Proceedings of Machine Learning Research vol 99:1–20, 2019 32nd Annual Conference on Learning Theory

Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression

Michał Dereziński MDEREZIN@BERKELEY.EDU
Department of Statistics, UC Berkeley

Kenneth L. Clarkson KLCLARKS@US.IBM.COM
IBM Research – Almaden

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
ICSI and Department of Statistics, UC Berkeley

Manfred K. Warmuth MANFRED@UCSC.EDU
Google Inc. Zürich & UC Santa Cruz

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

In experimental design, we are given a large collection of vectors, each with a hidden response value that we assume derives from an underlying linear model, and we wish to pick a small subset of the vectors such that querying the corresponding responses will lead to a good estimator of the model. A classical approach in statistics is to assume the responses are linear, plus zero-mean i.i.d. Gaussian noise, in which case the goal is to provide an unbiased estimator with smallest mean squared error (A-optimal design). A related approach, more common in computer science, is to assume the responses are arbitrary but fixed, in which case the goal is to estimate the least squares solution using few responses, as quickly as possible, for worst-case inputs. Despite many attempts, characterizing the relationship between these two approaches has proven elusive. We address this by proposing a framework for experimental design where the responses are produced by an arbitrary unknown distribution. We show that there is an efficient randomized experimental design procedure that achieves strong variance bounds for an unbiased estimator using few responses in this general model. Nearly tight bounds for the classical A-optimality criterion, as well as improved bounds for worst-case responses, emerge as special cases of this result. In the process, we develop a new algorithm for a joint sampling distribution called volume sampling, and we propose a new i.i.d. importance sampling method: inverse score sampling. A key novelty of our analysis is in developing new expected error bounds for worst-case regression by controlling the tail behavior of i.i.d. sampling via the jointness of volume sampling. Our result motivates a new minimax-optimality criterion for experimental design with unbiased estimators, which can be viewed as an extension of both A-optimal design and sampling for worst-case regression.

Keywords: A-optimality, worst-case, volume sampling, minimax, linear regression, least squares.

1. Introduction

Consider fixed design regression in d dimensions, with $n \gg d$ experiments parameterized by vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and the associated real random response variables $y_1, \ldots, y_n$. Suppose that each response variable is modeled as a linear function of the parameters plus i.i.d. Gaussian noise: $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma^2)$. Let $\mathbf{X}$ be the $n \times d$ matrix whose rows are $\mathbf{x}_i^\top$ (assumed to be full rank) and let $\mathbf{y}$ be the vector of the n random responses y_i. Under the above standard

statistical assumptions the *least squares estimator* $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X}) = \mathbf{X}^{\dagger}\mathbf{y}$ (where $\mathbf{X}^{\dagger} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}$ is the Moore-Penrose pseudo-inverse) is known to be the minimum variance unbiased estimator for $\mathbf{w}^* \in \mathbb{R}^d$. This implies that it satisfies $\mathbb{E}_{\mathbf{y}}[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})] = \mathbf{w}^*$, while achieving the smallest possible *mean squared error*: $\mathrm{MSE}[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})] = \mathbb{E}_{\mathbf{y}}\big[\|\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X}) - \mathbf{w}^*\|^2\big] = \sigma^2\phi$, where σ^2 is the magnitude of the noise and $\phi = \mathrm{tr}((\mathbf{X}^{\top}\mathbf{X})^{-1})$ captures the relevant spectral structure of $\mathbf{X}$. To compute this estimator exactly, we have to observe all n responses.

In the realm of experimental design (Fedorov, 1972), one asks: what if we are given all n vectors $\mathbf{x}_i$ but are allowed to query only $k \ll n$ of the responses? An unbiased estimator produced under this additional restriction will certainly be no better than $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})$ (in terms of its MSE). There are many experimental design criteria that have been considered. For example, we may wish to find a weight vector that minimizes the excess mean squared error resulting from the restricted access to the responses. This criterion is known as an A-optimal design. In this model, the problem reduces to finding a subset $S \subseteq [n]$ of k experiments for which the mean squared error of the least squares estimator is minimized. Its MSE then becomes $\min_S \sigma^2\mathrm{tr}((\mathbf{X}_S^{\top}\mathbf{X}_S)^{-1})$, where $\mathbf{X}_S$ is a submatrix with k rows selected by S. Other optimality criteria have been studied for selecting subset S, e.g., V-optimality (which we discuss below), as well as D- and E-optimality (which are not based on the variance of the estimator, therefore they are not as relevant to this discussion).

