Randomized Least Squares Regression: Combining Model- and Algorithm-Induced Uncertainties[*]

Jocelyn T. Chi[†] and Ilse C. F. Ipsen[‡]

Abstract. We analyze the uncertainties in the minimum norm solution of full-rank regression problems, arising from Gaussian linear models, computed by randomized (row-wise sampling and, more generally, sketching) algorithms. From a deterministic perspective our structural perturbation bounds imply that least squares problems are less sensitive to multiplicative perturbations than to additive perturbations. From a probabilistic perspective, our expressions for the total expectation and variance 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. The relative 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 associated with the original and the sketched problems. A simple example, based on uniform sampling with replacement, illustrates the statistical quantities.

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 regression problems arising from 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 multivariate normal distribution. For a fixed response vector $\mathbf{y} \in \mathbb{R}^n$, one can determine a unique maximum likelihood estimator of $\boldsymbol{\beta}_0$ by computing the unique solution $\hat{\boldsymbol{\beta}}$ of the least squares problem

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

Statistical quality measures include expectation and variance of $\hat{\boldsymbol{\beta}}$, and residual sum of squares $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2$ [13, Section 7.2]; while roundoff errors from a numerically stable method are bounded in terms of the condition number of $\mathbf{X}$ with respect to (left) inversion, and the least squares residual $\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}$ [7, Chapter 5], [8, Chapter 20].

Randomized algorithms try to reduce the time complexity by first "compressing" or "preconditioning" the least squares problem. They can be classified according to [23, Section 1]: Compression of rows [2, 5, 6, 12, 15, 16, 21]; or columns [1]; or both [17]. We consider row compression

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

[*]The work was supported in part by NSF grants DGE-1633587 and DMS-1760374.
[†]Department of Statistics, North Carolina State University, Raleigh, NC, jtchi@ncsu.edu
[‡]Department of Mathematics, North Carolina State University, Raleigh, NC, ipsen@ncsu.edu

where $\mathbf{S} \in \mathbb{R}^{r\times n}$ is a random sampling or, more generally, sketching matrix with $r \leq n$, and the minimum norm solution is $\tilde{\boldsymbol{\beta}}$. Matrix concentration inequalities are used to derive probabilistic bounds for the error due to randomization [1, 6], and for the condition number of $\mathbf{SX}$ [12]. From a practical perspective, bootstrapping can deliverfast error estimates [14].

The pioneering work [15, 16] was the first to combine the uncertainties from the Gaussian linear model with the algorithm-induced uncertainties from random sampling of rows. Here we extend the first-order expansions in [15, 16] in a number of ways.

1.1. Contributions.

1. Our main result presents *exact* expressions for the total expectation and variance of $\tilde{\boldsymbol{\beta}}$ with regard to both, model- and algorithm-induced uncertainties (Theorem 4.5).
2. Our expressions hold for general random matrices $\mathbf{S}$, including sketching matrices that perform projections prior to sampling. Furthermore, our expressions also hold for rank deficient matrices $\mathbf{SX}$.
3. To compare least squares problems of different dimensions, we introduce the *comparison hat matrix* $\mathbf{P} = \mathbf{X}(\mathbf{SX})^{\dagger}\mathbf{S}$, which reduces to the *traditional hat matrix* $\mathbf{XX}^{\dagger}$ when $\mathbf{S}$ is the identity (Lemma 3.1, Remark 3.2).
4. We quantify the relative change in the total uncertainty of $\tilde{\boldsymbol{\beta}}$ compared to that of the model problem (Corollary 4.6):
   (a) The total bias increases, in the relative sense, with the expected deviation of the random variable $\mathbf{SX}$ from having full column rank.
   (b) The relative difference between total variance and model variance increases with two terms: the expected deviation of $\mathbf{SX}$ from having full rank, plus the expected deviation of the random variable $\mathbf{P}$ being an orthogonal projector onto range($\mathbf{X}$).
5. We quantify the model-induced uncertainty of $\tilde{\boldsymbol{\beta}}$, conditioned on $\mathbf{S}$, compared to that of the model problem (Theorem 4.3, Corollary 4.4):
   (a) The bias increases, in the relative sense, with the deviation of $\mathbf{SX}$ from having full column rank.
   (b) The variance changes, in the relative sense, 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. In contrast, recovering the model variance requires reproducing all of range($\mathbf{X}$).
6. Our structural bounds improve existing bounds, and imply that the minimum norm solution $\tilde{\boldsymbol{\beta}}$ and its residual are less sensitive to multiplicative perturbations than to additive perturbations (Corollary 3.5).

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

1. Deterministic: 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 (Section 3).
2. Probabilistic: The matrix $\mathbf{S}$ is a matrix-valued random variable (1.3) and (1.3) is a randomized algorithm for solving the linear model (1.1), and we derive expressions for expectation and variance with regard to the model- and algorithm-induced uncertain-

ties (Section 4).

This is followed by a brief review of sketching matrices used in randomized least squares solvers (Section 5); a simple example, designed to illustrate the bounds in a way that is easy for readers to reproduce (Section 6); and finally the proofs (Appendix A).

2. Models for Least squares Regression.

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

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

We review the different incarnations of least squares regression: the Gaussian linear model (Section 2.1), the traditional computation (Section 2.2), and the randomized algorithm (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 $\epsilon \in \mathbb{R}^n$ 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 given $\mathbf{y}$ solve

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

where $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$ represents the two-norm and the superscript T the transpose.

