. The result implies that inference can be simpler than the full sample case if the sketching scheme is appropriately chosen.

1. Introduction

Big data sets can be costly to store and analyze, and one approach around the data bottlenecks is to work with a randomly chosen subset, or a sketch, of the data. Data privacy may also dictate that a sketch of the data be made available for public use. The early works of Sarlos (2006), Drineas et al. (2006) and Drineas et al. (2011) consider the algorithmic properties of the least squares estimator using sketched data. Subsequent work extends the analysis to ridge regression (e.g., Wang et al., 2018; Liu & Dobriban, 2020), and logistic regression (e.g., Wang, 2019). See, e.g., Woodruff (2014), Drineas & Mahoney (2018) and Martinsson & Tropp (2020) for a review. However, Ma et al. (2015), Raskuti

\(^1\)Equal contribution  \(^1\)Department of Economics, Columbia University, New York, USA, Correspondence to: Sokbae Lee <sli3841@columbia.edu>, Serena Ng <sn2294@columbia.edu>.

Proceedings of the 39\textsuperscript{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).
2. Sketched Least Squares Estimation with Heteroskedastic Errors

Given \( n \) observations \( \{(y_i, X_i, Z_i) : i = 1, \ldots, n\} \), we consider a linear regression model:
\[
y_i = X_i^T \beta_0 + e_i,
\]
where \( y_i \) is the scalar dependent variable, \( X_i \) is a \( p \times 1 \) vector of regressors, \( \beta_0 \) is a \( p \times 1 \) vector of unknown parameters. The innovation \( e_i \) is said to be (conditionally) homoskedastic if \( E[e_i^2 | X_i] = E[e_i^2] \). Otherwise, \( e_i \) is said to be heteroskedastic. The regressors are said to be exogenous if \( E[e_i, X_i] = 0 \). Otherwise it is endogenous. In that case, we assume a \( q \times 1 \) vector of instrumental variables, \( Z_i \), satisfying \( E[e_i, Z_i] = 0 \) are available. In matrix form, the model given in (1) can be written as
\[
y = X \beta_0 + e,
\]
where \( y \) and \( e \) are \( n \times 1 \) vectors whose \( i \)-th rows are \( y_i \) and \( e_i \), respectively, and \( X \) is the \( n \times p \) matrix of regressors whose \( i \)-th row is \( X_i^T \).

We first study the exogenous regressor case when \( E(e_i, X_i) = 0 \). The least squares estimator \( \hat{\beta}_{OLS} := (X^T X)^{-1} X^T y \) is \( \sqrt{n} \) consistent and asymptotically normal, i.e., \( \sqrt{n}(\hat{\beta}_{OLS} - \beta_0) \to_d N(0, V_1) \) as \( n \to \infty \), where
\[
V_1 := [E(X_i X_i^T)]^{-1} E(e_i e_i^T | X_i^T) [E(X_i X_i^T)]^{-1}
\]
is the heteroskedasticity-robust asymptotic variance. Under homoskedasticity, \( V_1 \) becomes
\[
V_0 := E(e_i^2)[E(X_i X_i^T)]^{-1}.
\]
The point estimates \( \hat{\beta} \) can be used to test hypothesis, say, \( H_0 : \beta_2 = \beta_2 \) using the \( t \) test \( \sqrt{n} \left( \hat{\beta}_2 - \beta_2 \right) / \sqrt{V_{22}} \), where \( \hat{V} \) is an estimate of either \( V_1 \) or \( V_0 \), \( \beta_2 \) is a specific element of \( \beta_0 \), and \( \sqrt{[\hat{V}]}_{22} \) the (2,2) diagonal element of \( \hat{V} \). The distribution of this test under the null hypothesis crucially depends on the correct standard error \( \sqrt{[\hat{V}]}_{22} \) being used. Using \( \sqrt{V_0} \) when the robust estimator \( \hat{V}_1 \) should have been used would lead to inaccurate inference, in the sense of rejecting the null hypothesis too often or not enough.

A sketch of the data \((y, X)\) is \((\bar{y}, \bar{X})\), where \( \bar{y} = \Pi y, \bar{X} = \Pi X \), and \( \Pi \) is usually an \( m \times n \) random matrix. The sketched least squares estimator is \( \hat{\beta}_{OLS} := (\bar{X}^T \bar{X})^{-1} \bar{X}^T \bar{y} \). Even though the sketched regression is based on a sample of size \( m, \bar{X}^T \bar{X} = X^T \Pi^T \Pi X \) and \( \bar{X}^T \bar{y} = X^T \Pi^T \Pi y \) can be seen as weighted moments in a sample of size \( n \). Thus let \( \bar{g}_n := X^T \bar{e}/n, \bar{g}_n := X^T e/n, \bar{A}_n := (\bar{X}^T \bar{X}/n)^{-1}, \) and \( \bar{A}_n := (\bar{X}^T \bar{X}/n)^{-1} \).

By the law of large numbers, \( \bar{A}_n - A \to_p (1) \), where \( A := \left[ E(X_i X_i^T) \right]^{-1} \), and by the central limit theorem, \( \bar{g}_n = O_p(n^{-1/2}) \). We show in Section 3 that for \( \Pi \) with subspace embedding property
\[
\bar{\beta}_{OLS} - \tilde{\beta}_{OLS} = A(\bar{g}_n - \bar{g}_n) + o_p(m^{-1/2}).
\]
We study \( \bar{\beta}_{OLS} \) under the following regularity conditions.

**Assumption 2.1.** (i) The data \( D_n := \{(y_i, X_i) \in \mathbb{R}^{1+r} : i = 1, \ldots, n\} \) are independent and identically distributed (i.i.d.), where \( p \) is fixed. Furthermore, \( X \) has singular value decomposition \( X = U \Sigma X V_X^T \).

(ii) \( E(y_i^2) < \infty, E(||X_i||^4) < \infty \), and \( E(X_i X_i^T)^2 \) has full rank \( p \).

(iii) The random matrix \( \Pi \) is independent of \( D_n \).

(iv) \( m = m_n \to \infty \) but \( m/n \to 0 \) as \( n \to \infty \).

Assumptions (i) and (ii) are standard. For (iii), note that for a general random \( \Pi \) whose \((k, i)\) element is \( \Pi_{ki} \), the difference between the full and the sketched moments such as \( \bar{g}_n - \bar{g}_n \) and \( \bar{A}_n - \bar{A}_n \) are of the form
\[
n^{-1} \left( U^T \Pi^T \Pi V - U^T V \right)
\]
\[
= n^{-1} \sum_{i=1}^n \psi_i U_i V_i + n^{-1} \sum_{i=1, j=1, j \neq i}^n U_i \varphi_{ij} V_j
\]
\[
= T_{n1} + T_{n2},
\]
where \( U \in \mathbb{R}^n \) and \( V \in \mathbb{R}^n \) are vectors of certain i.i.d. random variables \((U_i, V_i) \in \mathbb{R}^2 \) that are independent of \( \Pi \),
\[
\psi_i := \sum_{k=1}^{r \text{ dim}(\Pi)} \Pi_{ki}^2 - 1, \quad \varphi_{ij} := \sum_{k=1}^{r \text{ dim}(\Pi)} \Pi_{ki} \Pi_{kj},
\]
and \( r \text{ dim}(\Pi) \in \{m, n\} \) denotes the row dimension of \( \Pi \).

There are two classes of sketching schemes to consider. Random sampling schemes have \( \varphi_{ij} = 0 \) for all \( i \neq j \) because there is only one non-zero entry in each row of \( \Pi \). In such cases, \( T_{2n} \) is negligible and \( T_{1n} \) is the leading term. The second class is random projection schemes with which \( T_{1n} \) is asymptotically negligible and \( T_{2n} \) is the leading term.

To gain intuition, we first provide results for Bernoulli sampling (BS) from the first type and countsketch (CS) from the second type.
Theorem 2.2. Let Assumption 2.1 hold and $\mathbb{E}(e_i X_i) = 0$.

(i) Under BS, $m^{1/2}(\beta_{\text{OLS}} - \beta) \to_d N(0, V_1)$.

(ii) Under CS, $m^{1/2}(\beta_{\text{OLS}} - \beta) \to_d N(0, V_0)$.

Though Theorem 2.2 indicates that both sampling schemes yield asymptotically normal estimates, their variances are different, and normality holds for different reasons. The proof is given in the Appendix. Here, we sketch the main arguments.

First, under BS, the sampling probability is determined by i.i.d. Bernoulli random variables with success probability $m/n$. Thus, $\Pi = \sqrt{m/n} B$ is an $n \times n$ matrix (not $m \times n$), where $B$ is a diagonal sampling matrix. We have

$$
\frac{1}{n} \left( X^T \Pi^T (\frac{1}{n} X^T e) - \frac{1}{n} \right) X_1 e_1 = T_{n1}.
$$

Since the summands are i.i.d. with mean zero with variance $(\frac{m}{n} - 1)\mathbb{E}(e_i^2 X_i^T X_i^T)$, applying central limit theorem for i.i.d. observations yields the sandwich variance $V_1$.

Consider now CS. Each column of its $\Pi$ has one non-zero entry taking on value $\{+1, -1\}$ randomly drawn with equal probability and located uniformly at random. For such $\Pi$ and every nonzero $c \in \mathbb{R}^p$,

$$
c^T (X^T \Pi^T (\frac{1}{n} X^T e) - \frac{1}{n})/n = n^{-1} \sum_{i=1}^n X_i (c \sum_{k=1}^m \Pi_{ki} - 1) e_i + n^{-1} \sum_{i=1}^n \sum_{j=1, j \neq i}^n X_i (c) \Pi_{kj} \Pi_{ki} e_i = T_{n1} + T_{n2},
$$

where $X_i := \sum_{u=1}^q c_u X_{iu}$ is a weighted sum of elements of the $i$-th row of $X$. The term $T_{n1}$ is identical zero because there is only one non-zero entry per column of $\Pi$.

To analyze $T_{n2}$, let $W_i := (Y_i, X_i^T, \Pi_{1i}, \ldots, \Pi_{mi})^T$. Since the columns of $\Pi$ are i.i.d., $\{W_i : i = 1, \ldots, n\}$ are i.i.d. Now let $w = (y, x^T, \pi_1, \ldots, \pi_m)^T$ be a non-random index. Define $\tilde{H}(w_1, w_2) := \sum_{k=1}^m \tilde{X}_i (c) \pi_{ki} S_{ki} e_i$, and $H(w_1, w_2) := \tilde{H}(w_1, w_2) + H(w_2, w_1)$. We can write

$$
T_{n2} = n^{-1} \sum_{1 \leq i < j \leq n} H(W_i, W_j),
$$

noting that $H(w_1, w_2) = H(w_2, w_1)$, and $\mathbb{E}(H(W_1, W_2)|W_1) = \mathbb{E}(H(W_1, W_2)|W_2) = 0$. Importantly, $T_{n2}$ has now been represented as a degenerate $U$-statistic. In general, the asymptotic distribution of such statistics is either a weighted average of independent, centered chi-square random variables with complex weights, or a centered normal distribution. But if the conditions given in Hall (1984) are satisfied, the latter holds. Precisely,

$$
\left\{ \frac{1}{2} \mathbb{E}[H^2(W_1, W_2)] \right\}^{-1/2} T_{n2} \to_d N(0, 1).
$$

A sufficient condition for this result which we verify in the Appendix is

$$
\frac{\mathbb{E}[G^2(W_1, W_2)] + n^{-1} \mathbb{E}[H^4(W_1, W_2)]}{\mathbb{E}[H^2(W_1, W_2)]} \to 0 \quad \text{as} \quad n \to \infty,
$$

where $G(w_1, w_2) := \mathbb{E}(H(W_1, w_1) H(W_1, w_2))$. Furthermore, we also verify that for $W_i \neq W_j$,

$$
\frac{1}{2} \mathbb{E}[H^2(W_i, W_j)] = \frac{1}{m} \mathbb{E}(\tilde{X}_i^2(c)) \mathbb{E}(e_i^2).
$$

Note that $\mathbb{E}(e_i^2)$ appears separately from $\mathbb{E}(\tilde{X}_i^2(c))$. This is key to the claim in Theorem 2.2 that when $\beta$ is based on CS, $m^{1/2}(\beta - \beta) \to_d N(0, \mathbb{E}[e_i^2 A]) = N(0, V_0)$. Analogous arguments show that each entry of $(\tilde{X}^T X - X^T X)$ can also be written as a degenerate $U$-statistic and $A_m - \tilde{A}_m = o_p(1)$, which is needed for consistent estimation of $V_0$ and $V_1$.

