A GEOMETRIC ANALYSIS OF MODEL- AND ALGORITHM-INDUCED UNCERTAINTIES FOR RANDOMIZED LEAST SQUARES REGRESSION*

JOCELYN T. CHI† AND ILSE C. F. IPSEN‡

Abstract. For full-rank least squares regression problems under a Gaussian linear model, we analyze the uncertainties when the minimum-norm solution is computed by random row-sketching and, in particular random row-sampling. Our expressions for the total expectation and variance of the solution–with regard to both model- and algorithm-induced uncertainties– are exact; hold for general sketching matrices; and make no assumptions on the rank of the sketched matrix. They show that expectation and variance are governed by the rank-deficiency and spatial geometry induced by the sketching process, rather than by structural properties of specific sketching or sampling methods. In order to analyze the rank-deficient matrices from row-sketching, we introduce two projectors that connect least squares problems of different dimensions.

From a deterministic perspective, our structural perturbation bounds imply that least squares solutions are less sensitive to multiplicative perturbations than to additive perturbations. From a probabilistic perspective, we show that the differences between the total bias and variance on the one hand, and the model bias and variance on the other hand, are governed by two factors: (i) the expected rank deficiency of the sketched matrix, and (ii) the expected difference between projectors onto the spaces of the original and the sketched problems. Surprisingly, the matrix condition number has far less impact on the statistical quantities than it has on numerical errors.

Key words. Condition number with respect to inversion, projector, multiplicative perturbations, Moore Penrose inverse, expectation, variance, matrix valued random variable

AMS subject classification. 62J05, 62J10, 65F20, 65F22, 65F35, 68W20

1. Introduction.

We consider the randomized solution of least squares regression problems under the Gaussian linear model, and analyze the effect of both: the statistical noise in the model, as well as the error due to algorithmic randomization. Our analysis extends the pioneering work [15, 16] through rigorous validation in a general setting, and demonstrates that expectation and variance are governed by geometry rather than by structural properties of specific classes of sketching matrices: What matters is the rank deficiency induced by the sketching process, and the failure of the sketched matrix to reproduce the original column space.

1.1. Problem setting.

We start with a regression problem under the Gaussian linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{1.1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a given design matrix with $\operatorname{rank}(\mathbf{X}) = p$, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the true but unknown parameter vector, and the noise vector $\epsilon \in \mathbb{R}^n$ has a standard multivariate normal distribution. For a fixed response vector $\mathbf{y} \in \mathbb{R}^n$, the unique maximum likelihood estimator of $\boldsymbol{\beta}_0$ is the solution $\hat{\boldsymbol{\beta}}$ of the full-rank least squares problem[1]

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2. \tag{1.2}$$

*The work was supported in part by NSF grants DGE-1633587, DMS-1760374, and DMS-1745654.

†Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA, jtchi@ncsu.edu

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695, ipsen@ncsu.edu

[1]Here $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$ represents the Euclidean two-norm, and the superscript T the transpose.

Solution of this least squares problem via random row-sketching,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{S}(\mathbf{X}\boldsymbol{\beta}-\mathbf{y})\|_2, \tag{1.3}$$

is an effective approach in the highly over-constrained case [5, 6, 16, 22, 28] where observations far outnumber covariates, that is, $\mathbf{X}$ is tall and skinny with $n \gg p$. Here $\mathbf{S} \in \mathbb{R}^{r\times n}$ is a random sketching matrix with $r \leq n$, and the minimum norm solution is $\tilde{\boldsymbol{\beta}}$.

1.2. Existing work. Random sketching is a form of preconditioning and seems to have originated in [24]. By now, there are many variants which can be classified according to [26, Section 1]: Compression of rows [2, 5, 6, 13, 15, 16, 23, 28]; or columns [1]; or both [18]. Matrix concentration inequalities are used to derive probabilistic bounds for the error due to randomization [1, 6], and for the condition number of the sampled matrix [13]. From a practical perspective, bootstrapping can deliver fast error estimates [14].

Most of the randomized least squares work comes from theoretical computer science and numerical analysis and is mainly concerned with errors due to algorithmic randomization, while ignoring statistical noise in the model. The pioneering work [15, 16] is the first to quantify the total uncertainty from model-induced and algorithm-induced randomness. This being the first analysis of its kind, it started out with a few assumptions: the sampling matrices must preserve rank, and their expected value must be known; and the conditional expectation and variances must admit Taylor series. Thus, the resulting first-order expansions hold only approximately.

1.3. Specific Contributions. We extend the first-order expansions in [15, 16] as follows:

1. We derive *exact* expressions for the total expectation and variance of $\tilde{\boldsymbol{\beta}}$ with regard to model- and algorithm-induced uncertainties (Theorem 4.5). The expressions hold for *general* random sketching matrices $\mathbf{S}$, regardless of whether they preserve rank, and include sketching matrices that perform projections prior to sampling.
2. In contrast to most deterministic and randomized analyses, our expressions are not limited to full-rank matrices. We analyse the rank-deficient matrices in (1.3) by supplementing the *hat matrix* $\mathbf{P}_{\mathbf{x}} = \mathbf{X}\mathbf{X}^\dagger$, i.e. the orthogonal projector onto range($\mathbf{X}$), with two new projectors:
    (a) *Comparison hat matrix* $\mathbf{P} = \mathbf{X}(\mathbf{S}\mathbf{X})^\dagger\mathbf{S}$ (Lemma 3.1).
    This projector makes it possible to compare the model problem (1.1) with the lower-dimensional sketched problem (1.3). The difference $\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}$ quantifies the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).
    (b) *Bias projector* $\mathbf{P}_{\mathbf{0}} = (\mathbf{S}\mathbf{X})^\dagger(\mathbf{S}\mathbf{X})$ (Lemma 4.1).
    This projector captures the failure of $\mathbf{S}$ to preserve rank. The difference $\mathbf{I}-\mathbf{P}_{\mathbf{0}}$ quantifies the rank deficiency of the sketched matrix $\mathbf{S}\mathbf{X}$.
3. For the model-induced uncertainty of $\tilde{\boldsymbol{\beta}}$, conditioned on the sampling matrix $\mathbf{S}$, we show (Theorem 4.2, Corollary 4.3):
    (a) The conditional bias increases with the rank deficiency of $\mathbf{S}\mathbf{X}$.
    (b) The difference between conditional variance and model variance increases with the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).

    Thus, unbiasedness is easier to achieve because it only requires $\mathbf{S}\mathbf{X}$ to have full column rank. In contrast, recovering the model variance requires reproducing all of the original space range($\mathbf{X}$).

4. For the *total* uncertainty in the solution $\tilde{\boldsymbol{\beta}}$ we show (Theorem 4.5, Corollary 4.6):
   (a) The total bias increases with the expected rank deficiency of $\mathbf{SX}$.
   (b) The difference between total variance and model variance increases with two terms: the expected rank deficiency of $\mathbf{SX}$; and the expected deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).

   Thus, total expectation and variance are governed by the expected spatial geometry induced by the sketching process rather than by structural properties of specific $\mathbf{S}$. However, the condition number of $\mathbf{X}$ has far less impact than one would have expected based on numerical perturbation theory.
5. We show analogous results for norm-wise quantities (Theorem 4.8, Corollary 4.9). The total expectations of the *regression sum of squares* $\|\mathbf{P}\mathbf{y}\|_2^2$ and the *residual sum of squares* $\|(\mathbf{I}-\mathbf{P})\mathbf{y}\|_2^2$ depend on the norms of the projectors $\mathbf{P}$ and $\mathbf{I}-\mathbf{P}$, amplified by the model variance σ^2.
6. We present structural bounds that improve existing perturbation bounds (Corollary 3.5). They imply that the minimum norm solution $\tilde{\boldsymbol{\beta}}$ is less sensitive to multiplicative perturbations than to additive perturbations, because the dependence is only on the condition number, rather than on its square as in the case of additive perturbations.

The judicious design of numerical experiments that are representative and informative from both, numerical and statistical perspectives, is beyond this scope, and will be the subject of a separate paper.

1.4. Overview.

After reviewing the computational models for least squares regression (Section 2), we adopt two perspectives:

1. Deterministic (Section 3): The matrix $\mathbf{S}$ is fixed and the sketched problem (1.3) is a multiplicative perturbation of the deterministic problem (1.2), and we present structural perturbation bounds.
2. Probabilistic (Section 4): The matrix $\mathbf{S}$ is a matrix-valued random variable, and (1.3) is a randomized algorithm for solving the model problem (1.1), and we derive expressions for expectation and variance with regard to the model- and algorithm-induced uncertainties.

A brief discussion of our results (Section 5) ends the main part of the paper. Proofs are relegated to the Appendix (Section A), as are specific examples to provide insight for the geometry of the probabilistic results (Section B).

2. Models for Least Squares Regression.

Given is a fixed design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with rank($\mathbf{X}$) $= p$. Since $\mathbf{X}$ has full column rank, the Moore-Penrose inverse is a left inverse with

$$\mathbf{X}^\dagger = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \qquad \text{and} \qquad \mathbf{X}^\dagger\mathbf{X} = \mathbf{I}_p. \tag{2.1}$$

The two-norm condition number of $\mathbf{X}$ with regard to left inversion is

$$\kappa_2(\mathbf{X}) \equiv \|\mathbf{X}\|_2\|\mathbf{X}^\dagger\|_2.$$

We review the different incarnations of least squares regression: the Gaussian linear model (Section 2.1), the traditional computation (Section 2.2), and algorithmic leveraging (Section 2.3).