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 vector and the least squares residual vector are, respectively

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

In terms of the so-called *hat matrix* [3, 9, 24],

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

which is the orthogonal projector onto $\text{range}(\mathbf{X})$ along $\text{null}(\mathbf{X}^T)$, we can write

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

2.3. Randomized algorithm for least squares solution.

A randomized algorithm based on sketching, projecting or sampling of rows, is advantageous when $\mathbf{X}$ contains many redundant observations for a small set of variables, that is, $n \gg p$. From a deterministic perspective, this can be considered an extension of weighted least squares [7, Section 6.1] to rectangular weighting matrices.

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

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

Even if $\mathbf{S}$ has $r > p$ rows, $\operatorname{rank}(\mathbf{S}) < p$ is possible; and even if $\mathbf{S}$ does have full column rank, $\operatorname{rank}(\mathbf{S}\mathbf{X}) < p$ is still possible. Thus (2.7) can have infinitely many solutions, and one way to force uniqueness is to compute the solution of minimal two norm.

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 dimensions and cannot be directly compared; neither can their residuals. To remedy this, we follow previous work [5, 6, 20], and compare the predictions with regard to the *original* matrix,

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

3. Structural (deterministic perturbation) bounds.

Here $\mathbf{S}$ is a given, general matrix; and $\mathbf{S}\mathbf{X}$ is interpreted as a perturbation of $\mathbf{X}$. After deriving expressions for the solution, prediction and least squares residual of the perturbed problem (Section 3.1), we derive multiplicative perturbation bounds (Section 3.2), and discuss comparisons to existing work (Section 3.3).

3.1. The perturbed problem.

In analogy to the *hat matrix* $\mathbf{P}_{\mathbf{x}}$ in (2.5) for the original problem (2.3) we introduce a *comparison hat matrix* $\mathbf{P}$ for the perturbed problem (2.7), which allows a clean comparison between two least squares problems of different dimensions.

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

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

is an oblique projector where
1. $\mathbf{P}_{\mathbf{x}}\mathbf{P} = \mathbf{P}$.
2. $\mathbf{P} - \mathbf{P}_{\mathbf{x}}$ *reflects the difference between the spaces* $\operatorname{null}(\mathbf{P})$ *and* $\operatorname{null}(\mathbf{P}_{\mathbf{x}})$.
3. $\mathbf{P}\mathbf{X} = \mathbf{X}$ *if* $\operatorname{rank}(\mathbf{S}\mathbf{X}) = p$.

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* $\mathbf{P}_{\mathbf{x}}$ in (2.9).

If $\mathbf{S} = \mathbf{I}_n$, then $\mathbf{P} = \mathbf{P}_{\mathbf{x}}$. In general,

$$\operatorname{rank}(\mathbf{P}) = \operatorname{rank}(\mathbf{S}\mathbf{X}) \leq \operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{P}_{\mathbf{x}}) = p.$$

If $\operatorname{rank}(\mathbf{S}\mathbf{X}) = \operatorname{rank}(\mathbf{X})$, then $\mathbf{P}$ is an oblique version of $\mathbf{P}_{\mathbf{x}}$ with $\operatorname{range}(\mathbf{P}) = \operatorname{range}(\mathbf{P}_{\mathbf{x}})$, and the only difference is in their nullspaces. If $\operatorname{rank}(\mathbf{S}\mathbf{X}) < p$ then $\operatorname{rank}(\mathbf{P}) = \operatorname{rank}(\mathbf{S}\mathbf{X}) < p$, and $\mathbf{P}$ projects onto only a subspace of $\operatorname{range}(\mathbf{X})$. The example in Section 6.1 illustrates this.

Remark 3.2. The comparison hat matrix $\mathbf{P}$ generalizes the oblique projector $\mathbf{P_u}$ in [20, (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 $\operatorname{rank}(\mathbf{SX}) = p$, and equals $\mathbf{P_u} \equiv \mathbf{U}(\mathbf{SU})^\dagger\mathbf{S}$, where $\mathbf{U}$ is an orthonormal basis for $\operatorname{range}(\mathbf{X})$. In this case we have $\mathbf{P_u} = \mathbf{P}$. However, if $\operatorname{rank}(\mathbf{SX}) < \operatorname{rank}(\mathbf{X})$, 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 least squares residual* $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ *satisfy*

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

Proof. See Section A.2. ∎

Theorem 3.3 shows that the relations between perturbed and original least squares problems are governed by $\mathbf{P} - \mathbf{P_x}$, which reflects the difference between $\operatorname{null}(\mathbf{P})$ and $\operatorname{null}(\mathbf{P_x})$. Motivated by the ground breaking result [16, Lemma 1], reproduced in the lemma below, Theorem 3.3 strengthens it with explicit expressions for $\tilde{\boldsymbol{\beta}}$ that hold for general matrices $\mathbf{S}$ and do not require assumptions on $\operatorname{rank}(\mathbf{SX})$.

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*[1] $\mathbf{w} \equiv \operatorname{diag}(\mathbf{S}^T\mathbf{S}) \in \mathbb{R}^n$ *has a scaled multinomial distribution with expected value* $\mathbb{E}[\mathbf{w}] = \mathbb{1}$, $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, *and a Taylor series expansion around* $\mathbf{w}_0 = \mathbb{1}$ *of the solution* $\tilde{\boldsymbol{\beta}}(\mathbf{w})$ *of (2.7) exists with* $\tilde{\boldsymbol{\beta}}(\mathbf{w}_0) = \hat{\boldsymbol{\beta}}$, *then*

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

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 the problem (2.7) as a multiplicative perturbation, and derive norm-wise relative error bounds for the solution, prediction, and least squares residual; and compare them to existing bounds.

The vector two-norm induces the matrix norm $\|\mathbf{X}\|_2$, and the two-norm condition number of the full column-rank matrix $\mathbf{X}$ with regard to (left) inversion is

$$\kappa_2(\mathbf{X}) \equiv \|\mathbf{X}\|_2\|\mathbf{X}^\dagger\|_2 \geq 1.$$

[1]For a matrix $\mathbf{M}$, $\operatorname{diag}(\mathbf{M})$ represents the vector of diagonal elements.