How good (in terms of the MSE) can the A-optimal subset be in general? Not surprisingly, this will depend on the total noise of the responses, i.e. $\mathbb{E}\big[\|\boldsymbol{\xi}\|^2\big] = n\sigma^2$, where $\boldsymbol{\xi} \in \mathbb{R}^n$ is the vector of noise variables $\xi_1, \ldots, \xi_n$, as well as the structure of $\mathbf{X}$ described by $\phi = \mathrm{tr}((\mathbf{X}^{\top}\mathbf{X})^{-1})$. The following result from numerical linear algebra shows the existence of a subset S with a good A-optimality bound as a function of its size k which is known to be asymptotically tight for some matrices. The resulting experimental design given in the corollary can be computed efficiently.

Theorem 1 (Avron and Boutsidis, 2013) *For any full rank $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $d \le k \le n$, there is a subset $S \subseteq [n]$ of size k s.t. $\mathrm{tr}((\mathbf{X}_S^{\top}\mathbf{X}_S)^{-1}) \le \frac{n-d+1}{k-d+1}\mathrm{tr}((\mathbf{X}^{\top}\mathbf{X})^{-1})$.*

Although Theorem 1 was originally stated as a worst-case linear algebra statement, it easily leads to the following corollary regarding the statistical MSE. Here $\mathbf{y}_S$ denotes the vector of the selected random responses.

Corollary 2 *Given $\mathbf{X} \in \mathbb{R}^{n\times d}$ such that $\mathrm{tr}((\mathbf{X}^{\top}\mathbf{X})^{-1}) = \phi$ and $\epsilon > 0$, there is an experimental design $S \subseteq [n]$ of size $k \le d + \phi/\epsilon$ s.t. for any $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$, where $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{\xi}] = \sigma^2\mathbf{I}$,*

$$\mathbb{E}_{\mathbf{y}}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}_S|\mathbf{X}_S)\big] = \mathbf{w}^* \quad \textit{and} \quad \underbrace{\mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}_S|\mathbf{X}_S)\big]}_{\le \frac{n-d+1}{k-d+1}\sigma^2\phi} - \underbrace{\mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big]}_{\sigma^2\phi} \le \epsilon \cdot \underbrace{\mathbb{E}_{\mathbf{y}}\big[\|\boldsymbol{\xi}\|^2\big]}_{n\sigma^2}.$$

Note that the bound in Corollary 2 holds even without subtracting $\mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big]$, however we include it here for the sake of consistency with the later discussion.

1.1. Experimental design with arbitrary random responses

While noise $\boldsymbol{\xi}$ need not be i.i.d. Gaussian to show Corollary 2, it still has to be zero-mean, homoscedastic (same variances) and uncorrelated. In this section we show that there are experimental designs for which the MSE bound from Corollary 2 holds for any (even adversarial) noise. This will allow us to propose a new "minimax-optimality" criterion for experimental design (in Section 1.2) which can be viewed as a generalization of A-optimality to arbitrary random responses. From

now on, the only assumption we make on the random variables y_i is that they have a finite second moment. We next redefine the optimal linear predictor $\mathbf{w}^*$ and the vector of noise variables $\boldsymbol{\xi}$ as:[1]

$$\mathbf{w}^* \overset{def}{=} \operatorname*{argmin}_{\mathbf{w}} \mathbb{E}_{\mathbf{y}}\left[\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2\right] = \mathbf{X}^\dagger \mathbb{E}[\mathbf{y}], \qquad \boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}} \overset{def}{=} \mathbf{X}\mathbf{w}^* - \mathbf{y}.$$

Note that when the noise happens to have mean zero, i.e. $\mathbb{E}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}] = \mathbf{0}$, then this definition of $\mathbf{w}^*$ is consistent with the statistical setting. Having no knowledge of the response model means that we cannot commit to a particular fixed subset S because those responses could be adversarially noisy. To avoid this, we allow randomization in the design procedure.