As discussed in Charikar et al. (2004) and Clarkson & Woodruff (2013; 2017), the sparsity of $\Pi$ significantly reduces the run time required of the countsketch to compute $\Pi A$ to $O(nnz(A))$, where $nnz(A)$ is the number of non-zero entries of $A$. Another appeal of countsketch is that the sketches can be obtained by streaming without constructing $\Pi$. Here, we show that countsketch removes heteroskedasticity which is appealing because it simplifies inference. In the next section, we study the mean-squared sketching error and show that part (i) of Theorem 2.2 also holds for other $\Pi$s in the first class, while part (ii) holds for other $\Pi$s in the second class. Section 4 then shows that these results also hold when the regressors are not exogenous.

3. The Mean-Squared Sketching Error

For a random variable $G$, let $\text{MSE}(G) = [\mathbb{E}(G)]^2 + \text{Var}(G)$ denote the mean squared error. We now analyze the asymptotic behavior of mean squared sketching errors of $(U^T \Pi^T IV - U^T V)/n$, where $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^n$ denote vectors of i.i.d. random variables $U_i, V_i \in \mathbb{R}^2$ that are independent of $\Pi$, with $\mathbb{E}(U_i^2) < \infty$ and $\mathbb{E}(V_i^2) < \infty$. Recall that $m = m_n \to \infty$ but $m/n \to 0$ as $n \to \infty$.

3.1. Random Sampling with Replacement (RS)

For sketching by random sampling with replacement (RS), we suppose that for each $t = 1, \ldots, m$, we sample $k_t$ from $[n]$ with probability $p_t := \Pr(k_t = i)$ independently and
Least Squares Estimation using Sketched Data with Heteroskedastic Errors

with replacement. The random matrix \( \Pi \in \mathbb{R}^{m \times n} \) is then

\[
\Pi = \sqrt{\frac{n}{m}} (t_{k_1} \ldots t_{k_m})^T,
\]

where \( t_k \) is the \( k \)-th column vector of the \( n \times n \) dimensional identity matrix. Sketching schemes of the RS class have properties characterized by Lemma A.2 in the Appendix. Importantly, each \( \Pi \) in this class has \( T_{2n} = 0 \) and as a result, \( T_{n1} \) is the only term we need to consider. An important example in the RS class is uniform sampling with replacement with \( p_i = n^{-1} \).

**Theorem 3.1.** (i) If \( \Pi \) is a random matrix satisfying RS and \( \sum_{i=1}^{n} p_i^2 = o(m^{-1}) \), then, as \( n \to \infty \),

\[
\text{MSE} \left[ \sqrt{m} \left( U^T \Pi \Pi^T V - U^T V \right) / n \right] \to \text{Var}(U_i V_i).
\]

(ii) If \( \Pi \) is Bernoulli sampling matrix (BS), then, as \( n \to \infty \),

\[
\text{MSE} \left[ \sqrt{m} \left( U^T \Pi \Pi^T V - U^T V \right) / n \right] \to \mathbb{E}(U_i^2 V_i^2).
\]

The mean-squared errors for RS and BS are the same if \( U_i V_i \) is mean zero.

Theorem 3.1 is useful in two ways. First, let \( U = c^T X^T \) and \( V = X c. \) Under Assumption 2.1 and for all nonzero \( c \in \mathbb{R}^p \), Theorem 3.1 yields

\[
\text{MSE} \left[ \sqrt{m} (c^T \tilde{X}^T \tilde{X} c - c^T X^T X c) / n \right] = O(1).
\]

By Chebyshev’s Inequality, \( (\tilde{X}^T \tilde{X} - X^T X) / n = O_P(m^{-1/2}) \). Similarly, for \( U = X c \) and \( V = e \), the theorem implies \( \tilde{g}_n - \hat{g}_n = O_P(m^{-1/2}) \). Applying continuous mapping theorem gives the result stated earlier that \( \hat{\beta}_{OLS} - \hat{\beta}_{GLS} = A(\tilde{g}_n - \hat{g}_n) + o_P(m^{-1/2}) \).

It is known that the efficient estimator under heteroskedasticity is the generalized least squares (GLS), defined as

\[
\hat{\beta}_{GLS} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y,
\]

where \( \Omega \) is \( n \times n \) a diagonal matrix with \( \Omega_{ii} = \sigma_i^2 \). GLS weights each observation with \( \sigma_i^{-2} \) so that the errors in the weighted regression are homoskedastic. Now the OLS estimator applied to sketched data can be written as

\[
\tilde{\beta}_{OLS} = (X^T \Pi \Xi)^{-1} X^T \Pi \Xi Y.
\]

A question of interest is whether \( \Pi^T \Pi \) can play the role of \( \Omega^{-1} \). The theorem sheds light on this problem as its second use is to obtain the asymptotic variance of the sketched estimator. To this end, let again \( U_i = c^T X_i \) and \( V_i = e_i \). Assuming \( \mathbb{E}(g_i) = 0 \) where \( g_i = e_i X_i \), RS and BS imply:

\[
\text{MSE} \left[ \sqrt{m} e^T (\tilde{g}_n - \hat{g}_n) \right] = \mathbb{E}(e_i^2 e^T X_i X_i^T e).
\]

The asymptotic standard error is generally the expectation of a product of \( e_i^2 \) and \( (c^T X_i)^2 \) and becomes the product of two expectations only under homoskedasticity when \( \mathbb{E}(e_i^2 c^T X_i X_i^T e) = \mathbb{E}(e_i^2) \mathbb{E}(c^T X_i X_i^T e) \). Thus, under RS and BS, the asymptotic variance of \( \hat{\beta}_{OLS} \) is \( \sigma^2_0 \) only if homoskedasticity is explicitly imposed, implying that \( \Pi^T \Pi \) corresponding to random sampling will not homogenize error variance in the same way that \( \Omega^{-1} \) can.

It is noteworthy that even under homoskedasticity, we cannot always use a central limit theorem for i.i.d. data even if the full sample of data are i.i.d. because the sampling scheme may induce dependence in the sketched data. Thus the asymptotic normality result can only be analyzed on a case by case basis. Ma et al. (2020) confronts a similar problem when studying the asymptotic distribution of estimators in linear regressions under random sampling with replacement and homoskedastic errors. Let \( K_i \) and \( p_i \), respectively, denote the number of times and the probability that \( i^{th} \) observation is sampled. Their estimator has \( W = \text{diag} \{K_i/(mp_i)\}_{i=1}^n \) playing the role of \( \Pi^T \Pi \). Our Theorem 3.1 applies to their setup with uniform sampling where \( p_i = 1/n, n \geq m \), but it would not apply when \( p_i \) is data dependent. In this case, Ma et al. (2020) also cannot use a central limit theorem for i.i.d. data. Instead, they apply Hayék-Sidak central limit theorem and use Poissonization to account for dependence in the sketched data that arises from sampling.

### 3.2. Random Projection (RP)

Sketching schemes in the RP class have properties characterized by Lemma A.3 in the Appendix if \( \Pi \in \mathbb{R}^{m \times n} \) is a random matrix with the following properties:

**Assumption 3.2.** (i) \( \mathbb{E}[\Pi_{ki}] = 0, \mathbb{E}[\Pi_{ki}^2] = m^{-1} \) for all \( k \in [m] \) and all \( i \in [n] \), and \( \max_{k,i} [\mathbb{E}[\Pi_{ki}]] = O(m^{-1}) \);

(ii) \( \mathbb{E}[\Pi_{ki} \Pi_{kj}] = 0 \) and \( \mathbb{E}[(\Pi_{ki} \Pi_{kj})^2] = m^{-2} \) for all \( k \in [m] \) and all \( i \neq j \in [n] \);

(iii) \( \mathbb{E}[\Pi_{ki} \Pi_{kj} \Pi_{kp} \Pi_{pq}] = 0 \) for all \( k \neq \ell \in [m] \) and all \( i \neq j, p \neq q \in [n] \).

Under Assumption 3.2, \( T_{n1} = O_P(n^{-1/2}) \) is asymptotically negligible, and \( T_{n2} \) becomes the leading term for RP. As discussed above, the \( \Pi \) for the CS only has one non-zero entry in each column. Since \( \Pi_{ikj} = 0 \) for all \( k, i \neq j \), it is straightforward to check that the above conditions are satisfied. For Gaussian random projections,

\[
\text{GP} : \Pi_{ki} \sim N(0, m^{-1}).
\]

Since all elements of \( \Pi \) are i.i.d. with mean zero, variance \( m^{-1} \) and the fourth moment \( O(m^{-1}) \), the conditions are
also satisfied. The SRHT has
\[
\text{SRHT} : \Pi = \sqrt{\frac{n}{m}} SHD,
\]
where \(S \in \mathbb{R}^{m \times n}\) is a uniform sampling matrix with replacement, \(H \in \mathbb{R}^{n \times n}\) is a normalized Walsh-Hadamard transform matrix, and \(D \in \mathbb{R}^{n \times n}\) is a diagonal Rademacher matrix with i.i.d. entries of \(\pm 1\). The Appendix shows that the conditions for RP hold for SRHT.

The following theorem gives the asymptotic mean squared sketching errors of RP schemes.

**Theorem 3.3.** If \(\Pi\) is a random matrix satisfying RP, then, as \(n \to \infty\),

(i) \(\text{MSE} \left[ \sqrt{m} (U^T \Pi^T \Pi V - U^T V) / n \right] \to \mathbb{E}(U^2)\mathbb{E}(V^2) + \mathbb{E}(U^2 \Pi^T)\).

(ii) If, in addition, \(\mathbb{E}[e_i X_i] = 0\) and the columns of \(\Pi\) are i.i.d., then \(m^{1/2}(\hat{\beta}_{\text{OLS}} - \tilde{\beta}_{\text{OLS}}) \to_d N(0, V_0)\).

The limiting MSE of RP is simply the product between two marginal expectations when \(\mathbb{E}(U_i V_i) = 0\) (and not the expectation of the product). It implies
\[
\text{MSE} \left[ \sqrt{m} c^T (\tilde{g}_n - \tilde{g}_n) \right] = \mathbb{E}(c^T X_i X_i^T c)
\]
and is the reason why the asymptotic variance for \(\tilde{\beta}_{\text{OLS}}\) for RP schemes is \(V_0\).

If \(c^2_i\) and \((c^T X_i)^2\) are positively (respectively, negatively) correlated, the limiting MSE of RP is smaller (respectively, larger) than that of RS and BS. The limiting MSE is the same if \(c^2_i\) and \((c^T X_i)^2\) are uncorrelated.

Asymptotic normality of \(\tilde{\beta}\) can be established by applying a central limit theorem for degenerate \(U\)-statistic if the columns of \(\Pi\) are i.i.d., as reported in part (ii) of Theorem 3.3. The SRHT and SRFT are not covered by this result because the columns of their II matrix are not i.i.d. and requires a limit theorem for a particular type of mixing data. In general, establishing asymptotic normality of \(\tilde{\beta}\) based on SRHT or SRFT require different proof techniques. The approach taken in Ahfcock et al. (2020) is to condition on the data \(D_n\) and apply a central limit theorem for a triangular array of random variables. We do not condition on the data and appeal to the theory of degenerate \(U\)-statistics. Though deriving distribution theory for the SRHT and SRFT estimates is not straightforward, we will show in simulations that their finite sample properties are similar to those of CS.

### 4. Two-Stage Least Squares

The 2SLS estimator is appropriate when \(\mathbb{E}(X_i e_i) \neq 0\) but exogenous instruments \(Z_i\) satisfying \(\mathbb{E}(Z_i e_i) = 0\) are available. The 2SLS estimator is
\[
\hat{\beta}_{\text{2SLS}} = (X^T P_Z X)^{-1} X^T P_Z y,
\]
where \(P_Z := Z(Z^T Z)^{-1} Z^T\) is the projection matrix. The estimator first projects on \(Z\) to purge the variations in \(X\) correlated with \(e\), and in the second step replaces \(X\) with \(Z\).