2.1. Gaussian linear model. Let $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ denote the true but generally unknown parameter vector, and let the response vector $\mathbf{y} \in \mathbb{R}^n$ satisfy the Gauss-Markov assumptions,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n). \tag{2.2}$$

The noise vector $\boldsymbol{\epsilon} \in \mathbb{R}^n$ has a multivariate normal distribution whose mean is the vector of all zeros, $\mathbf{0} \in \mathbb{R}^n$, and whose covariance is a multiple $\sigma^2 > 0$ of the identity matrix $\mathbf{I}_n \in \mathbb{R}^{n\times n}$.

2.2. Traditional algorithm for least squares solution. For a fixed $\mathbf{y} \in \mathbb{R}^n$ solve

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2. \tag{2.3}$$

Since $\mathbf{X}$ has full column rank, (2.3) is well posed and has the unique solution

$$\hat{\boldsymbol{\beta}} \equiv \mathbf{X}^\dagger \mathbf{y}. \tag{2.4}$$

The prediction and the least squares residual are, respectively

$$\hat{\mathbf{y}} \equiv \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{and} \qquad \hat{\mathbf{e}} \equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}}.$$

In terms of the *hat matrix* [3, 10, 27],

$$\mathbf{P}_{\mathbf{x}} \equiv \mathbf{X}\mathbf{X}^\dagger = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \ \in \mathbb{R}^{n\times n}, \tag{2.5}$$

which is the orthogonal projector onto range($\mathbf{X}$) along null($\mathbf{X}^T$), the prediction and least squares residual can be expressed as

$$\hat{\mathbf{y}} = \mathbf{P}_{\mathbf{x}}\mathbf{y} \qquad \text{and} \qquad \hat{\mathbf{e}} = (\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{y}. \tag{2.6}$$

2.3. Random Row-Sketching. From a deterministic perspective, this can be considered an extension of weighted least squares [8, Section 6.1] to rectangular weighting matrices.

Given a sketching matrix $\mathbf{S} \in \mathbb{R}^{r\times n}$ with $1 \le r \le n$, solve

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{S}(\mathbf{X}\boldsymbol{\beta} - \mathbf{y})\|_2, \tag{2.7}$$

which has the minimum norm solution

$$\tilde{\boldsymbol{\beta}} \equiv (\mathbf{S}\mathbf{X})^\dagger\, \mathbf{S}\mathbf{y}. \tag{2.8}$$

This problem is generally ill-posed: Just because $\mathbf{S}$ has $r > p$ rows, this does not imply rank($\mathbf{S}$) $= p$; and even if $\mathbf{S}$ does have full column rank, rank($\mathbf{S}\mathbf{X}$) $< p$ is still possible.

By design, $\mathbf{S}$ has fewer rows than $\mathbf{X}$. Hence the corresponding predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{S}\mathbf{X}\tilde{\boldsymbol{\beta}}$ have different dimension and cannot be directly compared; neither can their residuals. To remedy this, we follow previous work [5, 6, 22], and compare the predictions with regard to the *original* matrix,

$$\tilde{\mathbf{y}} \equiv \mathbf{X}\tilde{\boldsymbol{\beta}} \qquad \text{and} \qquad \tilde{\mathbf{e}} \equiv \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{y} - \tilde{\mathbf{y}}. \tag{2.9}$$

Note that $\tilde{\mathbf{e}}$ is not a least squares residual; the least squares residual for (2.7) is $\mathbf{S}\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{S}\mathbf{y}$. However, we need $\tilde{\mathbf{e}}$ to assess the performance of $\tilde{\boldsymbol{\beta}}$ in the context of the original problem (2.3).

3. Structural (deterministic perturbation) bounds. Here $\mathbf{S}$ is a fixed, general matrix; and $\mathbf{SX}$ is interpreted as a perturbation of $\mathbf{X}$. We derive expressions for the quantities of interest from the perturbed problem (Section 3.1), followed by multiplicative perturbation bounds (Section 3.2).

3.1. The perturbed problem. We derive expressions for the solution, prediction and residual of the lower-dimensional problem (2.7). In order to relate them to the higher-dimensional original problem (2.3), we introduce (Lemma 3.1) a *comparison hat matrix* $\mathbf{P}$ for (2.7), which corresponds to the *hat matrix* $\mathbf{P_x}$ in (2.5) for the original problem (2.3). This makes it possible to express the solution, prediction, and residual of the perturbed problem in terms of the original problem (Theorem 3.3).

LEMMA 3.1 (Comparison hat matrix). *With the assumptions in Section 2,*

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{SX})^\dagger \mathbf{S}$$

is an oblique projector where
1. $\mathbf{P_x P} = \mathbf{P}$.
2. $\mathbf{P} - \mathbf{P_x}$ *reflects the difference between the spaces* $\mathrm{null}(\mathbf{P})$ *and* $\mathrm{null}(\mathbf{P_x})$.
3. $\mathbf{PX} = \mathbf{X}$ *if* $\mathrm{rank}(\mathbf{SX}) = p$.

Proof. See Section A.1 ∎

The name *comparison hat matrix* will become clear in Theorem 3.3, where $\mathbf{P}$ assumes the duties of the *hat matrix* $\mathbf{P_x}$ for the expressions in (2.9).

Remark 3.2. The following cases are possible.
- If $\mathbf{S} = \mathbf{I}_n$, then $\mathbf{P} = \mathbf{P_x}$.
- If $\mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X})$, then $\mathbf{P}$ is an oblique version of the orthogonal projector $\mathbf{P_x}$ with $\mathrm{range}(\mathbf{P}) = \mathrm{range}(\mathbf{P_x})$, but $\mathrm{null}(\mathbf{P}) \neq \mathrm{null}(\mathbf{P_x})$ in general.
- If
$$\mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{SX}) < \mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{P_x}) = p,$$
then $\mathbf{P}$ projects only onto a subspace of $\mathrm{range}(\mathbf{X})$.

The comparison hat matrix $\mathbf{P}$ generalizes the oblique projector $\mathbf{P_u}$ in [22, (11)], which was introduced to quantify *prediction efficiency* and *residual efficiency* of sketching algorithms in the statistical setting (2.2). This projector $\mathbf{P_u}$ is defined if $\mathrm{rank}(\mathbf{SX}) = p$, and equals $\mathbf{P_u} \equiv \mathbf{U}(\mathbf{SU})^\dagger \mathbf{S} = \mathbf{P}$, where $\mathbf{U}$ is an orthonormal basis for $\mathrm{range}(\mathbf{X})$. However, if $\mathrm{rank}(\mathbf{SX}) < \mathrm{rank}(\mathbf{X})$, then $\mathbf{P_u}$ is not sufficient in our context.

THEOREM 3.3 (Perturbed least squares problem). *With the assumptions in Section 2, the solution of (2.7) satisfies*

$$\tilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{P} \mathbf{y} = \hat{\boldsymbol{\beta}} + \mathbf{X}^\dagger (\mathbf{P} - \mathbf{P_x}) \mathbf{y}.$$

The prediction $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}}$ *and residual* $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ *satisfy*

$$\begin{aligned}
\tilde{\mathbf{y}} &= \mathbf{P}\mathbf{y} = \hat{\mathbf{y}} + (\mathbf{P} - \mathbf{P_x})\mathbf{y}, \\
\tilde{\mathbf{e}} &= (\mathbf{I} - \mathbf{P})\,\mathbf{y} = \hat{\mathbf{e}} + (\mathbf{P_x} - \mathbf{P})\mathbf{y}.
\end{aligned}$$

Proof. See Section A.2. ∎

Theorem 3.3 shows that the relation between perturbed and original least squares problems is governed by $\mathbf{P} - \mathbf{P}_{\mathbf{x}}$, which reflects the difference between the spaces $\mathrm{null}(\mathbf{P})$ and $\mathrm{null}(\mathbf{P}_{\mathbf{x}})$. The dependence on the sketching matrix is implicit, through the induced spaces.

With its explicit expressions for $\tilde{\boldsymbol{\beta}}$ that hold for general matrices $\mathbf{S}$ without assumptions on $\mathrm{rank}(\mathbf{SX})$, Theorem 3.3 also strengthens the ground breaking result [16, Lemma 1], reproduced in the lemma below.

LEMMA 3.4 (Lemma 1 in [15] and [16]). *If, in addition to the assumptions in Section 2, the matrix* $\mathbf{S}$ *in (2.7) has a single nonzero entry per row, the vector* $\mathbf{w} \equiv \mathrm{diag}(\mathbf{S}^T\mathbf{S}) \in \mathbb{R}^n$ *has a scaled multinomial distribution with expected value* $\mathbb{E}[\mathbf{w}] = \mathbb{1}$*, satisfies* $\mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X})$*, and admits a Taylor series expansion of the solution* $\tilde{\boldsymbol{\beta}}(\mathbf{w})$ *of (2.7) around* $\mathbf{w}_0 = \mathbb{1}$ *with* $\tilde{\boldsymbol{\beta}}(\mathbf{w}_0) = \hat{\boldsymbol{\beta}}$*, then*

$$\tilde{\boldsymbol{\beta}}(\mathbf{w}) = \hat{\boldsymbol{\beta}} + \mathbf{X}^\dagger \,\mathrm{diag}(\hat{\mathbf{e}})(\mathbf{w} - \mathbb{1}) + R(\mathbf{w}),$$

where $R(\mathbf{w})$ *is the remainder of the Taylor series expansion. The Taylor series expansion is valid if* $R(\mathbf{w}) = o(\|\mathbf{w} - \mathbf{w}_0\|_2)$ *with high probability.*

3.2. Multiplicative perturbation bounds. We consider (2.7) as a multiplicative perturbation of the original problem (2.3) and derive norm-wise relative perturbation bounds (Corollary 3.5), followed by comparisons to existing work.