Definition 3 *A "random experimental design" $(S, \widehat{\mathbf{w}})$ of size k consists of a **random** set $S \subseteq [n]$ of size at most k and a **random** function $\widehat{\mathbf{w}} : \mathbb{R}^{|S|} \to \mathbb{R}^d$, which returns an estimator $\widehat{\mathbf{w}}(\mathbf{y}_S)$.*

The mean squared error in this context is defined as: $\mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{y}}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2\big]$, so it is exactly the standard MSE, except with the expectation taken over the randomness of both the responses and the design. Our main result shows that when we allow the experimental design procedure to be randomized, the mean squared error bound given in Corollary 2 for homoscedastic noise can be recovered almost exactly for arbitrary random response vectors $\mathbf{y}$ (which includes deterministically chosen response vectors as a special case).

Theorem 4 *Given $\mathbf{X} \in \mathbb{R}^{n\times d}$ such that $\mathrm{tr}((\mathbf{X}^\top\mathbf{X})^{-1}) = \phi$ and $\epsilon > 0$, there is a random experimental design $(S, \widehat{\mathbf{w}})$ of size $k = O(d\log n + \phi/\epsilon)$ s.t. for **any** random response vector $\mathbf{y}$,*

$$\mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{y}}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \mathbf{w}^* \quad \textit{and} \quad \mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big] \le \epsilon \cdot \mathbb{E}_{\mathbf{y}}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big].$$

To put this result in context we consider several different response models to which it applies:

1. *A-optimal experimental design.* If we assume independent homoscedastic zero-mean noise, then our model matches the classical A-optimal experimental design, except for allowing the design procedure to be randomized. Despite the broadness of Theorem 4 it still offers sample complexity that is only a log factor away from that of Corollary 2.

2. *Heteroscedastic regression.* We let each response have zero-mean noise with some unknown variance $\mathrm{Var}[\xi_i] = \sigma_i^2$. In this case, unlike existing work such as Dereziński and Warmuth (2018), we bound the MSE in terms of $\sum_i \sigma_i^2$ rather than $n \cdot \max_i \sigma_i^2$. Our design achieves this without having to adaptively estimate the variances as done by Wiens and Li (2014).

3. *Bayesian regression.* Suppose that $\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{z}$, where $\mathbf{w} \sim D_{\mathbf{w}}$ is a random vector with a prior $D_{\mathbf{w}}$ and mean $\mathbf{w}^*$, whereas $\mathbf{z}$ is a zero-mean random noise. In this case we may wish to minimize MSE w.r.t. $\mathbf{w}$ (and not $\mathbf{w}^*$), i.e., $\mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{w},\mathbf{z}}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}\|^2\big]$. For this purpose we can still apply Theorem 4 to the response vector $\mathbf{y}$ conditioned on $\mathbf{w}$, obtaining:

$$\underbrace{\mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}\|^2\big] - \mathbb{E}\big[\|\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X}) - \mathbf{w}\|^2\big]}_{\mathbb{E}\left[\mathbb{E}\left[\|\widehat{\mathbf{w}}(\mathbf{y}_S)-\mathbf{w}\|^2 - \|\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})-\mathbf{w}\|^2 \,\middle|\, \mathbf{w}\right]\right]} \le \underbrace{\epsilon \cdot \mathrm{tr}\big(\mathrm{Var}[\mathbf{z}]\big)}_{\mathbb{E}\left[\epsilon \cdot \mathbb{E}[\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 \,|\mathbf{w}]\right]}.$$

While traditional Bayesian experimental design (see Chaloner and Verdinelli, 1995) focuses on i.i.d. Gaussian noise, our results apply to arbitrary zero-mean noise. A natural future direction is to extend Theorem 4 to biased estimators that take advantage of the prior information.

1. Using the fact that $\mathbb{E}_{\mathbf{y}}\big[\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2\big] = \mathbb{E}_{\mathbf{y}}\big[\|\mathbf{X}\mathbf{w}-\mathbb{E}[\mathbf{y}]\|^2\big] + \mathbb{E}\big[\|\mathbf{y}-\mathbb{E}[\mathbf{y}]\|^2\big]$.