Analyzing \(\beta_{2SLS} - \beta_0 = \hat{\beta}_0 \tilde{g}_n\) under Assumption 4.3 given below, as \(n \to \infty\), we have
\[
\sqrt{n}(\hat{\beta}_{2SLS} - \beta_0) \to_d N(0, W_1),
\]
where \(W_1 := A \mathbb{E}(c_i^2 Z_i Z_i^T) A^T\) with
\[
A := \mathbb{E}(X_i Z_i^T) \mathbb{E}(Z_i Z_i^T)^{-1} \mathbb{E}(Z_i X_i^T)^{-1}.
\]
Under homoskedasticity, \(\mathbb{E}(c_i^2 | Z_i) = \delta^2\) and \(W_1\) reduces to
\[
W_0 := \mathbb{E}(c_i^2) \mathbb{E}(X_i Z_i^T) \mathbb{E}(Z_i Z_i^T)^{-1} \mathbb{E}(Z_i X_i^T)^{-1}.
\]
A sketched version of the 2SLS estimator is
\[
\tilde{\beta}_{2SLS} := (\tilde{X}^T P_{\tilde{Z}} \tilde{X})^{-1} \tilde{X}^T P_{\tilde{Z}} \tilde{y}.
\]

Low level conditions for Assumption 4.1(i)-(iii) are given in Cohen et al. (2016), among others. Assumption 4.1(i) is equivalent to the statement that the all eigenvalues of \(U^T \Pi^T \Pi U\) are bounded between \(1 - \varepsilon_1, 1 + \varepsilon_1\). This ensures that \(Z^T \tilde{Z}\) is non-singular with probability at least \(1 - \delta\). Part (iv) strengthens non-singularity of \(X^T P_Z X\) to require that \(\sigma_{\min}^2(U^T U X)\) is strictly positive and bounded below by the constant \(2f_1(\varepsilon_1, \varepsilon_2)\).

**Theorem 4.2. Under Assumptions 4.1, the following holds with probability at least \(1 - \delta\):

\[
|\tilde{\beta}_{2SLS} - \hat{\beta}_{2SLS}| \leq \frac{f_2(\varepsilon_1, \varepsilon_2) + \varepsilon_1 |\hat{e}|}{\sigma_{\min}(X) \sigma_{\min}^2(U^T U X)} \left(1 + 2f_1(\varepsilon_2, \varepsilon_2)\right) + \frac{f_2(\varepsilon_1, \varepsilon_2)}{\sigma_{\min}^2(U^T U X)}
\]

where \(f_2(\varepsilon_1, \varepsilon_2) := \varepsilon_2 + \varepsilon_1(1 - \varepsilon_1) + \varepsilon_2 \varepsilon_1(1 - \varepsilon_1)\).
The sketched estimator $\hat{\beta}_{2SLS}$ involves, firstly, a regression of $\hat{X}$ on $Z$, and then a regression of $\hat{y}$ on the fitted values in the first step. The estimator thus depends on the quality of subspace approximation in both steps. Theorem 4.2 provides a worst-case bound for $\beta_{2SLS} - \hat{\beta}_{2SLS}$ with the data $D_n$ being fixed. It depends on (i) $\epsilon_j$, $j = 1, 2, 3$, (ii) variability of $\hat{\epsilon}_i^2$, (iii) the signal from $X$ as given by $\sigma_{\text{min}}(X)$, and (iv) instrument strength as given by $\sigma_{\text{min}}(U_Z^2 U_X)$. The sketched estimator can be arbitrarily close to the full sample estimate with high probability, provided that the subsample size $m$ is sufficiently large, $X$ is linearly independent, and the instrument $Z$ is sufficiently relevant for $X$.

Though 2SLS is a two step estimator, we can still write

$$\hat{\beta}_{2SLS} - \hat{\beta}_{2SLS} = (\hat{A}_n - \hat{A}_n)\hat{y}_n + (\hat{A}_n - \hat{A}_n)(\hat{g}_n - \hat{g}_n)$$

as in the OLS case, but now $\hat{y}_n := Z^T \hat{\epsilon}/n$, $\hat{A}_n := [(X^T Z/n)(Z^T Z/n)^{-1}(Z^T X/n)^{-1}(X^T Z/n)(Z^T Z/n)^{-1}]$. A statistical analysis of $\beta_{2SLS}$ requires additional assumptions.

**Assumption 4.3.** (i) The data $D_n := \{(y_i, X_i, Z_i) \in \mathbb{R}^{1+p+q} : i = 1, \ldots, n\}$ are i.i.d. with $p \leq q$. Furthermore, $X$ and $Z$ have singular value decomposition $X = U_X \Sigma_X V_X^T$ and $Z = U_Z \Sigma_Z V_Z^T$.

(ii) $E(y_i^2) < \infty$, $E(\|X_i\|^4) < \infty$, $E(\|Z_i\|^4) < \infty$, and $E(X_i X_i^T)$ and $E(Z_i Z_i^T)$ have full rank $p$.

(iii) The random matrix $\Pi$ is independent of $D_n$.

(iv) $m = m_n \rightarrow \infty$ but $m/n \rightarrow 0$ as $n \rightarrow \infty$, while $p$ and $q$ are fixed.

Arguments similar to those used to prove Theorem 2.2 lead to the following.

**Theorem 4.4.** Let Assumption 4.3 hold and $E(Z_i \epsilon_i) = 0$. If $\Pi$ is RP satisfying RP(ii)-(iii) with columns that are i.i.d.

(i) Under BS, $m^{1/2}(\hat{\beta}_{2SLS} - \hat{\beta}_{2SLS}) \rightarrow_d N(0, W_1)$.

(ii) Under RP, $m^{1/2}(\hat{\beta}_{2SLS} - \hat{\beta}_{2SLS}) \rightarrow_d N(0, W_0)$.

Theorem 4.4 provides statistical properties of the sketched 2SLS estimator in Theorem 4.2 to complement the algorithmic results.

Theorem 4.4 states that when the data are sketched by RP, $\hat{\beta}$ is asymptotically normally distributed with mean $\hat{\beta}$ and variance $W_0/m$. We can consistently estimate $W_0$ by $\hat{W}_0 := \hat{\epsilon}^T (\hat{X}^T Z (Z^T Z)^{-1} Z^T X)^{-1}$, where $\hat{\epsilon} := y - X \hat{\beta}$ (not the residuals from the second step).

Interestingly, the asymptotic variance $W_0$ is the same as if the errors in the full sample regression were homoskedastic. But the result follows from estimation using sketched data rather than by assumption. This is not the case when inference is based on the full sample estimates, or the estimates computed from sketched data of the RS type. In such cases, a homoskedastic covariance weighting matrix would be inefficient since $E(\epsilon_i^2 | Z_i) \neq E(\epsilon_i^2)$.

Our analysis can be extended to the two-sample 2SLS estimator analyzed in Angrist & Krueger (1992; 1995) and Inoue & Solon (2010). However, it is not pursued for brevity of the paper.

In the econometrics literature, the instruments are said to be relevant if $E[Z_i X_i^T] \neq 0$. The latter is formalized by the rank condition in Assumption 4.3(ii). Tests for instrument relevance usually require robust standard errors corresponding to the parameter estimates in a regression of $X$ on $Z$ unless homoskedasticity can be ruled out. An implication of our preceding analysis is that this is not necessary when the regression is estimated on data sketched by RP, as will be illustrated below.

### 5. Practical Inference

In applications, researchers would like to test a hypothesis about $\beta_0$ using a sketched estimate, and our results provide all the quantities required for inference. In the exogenous regressor case, we generically have

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

where the form of $\hat{V}_m$ depends on $\Pi$. For any $\Pi$ in BS or RP class, we can use White (1980)’s homoskedasticity-consistent estimator:

$$\hat{V}_m = \hat{V}_{1,m} = (\hat{X}^T \hat{X})^{-1}(\sum_{i=1}^{m} \hat{X}_i \hat{X}_i^T \hat{s}_i^2)(\hat{X}^T \hat{X})^{-1}.$$

For $\Pi$ in the RP class, we can let $s_{OLS}^2 := \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - \hat{X}_i^T \hat{\beta}_{OLS})^2$. Then without assuming homoskedasticity,

$$\hat{V}_m = \hat{V}_{0,m} = s_{OLS}^2 (\hat{X}^T \hat{X})^{-1}.$$

In the endogenous regressor case, $\hat{W}_m^{-1/2}(\hat{\beta}_{2SLS} - \beta_0) \approx N(0, I_p)$. For RP, we let $\hat{s}_{2SLS}^2 := \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - \hat{X}_i^T \hat{\beta}_{2SLS})^2$ and $\hat{W}_{0,m} := \hat{s}_{2SLS}^2 (\hat{X}^T \hat{Z} (\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T \hat{X})^{-1}$. For BS, we define $\hat{W}_{1,m}$ from $\hat{W}_{1}$. Sketching estimators require a choice of $m$. From the algorithmic perspective, $m$ needs to be chosen as small as
We first consider the simulation design for which all the regressors are exogenous. The regressors \( X_i = (1, X_{2,i}, \ldots, X_{p,i})^T \) consist of a constant term and a \((p-1)\)-dimensional random vector \((X_{2,i}, \ldots, X_{p,i})^T\) generated from a multivariate normal distribution with mean zero vector and the variance covariance matrix \( \Sigma_i \), whose \((i,j)\) component is \( \Sigma_{ij} = \rho^{i-j} \) with \( \rho = 0.5 \). The dependent variable is generated by

\[
y_i = X_i^T \beta_0 + \sigma(X_i) e_i,
\]

where \( \beta_0 = (0, 1, \ldots, 1)^T \), and \( e_i \) is generated from \( N(0,1) \) independently from \( X_i \). We consider two designs for \( \sigma(X_i) \): (i) homoskedastic design \( \sigma(X_i) = 1 \) for all \( i \) and (ii) heteroskedastic design \( \sigma(X_i) = \exp(X_{p,i}) \), where \( X_{p,i} \) is the \( p \)-th element of \( X_i \). Throughout the Monte Carlo experiment, we set \( n = 10^6 \), \( m = 500 \), and \( p = 6 \). There were 5,000 replications for each experiment. Six sketching methods are considered: (i) Bernoulli sampling, (ii) uniform sampling, (iii) leverage score sampling and reweighted regression as in Ma et al. (2020); (iv) countsketch, (v) SRHT, (vi) subsampled randomized Fourier transforms using the real part of fast discrete Fourier transform (SRFT).

Table 1 reports the empirical size and power of the t-test. The null and alternative hypotheses are that \( H_0 : c^T \beta_0 = 1 \) vs. \( H_1 : c^T \beta_0 \neq 1 \) with \( c^T = (0, \ldots, 0, 1) \). Equivalently, the null hypothesis is \( \beta_p = 1 \). The power is obtained for the null value \( c^T \beta_0 = 1.1 \) for the homoskedastic design and \( c^T \beta_0 = 1.4 \) for the heteroskedastic design, respectively. The nominal size is 0.05.

In column (1) in Table 1, we report the size of the test, namely, the probability of rejecting \( H_0 \) when the null value is true. In this column, the t-statistic is constructed using homoskedasticity-only standard errors S.E.0. Though many methods perform well, both Bernoulli and uniform sampling show substantial size distortions for the heteroskedastic design. Leverage score sampling combined with reweighted regression seems to account for heteroskedasticity to some extent, but not enough to remove all size distortions. In column (2) which reports results using robust standard errors.
S.E.1, all methods have satisfactory size. In column (3), we report the power of the test, i.e., the probability of rejecting $H_0$ when the null value is false. For the heteroskedastic design, the powers of the tests using homoskedastic standard errors S.E.0 are worse for Bernoulli, uniform and leverage samplings than those for countsketch, SRHT, and SRFT. The power loss of the RS schemes is much more pronounced when the robust standard errors are used in column (4). This efficiency loss is consistent with asymptotic theory developed in the paper because the squared regression error is positively correlated with one of the elements of squared $X_i$ under the heteroskedastic design. All RP schemes perform similarly, hinting that even though a formal proof awaits future research, asymptotic normality may also hold for both SRHT and SRFT, and not just countsketch.

### 6.2. When One of the Regressors is Endogenous

We now move to the case when the regressors are $X_i = (1, X_{2,i}, \ldots, X_{p-1,i}, X_{p,i})^T$, and $y_i$ is generated by

$$y_i = X_i^T \beta_0 + \sigma_2(Z_i)(\eta_i + \epsilon_i),$$

where $\epsilon_i \sim N(0,1)$ is randomly drawn independently from $X_i$ and $\eta_i$. The first $p-1$ regressors, including the intercept term, are exogenous, but

$$X_{p,i} = Z_i^T \zeta_0 + \sigma_1(Z_i)\eta_i,$$

where $\zeta_0 = (\zeta_{0,1}, \ldots, \zeta_{0,p})^T$, $\eta_i \sim N(0,1)$ independently from $Z_i = (1, Z_{2,i}, \ldots, Z_{q,i})^T$. The presence of $\eta_i$ in both (3) and (4) induces endogeneity of $X_{p,i}$.