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 of (2.7) satisfies

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

The least squares residual $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ *satisfies*

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

Proof. See Section A.3 ∎

For $\mathbf{S} = \mathbf{I}_n$, the bounds in Corollary 3.5 are zero and tight since $\mathbf{P} = \mathbf{P}_{\mathbf{x}}$.

Remark 3.6 (Sensitivity to multiplicative perturbations). Corollary 3.5 implies that least squares solutions $\tilde{\boldsymbol{\beta}}$ are insensitive to multiplicative perturbations if $\mathbf{X}$ is well conditioned with regard to inversion, and if $\mathbf{y}$ is close to $\mathrm{range}(\mathbf{X})$. The bound for $\tilde{\boldsymbol{\beta}}$ consists of two parts:

1. The perturbation $\|\mathbf{P} - \mathbf{P}_{\mathbf{x}}\|_2$ reflects the distance between the null spaces $\mathrm{null}(\mathbf{P})$ and $\mathrm{null}(\mathbf{P}_{\mathbf{x}})$. It is an absolute as well as a relative perturbation since $\|\mathbf{P}_{\mathbf{x}}\|_2 = 1$.
2. The amplifier can be bounded by [7, (5.3.16)]

$$\kappa_2(\mathbf{X}) \, \frac{\|\mathbf{y}\|_2}{\|\mathbf{X}\|_2 \|\hat{\boldsymbol{\beta}}\|_2} \leq \kappa_2(\mathbf{X}) \frac{\|\mathbf{y}\|_2}{\|\mathbf{X}\hat{\boldsymbol{\beta}}\|_2} = \frac{\kappa_2(\mathbf{X})}{\cos\theta}.$$

3.3. Comparison to existing work. In contrast to multiplicative perturbation bounds for eigenvalue and singular value problems [10, 11], we do not require $\mathbf{S}$ to be nonsingular or square. Weighted least squares problems [7, Section 6.1] employ nonsingular diagonal matrices $\mathbf{S}$ for regularization or scaling of discrepancies, and do not view them as a perturbation.

Remark 3.7 (Comparison to additive perturbations). Corollary 3.5 also implies that the minimum norm solution of (2.7) and its residual are less sensitive to multiplicative perturbations than to additive perturbations, which are reviewed below in Lemma 3.8.

In contrast to additive bounds [7, (5.3.12)], [8, (20.12)], the bound for the least squares residual $\tilde{\mathbf{e}}$ is not affected by $\kappa_2(\mathbf{X})$. Note that the $\sin\theta$ term in the denominator is also occurs in additive bounds [7, (5.3.12)] if the relative error is normalized by $\hat{\mathbf{e}}$ rather than $\mathbf{y}$.

In contrast to additive bounds [7, Section 5.3.6], [8, Section 20.1], [22, (3.4)], the bound for $\hat{\boldsymbol{\beta}}$ does not square the condition number and does not require $\mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X})$. This can be seen from Lemma 3.8 below, where the first summand corresponds to the bound for $\tilde{\boldsymbol{\beta}}$ in Corollary 3.5.

Lemma 3.8 (Theorem 5.3.1 in [7]). *With the assumptions in Section 2, let* $\mathbf{X} + \mathbf{E}$ *have* $\mathrm{rank}(\mathbf{X} + \mathbf{E}) = \mathrm{rank}(\mathbf{X})$ *and* $\eta \equiv \|\mathbf{E}\|_2 / \|\mathbf{X}\|_2$.

The solution $\bar{\boldsymbol{\beta}}$ *to* $\min_{\boldsymbol{\beta}} \|(\mathbf{X} + \mathbf{E})\boldsymbol{\beta} - \mathbf{y}\|_2$ *satisfies*

$$\frac{\|\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2}{\|\hat{\boldsymbol{\beta}}\|_2} \leq \kappa_2(\mathbf{X}) \left(\frac{\|\mathbf{y}\|_2}{\|\mathbf{X}\|_2 \|\hat{\boldsymbol{\beta}}\|_2} + 1 \right) \eta + \kappa(\mathbf{X})^2 \frac{\|\hat{\mathbf{e}}\|_2}{\|\mathbf{X}\|_2 \|\hat{\boldsymbol{\beta}}\|_2} \, \eta + \mathcal{O}(\eta^2).$$

Compared to existing structural bounds for randomized least squares algorithms, which are reproduced in Lemma 3.9, the bound for $\tilde{\boldsymbol{\beta}}$ in Corollary 3.5 is more general and tighter in the sense that it does not exhibit nonlinear dependences on the perturbations.

Lemma 3.9 (Theorem 1 in [6]). *In addition to Section 2, also assume that* $\|\mathbf{P_x y}\|_2 \geq \gamma \|\mathbf{y}\|_2$ *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 [19, 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). From that we derive the total expectation and variance (Section 4.3).

4.1. Model-induced uncertainty.

We view the model-induced randomness in (2.2) as a property of the response vector $\mathbf{y}$. That is, the noise vector $\boldsymbol{\epsilon}$ in (2.2) has mean and covariance

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

The well-known statistical properties of (2.3) are reviewed below.