4. *Worst-case regression*. We let $\mathbf{y}$ be some arbitrary fixed vector $\mathbf{y} \in \mathbb{R}^n$, i.e., $\mathrm{Var}[\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}] = \mathbf{0}$ (a well-studied problem; see, e.g., Drineas et al., 2006). Then $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X}) = \mathbf{w}^*$ and we get:

$$\mathbb{E}_{S,\widehat{\mathbf{w}}}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2\big] \leq \epsilon \cdot \|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2,$$

the first such bound that holds: (a) "in expectation" (rather than "with constant probability"), (b) for an unbiased estimator, (c) for sample size $O(\phi/\epsilon)$ (when ϵ is sufficiently small).

As a corollary to Theorem 4, we give an additional result which bounds the *mean squared prediction error* (MSPE) instead of MSE, defined as $\mathrm{MSPE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{y}}\big[\|\mathbf{X}(\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*)\|^2\big]$. In many tasks, the performance of an estimator is evaluated in terms of the prediction accuracy, in which case MSPE may be a natural metric. Note that here the sample complexity no longer depends on the spectral parameter ϕ (which is replaced by d), just as it happens when bounding MSPE in the classical homoscedastic setting.

Theorem 5 *Given a full rank $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\epsilon > 0$, there is a random experimental design $(S, \widehat{\mathbf{w}})$ of size $k = O(d\log n + d/\epsilon)$ such that for **any** random response vector $\mathbf{y}$,*

$$\mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{y}}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \mathbf{w}^* \quad \textit{and} \quad \mathrm{MSPE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSPE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big] \leq \epsilon \cdot \mathbb{E}_{\mathbf{y}}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big].$$

In the statistical setting, minimizing the MSPE is often referred to as V-optimal design (see Wiens and Li, 2014). On the other hand, in worst-case regression analysis (when responses form a fixed vector $\mathbf{y} \in \mathbb{R}^n$), the mean squared prediction error is often replaced by the "square loss": $L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$. Theorem 5 implies a bound on the expected square loss of the estimator $\widehat{\mathbf{w}}(\mathbf{y}_S)$:

$$\mathbb{E}_{S,\widehat{\mathbf{w}}}\big[L(\widehat{\mathbf{w}}(\mathbf{y}_S))\big] \overset{(*)}{=} \mathrm{MSPE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] + L(\mathbf{w}^*) \leq (1+\epsilon)\cdot L(\mathbf{w}^*). \tag{1}$$

where $(*)$ follows from the unbiasedness of $\widehat{\mathbf{w}}(\mathbf{y}_S)$ via the bias-variance decomposition of the expected square loss. The only *expected* loss bound of this kind known prior to this result required sample size $k = O(d^2/\epsilon)$ (Dereziński and Warmuth, 2017).

Since our experimental design is randomized, each evaluation may produce a different result. In fact this can go to our advantage: instead of using one design with a larger k we can choose to produce multiple independent designs with a small k, say $(S_1, \widehat{\mathbf{w}}_1), \ldots, (S_m, \widehat{\mathbf{w}}_m)$, and then average them. This strategy may be preferrable in distributed settings and when data privacy is a concern. Since all the designs are unbiased for the random responses $\mathbf{y}$, it follows that:

$$\mathrm{MSE}\bigg[\frac{1}{m}\sum_{t=1}^{m}\widehat{\mathbf{w}}_t(\mathbf{y}_{S_t})\bigg] - \mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big] = \frac{1}{m}\Big(\mathrm{MSE}\big[\widehat{\mathbf{w}}_1(\mathbf{y}_{S_1})\big] - \mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big]\Big),$$

with an analogous formula also holding for the MSPE.

1.2. Minimax-optimal experimental design with unbiased estimators

If we divide both sides of the inequality in Theorem 4 by the right-hand-side $\mathbb{E}_{\mathbf{y}}[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2]$, we see a ratio bounded above by ϵ for all $\mathbf{y}$. This ratio, or rather its maximum over all $\mathbf{y}$, can be considered a quality criterion for experimental designs, to be minimized instead of only bounded. We will call the optimum a *minimax-optimal design*. Similarly as in the standard setup, we minimize over all *unbiased* estimators. The key difference is that we allow the design to be randomized. Let $\mathcal{F}$ denote the family of *all* random vectors in $\mathbb{R}^n$ with finite second moment.