COROLLARY 3.5. *With the assumptions in Section 2, let* $0 < \theta < \pi/2$ *be the angle between* $\mathbf{y}$ *and* $\mathrm{range}(\mathbf{X})$*.*

The solution $\tilde{\boldsymbol{\beta}}$ *of (2.7) satisfies*

$$\frac{\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2}{\|\hat{\boldsymbol{\beta}}\|_2} \leq \kappa_2(\mathbf{X}) \,\frac{\|\mathbf{y}\|_2}{\|\mathbf{X}\|_2\|\hat{\boldsymbol{\beta}}\|_2}\, \|\mathbf{P} - \mathbf{P}_{\mathbf{x}}\|_2 \leq \kappa_2(\mathbf{X}) \,\frac{\|\mathbf{P} - \mathbf{P}_{\mathbf{x}}\|_2}{\cos\theta}.$$

The prediction $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}}$ *satisfies*

$$\frac{\|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|_2}{\|\hat{\mathbf{y}}\|_2} \leq \frac{\|\mathbf{P} - \mathbf{P}_{\mathbf{x}}\|_2}{\cos\theta}.$$

Proof. This is a direct consequence of Theorem 3.3, and of [8, (5.3.16)] which implies

$$\|\mathbf{y}\|_2/(\|\mathbf{X}\|_2\|\hat{\boldsymbol{\beta}}\|_2) \leq \|\mathbf{y}\|_2/\|\mathbf{X}\hat{\boldsymbol{\beta}}\|_2 = 1/\cos\theta.$$

For $\mathbf{S} = \mathbf{I}_n$, the bounds in Corollary 3.5 are zero and therefore tight. Corollary 3.5 implies that the sensitivity of the minimum norm least squares solution $\tilde{\boldsymbol{\beta}}$ to multiplicative perturbations depends on the distance between the spaces $\mathrm{null}(\mathbf{P})$ and $\mathrm{null}(\mathbf{P}_{\mathbf{x}})$, quantified by $\|\mathbf{P} - \mathbf{P}_{\mathbf{x}}\|_2$. This distance is amplified, as expected, by the conditioning of $\mathbf{X}$ is with regard to (left) inversion, and by the closeness of $\mathbf{y}$ to $\mathrm{range}(\mathbf{X})$. Corollary 3.5 is an absolute as well as a relative bound since $\|\mathbf{P}_{\mathbf{x}}\|_2 = 1$.

In contrast to multiplicative perturbation bounds for eigenvalue and singular value problems [11, 12], we do not require $\mathbf{S}$ to be nonsingular or square. Weighted least squares problems [8, Section 6.1] employ nonsingular diagonal matrices $\mathbf{S}$ for regularization or scaling of discrepancies, and do not view them as a perturbation.

In contrast to additive bounds [8, Section 5.3.6], [9, Section 20.1], [25, (3.4)], there is no squaring of the condition number and no need for requiring $\mathrm{rank}(\mathbf{SX}) =$

rank($\mathbf{X}$). This suggests that the minimum norm solution of (2.7) and its residual are less sensitive to multiplicative perturbations than to additive perturbations.

In contrast to existing structural bounds for randomized least squares algorithms [6, Theorem 1], such as the one in Lemma 3.6 below, the bound for $\tilde{\boldsymbol{\beta}}$ in Corollary 3.5 is more general and tighter because it does not exhibit nonlinear dependencies on the perturbations.

LEMMA 3.6 (Theorem 1 in [6]). *In addition to Section 2, assume* $\|\mathbf{P}_{\mathbf{x}}\mathbf{y}\|_2 \geq \gamma\,\|\mathbf{y}\|_2$ *for some* $0 < \gamma \leq 1$ *and* $\|\tilde{\mathbf{e}}\|_2 \leq (1+\eta)\,\|\hat{\mathbf{e}}\|_2$. *Then*

$$\frac{\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2}{\|\hat{\boldsymbol{\beta}}\|_2} \leq \kappa_2(\mathbf{X})\sqrt{\gamma^{-2}-1}\,\sqrt{\eta}.$$

4. Model-induced and randomized algorithm-induced uncertainty.

Under the linear model (2.2), the computed solution $\hat{\boldsymbol{\beta}}$ has nice statistical properties [20, Chapter 6], as it is an unbiased estimator of $\boldsymbol{\beta}_0$ and it has minimal variance among all linear unbiased estimators. We show how this changes with the addition of algorithm-induced uncertainty.

After briefly reviewing the uncertainty induced by the linear model (Section 4.1); we derive the expectation and variance of $\tilde{\boldsymbol{\beta}}$, conditioned on the algorithm-induced uncertainty (Section 4.2), and from that the total expectation and variance (Section 4.3), followed by the derivation of the conditional and total expectations for the regression sum of squares and the residual sum of squares (Section 4.4).

4.1. Model-induced uncertainty.

We view the model-induced randomness in (1.1) and (2.2) as a property of the response vector $\mathbf{y}$, so that

$$\mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}] = \mathbf{0}, \qquad \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\boldsymbol{\epsilon}] = \sigma^2\,\mathbf{I}_n.$$

As a consequence

$$\mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_0, \qquad \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \in \mathbb{R}^{p\times p}. \tag{4.1}$$

This implies that the computed solution $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}_0$, and it signals the well-known dependence of the variance on the conditioning of $\mathbf{X}$ [25, Section 5].

The difficulty in analyzing random row-sketching (2.7), coupled with general concern about first-order expansions like the ones in [15, 16], is the frequent occurrence of rank deficiency in the sketched matrix, that is, rank($\mathbf{SX}$) < rank($\mathbf{X}$). In this case $(\mathbf{SX})^\dagger$ cannot be expressed in terms of $\mathbf{SX}$ as in (2.1).

One can derive bounds [1, Theorem 3.2], [13, Theorems 4.1and 5.2] on the probability that rank($\mathbf{SX}$) = rank($\mathbf{X}$) for matrices $\mathbf{S}$ that perform uniform sampling and leverage score sampling. However, such bounds are not useful here, because we need the expected values to run over *all* instances of $\mathbf{SX}$.

We introduce a projector that quantifies the deviation of the columns of $\mathbf{SX}$ from being linearly independent.

LEMMA 4.1 (Bias projector). *With the assumptions in Section 2,*

$$\mathbf{P_0} \equiv (\mathbf{SX})^\dagger(\mathbf{SX}) \in \mathbb{R}^{p\times p}$$

is an orthogonal projector with

1. $\mathbf{PX} = \mathbf{XP_0}$

2. $\mathbf{P_0} = \mathbf{I}_p$ *if* $\operatorname{rank}(\mathbf{SX}) = p$.

As a consequence, $\mathbf{I}_p - \mathbf{P_0}$ *quantifies the rank deficiency of* $\mathbf{SX}$.

Proof. See Section A.3. ∎

If $\operatorname{rank}(\mathbf{SX}) < p$, then $\mathbf{P_0}$ characterizes the subspace of $\operatorname{range}(\mathbf{X})$ onto which $\mathbf{P}$ projects. The name *bias projector* will become apparent in Theorem 4.2, where $\mathbf{P_0}$ represents the bias in $\tilde{\boldsymbol{\beta}}$.

4.2. Model-induced uncertainty, conditioned on algorithm-induced uncertainty. We determine the conditional expectation and variance for the solution of (2.7), by assuming that the random sketching matrix $\mathbf{S}$ is fixed at a specific value $\mathbf{S_0}$. The expectation conditioned on $\mathbf{S}$ is abbreviated as

$$\mathbb{E}_{\mathbf{y}}\left[\cdot \,\middle|\, \mathbf{S}\right] \;\equiv\; \mathbb{E}_{\mathbf{y}}\left[\cdot \,\middle|\, \mathbf{S} = \mathbf{S}_0\right].$$

The exact expressions below for general matrices $\mathbf{S}$ extend the first-order expressions for specific sampling matrices in [16, Lemmas 2-6].

THEOREM 4.2 (Model-induced uncertainty conditioned on $\mathbf{S}$). *With the assumptions in Section 2, the solution* $\tilde{\boldsymbol{\beta}}$ *of (2.7) has the conditional expectation*

$$\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \,\middle|\, \mathbf{S}\right] = \mathbf{P_0}\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0 - (\mathbf{I} - \mathbf{P_0})\boldsymbol{\beta}_0,$$

where $\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \,\middle|\, \operatorname{rank}(\mathbf{SX}) = p\right] = \boldsymbol{\beta}_0$*; and the conditional variance*

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \,\middle|\, \mathbf{S}\right] &= \sigma^2\, \left(\mathbf{X}^\dagger\mathbf{P}\right)\left(\mathbf{X}^\dagger\mathbf{P}\right)^T \\ &= \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] + \sigma^2\,\mathbf{X}^\dagger\left(\mathbf{P}\mathbf{P}^T - \mathbf{P_x}\right)(\mathbf{X}^\dagger)^T,\end{aligned}$$

with $\mathbf{P}\mathbf{P}^T - \mathbf{P_x}$ *representing the deviation of* $\mathbf{P}$ *from being an orthogonal projector onto* $\operatorname{range}(\mathbf{X})$.

Proof. See Section A.4. ∎

Theorem 4.2 shows that the conditional bias and variance of $\tilde{\boldsymbol{\beta}}$ depend on the rank deficiency of $\mathbf{SX}$, and the ability of $\mathbf{P}$ to reproduce the original space $\operatorname{range}(\mathbf{X})$. The fixed sketching matrix $\mathbf{S}$ is involved only implicitly, through the spaces induced by the sketching process. Specifically, Theorem 4.2 shows:

1. The conditional bias of $\tilde{\boldsymbol{\beta}}$ is proportional to the deviation $\mathbf{I} - \mathbf{P_0}$ of $\mathbf{SX}$ from having full column rank. That is, the conditional bias becomes worse as the rank deficiency increases. If $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, then $\tilde{\boldsymbol{\beta}}$ is a conditional unbiased estimator of $\boldsymbol{\beta}_0$, regardless of the specific sketching class to which $\mathbf{S}$ belongs.
2. The conditional variance is close to the model variance $\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]$, if $\mathbf{P}$ is close to being an orthogonal projector onto $\operatorname{range}(\mathbf{X})$. In the extreme case $\mathbf{S} = \mathbf{I}_n$, the conditional variance is identical to the model variance.