Lemma 4.1 (Model-induced uncertainty for (2.3)). *With the assumptions in Section 2, the response vector (2.2), and the least squares prediction (2.6) and solution (2.4) satisfy*

$$\begin{aligned} \mathbb{E}_{\mathbf{y}}[\mathbf{y}] &= \mathbf{X}\boldsymbol{\beta}_0, & \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\mathbf{y}] &= \sigma^2 \mathbf{I}_n \\ \mathbb{E}_{\mathbf{y}}[\hat{\mathbf{y}}] &= \mathbf{X}\boldsymbol{\beta}_0, & \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\mathbf{y}}] &= \sigma^2 \mathbf{P_x} \in \mathbb{R}^{n\times n} \\ \mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}_0, & \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] &= \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} \in \mathbb{R}^{p\times p}. \end{aligned}$$

Proof. See Section A.4. ∎

Lemma 4.1 asserts that the computed solution $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}_0$, and points to the well known dependence of the variance on the conditioning of $\mathbf{X}$ [22, Section 5].

The difficulty in analyzing the sketched problem (2.7), coupled with general concern about the first-order expansions like the ones in [15, 16], is that there are instances of $\mathbf{S}$ for which $\mathrm{rank}(\mathbf{SX}) < \mathrm{rank}(\mathbf{X})$. In this case $(\mathbf{SX})^\dagger$ cannot be expressed in terms of $\mathbf{SX}$ as in (2.1), and the least squares problem (2.7) is ill-posed.

One can derive bounds [1, Theorem 3.2], [12, Theorems 4.1and 5.2] on the probability that $\mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X})$ for matrices $\mathbf{S}$ that perform uniform sampling and leverage score sampling. However, such bounds are not useful here, because expected values run over *all* instances of $\mathbf{SX}$.

We introduce a quantity that signals the deviation of the columns of $\mathbf{SX}$ from linear independence.

Lemma 4.2 (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 where

1. $\mathbf{PX} = \mathbf{XP_0}$
2. $\mathbf{P_0} = \mathbf{I}_p$ *if* $\operatorname{rank}(\mathbf{SX}) = p$.
3. $\mathbf{I} - \mathbf{P_0}$ *represents the deviation of* $\mathbf{SX}$ *from full-rank.*

Proof. See Section A.5. ■

The name *bias projector* will become apparent in Theorem 4.3, where $\mathbf{P_0}$ represents the bias in $\tilde{\boldsymbol{\beta}}$.

If $\mathbf{S} = \mathbf{I}_n$, then $\mathbf{P_0} = \mathbf{P_x}$. If $\operatorname{rank}(\mathbf{SX}) = p$, then Lemma 4.2 recovers $\mathbf{PX} = \mathbf{X}$ from Lemma 3.1, confirming that $\mathbf{P}$ is a projector onto $\operatorname{range}(\mathbf{X})$. However, if $\operatorname{rank}(\mathbf{SX}) < p$, then $\mathbf{P_0}$ characterizes the subspace of $\operatorname{range}(\mathbf{X})$ onto which $\mathbf{P}$ projects.

4.2. Model-induced uncertainty, conditioned on algorithm-induced uncertainty.

We determine the expectation and variance for the solution of (2.7) conditioned on $\mathbf{S}$. That is, we assume that the random sketching matrix $\mathbf{S}$ is fixed at a specific value $\mathbf{S_0}$ and use $\mathbb{E}_{\mathbf{y}}\left[\cdot \,\middle|\, \mathbf{S}\right]$ as an abbreviation for the conditional expectation $\mathbb{E}_{\mathbf{y}}\left[\cdot \,\middle|\, \mathbf{S} = \mathbf{S}_0\right]$.

Theorem 4.3 (Model-induced uncertainty in (2.7), conditioned on S). *With the assumptions in Section 2, the solution of (2.7) satisfies*

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

where $\mathbf{P}\mathbf{P}^T - \mathbf{P_x}$ *is the deviation of* $\mathbf{P}$ *from being an orthogonal projector onto* $\operatorname{range}(\mathbf{X})$. *Furthermore* $\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \,\middle|\, \operatorname{rank}(\mathbf{SX}) = p\right] = \boldsymbol{\beta}_0$.

Proof. See Section A.6. ■

The exact expressions for general sketching matrices $\mathbf{S}$ in Theorem 4.3 extend the first-order expressions for specific sampling matrices in [16, Lemmas 2-6]. The examples in Section 6.1 illustrate the effect of rank deficiency of $\mathbf{SX}$ on the quantities in Theorem 4.3.

Theorem 4.3 shows that the bias of $\tilde{\boldsymbol{\beta}}$ is proportional to the deviation $\mathbf{I} - \mathbf{P_0}$ of $\mathbf{SX}$ from having full column rank. In other words, the bias becomes worse as the rank deficiency increases. If $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, then $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}_0$. Theorem 4.3 also implies that the conditional variance is close to the model variance if $\mathbf{P}$ is close to being an orthogonal projector onto $\operatorname{range}(\mathbf{X})$.

The relevance of $\mathbf{I} - \mathbf{P_0}$ and $\mathbf{P}\mathbf{P}^T - \mathbf{P_x}$ becomes clear in the relative differences below.

Corollary 4.4 (Relative difference between conditional and model quantities). *With the assumptions in Theorem 4.3,*

$$\frac{\| \mathbb{V}\mathrm{ar}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}\right] - \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2}{\| \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2} \leq \|\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}\|_2.$$

If also $\boldsymbol{\beta}_0 \neq \mathbf{0}$, *then*

$$\frac{\| \mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}\right] - \boldsymbol{\beta}_0\|_2}{\|\boldsymbol{\beta}_0\|_2} \leq \|\mathbf{I} - \mathbf{P}_{\mathbf{0}}\|_2.$$

Proof. See Section A.7. ∎

Corollary 4.4 implies that the relative differences to unbiasedness and model variance are solely governed by the quantities $\mathbf{I} - \mathbf{P}_{\mathbf{0}}$ and $\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}$, respectively. Both of them are absolute as well as relative measures since $\|\mathbf{I}\|_2 = \|\mathbf{P}_{\mathbf{x}}\|_2 = 1$. Specifically, the relative difference between conditional and model variance increases with the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$); and the bias of $\tilde{\boldsymbol{\beta}}$ increases, in the relative sense, with the deviation of $\mathbf{SX}$ from full column rank.