In each of the 1000 replications, $(X_{2,i}, \ldots, X_{p-1,i})^T = (Z_{2,i}, \ldots, Z_{p-1,i})^T$, while the $(q-1)$-dimensional $Z_i$ is multivariate normal with mean zero and the variance $\Sigma$, whose $(i,j)$ component is $\Sigma_{ij} = \rho^{\mid i-j \mid}$ with $\rho = 0.5$.

We consider two designs for $\sigma_1(Z_i)$: (i) homoskedastic design $\sigma_1(Z_i) = 1$ for all $i$ and (ii) heteroskedastic design

$$\sigma_1(Z_i) = \exp \left( \frac{1}{q} \sum_{j=2}^{p} |Z_{j,i}| \right) / 100.$$  

As in the previous section, we set $n = 10^8$, $m = 500$, $p = 6$, and $q = 21$. We consider five sketching schemes and no longer include leverage score sampling since it is unclear how to implement it in the case of 2SLS. The nominal size is 0.05. Throughout, $(\zeta_{0,1}, \ldots, \zeta_{0,q}) = (0, 0.1, \ldots, 0.1)^T$, but values of $\zeta_{j,0}$ for $j \geq p$ depend on the context as explained below.

We first examine the so-called first-stage F-test for instrument relevance. In this case of a scalar endogenous regressor, the null hypothesis of irrelevant instruments amounts to a joint test of $H_0 : \zeta_{j,0} = 0$ for every $j = p, \ldots, q$ in (4). The size of the test is evaluated at $\zeta_{j,0} = 0$ and the power at $\zeta_{j,0} = 0.1$ for $j = p, \ldots, q$. The F-test statistic is constructed as

$$F = \frac{1}{q-p+1} \hat{\zeta}_{-(p-1)}^T \left( \tilde{V}_{-(p-1),-(p-1)} \right)^{-1} \hat{\zeta}_{-(p-1)},$$

where $\hat{\zeta}_{-(p-1)}$ is a $(q-p+1)$-dimensional vector of the OLS estimate $\hat{\zeta}$ of regressing $X_{p,i}$ on $Z_i$, excluding the first $(p-1)$ elements, and $[\tilde{V}]_{-(p-1),-(p-1)}$ is the corresponding submatrix of $\tilde{V}$.

In Table 2, we report the size and power of the F-test for $H_0 : \zeta_{p-1+1,0} = \ldots = \zeta_{q,0} = 0$ using homoskedasticity-only (V.0) and heteroskedasticity-consistent (V.1) asymptotic variance estimates, respectively. As in the previous subsection, Bernoulli and uniform sampling sketches suffer from size distortions in the heteroskedastic design but V.0 is used. Tests based on V.1 have good size without sacrificing much power when the F-test is constructed from data sketched by RP.

We now turn to 2SLS estimation of $\beta_0$. To ensure that the instruments are powerful enough to estimate $\beta_0$, we now set $\zeta_{j,0} = 0.5$ for $j = p, \ldots, q$ with

### Table 2. F test for Instrument Relevance: V.0 and V.1 refer to homoskedasticity-only and heteroskedasticity-consistent asymptotic variance estimates, respectively.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V.0</td>
<td>S.E.0</td>
<td>S.E.1</td>
<td>Power</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>0.047</td>
<td>0.063</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>(2)</td>
<td>0.049</td>
<td>0.063</td>
<td>0.997</td>
<td>0.999</td>
</tr>
<tr>
<td>COUNTSKETCH</td>
<td>0.040</td>
<td>0.058</td>
<td>1.000</td>
<td>0.999</td>
</tr>
<tr>
<td>SRHT</td>
<td>0.048</td>
<td>0.051</td>
<td>0.999</td>
<td>0.998</td>
</tr>
<tr>
<td>SRFT</td>
<td>0.050</td>
<td>0.052</td>
<td>1.000</td>
<td>0.999</td>
</tr>
</tbody>
</table>

### Table 3. 2SLS based t test for $H_0 : \beta_p = 1, H_1 : \beta_p \neq 1$

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V.0</td>
<td>S.E.0</td>
<td>S.E.1</td>
<td>Power</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td>0.065</td>
<td>0.067</td>
<td>0.687</td>
<td>0.695</td>
</tr>
<tr>
<td>(2)</td>
<td>0.056</td>
<td>0.057</td>
<td>0.686</td>
<td>0.693</td>
</tr>
<tr>
<td>COUNTSKETCH</td>
<td>0.055</td>
<td>0.060</td>
<td>0.698</td>
<td>0.705</td>
</tr>
<tr>
<td>SRHT</td>
<td>0.043</td>
<td>0.046</td>
<td>0.710</td>
<td>0.714</td>
</tr>
<tr>
<td>FFT</td>
<td>0.061</td>
<td>0.068</td>
<td>0.704</td>
<td>0.703</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V.0</td>
<td>0.274</td>
<td>0.050</td>
<td>0.844</td>
<td>0.648</td>
</tr>
<tr>
<td>UNIFORM</td>
<td>0.291</td>
<td>0.046</td>
<td>0.864</td>
<td>0.654</td>
</tr>
<tr>
<td>COUNTSKETCH</td>
<td>0.042</td>
<td>0.047</td>
<td>0.930</td>
<td>0.930</td>
</tr>
<tr>
<td>SRHT</td>
<td>0.052</td>
<td>0.056</td>
<td>0.941</td>
<td>0.944</td>
</tr>
<tr>
<td>FFT</td>
<td>0.055</td>
<td>0.055</td>
<td>0.933</td>
<td>0.942</td>
</tr>
</tbody>
</table>
\( \sigma_1(Z_i) = 1 \) for all \( i \). We set \( \beta_0 = (0, 1, \ldots, 1)^T \) and consider two designs for \( \sigma_2(Z_i) \): (i) homoskedastic design \( \sigma_2(Z_i) = 1 \) for all \( i \) and (ii) heteroskedastic design \( \sigma_2(Z_i) = \exp \left( \frac{3}{4} \sum_{j=2}^{\infty} |Z_{j,i}| \right) / 100. \)

As in the previous subsection, we test \( H_0 : \beta_p = 1 \) against \( H_1 : \beta_p \neq 1 \), or equivalently, \( c^2 = (0, \ldots, 0, 1) \). The power is obtained for \( \beta_p = 1.05 \) in the homoskedastic design and \( \beta_p = 1.10 \) for the heteroskedastic design, respectively. Table 3 reports results for nominal size of 0.05. Generally, the same patterns are observed as in the previous section. Thus, simulations support the theoretical result that robust standard errors are not needed for inference when estimation is based on sketched data using sketching schemes in the RP class.

### 7. An Empirical Illustration

An exemplary application of the 2SLS in economics is causal inference, such as to estimate the return to education. Suppose that \( y_i \) is the wages for worker \( i \) (typically in logs) and \( X_i \) contains educational attainment \( e \mu_i \) (say, years of schooling completed). Here, the unobserved random variable \( e_i \) includes worker \( i \)'s unobserved ability among other things. Then, \( e \mu_i \) will be correlated with \( e_i \) if workers with higher ability tends to attain higher levels of education. The least-squares estimator may not provide a consistent estimate of the return to schooling. To overcome this problem, economists use an instrumental variable that is uncorrelated with \( e_i \) but correlated with \( e \mu_i \).

We re-examine the OLS and 2SLS estimates of return to education in columns (1) and (2) of Table IV in Angrist & Krueger (1991). The dependent variable \( y \) is the log weekly wages, the covariates \( X \) include years of education, the intercept term and 9 year-of-birth dummies (\( p = 11 \)). Following Angrist & Krueger (1991), the instruments \( Z \) are the exogenous regressors (i.e., the intercept and year-of-birth dummies) and a full set of quarter-of-birth (one quarter omitted) times year-of-birth interactions (\( q = 1 + 9 + 3 \times 10 = 40 \)). Their idea was that season of birth is unlikely to be correlated with workers’ ability but can affect educational attainment because of compulsory schooling laws. The full sample size is \( n = 247, 199 \).

To construct sketched data, we need to choose the sketch size \( m \). We use a data-oblivious sketch size \( m_3 \) defined in (2) with target size set to \( \alpha = 0.05 \) and target power to \( \gamma = 0.8 \), giving \( S_2^2(\alpha, \gamma) = 6.18 \). It remains to specify \( \tau_2(\infty) \), which can be interpreted as the value of t-statistic when the sample size is really large.

In the OLS case, we take \( \tau_2(\infty) = 10 \) resulting in \( m = 15, 283 \) (about 6% of \( n \)). Table 4 reports empirical results for the OLS estimates. For each sketching scheme, only one random sketch is drawn; hence, the results can change if we redraw sketches. Remarkably, all sketched estimates are 0.08, reproducing the full sample estimate up to the second digit. The sketched homoskedastic-only standard errors are also very much the same across different methods. The Eicker-Huber-White standard error S.E.1 is a bit larger than the homoskedastic standard error S.E.0 with the full sample.

As expected, the same pattern is observed for Bernoulli and uniform sampling, as these sampling schemes preserve conditional heteroskedasticity.

For 2SLS, as it is more demanding to achieve good precision, we take \( \tau_2(\infty) = 5 \), resulting in \( m = 61, 132 \) (about 25% of \( n \)). Table 5 reports empirical results for the 2SLS estimates. The sketched estimates vary from 0.053 to 0.115, reflecting that the 2SLS estimates are less precisely estimated than the OLS estimates. Both types of standard errors are almost identical across all sketches for 2SLS, suggesting that heteroskedasticity is not an issue in this data.

### Acknowledgements

We would like to thank anonymous reviewers for helpful comments. The first author would like to thank the European Research Council for financial support (ERC-2014-CoG646917-ROMIA) and the UK Economic and Social Research Council for research grant (ES/P008909/1) to the CeMMAP. The second author would like to thank the National Science Foundation for financial support (SES 2018369).
References


Sarlos, T. Improved approximation algorithms for large matrices via random projections. *Proceedings of the 47


A. Appendix: Proofs for OLS

Recall that $\hat{\beta}_{OLS} - \beta_0 = (X^T X)^{-1} X^T e$ and $\tilde{\beta}_{OLS} - \beta_0 = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{e}$. Thus

$$\tilde{\beta}_{OLS} - \hat{\beta}_{OLS} = \left((\tilde{X}^T \tilde{X})^{-1} - (X^T X)^{-1}\right) X^T e + (\tilde{X}^T \tilde{X})^{-1} \left(\tilde{X}^T \tilde{e} - X^T e\right)$$

$$= (\tilde{A}_n - \tilde{A}_n) \tilde{g}_n + \tilde{A}_n (\tilde{g}_n - \hat{g}_n) + (\tilde{A}_n - \hat{A}_n)(\hat{g}_n - \tilde{g}_n),$$

where $\tilde{g}_n := \tilde{X}^T \tilde{e}/n$, $\hat{g}_n := X^T e/n$, $\tilde{A}_n := (\tilde{X}^T \tilde{X}/n)^{-1}$, and $\hat{A}_n := (X^T X/n)^{-1}$. By the law of large numbers and the continuous mapping theorem, $\tilde{A}_n - A = o_p(1)$ and by the central limit theorem, $\tilde{g}_n = O_p(n^{-1/2})$. Furthermore, by repeated applications of Theorem 3.1,

$$\text{MSE}[(\tilde{X}^T \tilde{e} - X^T e)/n] = O(m^{-1}) \quad \text{and} \quad \text{MSE}[(\tilde{X}^T \tilde{X} - X^T X)/n] = O(m^{-1}),$$

and by Chebyshev’s inequality, $(\tilde{X}^T \tilde{e} - X^T e)/n = O_p(m^{-1/2})$ and $(\tilde{X}^T \tilde{X} - X^T X)/n = O_p(m^{-1/2})$. The latter combined with the continuous mapping theorem yields that $\tilde{A}_n - \tilde{A}_n = O_p(m^{-1/2})$. Thus,

$$\tilde{\beta}_{OLS} - \hat{\beta}_{OLS} = A(\tilde{g}_n - \hat{g}_n) + (\tilde{A}_n - A)(\tilde{g}_n - \hat{g}_n) + O_p(m^{-1/2} n^{-1/2} + m^{-1})$$

$$= A(\tilde{g}_n - \hat{g}_n) + o_p(m^{-1/2}).$$

We start with asymptotic normality for Bernoulli sampling.