The relevance of $\mathbf{I} - \mathbf{P_0}$ and $\mathbf{P}\mathbf{P}^T - \mathbf{P_x}$ is further corroborated below.

COROLLARY 4.3 (Relative differences between conditional and model uncertainties). *With the assumptions in Theorem 4.2,*

$$\begin{aligned}\|\,\mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}}\,|\,\mathbf{S}] - \boldsymbol{\beta}_0\|_2 &\le \|\mathbf{I} - \mathbf{P_0}\|_2\,\|\boldsymbol{\beta}_0\|_2 \\ \frac{\|\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}}\,|\,\mathbf{S}] - \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2}{\|\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2} &\le \|\mathbf{P}\mathbf{P}^T - \mathbf{P_x}\|_2.\end{aligned}$$

Proof. See Section A.5. ∎

Corollary 4.3 implies that the relative differences to conditional unbiasedness and model variance are solely governed by the quantities $\mathbf{I}-\mathbf{P_0}$ and $\mathbf{PP}^T-\mathbf{P_x}$, respectively. Somewhat surprisingly, the condition number of the model variance $\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]$ in (4.1) is not explicitly present. Instead, the conditional bias of $\tilde{\boldsymbol{\beta}}$ increases with the rank deficiency of $\mathbf{SX}$, while the relative difference between conditional and model variances increases with the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$). Thus, unbiasedness is easier to achieve because it only requires $\mathbf{SX}$ to have full column rank, while recovering the model variance requires reproducing all of range($\mathbf{X}$).

The examples in Section B.2.1 illustrate the effect of rank deficiency in Theorem 4.2 and Corollary 4.3.

Remark 4.4 (Sampling versus sketching). To confirm the importance of the induced spaces and the peripheral role of the particular structure of $\mathbf{S}$, we perform sketching by first applying row-mixing [1, Section 3.2] with a unitary transform $\mathbf{F}\in\mathbb{R}^{n\times n}$ prior to sampling,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\mathbf{S}(\mathbf{X}\boldsymbol{\beta}-\mathbf{y})\|_2 \qquad \text{where} \quad \mathbf{S}\equiv\mathbf{S}_1\mathbf{F}, \tag{4.2}$$

where $\mathbf{F}^T\mathbf{F}=\mathbf{FF}^T=\mathbf{I}_n$, and $\mathbf{S}_1\in\mathbb{R}^{p\times n}$ is a sampling matrix. The row-mixed problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\mathbf{F}(\mathbf{X}\boldsymbol{\beta}-\mathbf{y})\|_2$$

is equivalent to the original problem (2.3), since it has the same solution, and the same comparison hat matrix and bias projector,

$$\begin{aligned}\mathbf{X}(\mathbf{FX})^\dagger\mathbf{F}&=\mathbf{XX}^\dagger=\mathbf{P_x}\\(\mathbf{FX})^\dagger\,(\mathbf{FX})&=\mathbf{X}^\dagger\mathbf{X}=\mathbf{I}_n.\end{aligned}$$

Thus, any damaging effect on the conditional bias and variance comes from the possible rank deficiency and the spaces induced by the sampling process.

4.3. Combined algorithm-induced and model-induced uncertainty. We determine the total expectation and the total variance for the solution from (2.7) when $\mathbf{S}$ is a random sketching matrix, that is, $\mathbf{S}$ is a matrix-valued random variable.

The algorithm-induced uncertainty of the random matrix $\mathbf{S}$ is represented by the expectation $\mathbb{E}_{\mathbf{s}}[\cdot]$ and the variance $\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\cdot]$, while the total mean and variance of the combined uncertainty are denoted by $\mathbb{E}[\cdot]$ and $\mathbb{V}\mathrm{ar}[\cdot]$. The total mean is computed by conditioning on the algorithm-induced randomness

$$\mathbb{E}\left[\cdot\right]\;=\;\mathbb{E}_{\mathbf{s}}\left[\mathbb{E}_{\mathbf{y}}\left[\cdot\,\middle|\,\mathbf{S}\right]\right]. \tag{4.3}$$

Since $\mathbf{S}$ is a matrix-valued random variable, so are the projectors $\mathbf{P}$ and $\mathbf{P_0}$.

The exact expressions below for general random matrices $\mathbf{S}$ extend the first order approximations for specific sampling matrices in [16, Lemmas 2-6].

THEOREM 4.5 (Total uncertainty). *With the assumptions in Section 2, let $\mathbf{S}$ be a random sketching matrix. The solution $\tilde{\boldsymbol{\beta}}$ of (2.7) has total expectation and variance*

$$\begin{aligned}\mathbb{E}[\tilde{\boldsymbol{\beta}}]&=\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0]=\boldsymbol{\beta}_0+\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}-\mathbf{I}]\boldsymbol{\beta}_0\\ \mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}]&=\sigma^2\,\mathbf{X}^\dagger\,\mathbb{E}_{\mathbf{s}}\left[\mathbf{PP}^T\right](\mathbf{X}^\dagger)^T+\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0]\\ &=\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]+\sigma^2\,\mathbf{X}^\dagger\,\mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T-\mathbf{P_x}]\,(\mathbf{X}^\dagger)^T+\mathbb{V}\mathrm{ar}_{\mathbf{s}}[(\mathbf{P_0}-\mathbf{I})\boldsymbol{\beta}_0],\end{aligned}$$

where

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0] &= \mathbb{E}_{\mathbf{s}}\left[(\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T\right] - (\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0])(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0])^T \\ &= \mathbb{V}\mathrm{ar}_{\mathbf{s}}[(\mathbf{P_0}-\mathbf{I})\boldsymbol{\beta}_0].\end{aligned}$$

Proof. See Section A.6. ∎

Theorem 4.5 shows that total expectation and variance are governed by the representation of spaces associated with the original problem (2.3) and the sketched problem (2.7), rather than the specific class of sketching matrices over which $\mathbb{E}_{\mathbf{s}}$ and $\mathbb{V}\mathrm{ar}_{\mathbf{s}}$ range. Specifically,

1. The total bias of $\tilde{\boldsymbol{\beta}}$ is proportional to the expected deviation of the matrix-valued random variable $\mathbf{SX}$ from having full column rank. Note that the expectation $\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]$ of a projector $\mathbf{P_0}$ is in general not a projector, as the example in Section B.2.3 illustrates.
2. The total variance of $\tilde{\boldsymbol{\beta}}$ is proportional to the expected rank deficiency of $\mathbf{SX}$, plus the expected deviation of the matrix-valued random variable $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).

COROLLARY 4.6 (Relative differences between total and model uncertainties). *With the assumptions in Theorem 4.5,*

$$\begin{aligned}\|\,\mathbb{E}[\tilde{\boldsymbol{\beta}}]-\boldsymbol{\beta}_0\|_2 &\le \|\,\mathbb{E}_{\mathbf{s}}[\mathbf{I}-\mathbf{P_0}]\|_2\,\|\boldsymbol{\beta}_0\|_2 \\ \frac{\|\,\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}]-\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2}{\|\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2} &\le \|\,\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T-\mathbf{P_x}]\|_2 + \frac{\|\,\mathbb{V}\mathrm{ar}_{\mathbf{s}}[(\mathbf{I}-\mathbf{P_0})\boldsymbol{\beta}_0]\|_2}{\|\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2}.\end{aligned}$$

Proof. See Section A.7. ∎

Corollary 4.6 implies that the relative differences to unbiasedness and model variance are solely governed by the quantities $\mathbb{E}_{\mathbf{s}}[\mathbf{I}-\mathbf{P_0}]$ and $\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T-\mathbf{P_x}]$. Specifically, the total bias of $\tilde{\boldsymbol{\beta}}$ increases with the expected rank deficiency of $\mathbf{SX}$, while the relative difference between total and model variances increases with the expected deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$), and the expected rank deficiency of $\mathbf{SX}$.

The examples in Sections B.2.3-B.2.5 illustrate the effect of expected rank deficiency in Theorem 4.5 and Corollary 4.6.

4.4. Regression and residual sums of squares. Two quantities from the original least squares problem (2.3) play a key role in hypothesis testing, regression diagnostics, and model selection metrics, such as the (adjusted) R^2 statistic, *Mallows's* C_p, the *Akaike information criterion*, and the *Bayesian information criterion* [7, 17, 20, 21].