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

4.3. Combined algorithm-induced and model-induced uncertainty.

We determine the total expectation and variance for the solution in (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 $\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\cdot]$. The total mean and variance of the combined uncertainty are denoted by $\mathbb{E}[\cdot]$ and $\mathbb{V}\mathrm{ar}[\cdot]$, and computed by conditioning on the algorithm-induced randomness,

$$\mathbb{E}\left[\cdot\right] = \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}_{\mathbf{y}}\left[\cdot \,\middle|\, \mathbf{S}\right]\right]. \tag{4.1}$$

Since $\mathbf{S}$ is a matrix-valued random variable, so are the projectors $\mathbf{P}$ and $\mathbf{P}_{\mathbf{0}}$. Examples of $\mathbf{S}$ can be found in Sections 5 and 6.

Theorem 4.5 (Total mean and variance for (2.7)). *With the assumptions in Section 2, let* $\mathbf{S}$ *be a random sketching matrix. The solution of (2.7) satisfies*

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

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.8. ∎

Theorem 4.5 presents exact expressions for general random matrices $\mathbf{S}$, thereby extending the first order approximations for specific sampling matrices in [16, Lemmas 2-6], and shows:

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.
   The expectation $\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]$ of a projector $\mathbf{P_0}$ is not a projector, in general, as the example in Section 6.3 illustrates.
2. The total variance of $\tilde{\boldsymbol{\beta}}$ is proportional to the expected deviation of $\mathbf{SX}$ from full column rank, plus the expected deviation of the matrix-valued random variable $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).

The importance of the expected deviations of the projectors appears in the analog of Corollary 4.4 below.

Corollary 4.6. *With the assumptions in Theorem 4.5,*

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

If also $\boldsymbol{\beta}_0 \neq \mathbf{0}$, *then*

$$\frac{\|\,\mathbb{E}_{\mathbf{s}}[\tilde{\boldsymbol{\beta}}] - \boldsymbol{\beta}_0\|_2}{\|\boldsymbol{\beta}_0\|_2} \leq \|\,\mathbb{E}_{\mathbf{s}}[\mathbf{I} - \mathbf{P_0}]\|_2.$$

Corollary 4.6 implies that the bias of $\tilde{\boldsymbol{\beta}}$ increases, in the relative sense, with the expected deviation of $\mathbf{SX}$ from full rank; and that the relative difference from total variance to model variance increases with (i) the expected deviation of $\mathbf{P}$ being an orthogonal projector onto range($\mathbf{X}$), plus (ii) the expected deviation of $\mathbf{SX}$ from full rank.

5. Random sketching matrices in least squares.

We present a few examples of sketching matrices from the randomized least squares solvers [1, 2, 5, 6, 14, 15, 16, 17, 21].

Uniform sampling with replacement. This is the *EXACTLY(c)* algorithm [6, Algorithm 3] with uniform probabilities, which is used for row-wise compression of direct methods for the solution of full column rank least squares in [6, Algorithm 3], see also the *BasicMatrixMultiplication Algorithm* [4, Fig. 2], [12, Algorithm 3.2], [14, Algorithms 1 and 2], and [16, UNIF].

Algorithm 5.1 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
  $\mathbf{S} = \sqrt{\frac{n}{r}} \begin{pmatrix} \mathbf{e}_{k_1} & \ldots & \mathbf{e}_{k_r} \end{pmatrix}^T$

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

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 [17, 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: $\mathbf{S} = \texttt{randn}(r, n)$.

6. Example.

We illustrate the projectors in Corollary 3.5, Theorem 4.3, Corollary 4.4. Theorem 4.5, and Corollary 4.6 in a way that is easy for readers to reproduce. For a small example matrix, we illustrate the effects of rank deficiency $\mathbf{SX}$ (Section 6.1); perform uniform sampling with replacement (Section 6.2); compute the expectations for $\mathbf{P_0}$ (Section 6.3) and $\mathbf{PP}^T$ (Section 6.4); and put this into context with two matrices at opposite ends of sampling performance (Section 6.5).

Consider the full column rank matrix

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

$\mathrm{rank}(\mathbf{X}) = 2$,

$$\mathbf{P_x} = \mathbf{X}\mathbf{X}^\dagger = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}}] = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}, \tag{6.1}$$

and

$$\mathrm{null}(\mathbf{P_x}) = \mathrm{range}(\mathbf{I} - \mathbf{P_x}) = \mathrm{range} \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

6.1. Effects of rank deficiency. We illustrate the effect of rank deficiency on the quantities in Corollary 3.5, Theorem 4.3 and Corollary 4.4 by choosing two different matrices $\mathbf{S}$ with $\mathrm{rank}(\mathbf{S}) = 2$.