**Proof of Theorem 2.2(i).** In view of the Cramer-Wold device, it suffices to show that for any nonzero constant vector $c \in \mathbb{R}^p$,

$$m^{1/2} \left[c^T \mathbb{E}(e_i^2 X_i X_i^T) c\right]^{-1/2} c^T (\tilde{g}_n - \hat{g}_n) \to_d N(0, 1).$$

Write

$$c^T (\tilde{g}_n - \hat{g}_n) = n^{-1} \sum_{i=1}^n \left(\frac{n}{m} B_{ii} - 1\right) e_i X_i^T c.$$

Because the summands are i.i.d. with mean zero and finite variance, the central limit theorem yields the desired result immediately. \qed

**Proof of Theorem 2.2(ii).** This result is a special case of Theorem 3.3(ii) and we prove Theorem 3.3(ii) below. \qed

In what follows, we focus on the instance that $r\dim(\Pi) = m$. Recall that

$$n^{-1} \left(U^T \Pi^T V U - U^T V\right) = n^{-1} \sum_{i=1}^n \psi_i U_i V_i + n^{-1} \sum_{i=1}^n \sum_{j=1, j \neq i}^n U_i \varphi_{ij} V_j =: T_{n1} + T_{n2},$$

where $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^n$ are vectors of certain i.i.d. random variables $(U_i, V_i) \in \mathbb{R}^2$ that are independent of $\Pi$,

$$\psi_i := \sum_{k=1}^m \Pi_{ki}^2 - 1, \quad \varphi_{ij} := \sum_{k=1}^m \Pi_{ki} \Pi_{kj}.$$ 

There are two important cases. In case (i), $T_{n1}$ is the leading term and $T_{n2}$ is identically zero. The latter is true if $\varphi_{ij} = 0$ for all $i \neq j$. Methods in this class generate sketches by sampling from the full data matrix using deterministic or data dependent probabilities. The case includes random sampling with replacement (RS).
In case (ii), $T_{n2}$ is the leading term and $T_{n1}$ is identically zero or asymptotically negligible. Recall that for $w = (u, v, \pi_1, \ldots, \pi_m)^T$,

$$
\tilde{H}(w_1, w_2) := \sum_{k=1}^{m} u_1 \pi_{k1} \pi_{k2} v_2
$$

and $H(w_1, w_2) := \tilde{H}(w_1, w_2) + \tilde{H}(w_2, w_1)$. Then,

$$
T_{n2} = n^{-1} \sum_{1 \leq i < j \leq n} H(W_i, W_j).
$$

(5)

Note that $H(W_i, W_j)$ is symmetric, i.e., $H(W_i, W_j) = H(W_j, W_i)$. The canonical form of $\Pi$ we consider for case (ii) is random projection whose properties are given in the main text.

Before proving Theorems 3.1 and 3.3, we first prove Lemma A.1 and establish some useful lemmas.

**Lemma A.1.** GP, CS, and SRHT satisfy the conditions for RP.

**Proof of Lemma A.1.** For GP, it is straightforward to check all the conditions as all elements of $\Pi$ are i.i.d. We omit the details. For CS, note that the columns of $\Pi$ are i.i.d. and $\Pi$ has only one non-zero entry in each column, hence implying that $\Pi_k, \Pi_j = 0$ for all $k, i \neq j$. Then, it is easy to see that all the conditions are satisfied. It is more involving to check the conditions for SRHT. To do so, write

$$
\Pi_k = \sqrt{\frac{n}{m}} \sum_{j=1}^{n} S_{kj} H_{ji} D_{ii}
$$

Using the fact that for each $k$, $S_{k\ell_1}, S_{k\ell_2} = 0$ whenever $\ell_1 \neq \ell_2$ (the property of uniform sampling), further write

$$
\Pi_k \Pi_k = \frac{n}{m} \sum_{\ell_1, \ell_2=1}^{n} S_{k\ell_1} H_{\ell_1i} D_{ii} S_{k\ell_2} H_{\ell_2j} D_{jj}
$$

$$
= \frac{n}{m} \sum_{j=1}^{n} S_{kj} H_{ji} H_{ji} D_{jj}
$$

and

$$
\Pi_k \Pi_k \Pi_{\ell p} \Pi_{\ell q} = \frac{n^2}{m^2} \sum_{\ell_1, \ell_2=1}^{n} S_{k\ell_1} H_{\ell_1i} D_{ii} H_{\ell_2j} D_{jj} S_{\ell_1p} H_{\ell_2p} D_{pp} H_{\ell_3q} D_{qq}
$$

Using the facts that $E(S_{kj}) = n^{-1}, E(D_{ii}) = 0$, $\sum_{j=1}^{n} H_{ji}^2 = 1$, and $|H_{ji}| = n^{-1/2}$, we have

$$
E(\Pi_k) = \sqrt{\frac{n}{m}} \sum_{j=1}^{n} E(S_{kj}) H_{ji} E(D_{ii}) = 0,
$$

$$
E(\Pi_k^2) = \frac{n}{m} \sum_{j=1}^{n} E(S_{kj}) H_{ji}^2 = \frac{1}{m} \sum_{j=1}^{n} H_{ji}^2 = \frac{1}{m},
$$

$$
E(\Pi_k^2 \Pi_{\ell j}) = \frac{n^2}{m^2} \sum_{\ell_1}^{n} E(S_{k\ell_1}) H_{\ell_1i}^2 H_{\ell_2j}^2 = \frac{n}{m^2} \sum_{\ell_1}^{n} H_{\ell_1i}^2 H_{\ell_2j}^2 = \frac{1}{m^2},
$$

$$
E(\Pi_k^4) = \frac{n^2}{m^2} \sum_{\ell_1}^{n} E(S_{k\ell_1}) H_{\ell_1i}^4 = \frac{n}{m^2} \sum_{\ell_1}^{n} H_{\ell_1i}^4 = \frac{1}{m^2}.
$$

Furthermore, note that the diagonal elements of $D$ are i.i.d. and the rows of $S$ are i.i.d. Then, we have that $E[\Pi_k \Pi_{kj}] = 0$ for all $k, i \neq j$ and $E[\Pi_k \Pi_{kj} \Pi_{\ell p} \Pi_{\ell q}] = 0$ for all $k \neq \ell, i \neq j, p \neq q$. Therefore, we have verified all the required conditions.
Lemma A.2. If $\Pi$ is a random matrix satisfying RS, then,

$$
\mathbb{E} \left[ n^{-1} (U^T \Pi^T \Pi V - U^T V) \right] = 0,
$$

$$
\text{Var} \left[ n^{-1} (U^T \Pi^T \Pi V - U^T V) \right] = \left\{ \frac{1}{m} - \frac{1}{n} + \left( 1 - \frac{1}{m} \right) \sum_{i=1}^{n} p_i^2 \right\} \text{Var}(U_i V_i).
$$

In particular, when $p_i = n^{-1}$, the variance is reduced to $\frac{n-1}{m} \text{Var}(U_i V_i)$.

**Proof of Lemma A.2.** Using the property of RS, we have that

$$
\mathbb{E}[\psi_i] = \sum_{k=1}^{m} \mathbb{E}[\Pi_{ki}^2] - 1 = np_i - 1,
$$

$$
\mathbb{E}[\psi_i^2] = \sum_{k=1}^{m} \sum_{\ell=1}^{m} \mathbb{E}[\Pi_{ki}^2 \Pi_{\ell i}^2] - 2 \sum_{k=1}^{m} \mathbb{E}[\Pi_{ki}^2] + 1
$$

$$
= \sum_{k=1}^{m} \mathbb{E}[\Pi_{ki}^2] + \sum_{k=1}^{m} \sum_{\ell=1, \ell \neq k}^{m} \mathbb{E}[\Pi_{ki}^2 \Pi_{\ell i}^2] - 2 \sum_{k=1}^{m} \mathbb{E}[\Pi_{ki}^2] + 1
$$

$$
= \frac{n^2}{m} p_i + \frac{n^2 m (m-1)}{m^2} p_i^2 - 2np_i + 1,
$$

and for $i \neq j$, using the fact that $\Pi_{ki} \Pi_{kj} = 0$ whenever $i \neq j$,

$$
\mathbb{E}[\psi_i \psi_j] = \sum_{k=1}^{m} \sum_{\ell=1}^{m} \mathbb{E}[\Pi_{ki}^2 \Pi_{\ell j}^2] - \sum_{k=1}^{m} \mathbb{E}[\Pi_{ki}^2] - \sum_{\ell=1}^{m} \mathbb{E}[\Pi_{\ell j}^2] + 1
$$

$$
= \sum_{k=1}^{m} \mathbb{E}[\Pi_{ki}^2] + \sum_{k=1}^{m} \sum_{\ell=1, \ell \neq k}^{m} \mathbb{E}[\Pi_{ki}^2 \Pi_{\ell j}^2] - \sum_{k=1}^{m} \mathbb{E}[\Pi_{ki}^2] - \sum_{\ell=1}^{m} \mathbb{E}[\Pi_{\ell j}^2] + 1
$$

$$
= \frac{n^2 m (m-1)}{m^2} p_i p_j - np_i - np_j + 1.
$$

Note that $T_{n2} = 0$ because $\varphi_{ij} = 0$. Hence, it suffices to compute the mean and variance of $T_{n1}$. Write

$$
\mathbb{E}(T_{n1}) = n^{-1} \sum_{i=1}^{n} \mathbb{E}(\psi_i) \mathbb{E}(U_i V_i) = n^{-1} \sum_{i=1}^{n} (np_i - 1) \mathbb{E}(U_i V_i) = 0,
$$

$$
\text{Var}(T_{n1}) = n^{-2} \sum_{i=1}^{n} \mathbb{E}(\psi_i^2) \mathbb{E}(U_i^2 V_i^2) + n^{-2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{E}(\psi_i \psi_j) \mathbb{E}(U_i V_i) \mathbb{E}(U_j V_j)
$$

$$
= n^{-2} \sum_{i=1}^{n} \left( \frac{n^2}{m} p_i + \frac{n^2 m (m-1)}{m^2} p_i^2 - 2np_i + 1 \right) \mathbb{E}(U_i^2 V_i^2)
$$

$$
+ n^{-2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left( \frac{n^2 m (m-1)}{m^2} p_i p_j - np_i - np_j + 1 \right) \mathbb{E}(U_i V_i) \mathbb{E}(U_j V_j)
$$

$$
= \left\{ \frac{1}{m} + \left( 1 - \frac{1}{m} \right) \sum_{i=1}^{n} p_i^2 - \frac{1}{n} \right\} \mathbb{E}(U_i^2 V_i^2)
$$

$$
+ \left\{ \left( 1 - \frac{1}{m} \right) \left( 1 - \sum_{i=1}^{n} p_i^2 \right) - \frac{n-1}{n} \right\} \mathbb{E}(U_i V_i) \mathbb{E}(U_j V_j)
$$

$$
= \left\{ \frac{1}{m} - \frac{1}{n} + \left( 1 - \frac{1}{m} \right) \sum_{i=1}^{n} p_i^2 \right\} \left\{ \mathbb{E}(U_i^2 V_i^2) - [\mathbb{E}(U_i V_i)]^2 \right\}.
$$

Therefore, we have proved the lemma. \qed
Lemma A.3. If $\Pi$ is a random matrix satisfying RP, then,

$$E(T_{n1}) = E(T_{n2}) = 0, \quad \text{Var}(T_{n1}) = O(n^{-1}), \quad \text{and} \quad \text{Var}(T_{n2}) = \frac{n-1}{n} \{E(U_i^2)E(V_i^2) + [E(U_iV_i)]^2\}.$$

Proof of Lemma A.3. As in the proof of Lemma A.2, we have that

$$E[\psi_i] = \sum_{k=1}^{m} E[\Pi_{ki}^2] - 1 = 0,$$

$$E[\psi_i^2] = \sum_{k=1}^{m} E[\Pi_{ki}^2] + \sum_{k=1}^{m} \sum_{\ell=1, \ell \neq k}^{m} E[\Pi_{ki}^2 \Pi_{\ell j}^2] - 2 \sum_{k=1}^{m} E[\Pi_{ki}^2] + 1 = O(1) \quad \text{using the assumption that} \ E[\Pi_{ki}^2] = O(m^{-1}),$$

and for $i \neq j$,

$$E[\psi_i\psi_j] = \sum_{k=1}^{m} E[\Pi_{ki}^2 \Pi_{\ell j}^2] + \sum_{k=1}^{m} \sum_{\ell=1, \ell \neq k}^{m} E[\Pi_{ki}^2 \Pi_{\ell j}^2] - \sum_{k=1}^{m} E[\Pi_{ki}^2] - \sum_{\ell=1}^{m} E[\Pi_{\ell j}^2] + 1 = \frac{1}{m} + \frac{m(m-1)}{m^2} - 2 + 1 = 0.$$

Now write

$$E(T_{n1}) = n^{-1} \sum_{i=1}^{n} E(\psi_i)E(U_iV_i) = 0,$$

$$\text{Var}(T_{n1}) = n^{-2} \sum_{i=1}^{n} E(\psi_i^2)E(U_i^2V_i^2) + n^{-2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} E(\psi_i\psi_j)E(U_iV_i)E(U_jV_j) = O(n^{-1}).$$

Furthermore, $E[H(W_i, W_j)] = E[E[H(W_i, W_j)|U_i, U_j, V_i, V_j]] = 0$ by RP(ii), specifically, $E[\Pi_{ki}\Pi_{\ell j}] = 0$. Hence, $E(T_{n2}) = 0$.