- *Regression sum of squares*, i.e. the squared norm of the prediction,

$$\mathrm{SSR}_{\mathrm{ols}} \equiv \mathbf{y}^T\mathbf{P_x}\mathbf{y} = \mathbf{y}^T{\mathbf{P_x}}^T\mathbf{P_x}\mathbf{y} = \|\hat{\mathbf{y}}\|_2^2.$$

- *Residual sum of squares*, i.e. the squared norm of the least squares residual,

$$\mathrm{RSS}_{\mathrm{ols}} = \mathbf{y}^T(\mathbf{I}-\mathbf{P_x})\mathbf{y} = \mathbf{y}^T(\mathbf{I}-\mathbf{P_x})^T(\mathbf{I}-\mathbf{P_x})\mathbf{y} = \|\hat{\mathbf{e}}\|_2^2,$$

From $\hat{\mathbf{y}}^T\hat{\mathbf{e}} = 0$ follows

$$\|\mathbf{y}\|_2^2 = \|\hat{\mathbf{y}}\|_2^2 + \|\hat{\mathbf{e}}\|_2^2 = \mathrm{SSR}_{\mathrm{ols}} + \mathrm{RSS}_{\mathrm{ols}},$$

which decomposes the observation into a portion that is explained by the model; and a portion that represents the error in the model. The corresponding quantities for random row-sketching are

$$\begin{aligned}\mathrm{SSR} &\equiv \mathbf{y}^T\mathbf{P}^T\mathbf{P}\mathbf{y} = \|\tilde{\mathbf{y}}\|_2^2\\ \mathrm{RSS} &\equiv \mathbf{y}^T(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})\mathbf{y} = \|\tilde{\mathbf{e}}\|_2^2.\end{aligned}$$

They relate to their counter parts in the original problem (2.3) via the two-norm version of Theorem 3.3,

$$\begin{aligned}\mathrm{SSR} &= \mathrm{SSR}_{\mathrm{ols}} + \mathbf{y}^T(\mathbf{P}^T\mathbf{P}-\mathbf{P}_{\mathbf{x}})\mathbf{y}\\ \mathrm{RSS} &= \mathrm{RSS}_{\mathrm{ols}} + \|(\mathbf{P}-\mathbf{P}_{\mathbf{x}})\mathbf{y}\|_2^2.\end{aligned}$$

Since RSS evaluates the solution $\tilde{\boldsymbol{\beta}}$ of (2.7) in the context of the original problem, $\tilde{\boldsymbol{\beta}}$ is not a minimizer of (2.3), so clearly $\mathrm{RSS} \geq \mathrm{RSS}_{\mathrm{ols}}$. The difference between the quantities from random sketching and their deterministic counterparts is governed by the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).

THEOREM 4.7 (Model-induced uncertainty conditioned on $\mathbf{S}$). *With the assumptions in Section 2,*

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}[\mathrm{SSR}\,|\,\mathbf{S}] &= \|\mathbf{P}\mathbf{X}\boldsymbol{\beta}_0\|_2^2 + \sigma^2\,\|\mathbf{P}\|_F^2\\ \mathbb{E}_{\mathbf{y}}[\mathrm{RSS}\,|\,\mathbf{S}] &= \|(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta}_0\|_2^2 + \sigma^2\,\|\mathbf{I}-\mathbf{P}\|_F^2.\end{aligned}$$

Proof. See Section A.8. ∎

The total expectations follow immediately from Theorem 4.7.

THEOREM 4.8 (Total uncertainty). *With the assumptions in Section 2,*

$$\begin{aligned}\mathbb{E}[\mathrm{SSR}] &= (\mathbf{X}\boldsymbol{\beta}_0)^T\,\mathbb{E}_{\mathbf{s}}[\mathbf{P}^T\mathbf{P}](\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2\ \mathrm{trace}\left(\mathbb{E}_{\mathbf{s}}[\mathbf{P}^T\mathbf{P}]\right)\\ \mathbb{E}[\mathrm{RSS}] &= (\mathbf{X}\boldsymbol{\beta}_0)^T\,\mathbb{E}_{\mathbf{s}}[(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})](\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2\ \mathrm{trace}\left(\mathbb{E}_{\mathbf{s}}[(\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P})]\right).\end{aligned}$$

At last we show that the difference between combined and model uncertainties is governed by the expected deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$), and the expected deviation of $\mathbf{I}-\mathbf{P}$ from being an orthogonal projector onto range$(\mathbf{X})^\perp$, both amplified by the model variance σ^2.

COROLLARY 4.9 (Difference between total and model uncertainty). *With the assumptions in Section 2,*

$$\begin{aligned}\mathbb{E}[\mathrm{SSR}] - \mathbb{E}_{\mathbf{y}}[\mathrm{SSR}_{\mathrm{ols}}] &= (\mathbf{X}\boldsymbol{\beta}_0)^T\,\mathbb{E}_{\mathbf{s}}[\boldsymbol{\Gamma}](\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2\ \mathrm{trace}\left(\mathbb{E}_{\mathbf{s}}[\boldsymbol{\Gamma}]\right)\\ \mathbb{E}[\mathrm{RSS}] - \mathbb{E}_{\mathbf{y}}[\mathrm{RSS}_{\mathrm{ols}}] &= (\mathbf{X}\boldsymbol{\beta}_0)^T\,\mathbb{E}_{\mathbf{s}}[\boldsymbol{\Gamma}_\perp](\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2\ \mathrm{trace}\left(\mathbb{E}_{\mathbf{s}}[\boldsymbol{\Gamma}_\perp]\right),\end{aligned}$$

where we abbreviate

$$\boldsymbol{\Gamma} \equiv \mathbf{P}^T\mathbf{P} - \mathbf{P}_{\mathbf{x}}, \qquad \boldsymbol{\Gamma}_\perp \equiv (\mathbf{I}-\mathbf{P})^T(\mathbf{I}-\mathbf{P}) - (\mathbf{I}-\mathbf{P}_{\mathbf{x}}).$$

Proof. See Section A.9. ∎

5. Discussion. We considered the randomized solution of least squares regression problems

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{S}(\mathbf{X}\boldsymbol{\beta}-\mathbf{y})\|_2$$

arising from a standard Gaussian linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

and analyzed the effect on the solution $\tilde{\boldsymbol{\beta}}$ of the combined uncertainties from algorithmic randomization and statistical noise.

Our results show that the expectation and variance of $\tilde{\boldsymbol{\beta}}$ are governed by the spatial geometry of the sketching process, rather than by structural properties of specific sketching matrices. Surprisingly, the condition number $\kappa_2(\mathbf{X})$ with respect to (left) inversion has far less impact on the statistical measures than it has on the numerical errors. Even from the deterministic view of the sampled problem as a multiplicative perturbation, the relative accuracy of $\tilde{\boldsymbol{\beta}}$ depends only on $\kappa_2(\mathbf{X})$ –rather than on the larger factor $\kappa_2(\mathbf{X})^2$ typical for additive perturbations.

The natural next step is the illustration of our analytical results through numerical experiments that are representative and informative from both, numerical and statistical perspectives.

Appendix A. Proofs.

We present the proofs for Sections 3 and 4.

Our results depend on projectors constructed from the possibly rank-deficient matrix $\mathbf{SX}$. In this case, the Moore-Penrose inverse cannot be expressed in terms of the matrix $\mathbf{SX}$ proper, so we rely on the four conditions [8, Section 5.5.2] that uniquely characterize the Moore-Penrose inverse,

$$\begin{aligned} (\mathbf{SX})(\mathbf{SX})^\dagger(\mathbf{SX}) &= \mathbf{SX}, \qquad & \left((\mathbf{SX})(\mathbf{SX})^\dagger\right)^T &= (\mathbf{SX})(\mathbf{SX})^\dagger \\ (\mathbf{SX})^\dagger(\mathbf{SX})(\mathbf{SX})^\dagger &= (\mathbf{SX})^\dagger, \qquad & \left((\mathbf{SX})^\dagger(\mathbf{SX})\right)^T &= (\mathbf{SX})^\dagger(\mathbf{SX}). \end{aligned} \tag{A.1}$$

A.1. Proof of Lemma 3.1.

The Moore-Penrose conditions [8, Section 5.5.2] imply $\mathbf{P}^2 = \mathbf{P}$ for the generally nonsymmetric matrix $\mathbf{P}$.

1. This follows from the Moore-Penrose conditions (A.1).
2. Use the fact [19, Problem 5.9.12] that $\mathrm{null}(\mathbf{P}) = \mathrm{null}(\mathbf{P_x})$ if and only if $\mathbf{P}\mathbf{P_x} - \mathbf{P} = \mathbf{0}$ and $\mathbf{P_x}\mathbf{P} - \mathbf{P_x} = \mathbf{0}$. With item 1, this implies $\mathbf{P_x}\mathbf{P} - \mathbf{P_x} = \mathbf{P} - \mathbf{P_x}$. Thus $\mathbf{P} - \mathbf{P_x}$ can be interpreted as a measure for the distance between $\mathrm{null}(\mathbf{P})$ and $\mathrm{null}(\mathbf{P_x})$.
3. This follows from (2.1).

A.2. Proof of Theorem 3.3.

The first expression for $\tilde{\boldsymbol{\beta}}$ follows from (2.1), (2.8), and Lemma 3.1. The second expression follows from adding and subtracting in the first expression the term $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{y} = \mathbf{X}^\dagger\mathbf{P_x}\mathbf{y}$.

The first expression for $\tilde{\mathbf{y}}$ follows from (2.8) and Lemma 3.1. The second expression follows from adding and subtracting in the first expression the first term in (2.6).

The first expression for $\tilde{\mathbf{e}}$ follows from (2.9), (2.8) and Lemma 3.1. The second expression for $\tilde{\mathbf{e}}$ follows from adding and subtracting in the first expression the second term in (2.6).

A.3. Proof of Lemma 4.1.

The Moore-Penrose conditions (A.1) imply $(\mathbf{P_0})^2 = \mathbf{P_0}$ and $(\mathbf{P_0})^T = \mathbf{P_0}$, confirming that $\mathbf{P_0}$ is an orthogonal projector.

1. This follows from Lemma 3.1.
2. If $\mathrm{rank}(\mathbf{SX}) = p$, then (2.1) implies that $(\mathbf{SX})^\dagger$ is a left-inverse.

A.4. Proof of Theorem 4.2.

The conditional expectation follows from Theorem 3.3, (4.1), Lemma 4.1, and (2.1).