Full column rank SX. Here

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

$\mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X}) = 2$, and $\mathrm{range}(\mathbf{P}) = \mathrm{range}(\mathbf{X})$. This gives the projectors

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

with

$$\mathrm{null}(\mathbf{P}) = \mathrm{range}(\mathbf{I} - \mathbf{P}) = \mathrm{range} \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This shows:

- $\mathbf{P}$ is not an orthogonal projector, since it is not symmetric.
- The solution $\tilde{\boldsymbol{\beta}}$ in Theorem 4.3 is an unbiased estimator.
- The conditional variance in Theorem 4.3 has increased compared to (6.1), since

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

Rank deficient SX. Here

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

$\mathrm{rank}(\mathbf{SX}) = 1 < \mathrm{rank}(\mathbf{X})$, and $\mathrm{range}(\mathbf{P}) \subset \mathrm{range}(\mathbf{X})$. This gives the projectors

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

with

$$\mathrm{null}(\mathbf{P}) = \mathrm{range}(\mathbf{I} - \mathbf{P}) = \mathrm{range} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This shows:

- The rank deficiency of $\mathbf{SX}$ causes the dimension of $\mathrm{null}(\mathbf{P})$ to increase.
- The solution $\tilde{\boldsymbol{\beta}}$ in Theorem 4.3 is a biased estimator since $\mathbf{P_0} \neq \mathbf{I}_2$.
- The conditional variance in Theorem 4.3 has become singular, since

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

6.2. Uniform sampling with replacement. Algorithm 5.1 with $n = 4$ and $r = 2$ produces a sampling matrix $\mathbf{S} \in \mathbb{R}^{2\times 4}$, which has $n^2 = 16$ instances

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

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,

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

6.3. Expected deviation of SX from rank deficiency. We compute the expectation of $\mathbf{P_0} \in \mathbb{R}^{2\times 2}$ in Theorem 4.5 and Corollary 4.6,

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

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, 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 $\mathbf{S}_{32}\mathbf{X}$. This shows

- $\mathbb{E}_{\mathbf{s}}[\mathbf{P}_{\mathbf{0}}]$ is not a projector, since it is not idempotent.
- The solution $\tilde{\boldsymbol{\beta}}$ in Theorem 4.5 is a biased estimator, since $\mathbb{E}_{\mathbf{s}}[\mathbf{P}_{\mathbf{0}}] \neq \mathbf{I}_2$.
- The relative difference of $\tilde{\boldsymbol{\beta}}$ from unbiasedness in Corollary 4.6 can exceed 50 percent, since it is bounded by $\|\,\mathbb{E}_{\mathbf{s}}[\mathbf{I}-\mathbf{P}_{\mathbf{0}}]\|_2 = \frac{9}{16}$, where

$$\mathbb{E}_{\mathbf{s}}[\mathbf{I}-\mathbf{P}_{\mathbf{0}}] = \tfrac{1}{16}\begin{pmatrix}4 & 0\\ 0 & 9\end{pmatrix}.$$

6.4. Expected deviation of $\mathbf{P}$ from being an orthogonal projector. We compute the expectation of $\mathbf{P}\mathbf{P}^T \in \mathbb{R}^{4\times 4}$ in Theorem 4.5 and Corollary 4.6. Since the trailing column of $\mathbf{X}$ is zero, and

$$\mathbf{P}\mathbf{P}^T = \mathbf{X}(\mathbf{S}\mathbf{X})^\dagger\mathbf{S}\mathbf{S}^T((\mathbf{S}\mathbf{X})^\dagger)^T\mathbf{X}^T,$$

the trailing row and columns of all instances of $\mathbf{P}\mathbf{P}^T$ and $\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T]$ are, too. Thus

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

This shows

- $\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T]$ is not a projector since it is not idempotent.
- The expected deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$) in Corollary 4.6 can exceed 50 percent, since it is bounded by $\left\|\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}]\right\|_2 = \frac{9}{16}$, where with the hat matrix $\mathbf{P}_{\mathbf{x}}$ in (6.1)

$$\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}] = \tfrac{1}{16}\begin{pmatrix}3 & 0 & 3 & 0\\ 0 & -9 & 0 & 0\\ 3 & 0 & 3 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}.$$

6.5. 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 6.2.

Columns of the Hadamard matrix. With its mass spread uniformly spread, which is quantified by minimal coherence and uniform leverage scores [12, 16], this matrix is optimal for uniform sampling,

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

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

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

The expected deviations of $\mathbf{SX}$ from full column rank and of $\mathbf{P}$ from being an orthogonal projector are clearly lower, thus better, than the respective ones in Sections 6.3 and 6.4,

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

Columns of the identity matrix. With its concentrated mass spread, which is quantified by maximal coherence and widely differing leverage scores [12, 16], this matrix presents a worst case for a 4×2 matrix of full column rank.

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

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 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.

Appendix A. Proofs. We present the proofs for Sections 3.1 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 [7, Section 5.5.2] that uniquely characterize the Moore-Penrose inverse,

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

A.1. Proof of Lemma 3.1. The Moore-Penrose conditions (A.1) imply

$$\mathbf{P}^2 = \mathbf{X}\underbrace{(\mathbf{SX})^\dagger\mathbf{S}\,\mathbf{X}(\mathbf{SX})^\dagger}_{(\mathbf{SX})^\dagger}\mathbf{S} = \mathbf{X}(\mathbf{SX})^\dagger\mathbf{S} = \mathbf{P}.$$

Since $\mathbf{P}^2 = \mathbf{P}$, but $\mathbf{P}$ is not symmetric in general, it is an oblique projector.