For the final result, write

$$E(\{H(W_i, W_j)\}^2) = E\left\{ \left[ \bar{H}(W_i, W_j) + \bar{H}(W_j, W_i) \right] \left[ \bar{H}(W_i, W_j) + \bar{H}(W_j, W_i) \right] \right\}.$$

Write

$$\{H(W_i, W_j)\}^2 = T_{ij1} + T_{ij2} + 2T_{ij3}, \quad (6)$$

where

$$T_{ij1} := \bar{H}(W_i, W_j)\bar{H}(W_i, W_j),$$

$$T_{ij2} := \bar{H}(W_j, W_i)\bar{H}(W_j, W_i),$$

$$T_{ij3} := \bar{H}(W_i, W_j)\bar{H}(W_j, W_i).$$

Use RP(ii)-(iii), in particular, $E[\Pi_{ki}^2 \Pi_{\ell j}^2] = m^{-2}$ and $E[\Pi_{ki}\Pi_{\ell j}\Pi_{k'i}\Pi_{\ell'j}] = 0$ whenever $k \neq \ell, i \neq j$, to obtain

$$E[T_{ij1}|U_i, V_i, U_j, V_j] = \sum_{k=1}^{m} U_i^2V_j^2E[\Pi_{ki}^2 \Pi_{\ell j}^2] = \frac{1}{m} U_i^2V_j^2,$$

$$E[T_{ij2}|U_i, V_i, U_j, V_j] = \sum_{k=1}^{m} U_j^2V_i^2E[\Pi_{ki}^2 \Pi_{\ell j}^2] = \frac{1}{m} U_j^2V_i^2,$$

$$E[T_{ij3}|U_i, V_i, U_j, V_j] = \sum_{k=1}^{m} U_iU_jV_iV_jE[\Pi_{ki}^2 \Pi_{\ell j}^2] = \frac{1}{m} U_iU_jV_iV_j.$$
Assume $H$.

Then, the desired result follows immediately.

Thus,
\[
E[\{H(W_i, W_j)\}^2] = E[E[\{H(W_i, W_j)\}^2 | U_i, V_i, U_j, V_j]]
\]
\[
= \frac{1}{m} E \left( U_i^2 V_j^2 + U_j^2 V_i^2 + U_i U_j V_i V_j \right)
\]
\[
= \frac{2}{m} (E[U_i^2]E[V_i^2] + [E(U_i V_i)]^2).
\]

Then, $\text{Var} (T_{n,2})$ can be obtained by combining the U statistic formula given in (5) with RP(iii).

**Proof of Theorem 3.1.** The theorem for RS follows immediately from Lemma A.2 under the condition that $\sum_{i=1}^n p_i^2 = o(m^{-1})$. The theorem for RP follows straightforwardly from Lemma A.3 using Cauchy–Schwarz inequality. We now consider BS. Recall that
\[
n^{-1} (U^T \Pi^T \Pi V - U^T V) = n^{-1} \sum_{i=1}^n \left( \frac{n}{m} B_{ii} - 1 \right) U_i V_i
\]
and that the summands are i.i.d.,
\[
E \left[ \left( \frac{n}{m} B_{ii} - 1 \right) U_i V_i \right] = 0, \quad \text{and} \quad \text{Var} \left[ \left( \frac{n}{m} B_{ii} - 1 \right) U_i V_i \right] = \left( \frac{n}{m} - 1 \right) E[ U_i^2 V_i^2 ].
\]

Then, the desired result follows immediately.

In order to prove Theorem 3.3(ii), we use the central limit theorem for degenerate $U$-statistics of Hall (1984). For the sake of easy referencing, we reproduce it below.

**Lemma A.4 (Theorem 1 of Hall (1984)).** Assume that $\{W_1, \ldots, W_n\}$ are independent and identically distributed random vectors. Define
\[
U_n := \sum_{1 \leq i < j \leq n} H_n(W_i, W_j).
\]
Assume $H_n$ is symmetric, $E[H_n(W_1, W_2)|W_1] = 0$ almost surely and $E[H_n^2(W_1, W_2)] < \infty$ for each $n$. Let $G_n(w_1, w_2) := E[H_n(W_1, w_1)H_n(W_1, w_2)]$. If
\[
\frac{E[G_n^2(W_1, W_2)] + n^{-1} E[H_n^4(W_1, W_2)]}{\{E[H_n^2(W_1, W_2)]\}^2} \to 0
\]
as $n \to \infty$, then
\[
\sqrt{n} \frac{U_n}{n} \to_d N(0, 1),
\]
where $V_n := \frac{1}{2} E[H_n^4(W_1, W_2)]$.

**Lemma A.5.** Let $\Pi$ be a random matrix satisfying RP. Let $G(w_1, w_2) := E[H(W_1, w_1)H(W_1, w_2)]$. Then,
\[
E[G^2(W_i, W_j)] = O(m^{-3}).
\]

**Proof.** First, write
\[
G(w_1, w_2) = E \left\{ \left[ \tilde{H}(W_i, w_1) + \tilde{H}(w_1, W_i) \right] \left[ \tilde{H}(W_i, w_2) + \tilde{H}(w_2, W_i) \right] \right\}.
\]
Because $E[\Pi_k^2] = m^{-1}$ and $E[\Pi_k \Pi_k^T] = 0$ whenever $k \neq l$ for each $i$, we have that
\[
E[\tilde{H}(W_i, w_1) \tilde{H}(W_i, w_2)|U_i, V_i, U_j, V_j] = \sum_{k=1}^m \sum_{j=1}^m U_i^2 v_1 v_2 E[\Pi_k \Pi_j] \pi_{k_1} \pi_{j_2}
\]
\[
= \sum_{k=1}^m U_i^2 v_1 v_2 E[\Pi_k^2] \pi_{k_1} \pi_{k_2}
\]
\[
= m^{-1} \sum_{k=1}^m U_i^2 v_1 v_2 \pi_{k_1} \pi_{k_2}.
\]
Similarly,
\[
\mathbb{E}[\tilde{H}(w_1, W_i) \tilde{H}(w_2, W_i)|U_i, V_i, U_j, V_j] = m^{-1} \sum_{k=1}^{m} u_1 u_2 V_i^2 \sigma_{k1} \sigma_{k2},
\]
\[
\mathbb{E}[\tilde{H}^2(W_s, w_1) \tilde{H}(w_2, W_i)|U_i, V_i, U_j, V_j] = m^{-1} \sum_{k=1}^{m} U_i u_2 v_1 V_i \pi_{k1} \pi_{k2},
\]
\[
\mathbb{E}[\tilde{H}(w_1, W_i) \tilde{H}(W_i, w_2)|U_i, V_i, U_j, V_j] = m^{-1} \sum_{k=1}^{m} u_1 U_i v_2 \pi_{k1} \pi_{k2}.
\]

Then, by simple algebra,
\[
G(W_i, W_j) = m^{-1} \sum_{k=1}^{m} \{ \mathbb{E}[U_i^2] V_i V_j + \mathbb{E}[V_i^2] U_i U_j + \mathbb{E}[U_i V_i] U_i V_j + \mathbb{E}[U_i V_i] U_i V_j \} \Pi_{ki} \Pi_{kj}.
\]

Using RP(ii)-(iii), write
\[
\mathbb{E}[G^2(W_i, W_j)] = \mathbb{E}[\mathbb{E}[G^2(W_i, W_j)|U_i, V_i, U_j, V_j]]
\]
\[
= m^{-2} \sum_{k=1}^{m} \mathbb{E} \left[ \left( \mathbb{E}[U_i^2] V_i V_j + \mathbb{E}[V_i^2] U_i U_j + \mathbb{E}[U_i V_i] U_i V_j + \mathbb{E}[U_i V_i] U_i V_j \right)^2 \right] \mathbb{E}[\Pi_{ki}^2 \Pi_{kj}^2]
\]
\[
= m^{-4} \sum_{k=1}^{m} \mathbb{E} \left[ \left( \mathbb{E}[U_i^2] V_i V_j + \mathbb{E}[V_i^2] U_i U_j + \mathbb{E}[U_i V_i] U_i V_j + \mathbb{E}[U_i V_i] U_i V_j \right)^2 \right]
\]
\[
= O(m^{-3}),
\]
which proves the lemma.

**Lemma A.6.** Let \( \Pi \) be a random matrix satisfying RP. Furthermore, assume that the columns of \( \Pi \) are i.i.d. Then, for \( i \neq j \), \( \mathbb{E}[(H(W_i, W_j))^4] = O(m^{-1}) \).

**Proof of Lemma A.6.** Using (6), write
\[
\{H(W_i, W_j)\}^4 = (T_{ij1} + T_{ij2} + 2T_{ij3})(T_{ij1} + T_{ij2} + 2T_{ij3}).
\]

We expand the right-hand side of the equation above. The first term has the form
\[
T_{ij1} T_{ij1} = \sum_{k_1=1}^{m} \sum_{k_2=1}^{m} \sum_{k_3=1}^{m} \sum_{k_4=1}^{m} U_i^4 V_j^4 \Pi_{k1} \Pi_{k1} \Pi_{k2} \Pi_{k2} \Pi_{k3} \Pi_{k3} \Pi_{k4} \Pi_{k4} \Pi_{k4}.
\]

Combining RP with the additional assumption that the columns of \( \Pi \) are i.i.d., we have that \( \mathbb{E}[\Pi_{k1}^2 \Pi_{k2}^2] = O(m^{-2}) \) uniformly. Also, \( \mathbb{E}[\Pi_{k1} \Pi_{k1} \Pi_{k2} \Pi_{k2} \Pi_{k3} \Pi_{k3} \Pi_{k4} \Pi_{k4}] \) is nonzero only if all four indices are the same \( (k_1 = k_2 = k_3 = k_4), \) or two pairs of the indices are the same \( (e.g., k_1 = k_2, k_3 = k_4). \) This implies that
\[
\mathbb{E}[T_{ij1} T_{ij1}] = \mathbb{E} \left( U_i^4 V_j^4 \right) \sum_{k=1}^{m} \mathbb{E}[\Pi_{k1}^2 \Pi_{k2}^2] + 6 \sum_{k=1}^{m} \sum_{\ell=1, \ell \neq k}^{m} \mathbb{E}[\Pi_{k1}^2 \Pi_{k2}^2 \Pi_{k1}^2 \Pi_{k2}^2] = O(m^{-1}).
\]

Moreover, using similar arguments, we can show that all other terms \( \mathbb{E}[T_{ijk} T_{ij\ell}] = O(m^{-1}) \), where \( k, \ell \in \{1, 2, 3\} \). Therefore, we have proved the lemma.