The definition of variance, Theorem 3.3, and the above imply

$$\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,\big|\, \mathbf{S}] = \mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T \,\big|\, \mathbf{S}] - \mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,\big|\, \mathbf{S}]\, \mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,\big|\, \mathbf{S}]^T$$
$$= \left(\mathbf{X}^\dagger\mathbf{P}\right) \mathbb{E}_{\mathbf{y}}[\mathbf{y}\mathbf{y}^T] \left(\mathbf{X}^\dagger\mathbf{P}\right)^T - (\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T.$$

The middle term in the first summand equals

$$\mathbb{E}_{\mathbf{y}}[\mathbf{y}\mathbf{y}^T] = (\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{X}\boldsymbol{\beta}_0)^T + \mathbf{X}\boldsymbol{\beta}_0\, \mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}]^T + \mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}](\mathbf{X}\boldsymbol{\beta}_0)^T + \mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$$
$$= (\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{X}\boldsymbol{\beta}_0)^T + \sigma^2\mathbf{I}_n. \tag{A.2}$$

To obtain the first expression for the conditional variance, insert (A.2) into the conditional variance above, and apply Lemma 4.1 to cancel out the expressions with $\mathbf{P_0}$.

For the second expression, use (2.1) and (2.5) to write the model variance in (4.1) as

$$\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] = \sigma^2\, \mathbf{X}^\dagger\mathbf{P}_{\mathbf{x}}(\mathbf{X}^\dagger)^T.$$

Then add and subtract this term in the first expression for the conditional variance.

If $\mathbf{P}$ were an orthogonal projector onto range($\mathbf{X}$), then $\mathbf{P}^T\mathbf{P} = \mathbf{P} = \mathbf{P}_{\mathbf{x}}$. Thus, $\mathbf{P}^T\mathbf{P} - \mathbf{P}_{\mathbf{x}}$ represents the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).

A.5. Proof of Corollary 4.3. The bound for the conditional expectation follows from (4.1), and the second expression for the expectation in Theorem 4.2. The second expression for the conditional variance in Theorem 4.2 implies

$$\|\, \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,|\, \mathbf{S}] - \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2 \le \sigma^2\, \|\mathbf{X}^\dagger\|_2\, \|\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}\|_2 \|(\mathbf{X}^\dagger)^T\|_2.$$

Now apply $\|\mathbf{M}\|_2\, \|\mathbf{M}^T\|_2 = \|\mathbf{M}\mathbf{M}^T\|_2$, and $\mathbf{M}^\dagger(\mathbf{M}^\dagger)^T = (\mathbf{M}^T\mathbf{M})^{-1}$ for a full column-rank matrix $\mathbf{M}$ to deduce

$$\sigma^2\, \|\mathbf{X}^\dagger\|_2\, \|(\mathbf{X}^\dagger)^T\|_2 = \|\, \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2, \tag{A.3}$$

where $\|\, \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2 \neq 0$ by assumption in Section 2.1.

A.6. Proof of Theorem 4.5. Apply the iterated expectation (4.3), followed by Theorem 4.2 to obtain the mean,

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \,\Big|\, \mathbf{S}\right]\right] = \mathbb{E}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0] = \mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0.$$

Insert this into the definition of the variance, and apply again (4.3),

$$\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}] = \mathbb{E}[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T] - \mathbb{E}[\tilde{\boldsymbol{\beta}}]\, \mathbb{E}[\tilde{\boldsymbol{\beta}}]^T$$
$$= \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T \,\Big|\, \mathbf{S}\right]\right] - \left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)\left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)^T.$$

Treat the first summand as in the proof of Theorem 4.2 in Section A.4 to deduce

$$\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T \,\Big|\, \mathbf{S}\right] = \sigma^2\mathbf{X}^\dagger\mathbf{P}\mathbf{P}^T(\mathbf{X}^\dagger)^T + (\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T.$$

Condition this on $\mathbf{y}$ and then insert it into the above expression for the variance,

$$\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}] = \sigma^2\mathbf{X}^\dagger\, \mathbb{E}_{\mathbf{s}}\left[\mathbf{P}\mathbf{P}^T\right](\mathbf{X}^\dagger)^T$$
$$+ \underbrace{\mathbb{E}_{\mathbf{s}}\left[(\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T\right] - \left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)\left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)^T}_{\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0]}.$$

The second expression for $\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0]$ follows from adding and subtracting

$$\boldsymbol{\beta}_0\boldsymbol{\beta}_0^T - \boldsymbol{\beta}_0(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0)^T - (\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0)\boldsymbol{\beta}_0^T,$$

in other words from $\boldsymbol{\beta}_0$ having zero variance.

A.7. Proof of Corollary 4.6. The bound for the total expectation follows from (4.1), and the second expression for the expectation in Theorem 4.5. The bound for the total variance follows from the second expression for the variance in Theorem 4.5, and from (A.3).

A.8. Proof of Theorem 4.7. We need the following auxiliary result about expectations of quadratic forms.

LEMMA A.1. *With the assumptions in Section 2, if* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is a constant matrix, then,*

$$\mathbb{E}[\mathbf{y}^T\mathbf{A}\mathbf{y}] = (\mathbf{X}\boldsymbol{\beta}_0)^T\mathbf{A}(\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2\,\mathrm{trace}(\mathbf{A}).$$

Proof. This follows from $\mathbf{y}^T\mathbf{A}\mathbf{y}$ being a real scalar, the circular commutativity of the trace, the interchangeability of the trace and expectation since both are sums, and (A.2) as follows,

$$\begin{aligned}\mathbb{E}[\mathbf{y}^T\mathbf{A}\mathbf{y}] &= \mathbb{E}\left[\mathrm{trace}(\mathbf{y}^T\mathbf{A}\mathbf{y})\right] = \mathbb{E}\left[\mathrm{trace}(\mathbf{A}\mathbf{y}\mathbf{y}^T)\right] = \mathrm{trace}\left(\mathbf{A}\,\mathbb{E}[\mathbf{y}\mathbf{y}^T]\right)\\ &= \mathrm{trace}\left(\mathbf{A}(\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{X}\boldsymbol{\beta}_0)^T + \sigma^2\mathbf{A}\right) = (\mathbf{X}\boldsymbol{\beta}_0)^T\mathbf{A}(\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2\,\mathrm{trace}(\mathbf{A}).\end{aligned}$$

∎

Proof of the Theorem. The expression for SSR follows from $\tilde{\mathbf{y}} = \mathbf{P}\mathbf{y}$ in Theorem 3.3, Lemma A.1, and $\mathrm{trace}(\mathbf{P}^T\mathbf{P}) = \|\mathbf{P}\|_F^2$. Analogously, the expression for RSS follows from $\tilde{\mathbf{e}} = (\mathbf{I}-\mathbf{P})\mathbf{y}$.

A.9. Proof of Corollary 4.9. From (2.6) and Lemma A.1 follows

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}[\mathbf{y}^T\mathbf{P}_{\mathbf{x}}\mathbf{y}] &= (\mathbf{X}\boldsymbol{\beta}_0)^T\mathbf{P}_{\mathbf{x}}(\mathbf{X}\boldsymbol{\beta}_0) + \sigma^2\ \mathrm{trace}(\mathbf{P}_{\mathbf{x}})\\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}^T(\mathbf{I}-\mathbf{P}_{\mathbf{x}})\mathbf{y}] &= \underbrace{(\mathbf{X}\boldsymbol{\beta}_0)^T(\mathbf{I}-\mathbf{P}_{\mathbf{x}})(\mathbf{X}\boldsymbol{\beta}_0)}_{0} + \sigma^2\ \mathrm{trace}(\mathbf{I}-\mathbf{P}_{\mathbf{x}}).\end{aligned}$$

Add and subtract these to the respective expressions in Theorem 4.8.

Appendix B. Examples with uniform row sampling. We start with a brief review of sketching matrices for least squares problems (Section B.1), before presenting examples that give insight into the results of Section 4 and the detrimental effects of rank deficiency (Section B.2).

B.1. Random sketching matrices in least squares. We present a few examples of sketching matrices used by the randomized least squares solvers [1, 2, 5, 6, 14, 15, 16, 18, 23].

Uniform sampling with replacement. This is the *EXACTLY(c)* algorithm [6, Algorithm 3] with uniform probabilities, which performs row-wise compression for direct methods for the solution of full column rank least squares in [6, Algorithm 3], see also the *BasicMatrixMultiplication Algorithm* [4, Fig. 2], [13, Algorithm 3.2], [14, Algorithms 1 and 2], and the *Uniform Sampling Estimator* [16, Section 2.2].

The probability of a particular instance of $\mathrm{diag}(\mathbf{S}^T\mathbf{S})$, and therefore $\mathbf{S}$ is given by a scaled multinomial distribution [16, Section 3.1].

Algorithm B.1 Uniform sampling with replacement

Input: Integers $n \geq 1$ and $1 \leq r \leq n$
Output: Sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times n}$ with $\mathbb{E}_{\mathbf{s}}[\mathbf{S}^T\mathbf{S}] = \mathbf{I}_n$
  for $t = 1 : r$ **do**
    Sample k_t from $\{1, \ldots, n\}$ with probability $1/n$,
    independently and with replacement
  end for
  $\mathbf{S} = \sqrt{\frac{n}{r}}\begin{pmatrix}\mathbf{e}_{k_1} & \ldots & \mathbf{e}_{k_r}\end{pmatrix}^T$

Random orthogonal sketching. This is used in *Blendenpik* [1, Algorithm 1] to compute randomized preconditioners for the iterative solution of full column rank least squares problems.

Here $\mathbf{S} = \mathbf{BTD} \in \mathbb{R}^{n \times n}$, where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal elements are independent Rademacher random variables, equaling ± 1 with equal probability; $\mathbf{T} \in \mathbb{R}^{n \times n}$ is a unitary matrix, such as a Walsh-Hadamard, discrete cosine, or discrete Hartley transform; and $\mathbf{B}$ is a diagonal matrix whose diagonal elements are Bernoulli variables, equaling 1 with probability $\gamma p/n$ for some $\gamma > 0$, and 0 otherwise.