Definition 6 (unbiased estimators) *Given matrix* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *and budget* $k \in \{d, \dots, n\}$, *let* $\mathcal{W}_k(\mathbf{X})$ *be the family of all random experimental designs* $(S, \widehat{\mathbf{w}})$ *of size* k *such that:*

$$\mathbb{E}_{S,\widehat{\mathbf{w}},\mathbf{y}}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \operatorname*{argmin}_{\mathbf{w}} \mathbb{E}\big[\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2\big] = \mathbf{X}^\dagger\mathbb{E}[\mathbf{y}] \quad \textit{for all } \mathbf{y} \in \mathcal{F}.$$

In Appendix A we show that the least squares estimator $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X}) = \mathbf{X}^\dagger\mathbf{y}$ is the *minimum variance unbiased estimator* (MVUE) among all such estimators with an unrestricted budget, i.e., $\mathcal{W}_n(\mathbf{X})$.

Proposition 7 *Given any full rank matrix* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *and any random function* $\widehat{\mathbf{w}} : \mathbb{R}^n \to \mathbb{R}^d$,

$$\textit{if} \quad \mathbb{E}_{\mathbf{y},\widehat{\mathbf{w}}}[\widehat{\mathbf{w}}(\mathbf{y})] = \mathbf{X}^\dagger\mathbb{E}[\mathbf{y}] \quad \forall_{\mathbf{y}\in\mathcal{F}}, \quad \textit{then} \quad \mathrm{Var}\big[\widehat{\mathbf{w}}(\mathbf{y})\big] \succeq \mathrm{Var}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big] \quad \forall_{\mathbf{y}\in\mathcal{F}}.$$

It is thus natural to minimize the excess mean squared error incurred by an unbiased estimator with a restricted budget compared to that of $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})$. Note that if we did not introduce the unbiasedness constraint it would be unclear what comparator should be used in place of $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})$, which makes that constraint particularly important here. Since we take a maximum over all response vectors in $\mathcal{F}$, we normalize the error by the noise $\mathbb{E}_{\mathbf{y}}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big]$ (equal to $n\sigma^2$ in the classical setting). To avoid division by zero, we exclude all fixed vectors in the column span of $\mathbf{X}$, denoted $\mathrm{Sp}(\mathbf{X}) \subseteq \mathbb{R}^n$.

Definition 8 (minimax-optimal value) *For a full-rank* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *and* $d \le k \le n$, *define:*

$$R_k^*(\mathbf{X}) \overset{def}{=} \min_{(S,\widehat{\mathbf{w}})\in\mathcal{W}_k(\mathbf{X})} \max_{\mathbf{y}\in\mathcal{F}\setminus\mathrm{Sp}(\mathbf{X})} \frac{\mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big]}{\mathbb{E}_{\mathbf{y}}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big]},$$

where $\mathrm{MSE}\big[\widehat{\mathbf{w}}\big]$ *for any unbiased estimator* $\widehat{\mathbf{w}}$ *denotes* $\mathbb{E}\big[\|\widehat{\mathbf{w}} - \mathbb{E}[\widehat{\mathbf{w}}]\|^2\big]$.

Proposition 9 (matching upper and lower bounds) *Let* $\mathbf{X}$ *denote a full rank* $n \times d$ *matrix and* $\phi = \mathrm{tr}((\mathbf{X}^\top\mathbf{X})^{-1})$. *Then, for any* $d \le k \le n$, *we have* $0 \le R_k^*(\mathbf{X}) < \infty$. *Furthermore:*

1. (upper) *There is* $C > 0$ *s.t. for any* $\mathbf{X}$ *and* $k \ge C \cdot d\log n$, *we have* $R_k^*(\mathbf{X}) \le C \cdot \phi/k$;
2. (lower) *For any* n, d *and* $\epsilon \in (0,1)$, *there is* $\mathbf{X}$ *s.t. if* $k^2 < \epsilon nd/3$ *then* $R_k^*(\mathbf{X}) \ge (1-\epsilon)\cdot\phi/k$.