**Lemma A.7.** Let \( \Pi \) be a random matrix satisfying RP. Furthermore, assume that the columns of \( \Pi \) are i.i.d. Then, as \( n \to \infty \),
\[
\sqrt{m} T_{n2} \to_d N[0, \{ \mathbb{E}(U_i^2) \mathbb{E}(V_i^2) + \mathbb{E}(U_i V_i)^2 \}],
\]

Least Squares Estimation using Sketched Data with Heteroskedastic Errors
Proof of Lemma A.7. Note that $H(w_1, w_2) = H(w_2, w_1)$ and $\mathbb{E}(H(W_1, W_2)|W_1) = \mathbb{E}(H(W_1, W_2)|W_2) = 0$. Thus, $T_{n_2}$ is a degenerate $U$-statistic. By Lemmas A.3, A.5 and A.6, we have that
\[
\frac{\mathbb{E}[G^2(W_1, W_2)] + n^{-1}\mathbb{E}[H^4(W_1, W_2)]}{\{\mathbb{E}[H^2(W_1, W_2)]\}^2} = O(m^{-1} + n^{-1}m) = o(1).
\]
Then, the conclusion of Lemma A.7 follows directly by applying Lemma A.4 along with Lemma A.3.

Proof of Theorem 3.3(ii). Recall that
\[
\tilde{\beta}_{OLS} - \hat{\beta}_{OLS} = \Lambda(\tilde{g}_n - \hat{g}_n) + o_p(m^{-1/2}).
\]
Then, the theorem follows immediately by applying Lemma A.7 with $U_i = c^T X_i$ and $V_i = e_i$ for each constant vector $c \in \mathbb{R}^p$.

B. Appendix: Proofs for 2SLS

B.1. Proof of Theorem 4.2

Recall that using the singular value decomposition of $X$ and $Z$, we write $X = U_X \Sigma_X V_X^T$ and $Z = U_Z \Sigma_Z V_Z^T$. Define
\[
\tilde{\theta} := (U_X^T U_Z U_Z^T U_X)^{-1} U_X^T U_Z^T U_X y \tag{7}
\]
and
\[
\hat{\theta} := (U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi U_X)^{-1} U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi y. \tag{8}
\]
It would be convenient to work with $U_X \tilde{\theta}$ and $U_X \hat{\theta}$ in order to analyze algorithmic properties of sketched 2SLS estimators because $U_X$ is an orthonormal matrix. The following lemma establishes the equivalence between $X \tilde{\beta}$ and $U_X \hat{\theta}$.

Lemma B.1. Let Assumption 4.1 hold. Then, $X \tilde{\beta} = U_X \hat{\theta}$.

Proof. By the singular value decomposition of $X$ and $Z$, we have that
\[
Z^T Z = V_Z \Sigma_Z^2 V_Z^T,
\]
\[
Z(Z^T Z)^{-1} Z^T = U_Z U_Z^T,
\]
\[
X^T Z(Z^T Z)^{-1} Z^T X = V_X \Sigma_X U_X^T U_Z U_Z^T U_X \Sigma_X V_X^T,
\]
\[
(X^T Z(Z^T Z)^{-1} Z^T X)^{-1} = V_X \Sigma_X^{-1} (U_X^T U_Z U_Z^T U_X)^{-1} \Sigma_X^{-1} V_X^T,
\]
\[
X^T Z(Z^T Z)^{-1} Z^T y = V_X \Sigma_X U_X^T U_Z U_Z^T y.
\]
Therefore,
\[
\tilde{\beta} = V_X \Sigma_X^{-1} (U_X^T U_Z U_Z^T U_X)^{-1} U_X^T U_Z^T U_Z^T y,
\]
\[
X \tilde{\theta} = U_X (U_X^T U_Z U_Z^T U_X)^{-1} U_X^T U_Z^T U_Z^T y,
\]
which in turn implies the conclusion in view of the definition of $\tilde{\theta}$ in (7).

As in Lemma B.1, the equivalence between $X \tilde{\beta}$ and $U_X \hat{\theta}$ holds.

Lemma B.2. Assume that (i) $\tilde{Z}^T \tilde{Z}$ is non-singular and (ii) $X^T \tilde{Z}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \tilde{X}$ is non-singular. Then, $X \tilde{\beta} = U_X \hat{\theta}$.
Proof. As in the proof of Lemma B.1, we have that

\[ \begin{align*}
\tilde{Z}^T \tilde{Z} &= V_Z \Sigma_Z U_Z^T \Pi^T \Pi U_Z \Sigma_Z^T, \\
(\tilde{Z}^T \tilde{Z})^{-1} &= V_Z \Sigma_Z^{-1} (U_Z^T \Pi^T \Pi U_Z)^{-1} \Sigma_Z^{-1} V_Z^T, \\
\tilde{Z}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T &= \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T, \\
\tilde{X}^T \tilde{Z}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{X} &= X^T \Sigma_X U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi U_X \Sigma_X V_X^T, \\
\left(\tilde{X}^T \tilde{Z}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{X}\right)^{-1} &= V_X \Sigma_X^{-1} \left(U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi U_X \right)^{-1} \Sigma_X^{-1} V_X^T, \\
\tilde{X}^T \tilde{Z}(\tilde{Z}^T \tilde{Z})^{-1} \tilde{X}^T \tilde{y} &= V_X \Sigma_X U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi y.
\end{align*}\]

Therefore,

\[ \begin{align*}
\tilde{\beta} &= V_X \Sigma_X^{-1} \left(U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi U_X \right)^{-1} U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi y, \\
X\tilde{\beta} &= U_X \left(U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi U_X \right)^{-1} U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi y.
\end{align*}\]

which again implies the conclusion in view of the definition of \( \tilde{\beta} \) in (8). \( \square \)

Abusing the notation a bit, define now

\[ \begin{align*}
\tilde{A} := U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi U_X, \\
\hat{A} := U_X^T U_Z U_Z^T U_X, \\
\tilde{B} := U_X^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi \hat{\epsilon}, \\
\hat{B} := U_X^T U_Z U_Z^T \hat{\epsilon}.
\end{align*}\]

Lemma B.3. Let Assumption 4.1 hold. Then, \( \hat{A}^{-1} \hat{B} = 0 \).

Proof. Note that

\[ \begin{align*}
\hat{A}^{-1} \hat{B} &= (U_X^T U_Z U_Z^T U_X)^{-1} U_X^T U_Z U_Z^T \hat{\epsilon} \\
&= (U_X^T U_Z U_Z^T U_X)^{-1} U_X^T U_Z U_Z^T (X\tilde{\beta} - (U_X^T U_Z U_Z^T U_X)^{-1} U_X^T U_Z U_Z^T y) \\
&= 0,
\end{align*}\]

since \( X\tilde{\beta} = U_X (U_X^T U_Z U_Z^T U_X)^{-1} U_X^T U_Z U_Z^T y \). \( \square \)

Under Assumption 4.1, we first obtain the following lemma.

Lemma B.4. Let Assumption 4.1 hold. Then, the following holds jointly with probability at least \( 1 - \delta \):

\[ \begin{align*}
\|\tilde{A} - \hat{A}\|_2 &\leq f_1(\varepsilon_1, \varepsilon_2), \\
\|\tilde{B} - \hat{B}\|_2 &\leq \varepsilon_3 \|\hat{\epsilon}\| + f_2(\varepsilon_1, \varepsilon_2) \left[ 1 + \varepsilon_3 \|\hat{\epsilon}\| \right].
\end{align*}\]
Proof. Let $\tilde{A}_1 := U_Z^T \Pi^T \Pi U_X$, $\tilde{A}_2 := (U_Z^T \Pi^T \Pi U_Z)^{-1}$, $\hat{A}_1 := U_Z^T U_X$, and $\hat{A}_2 := I$. Then we have that

$$
\tilde{A} - \hat{A} = \tilde{A}_1^T \tilde{A}_2 \tilde{A}_1 - \tilde{A}_1^T \tilde{A}_2 \hat{A}_1
$$

$$
= (\tilde{A}_1 - \hat{A}_1)^T \tilde{A}_2 (\tilde{A}_1 - \hat{A}_1) + \tilde{A}_1^T \tilde{A}_2 (\tilde{A}_1 - \hat{A}_1) + (\tilde{A}_1 - \hat{A}_1)^T \tilde{A}_2 (\tilde{A}_2 - \hat{A}_2) \tilde{A}_1
$$

$$
= (\tilde{A}_1 - \hat{A}_1)^T \tilde{A}_2 (\tilde{A}_1 - \hat{A}_1) + (\tilde{A}_1 - \hat{A}_1)^T (\tilde{A}_2 - \hat{A}_2) (\tilde{A}_1 - \hat{A}_1)
$$

$$
+ \tilde{A}_1^T \tilde{A}_2 (\tilde{A}_1 - \hat{A}_1) + \tilde{A}_1^T (\tilde{A}_2 - \hat{A}_2) (\tilde{A}_1 - \hat{A}_1)
$$

$$
+ \tilde{A}_1^T (\tilde{A}_2 - \hat{A}_2) \tilde{A}_1.
$$

It is straightforward to show that $\|\tilde{A}_2 - \hat{A}_2\|_2 \leq \varepsilon_1/(1 - \varepsilon_1)$ using Assumption 4.1(i). Since $\|\tilde{A}_1\|_2 \leq \|U_Z\|_2 \|U_X\|_2 = 1$ and $\|\tilde{A}_2\|_2 = 1$, we have that

$$
\|\tilde{A} - \hat{A}\|_2 \leq \varepsilon_2^2 + \varepsilon_2^2 \varepsilon_1/(1 - \varepsilon_1) + 2\varepsilon_2 + 2\varepsilon_2 \varepsilon_1/(1 - \varepsilon_1) + \varepsilon_1/(1 - \varepsilon_1)
$$

$$
= \varepsilon_1 + \varepsilon_2 (\varepsilon_2 + 2)/(1 - \varepsilon_1) = f_1(\varepsilon_1, \varepsilon_2),
$$

using Assumption 4.1. This proves the first desired result.

Now let $\tilde{B}_1 := U_Z^T \Pi^T \Pi \hat{e}$ and $\tilde{B}_1 := U_Z^T \hat{e}$. Consider

$$
\tilde{B} - \hat{B} = U_Z^T \Pi^T \Pi U_Z (U_Z^T \Pi^T \Pi U_Z)^{-1} U_Z^T \Pi^T \Pi \hat{e} - U_Z^T U_Z U_Z^T \hat{e}
$$

$$
= \tilde{A}_1^T \tilde{A}_2 \tilde{B}_1 - \tilde{A}_1^T \tilde{A}_2 \hat{B}_1
$$

$$
= (\tilde{A}_1 - \hat{A}_1)^T \tilde{A}_2 (\tilde{B}_1 - \hat{B}_1) + \tilde{A}_1^T \tilde{A}_2 (\tilde{B}_1 - \hat{B}_1) + (\tilde{A}_1 - \hat{A}_1)^T (\tilde{A}_2 - \hat{A}_2) (\tilde{B}_1 - \hat{B}_1)
$$

$$
+ \tilde{A}_1^T (\tilde{A}_2 - \hat{A}_2) (\tilde{B}_1 - \hat{B}_1)
$$

$$
+ (\tilde{A}_1 - \hat{A}_1)^T \tilde{A}_2 \tilde{B}_1 + (\tilde{A}_1 - \hat{A}_1)^T (\tilde{A}_2 - \hat{A}_2) \tilde{B}_1
$$

$$
+ \tilde{A}_1^T (\tilde{A}_2 - \hat{A}_2) \tilde{B}_1.
$$

Since $\|\tilde{B}_1\|_2 = \|U_Z^T \hat{e}\|_2 \leq \|U_Z\|_2 \|\hat{e}\| \leq \|\hat{e}\|$, we have that

$$
\|\tilde{B} - \hat{B}\|_2 \leq \varepsilon_4 \|\hat{e}\| + \varepsilon_4 \varepsilon_1/(1 - \varepsilon_1) + \varepsilon_2 \varepsilon_1/(1 - \varepsilon_1) \|\hat{e}\| + \varepsilon_3 \|\hat{e}\|
$$

$$
= \varepsilon_4 \|\hat{e}\| + f_2(\varepsilon_1, \varepsilon_2) \|\hat{e}\|, \varepsilon_4 \|\hat{e}\| + f_2(\varepsilon_1, \varepsilon_2) \|\hat{e}\|,
$$

again using Assumption 4.1. This proves the second desired result.

\[ \square \]

Lemma B.5. Let Assumptions 4.1 hold. Then, the following holds with probability at least $1 - \delta$:

$$
\sigma_{\min}(\tilde{A}) \geq \frac{1}{2} \sigma_{\min}^2(U_Z^T U_X).
$$

Proof. Use the fact that for real matrices $C$ and $D,$

$$
\sigma_{\min}(C + D) \geq \sigma_{\min}(C) - \sigma_{\max}(D)
$$

to obtain

$$
\sigma_{\min}(\tilde{A}) \geq \sigma_{\min}(\tilde{A}) - \sigma_{\max}(\tilde{A} - \tilde{A}).$$

Least Squares Estimation using Sketched Data with Heteroskedastic Errors
Then the desired result follows from the first conclusion of Lemma B.4, since
\[
\sigma_{\min}(\hat{A}) = \sigma_{\min}(U_X^2 U_Z U_Z^T U_X) = \sigma_{\min}^2(U_Z^T U_X) \quad \text{and} \quad \sigma_{\max}(\hat{A} - \hat{A}) \leq \|\hat{A} - \hat{A}\|_2^2.
\]

Lemma B.5 implies that \(\hat{A}^{-1}\) is well defined with probability at least \(1 - \delta\).