Gaussian sketching. This is used in to compute randomized preconditioners for the iterative solution of general least squares problems [18, Algorithms 1 and 2].

Here the elements of $\mathbf{S} \in \mathbb{R}^{r \times n}$ are independent $\mathcal{N}(0, 1)$ random variables. In Matlab: $\mathbf{S} = \texttt{randn}(r, n)$.

B.2. Examples. The purpose is to provide insight for Theorem 4.2, Corollary 4.3, Theorem 4.5 and Corollary 4.6 in a way that is easy to reproduce. For a small example matrix, we illustrate the effect of rank deficiency $\mathbf{SX}$ (Section B.2.1); perform uniform sampling with replacement (Section B.2.2); compute the expectations for $\mathbf{P_0}$ (Section B.2.3) and $\mathbf{PP}^\mathsf{T}$ (Section B.2.4); and put this into context with two matrices $\mathbf{S}$ at opposite ends of sampling performance (Section B.2.5).

Our example is the full column-rank matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{4 \times 2} \qquad \text{with} \qquad \mathbf{X}^\dagger = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

and $\operatorname{rank}(\mathbf{X}) = 2$. The hat matrix (2.5) and its null space are

$$\mathbf{P_x} = \mathbf{XX}^\dagger = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \operatorname{null}(\mathbf{P_x}) = \operatorname{range}\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ -1 & 0 \\ 0 & 1 \end{pmatrix}$$

while the model variance (4.1) is

$$\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \sigma^2\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}. \tag{B.1}$$

B.2.1. Effect of rank deficiency in Theorem 4.2 and Corollary 4.3. We choose two different matrices $\mathbf{S}$ with full row-rank $\operatorname{rank}(\mathbf{S}) = 2$, one producing a full rank $\mathbf{SX}$, and the other one a rank-deficient $\mathbf{SX}$.

1. Full column-rank $\mathbf{SX}$. The sketching matrix is

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{where} \quad \mathbf{SX} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = (\mathbf{SX})^\dagger = \mathbf{I}_2,$$

$\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X}) = 2$. The comparison hat matrix in Lemma 3.1 and the bias projector in Lemma 4.1 are

$$\mathbf{P} = \mathbf{X}(\mathbf{SX})^\dagger\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{P_0} = (\mathbf{SX})^\dagger(\mathbf{SX}) = \mathbf{I}_2.$$

Thus $\operatorname{range}(\mathbf{P}) = \operatorname{range}(\mathbf{X})$. The deviation of $\mathbf{P}$ from being an orthogonal projector onto $\operatorname{range}(\mathbf{X})$ is

$$\mathbf{PP}^T - \mathbf{P_x} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \|\mathbf{PP}^T - \mathbf{P_x}\|_2 = 1.$$

Thus, the solution $\tilde{\boldsymbol{\beta}}$ of (2.7) is an unbiased estimator, but with increased variance. Specifically,

- $\mathbf{P}$ is a projector onto $\operatorname{range}(\mathbf{X})$, but it is not an orthogonal projector, since $\mathbf{P}$ is not symmetric.
- The conditional expectation of $\tilde{\boldsymbol{\beta}}$ is $\mathbb{E}_\mathbf{y}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}] = \boldsymbol{\beta}_0$, since $\mathbf{P_0} = \mathbf{I}_2$, and the corresponding bound in Corollary 4.3 holds with equality.
- The conditional variance has increased compared to (B.1), because

$$\mathbb{V}\mathrm{ar}_\mathbf{y}\left[\tilde{\boldsymbol{\beta}} \,\middle|\, \mathbf{S}\right] = \sigma^2\,\mathbf{X}^\dagger\mathbf{PP}^T(\mathbf{X}^\dagger)^T = \sigma^2\,\mathbf{I}_2.$$

  In the worst case, it has zero norm-wise relative accuracy since

$$\|\,\mathbb{V}\mathrm{ar}_\mathbf{y}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}] - \mathbb{V}\mathrm{ar}_\mathbf{y}[\hat{\boldsymbol{\beta}}]\|_2/\|\,\mathbb{V}\mathrm{ar}_\mathbf{y}[\hat{\boldsymbol{\beta}}]\|_2 = \frac{1}{2} \leq \|\mathbf{PP}^T - \mathbf{P_x}\|_2 = 1.$$

2. Rank deficient $\mathbf{SX}$. The sketching matrix is

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{where} \quad \mathbf{SX} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = (\mathbf{SX})^\dagger,$$

$\operatorname{rank}(\mathbf{SX}) = 1 < \operatorname{rank}(\mathbf{X})$, and $\operatorname{range}(\mathbf{P}) \subset \operatorname{range}(\mathbf{X})$. The comparison hat matrix in Lemma 3.1 and the bias projector in Lemma 4.1 are

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{P_0} = (\mathbf{SX})^\dagger(\mathbf{SX}) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The deviation of $\mathbf{P}$ from being an orthogonal projector onto $\operatorname{range}(\mathbf{X})$ is

$$\mathbf{PP}^T - \mathbf{P_x} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & -1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \|\mathbf{PP}^T - \mathbf{P_x}\|_2 = 1,$$

and the rank deficiency of $\mathbf{SX}$ is represented by $\|\mathbf{I} - \mathbf{P_0}\|_2 = 1$. Thus, the solution $\tilde{\boldsymbol{\beta}}$ of (2.7) is a biased estimator with a conditional variance that is singular. Specifically,

- Although $\mathbf{P}$ is a projector, it is not an orthogonal projector onto range($\mathbf{X}$), since $\mathbf{P}$ is not symmetric and it projects only onto a lower-dimensional subspace of range($\mathbf{X}$).
- The conditional expectation of $\tilde{\boldsymbol{\beta}}$ is $\mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,|\, \mathbf{S}] \neq \boldsymbol{\beta}_0$, since $\mathbf{P_0} \neq \mathbf{I}_2$, and the relative distance to unbiasedness can be maximal in the worst case, since $\| \mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,|\, \mathbf{S}] - \boldsymbol{\beta}_0\|_2 \leq \|\boldsymbol{\beta}_0\|_2$.
- The conditional variance has become singular,

$$\mathbb{V}\mathrm{ar}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \,\middle|\, \mathbf{S}\right] = \sigma^2\, \mathbf{X}^\dagger \mathbf{P}\mathbf{P}^\mathsf{T} (\mathbf{X}^\dagger)^T = \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

  with zero norm-wise relative accuracy, and the corresponding bound holds with equality,

$$\| \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,|\, \mathbf{S}] - \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2 / \| \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2 = 1 = \|\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}\|_2.$$

B.2.2. Uniform sampling with replacement. Algorithm B.1 with $n = 4$ and $r = 2$ produces a sampling matrix $\mathbf{S} \in \mathbb{R}^{2\times 4}$, which has $n^2 = 16$ instances

$$\mathbf{S}_{ij} = \sqrt{2} \begin{pmatrix} \mathbf{e}_i^T \\ \mathbf{e}_j^T \end{pmatrix}, \qquad 1 \leq i, j \leq n,$$

each occurring with probability $1/n^2$. For instance,

$$\mathbf{S}_{11} = \sqrt{2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{S}_{42} = \sqrt{2} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The expectation of the Gram product is an unbiased estimator of the identity, since

$$\mathbb{E}_{\mathbf{s}}[\mathbf{S}^T\mathbf{S}] = \sum_{i=1}^{4}\sum_{j=1}^{4} \tfrac{1}{16} \mathbf{S}_{ij}^T \mathbf{S}_{ij} = \sum_{i=1}^{4}\sum_{j=1}^{4} \tfrac{1}{16} (\mathbf{e}_i\mathbf{e}_i^T + \mathbf{e}_j\mathbf{e}_j^T) = \mathbf{I}_4.$$

B.2.3. Expected rank deficiency in Theorem 4.5 and Corollary 4.6. The total expectation of $\mathbf{P_0} \in \mathbb{R}^{2\times 2}$ is

$$\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}] = \sum_{i=1}^{4}\sum_{j=1}^{4} \tfrac{1}{16} (\mathbf{S}_{ij}\mathbf{X})^\dagger (\mathbf{S}_{ij}\mathbf{X}) = \mathbb{E}_{\mathbf{s}}[\mathbf{P_0}] = \tfrac{1}{16} \begin{pmatrix} 12 & 0 \\ 0 & 7 \end{pmatrix}.$$

For instance, representative summands include

$$(\mathbf{S}_{13}\mathbf{X})^\dagger = \sqrt{\tfrac{1}{2}} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}^\dagger = \sqrt{\tfrac{1}{2}} \begin{pmatrix} 1/2 & 1/2 \\ 0 & 0 \end{pmatrix}, \qquad (\mathbf{S}_{13}\mathbf{X})^\dagger (\mathbf{S}_{13}\mathbf{X}) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

$$(\mathbf{S}_{32}\mathbf{X})^\dagger = \sqrt{\tfrac{1}{2}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^\dagger = \sqrt{\tfrac{1}{2}} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \sqrt{\tfrac{1}{2}}\, \mathbf{I}_2, \qquad (\mathbf{S}_{32}\mathbf{X})^\dagger (\mathbf{S}_{32}\mathbf{X}) = \mathbf{I}_2,$$

$$(\mathbf{S}_{44}\mathbf{X})^\dagger = \sqrt{\tfrac{1}{2}} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}^\dagger = \mathbf{0}, \qquad (\mathbf{S}_{44}\mathbf{X})^\dagger (\mathbf{S}_{44}\mathbf{X}) = \mathbf{0}.$$

Among the sketched matrices $\mathbf{SX}$, 75 percent are rank deficient. The ones with full column rank are $\mathbf{S}_{12}\mathbf{X}$, $\mathbf{S}_{21}\mathbf{X}$, $\mathbf{S}_{23}\mathbf{X}$, and $\mathbf{S}_{32}\mathbf{X}$. The expected rank deficiency of $\mathbf{SX}$ equals

$$\mathbb{E}_{\mathbf{s}}[\mathbf{I} - \mathbf{P_0}] = \tfrac{1}{16} \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix} \quad \text{with} \quad \| \mathbb{E}_{\mathbf{s}}[\mathbf{I} - \mathbf{P_0}]\|_2 = \tfrac{9}{16}.$$

Thus, the solution $\tilde{\boldsymbol{\beta}}$ of (2.7) is a biased estimator. Specifically,

- $\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]$ is not a projector, since it is not idempotent.
- The total expectation of $\tilde{\boldsymbol{\beta}}$ equals $\mathbb{E}_{\mathbf{s}}[\tilde{\boldsymbol{\beta}}] \neq \boldsymbol{\beta}_0$, since $\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}] \neq \mathbf{I}_2$. and the relative distance to unbiasedness can be large, since $\| \mathbb{E}[\tilde{\boldsymbol{\beta}}] - \boldsymbol{\beta}_0\|_2 \leq \frac{9}{16} \|\boldsymbol{\beta}_0\|_2$.