1. From (2.1) follows

$$\mathbf{P_x P} = \mathbf{XX}^\dagger\mathbf{P} = \underbrace{\mathbf{XX}^\dagger\mathbf{X}}_{\mathbf{X}}(\mathbf{SX})^\dagger\mathbf{S} = \mathbf{X}(\mathbf{SX})^\dagger\mathbf{S} = \mathbf{P}.$$

2. Use the fact [18, Problem 5.9.12] that $\text{null}(\mathbf{P}) = \text{null}(\mathbf{P_x})$ if and only if $\mathbf{PP_x} - \mathbf{P} = \mathbf{0}$ and $\mathbf{P_x P} - \mathbf{P_x} = \mathbf{0}$. For the latter, the above implies $\mathbf{P_x P} - \mathbf{P_x} = \mathbf{P} - \mathbf{P_x}$. Thus we can interpret $\mathbf{P} - \mathbf{P_x}$ as a measure for the distance between $\text{null}(\mathbf{P})$ and $\text{null}(\mathbf{P_x})$.
3. If $\text{rank}(\mathbf{SX}) = p$ then we can express the Moore-Penrose inverse as in (2.1),

$$\mathbf{PX} = \mathbf{X}\underbrace{\left((\mathbf{SX})^T\mathbf{SX}\right)^{-1}(\mathbf{SX})^T}_{(\mathbf{SX})^\dagger}\mathbf{SX} = \mathbf{X}.$$

A.2. Proof of Theorem 3.3. The first expression for the least squares solution follows from (2.1), (2.8), Lemma 3.1, and

$$\tilde{\boldsymbol\beta} = \mathbf{X}^\dagger\mathbf{X}\tilde{\boldsymbol\beta} = \mathbf{X}^\dagger\ \mathbf{X}(\mathbf{SX})^\dagger\mathbf{Sy} = \mathbf{X}^\dagger\mathbf{Py}.$$

Adding $\hat{\boldsymbol\beta} - \mathbf{X}^\dagger\mathbf{y} = \mathbf{0}$ from (2.4) to the above gives the second expression

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

where the last equality is due to (A.1) and

$$\mathbf{X}^\dagger = \mathbf{X}^\dagger\mathbf{XX}^\dagger = \mathbf{X}^\dagger\mathbf{P_x}.$$

Regarding the least squares residual, from (2.9), the first expression for $\tilde{\boldsymbol\beta}$, (2.5) and Lemma 3.1 follows

$$\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol\beta} = \mathbf{XX}^\dagger\mathbf{Py} = \mathbf{P_x Py} = \mathbf{Py}.$$

Adding $\hat{\mathbf{y}} - \mathbf{P_x y} = \mathbf{0}$ from (2.6) gives

$$\tilde{\mathbf{y}} = \hat{\mathbf{y}} + \mathbf{P y} - \mathbf{P_x y} = \hat{\mathbf{y}} + (\mathbf{P} - \mathbf{P_x})\mathbf{y}.$$

As for the predictor, (2.9) and the above expression for $\tilde{\mathbf{y}}$ imply

$$\tilde{\mathbf{e}} = \mathbf{y} - \tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{P})\mathbf{y}.$$

Adding and subtracting $\hat{\mathbf{e}} - (\mathbf{I} - \mathbf{P_x})\mathbf{y} = \mathbf{0}$ from (2.6) gives

$$\tilde{\mathbf{e}} = \hat{\mathbf{e}} + (\mathbf{I} - \mathbf{P})\mathbf{y} - (\mathbf{I} - \mathbf{P_x})\mathbf{y} = \hat{\mathbf{e}} + (\mathbf{P_x} - \mathbf{P})\mathbf{y}.$$

A.3. Proof of Corollary 3.5. The bounds are a direct consequence of Theorem 3.3.

From [7, Theorem 5.3.1] follows that $\|\hat{\mathbf{e}}\|_2/\|\mathbf{y}\|_2 = \sin\theta$. The assumption $\theta < \pi/2$ implies $\sin\theta < 1$, hence $\|\hat{\mathbf{e}}\|_2 < \|\mathbf{y}\|_2$ and therefore $\hat{\boldsymbol{\beta}} \neq \mathbf{0}$. The assumption $\theta > 0$ implies $\mathbf{y} \notin \text{range}(\mathbf{X})$, thus $\hat{\mathbf{e}} \neq \mathbf{0}$. Therefore we can divide by the appropriate quantities. In the bound for $\tilde{\mathbf{e}}$, write [7, Theorem 5.3.1]

$$\|\mathbf{y}\|_2/\|\hat{\mathbf{e}}\|_2 = 1/\sin\theta.$$

A.4. Proof of Lemma 4.1. The linearity of the mean and (2.2) imply

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\mathbf{y}] &= \mathbb{E}_{\mathbf{y}}[\mathbf{X}\,\boldsymbol{\beta}_0] + \mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}] = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{0} = \mathbf{X}\boldsymbol{\beta}_0\\
\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\mathbf{y}] &= \mathbb{V}\mathrm{ar}_{\mathbf{y}}\,[\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}] = \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}_n.
\end{aligned}$$

From (2.6), the above, and $(\mathbf{P_x})^2 = \mathbf{P_x}$ follows

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\hat{\mathbf{y}}] &= \mathbb{E}_{\mathbf{y}}[\mathbf{P_x y}] = \mathbf{P_x}\,\mathbb{E}_{\mathbf{y}}[\mathbf{y}] = \mathbf{P_x X}\boldsymbol{\beta}_0 = \mathbf{X}\boldsymbol{\beta}_0\\
\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\mathbf{y}}] &= \mathbb{V}\mathrm{ar}_{\mathbf{y}}\,[\mathbf{P_x y}] = \mathbf{P_x}\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\mathbf{y}]\mathbf{P_x} = \sigma^2\mathbf{P_x}.
\end{aligned}$$

From the above, (2.4), and (2.1) follows

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}_{\mathbf{y}}[\mathbf{X}^\dagger\mathbf{y}] = \mathbf{X}^\dagger\,\mathbb{E}_{\mathbf{y}}[\mathbf{y}] = \mathbf{X}^\dagger\mathbf{X}\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0\\
\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] &= \mathbb{V}\mathrm{ar}_{\mathbf{y}}\left[\mathbf{X}^\dagger\mathbf{y}\right] = \mathbf{X}^\dagger\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\mathbf{y}](\mathbf{X}^\dagger)^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned}$$

A.5. Proof of Lemma 4.2. The Moore-Penrose conditions (A.1) imply

$$(\mathbf{P_0})^2 = (\mathbf{SX})^\dagger\underbrace{(\mathbf{SX})(\mathbf{SX})^\dagger(\mathbf{SX})}_{\mathbf{SX}} = (\mathbf{SX})^\dagger(\mathbf{SX}) = \mathbf{P_0},$$

and $(\mathbf{P_0})^T = \mathbf{P_0}$, confirming that $\mathbf{P_0}$ is an orthogonal projector.