**Lemma B.6.** Let Assumptions 4.1 hold. Then, the following holds with probability at least \(1 - \delta\):
\[
\left\|\hat{A}^{-1} - \hat{A}^{-1}\right\|_2 \leq \frac{2f_1(\varepsilon_1, \varepsilon_2)}{\sigma_{\min}^4(U_Z^T U_X)}.
\]

**Proof.** Write
\[
\hat{A}^{-1} - \hat{A}^{-1} = \hat{A}^{-1} (\hat{A} - \hat{A}) \hat{A}^{-1}.
\]
Thus,
\[
\left\|\hat{A}^{-1} - \hat{A}^{-1}\right\|_2 \leq \left\|\hat{A}^{-1}\right\|_2 \left\|\hat{A} - \hat{A}\right\|_2 \left\|\hat{A}^{-1}\right\|_2
\]
\[
\leq \frac{2f_1(\varepsilon_1, \varepsilon_2)}{\sigma_{\min}^4(U_Z^T U_X)}
\]
since \(\|\hat{A}^{-1}\|_2 = [\sigma_{\min}^2(U_Z^T U_X)]^{-1}\), by Lemma B.4, \(\|\hat{A} - \hat{A}\|_2 \leq f_1(\varepsilon_1, \varepsilon_2)\) and, by Lemma B.5, \(\|\hat{A}^{-1}\|_2 \leq 2 [\sigma_{\min}^2(U_Z^T U_X)]^{-1}\) with probability at least \(1 - \delta\).

**Proof of Theorem 4.2.** By Lemmas B.1 and B.2,
\[
X(\hat{\beta} - \beta) = U_X(\hat{\theta} - \hat{\theta}),
\]
so that
\[
\sigma_{\min}(X) \left\|\hat{\beta} - \beta\right\| \leq \left\|\hat{\theta} - \hat{\theta}\right\|.
\]
Thus, it suffices to bound \(\left\|\hat{\theta} - \hat{\theta}\right\|\). To do so, write
\[
\hat{y} = \Pi \left(X\hat{\beta} + \hat{\epsilon}\right) = \hat{X}\hat{\beta} + \hat{\epsilon} = \Pi U_X \hat{\theta} + \hat{\epsilon},
\]
where \(\hat{\epsilon} = \Pi \hat{\epsilon}\). Plugging (9) into (8) yields
\[
\hat{\theta} - \hat{\theta} = \hat{A}^{-1} \hat{B}.
\]
Then, by Lemma B.3, we have that \(\hat{\theta} - \hat{\theta} = \hat{A}^{-1} \hat{B} = \hat{A}^{-1} \hat{B} - \hat{A}^{-1} \hat{B}\). Further, write
\[
\hat{\theta} - \hat{\theta} = (\hat{A}^{-1} - \hat{A}^{-1}) \hat{B} + \hat{A}^{-1} (\hat{B} - \hat{B}) + (\hat{A}^{-1} - \hat{A}^{-1}) (\hat{B} - \hat{B}).
\]
Thus,
\[
\left\|\hat{\theta} - \hat{\theta}\right\| = \left\|\hat{A}^{-1} \hat{B} - \hat{A}^{-1} \hat{B}\right\|_2
\]
\[
\leq \left\|\hat{A}^{-1} - \hat{A}^{-1}\right\|_2 \left\|\hat{B}\right\|_2 + \left\|\hat{A}^{-1}\right\| \left\|\hat{B} - \hat{B}\right\|_2 + \left\|\hat{A}^{-1} - \hat{A}^{-1}\right\| \left\|\hat{B} - \hat{B}\right\|_2
\]
\[
\leq \frac{2f_1(\varepsilon_1, \varepsilon_2)}{\sigma_{\min}^4(U_Z^T U_X)} \|\hat{\epsilon}\| + \frac{\varepsilon_3 \|\hat{\epsilon}\| + f_2(\varepsilon_1, \varepsilon_2) [1 + \varepsilon_3 \|\hat{\epsilon}\|]}{\sigma_{\min}^4(U_Z^T U_X)}
\]
\[
+ \frac{2f_1(\varepsilon_1, \varepsilon_2)}{\sigma_{\min}^4(U_Z^T U_X)} \left\{\varepsilon_3 \|\hat{\epsilon}\| + f_2(\varepsilon_1, \varepsilon_2) [1 + \varepsilon_3 \|\hat{\epsilon}\|]\right\}
\]
\[
= \frac{f_2(\varepsilon_1, \varepsilon_2) + \varepsilon_3 \|\hat{\epsilon}\| [1 + f_2(\varepsilon_1, \varepsilon_2)]}{\sigma_{\min}^4(U_Z^T U_X)} \left[1 + \frac{2f_1(\varepsilon_1, \varepsilon_2)}{\sigma_{\min}^4(U_Z^T U_X)}\right],
\]
where the last inequality follows from Assumption 4.1.
Least Squares Estimation using Sketched Data with Heteroskedastic Errors

We now specialize Theorem 4.2 for countsketch.

**Theorem B.7.** Let data $D_n$ be fixed, $Z^TZ$ and $X^TP_ZX$ are non-singular. Let $\Pi \in \mathbb{R}^{m \times n}$ be countsketch with $m \geq \max\{q(q + 1), 2pq\}/(\varepsilon^2\delta)$ for some $\varepsilon \in (0, 1/3]$. Suppose that $\sigma_{min}^2(U_Z^TU_X) \geq \frac{16\varepsilon(1+\varepsilon)}{1-\varepsilon}$. and let $\sigma^* = \left[\sigma_{min}(X)\sigma_{min}^2(U_Z^TU_X)\right]^{-1}$. Then, the following holds with probability at least $1 - \delta$:

$$\|\hat{\beta}_{2SLS} - \beta_{2SLS}\| \leq \frac{4\varepsilon}{1-\varepsilon} \left[2 + \frac{3\|\tilde{e}\|}{\sqrt{p}}\right] \sigma^*$$

To establish Lemma B.10 given below, we first state the following known results in the literature.

**Lemma B.8** (Theorem 6.2 of Kane & Nelson (2014)). Distribution $D$ over $\mathbb{R}^{m \times n}$ is defined to have $(\varepsilon, \delta, 2)$-JL (Johnson-Lindenstrauss) moments if for any $x \in \mathbb{R}^n$ such that $\|x\| = 1$,

$$E_{\Pi \sim D}\left[\|\Pi x\|^2 - 1\right]^2 \leq \varepsilon^2 \delta.$$

Given $\varepsilon, \delta \in (0, 1/2)$, let $D$ be any distribution over matrices with $n$ columns with the $(\varepsilon, \delta, 2)$-JL moment property. Then, for any $A$ and $B$ real matrices each with $n$ rows,

$$P_{\Pi \sim D}\left(\|A^T\Pi^T\Pi B - A^T B\|_F > 3\varepsilon\|A\|_F\|B\|_F\right) < \delta.$$

**Lemma B.9** (Theorem 2.9 of Woodruff (2014)). Let $\Pi \in \mathbb{R}^{m \times n}$ be countsketch with $m \geq 2/((\varepsilon^2\delta))$. Then, $\Pi$ satisfies the $(\varepsilon, \delta, 2)$-JL moment property.

**Lemma B.10.** Let $\Pi \in \mathbb{R}^{m \times n}$ be countsketch with $m \geq \max\{q(q + 1), 2pq\}/(\varepsilon^2\delta)$ for some $\varepsilon \in (0, 1/2)$. Then, Assumption 4.1 holds with $\varepsilon_1 = \varepsilon, \varepsilon_2 = 3\varepsilon, \varepsilon_3 = 3\varepsilon^{p^{-1/2}}$.

**Proof of Lemma B.10.** As shown in the proof of Theorem 2 of Nelson & Nguyễn (2013),

$$P_{\Pi \sim D}\left(\|U_Z^T\Pi^T\Pi U_Z - I_q\|_2 > \varepsilon\right) < \delta,$$

provided that $m \geq q(q + 1)/(\varepsilon^2\delta)$. This verifies the first condition of Assumption 4.1.

Now to verify conditions (ii) and (iii) of Assumption 4.1, note that since countsketch with $m \geq 2/((\varepsilon^2\delta))$ satisfies the $(\varepsilon, \delta, 2)$-JL moment property, we have, for any $A$ and $B$ real matrices each with $n$ rows,

$$P_{\Pi \sim D}\left(\|A^T\Pi^T\Pi B - A^T B\|_2 > 3\varepsilon\|A\|_F\|B\|_F\right) < \delta.$$

Since $\|U_X\|_F^2 = p, \|U_Z\|_F^2 = q$ and $\|\tilde{e}\|_F = \|\tilde{e}\|$, we have that

$$P_{\Pi \sim D}\left(\|U_Z^T\Pi^T\Pi U_X - U_Z^T U_X\|_2 > 3\varepsilon\sqrt{pq}\right) < \delta,$$

provided that $m \geq 2/((\varepsilon^2\delta))$. Replacing $\varepsilon$ with $\varepsilon/\sqrt{pq}$ yields that

$$P_{\Pi \sim D}\left(\|U_Z^T\Pi^T\Pi U_X - U_Z^T U_X\|_2 > 3\varepsilon\right) < \delta,$$

provided that $m \geq 2pq/(\varepsilon^2\delta)$. Thus, we have proved Lemma B.10.

**Proof of Theorem B.7.** In view of Lemma B.10, this theorem follows directly from applying Theorem 4.2 to the case when $\Pi$ is a countsketch.
B.2. Proof of Theorem 4.4

Proof of Theorem 4.4. It follows from the definition of the estimator that
\[
\tilde{\beta} = \left( \tilde{X}^T \tilde{Z} (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \right)^{-1} \tilde{X}^T \tilde{Z} (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \left( \tilde{X} \beta_0 + \tilde{e} \right).
\]
Thus,
\[
\hat{\beta} = \beta_0 + \left( (X^T Z/n) (Z^T Z/n)^{-1} (Z^T X/n) \right)^{-1} (X^T Z/n) (Z^T Z/n)^{-1} (Z^T e/n).
\]
Write
\[
\tilde{\beta} - \hat{\beta} = (A_n - A) g_n + A (\tilde{g}_n - g_n) + (\tilde{A}_n - A) (\tilde{g}_n - g_n),
\]
where \( \tilde{g}_n = \tilde{Z}^T e/n, g_n = Z^T e/n, \)
\[
\tilde{A}_n = \left( (X^T Z/n) (Z^T Z/n)^{-1} (Z^T X/n) \right)^{-1} (X^T Z/n) (Z^T Z/n)^{-1},
\]
\[
A = \left[ E(X_i Z_i^T) \right] \left[ E(Z_i Z_i^T) \right]^{-1} E(Z_i X_i^T) \left[ E(Z_i Z_i^T) \right]^{-1}.
\]
As in the proof for OLS, \( \tilde{A}_n - A = o_p(1) \) and \( \tilde{g}_n - g_n = O_p(m^{-1/2}) \). By the central limit theorem, \( g_n = O_p(n^{-1/2}) \). Hence,
\[
\tilde{\beta} - \hat{\beta} = A (\tilde{g}_n - g_n) + o_p \left( n^{-1/2} + m^{-1/2} \right).
\]
Moreover, by Lemma A.7 that
\[
m^{1/2} (\tilde{g}_n - g_n) \rightarrow_d N[0, \mathbb{E}(e_i^2 Z_i Z_i^T)].
\]
Combining all the arguments above yields
\[
m^{1/2} (\bar{\beta} - \hat{\beta}) \rightarrow_d N[0, A \mathbb{E}(e_i^2 Z_i Z_i^T) A^T],
\]
which gives the conclusion of the theorem. \( \square \)