B.2.4. Expected deviation of P from being an orthogonal projector in Theorem 4.5 and Corollary 4.6. To the expectation of $\mathbf{PP}^T \in \mathbb{R}^{4\times 4}$, note that the trailing column of $\mathbf{X}$ is zero, and

$$\mathbf{PP}^T = \mathbf{X}(\mathbf{SX})^\dagger \mathbf{S}\,\mathbf{S}^T \left((\mathbf{SX})^\dagger\right)^T \mathbf{X}^T,$$

the trailing row and column of all instances of $\mathbf{PP}^T$ and $\mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T]$ are zero as well, and

$$\mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T] = \sum_{i=1}^{4}\sum_{j=1}^{4} \tfrac{1}{16}\, \mathbf{X}(\mathbf{S}_{ij}\mathbf{X})^\dagger \mathbf{S}_{ij}\mathbf{S}_{ij}^T \left((\mathbf{S}_{ij}\mathbf{X})^\dagger\right)^T \mathbf{X}^T = \tfrac{1}{16} \begin{pmatrix} 11 & 0 & 11 & 0 \\ 0 & 7 & 0 & 0 \\ 11 & 0 & 11 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Thus, $\mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T]$ is not a projector since it is not idempotent, and the expected deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$) can be larger than 50 percent, since

$$\mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T - \mathbf{P_x}] = \tfrac{1}{16} \begin{pmatrix} 3 & 0 & 3 & 0 \\ 0 & -9 & 0 & 0 \\ 3 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{with} \quad \| \mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T - \mathbf{P_x}]\|_2 = \tfrac{9}{16}.$$

B.2.5. Extreme examples. We consider two more 4×2 matrices, both with orthogonal columns, but at the opposite ends in terms of the performance for uniform sampling in Section B.2.2.

Columns of the Hadamard matrix. With its mass spread uniformly spread, which is quantified by minimal coherence and uniform leverage scores [13, 16], this matrix is optimal for uniform row sampling,

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \in \mathbb{R}^{4\times 2}, \qquad \mathbf{P_x} = \mathbf{XX}^\dagger = \tfrac{1}{2} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Half of the sketched matrices $\mathbf{SX}$ have full column rank. The expectations for the projectors are

$$\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}] = \tfrac{12}{16}\mathbf{I}_2, \qquad \mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T] = \tfrac{11}{16} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Thus the expected deviation of $\mathbf{SX}$ from full column-rank rank, and the expected deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$) are

$$\|\mathbb{E}_{\mathbf{s}}[\mathbf{I} - \mathbf{P_0}]\|_2 = \tfrac{4}{16}, \qquad \left\|\mathbb{E}_{\mathbf{s}}[\mathbf{PP}^T - \mathbf{P_x}]\right\|_2 = \tfrac{3}{16},$$

and clearly lower, and therefore better than the respective ones in Sections B.2.3 and B.2.4.

Columns of the identity matrix. With its concentrated mass spread, which is quantified by maximal coherence and widely differing leverage scores [13, 16], this matrix presents a worst case for uniform row sampling of 4×2 a full column-rank matrix,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{4\times 2}, \qquad \mathbf{P_x} = \mathbf{X}\mathbf{X}^\dagger = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Only two among the 16 sketched matrices $\mathbf{SX}$ have full column rank, $\mathbf{S}_{12}\mathbf{X}$ and $\mathbf{S}_{21}\mathbf{X}$. The expectations for the projectors are

$$\mathbb{E}_\mathbf{s}[\mathbf{P_0}] = \tfrac{7}{16}\mathbf{I}_2, \qquad \mathbb{E}_\mathbf{s}[\mathbf{P}\mathbf{P}^T] = \tfrac{7}{16}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The expected deviations of $\mathbf{SX}$ from full column-rank and of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$) are

$$\|\mathbb{E}_\mathbf{s}[\mathbf{I} - \mathbf{P_0}]\|_2 = \tfrac{9}{16}, \qquad \left\|\mathbb{E}_\mathbf{s}[\mathbf{P}\mathbf{P}^T - \mathbf{P_x}]\right\|_2 = \tfrac{9}{16},$$

thus clearly worse than those for the Hadamard matrix.

Acknowledgements. We are grateful to Dennis Boos and Chris Waddell for many helpful discussions.

REFERENCES

[1] H. Avron, P. Maymounkov, and S. Toledo, *Blendenpik: supercharging Lapack's least-squares solver*, SIAM J. Sci. Comput., 32 (2010), pp. 1217–1236.

[2] C. Boutsidis and P. Drineas, *Random projections for the nonnegative least-squares problem*, Linear Algebra Appl., 431 (2009), pp. 760–771.

[3] S. Chatterjee and A. S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Statist. Sci., 1 (1986), pp. 379–416. With discussion.

[4] P. Drineas, R. Kannan, and M. W. Mahoney, *Fast Monte Carlo Algorithms for Matrices. I: Approximating Matrix Multiplication*, SIAM J. Comput., 36 (2006), pp. 132–157.

[5] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Sampling algorithms for l_2 regression and applications*, in Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 2006, pp. 1127–1136.

[6] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math., 117 (2011), pp. 219–249.

[7] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics New York, 2001.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, fourth ed., 2013.

[9] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, second ed., 2002.

[10] D. C. Hoaglin and R. E. Welsch, *The Hat matrix in regression and ANOVA*, Amer. Statist., 32 (1978), pp. 17–22.

[11] I. C. F. Ipsen, *Relative perturbation results for matrix eigenvalues and singular values*, in Acta Numerica 1998, vol. 7, Cambridge University Press, Cambridge, 1998, pp. 151–201.

[12] I. C. F. Ipsen, *An overview of relative* $\sin\Theta$ *theorems for invariant subspaces of complex matrices*, J. Comput. Appl. Math., 123 (2000), pp. 131–153. Invited Paper for the special issue *Numerical Analysis 2000: Vol. III – Linear Algebra*.

[13] I. C. F. Ipsen and T. Wentworth, *The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 1490–1520.

[14] M. E. Lopes, S. Wang, and M. W. Mahoney, *Error estimation for randomized least-squares algorithms via the bootstrap*, in Proc. 35th International Conference on Machine Learning, vol. 80, PMLR, 2018, pp. 3217–3226.

[15] P. Ma, M. W. Mahoney, and B. Yu, *A statistical perspective on algorithmic leveraging*, in Proceedings of the 31st International Conference on International Conference on Machine Learning, vol. 32 of ICML'14, JMLR.org, 2014, pp. I–91–I–99.

[16] P. Ma, M. W. Mahoney, and B. Yu, *A statistical perspective on algorithmic leveraging*, J. Mach. Learn. Res., 16 (2015), pp. 861–911.

[17] C. L. Mallows, *Some comments on c p*, Technometrics, 15 (1973), pp. 661–675.

[18] X. Meng, M. A. Saunders, and M. W. Mahoney, *LSRN: a parallel iterative solver for strongly over- or underdetermined systems*, SIAM J. Sci. Comput., 36 (2014), pp. C95–C118.

[19] C. D. Meyer, *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[20] J. F. Monahan, *A primer on linear models*, Texts in Statistical Science Series, Chapman & Hall/CRC, Boca Raton, FL, 2008.

[21] D. Posada and T. R. Buckley, *Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests*, Systematic biology, 53 (2004), pp. 793–808.

[22] G. Raskutti and M. W. Mahoney, *A statistical perspective on randomized sketching for ordinary least-squares*, J. Mach. Learn. Res., 17 (2016), pp. Paper No. 214, 31.

[23] V. Rokhlin and M. Tygert, *A fast randomized algorithm for overdetermined linear least-squares regression*, Proc. Natl. Acad. Sci. USA, 105 (2008), pp. 13212–13217.

[24] T. Sarlós, *Improved Approximation Algorithms for Large Matrices via Random Projections*, in 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06), IEEE, Oct 2006, pp. 143–152.

[25] G. W. Stewart, *Collinearity and least squares regression*, Statist. Sci., 2 (1987), pp. 68–100. With discussion.

[26] G.-A. Thanei, C. Heinze, and N. Meinshausen, *Random Projections For Large-Scale Regression*, 2017, https://arxiv.org/abs/1701.05325.

[27] P. F. Velleman and R. E. Welsch, *Efficient computing of regression diagnostics*, Amer. Statist., 35 (1981), pp. 234–242.

[28] H. Wang, R. Zhu, and P. Ma, *Optimal Subsampling for Large Scale Logistic Regression*, J. Amer. Stat. Assoc., 113 (2018), pp. 829–844.