1. Lemma 3.1 implies $\mathbf{PX} = \mathbf{X}(\mathbf{SX})^\dagger\mathbf{S}\,\mathbf{X} = \mathbf{XP_0}$.
2. If $\text{rank}(\mathbf{SX}) = p$, then $(\mathbf{SX})^\dagger$ is a left-inverse, see (2.1), so that $\mathbf{P_0} = \mathbf{I}_p$.

A.6. Proof of Theorem 4.3. The expectation follows from Theorem 3.3, Lemma 4.1, Lemma 4.2, and (2.1),

$$\mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,\big|\, \mathbf{S}] = \mathbf{X}^{\dagger}\mathbf{P}\,\mathbb{E}_{\mathbf{y}}[\mathbf{y}] = \mathbf{X}^{\dagger}\underbrace{\mathbf{P}\mathbf{X}}_{\mathbf{X}\mathbf{P_0}}\boldsymbol{\beta}_0 = \mathbf{X}^{\dagger}\mathbf{X}\mathbf{P_0}\boldsymbol{\beta}_0 = \mathbf{P_0}\boldsymbol{\beta}_0. \tag{A.2}$$

From the definition of variance, Theorem 3.3, and the above follows

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

For the middle term in first summand, Lemma 4.1 implies

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}[\mathbf{y}\mathbf{y}^T] &= (\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{X}\boldsymbol{\beta}_0)^T + \mathbf{X}\boldsymbol{\beta}_0\,\mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}]^T + \mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}](\mathbf{X}\boldsymbol{\beta}_0)^T + \mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \\ &= (\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{X}\boldsymbol{\beta}_0)^T + \sigma^2\mathbf{I}_n,\end{aligned} \tag{A.3}$$

and when inserting this into the leading half of the first summand, one obtains as in (A.2) that

$$\mathbf{X}^{\dagger}\mathbf{P}\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{P_0}\boldsymbol{\beta}_0. \tag{A.4}$$

This gives the first expression for the conditional variance,

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,\big|\, \mathbf{S}] &= (\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T + \sigma^2\,\mathbf{X}^{\dagger}\mathbf{P}\mathbf{P}^T(\mathbf{X}^{\dagger})^T - (\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T \\ &= \sigma^2\,\mathbf{X}^{\dagger}\mathbf{P}\mathbf{P}^T(\mathbf{X}^{\dagger})^T.\end{aligned}$$

To obtain the second expression, multiply the model variance from Lemma 4.1 by $\mathbf{I} = (\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}$,

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] = \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1} &= \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1}\,(\mathbf{X}^T\mathbf{X})\,(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{P_x}\mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\,\mathbf{X}^{\dagger}\mathbf{P_x}(\mathbf{X}^{\dagger})^T,\end{aligned}$$

where the remaining equalities follow from $\mathbf{X} = \mathbf{P_x}\mathbf{X}$ in (2.5) and from (2.1). Now add $\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] - \sigma^2\,\mathbf{X}^{\dagger}\mathbf{P_x}(\mathbf{X}^{\dagger})^T = \mathbf{0}$ in the first expression for the variance.

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

A.7. Proof of Corollary 4.4. The second expression for the variance in Theorem 4.3 and submultiplicativity imply

$$\begin{aligned}\|\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,\big|\, \mathbf{S}] - \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2 &\leq \sigma^2\,\|\mathbf{X}^{\dagger}\|_2\,\|\mathbf{P}\mathbf{P}^T - \mathbf{P_x}\|_2\|(\mathbf{X}^{\dagger})^T\|_2 \\ &= \|\mathbf{P}\mathbf{P}^T - \mathbf{P_x}\|_2\|\,\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]\|_2,\end{aligned}$$

where the equality follows from $\|\mathbf{M}\|_2\|\mathbf{M}^T\|_2 = \|\mathbf{M}\mathbf{M}^T\|_2$, and for any full-column rank matrix $\mathbf{M}$,

$$\mathbf{M}^{\dagger}(\mathbf{M}^{\dagger})^T = (\mathbf{M}^T\mathbf{M})^{-1}\,\mathbf{M}^T\mathbf{M}\,(\mathbf{M}^T\mathbf{M})^{-1} = (\mathbf{M}^T\mathbf{M})^{-1}.$$

The second expression for the expectation in Theorem 4.3 and submultiplicativity imply

$$\|\,\mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \,|\, \mathbf{S}] - \boldsymbol{\beta}_0\|_2 \leq \|\mathbf{I} - \mathbf{P_0}\|_2\|\boldsymbol{\beta}_0\|_2.$$

A.8. Proof of Theorem 4.5. The expectation follows from sequential conditioning (4.1) and Lemma 4.3,

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

Insert this expression for the mean into the definition of the variance, and apply sequential conditioning (4.1),

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

From Theorem 3.3, (A.3) and (A.3) follows

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

Conditioning this on $\mathbf{S}$ gives

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

Put everything together to obtain the first expression for the variance,

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

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

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

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

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

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

[10] 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.

[11] 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*.

[12] 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.

[13] K. Lange, *Numerical analysis for statisticians*, Statistics and Computing, Springer, New York, second ed., 2010.

[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] 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.

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

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

[20] 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.

[21] 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.

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

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

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