Part 1 of Proposition 9 follows from our main result (Theorem 4), whereas part 2 is an application of a matrix inequality of Avron and Boutsidis (2013), see details in Appendix A. Note that if we defined $\mathcal{F}$ as the family of all random vectors $\mathbf{y}$ such that the noise $\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}$ is i.i.d. centered Gaussian (with any variance), then in this case the least squares estimator would also be the MVUE, and the above definition would become equivalent to the classical A-optimality criterion. Even in this special case, finding an exactly optimal design is hard (to our knowledge, NP-hardness has not been established), although efficient approximation algorithms exist for A-optimality (see Section 2). Similar questions can be asked about minimax-optimal design, however without any restrictions on the design procedure, this task appears daunting. In Section 3 we present one such restriction based on "volume sampling" which leads to a family of efficient unbiased estimators that we used in Theorems 4 and 5.

1.3. Construction and efficiency of random experimental designs

The random experimental design used in Theorems 4 and 5 consists of two primary components:

1. *volume sampling:* the initial few experiments are drawn from a joint sampling distribution over sets $S \subseteq [n]$ of size d such that $\Pr(S) \propto \det(\mathbf{X}_S)^2$;

2. *i.i.d. sampling:* the remaining $k-d$ experiments are sampled independently from a carefully chosen distribution $q = (q_1, \ldots, q_n)$.

While it is mainly the i.i.d. sampling that is responsible for bounding the sample size k, volume sampling is necessary for establishing both the unbiasedness and the expected bounds. The key novelty of our analysis is using volume sampling to control the MSE in the *tail* of the distribution, and using the concentration properties of i.i.d. sampling to bound it in the *bulk* of the distribution (see Section 4). The i.i.d. sampling distribution q used in the proof is a mixture of uniform distribution with two importance sampling techniques:

1. *Leverage score sampling*: $\Pr(i) = p_i^{\text{lev}} \overset{def}{=} \frac{1}{d}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i$ for $i \in [n]$. This is a standard sampling method which has been used in obtaining bounds for worst-case linear regression.

2. *Inverse score sampling*: $\Pr(i) = p_i^{\text{inv}} \overset{def}{=} \frac{1}{\phi}\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-2}\mathbf{x}_i$ for $i \in [n]$. This is a novel sampling technique which is essential for achieving $O(\phi/\epsilon)$ sample size for small ϵ.

As discussed earlier, having chosen a design, we may wish to produce multiple independent samples of it, for example to construct an averaged estimator. Thus, we break down the computational cost into the preprocessing cost (incurred once per given matrix $\mathbf{X}$) and sampling/estimation cost (incurred every time a new estimator is produced). The estimation step simply requires computing a least squares estimator from k samples, which costs $O(kd^2)$. The preprocessing involves all the calculations necessary to construct the sampling distributions. Both of the above importance sampling distributions can be computed exactly in time $O(nd^2)$ or approximately in time $O(nd\log n + d^3\log d)$ using standard sketching techniques (see Drineas et al., 2012). Once they are obtained, the sampling cost is negligible. On the other hand, for volume sampling both preprocessing and sampling cost can be significant. Dereziński et al. (2018) showed that a volume sampled set of size d can be generated in time $O(d^4)$ by selecting it from a sequence of $O(d^2)$ i.i.d. samples from the leverage score distribution p^{lev} (see Theorem 6 there). Below, we improve on that result.

Theorem 10 *For any $\mathbf{X}$ and q such that $q_i \geq \frac{1}{2}p_i^{\text{lev}}$, there is an algorithm which, given matrix $\mathbf{X}^\top\mathbf{X}$ and a stream of i.i.d samples from q, returns a set S s.t. $\Pr(S) \propto \det(\mathbf{X}_S)^2$, and w.p. at least $1-\delta$ it runs in time $O\big(d^3\log d\log\frac{1}{\delta}\big)$ using $O(d\log d\log\frac{1}{\delta})$ i.i.d. samples.*

Our algorithm improves on the best known sampling time for volume sampling from $O(d^4)$ to $O(d^3\log d)$, which has important implications for other applications of this distribution such as determinantal point processes (see Section 3 for the proof and further discussion). To establish correctness of the sampling, this algorithm requires the exact computation of matrix $\mathbf{X}^\top\mathbf{X}$, which typically costs $O(nd^2)$. The algorithm of Dereziński et al. (2018) only requires an approximation of this matrix, which can be computed in time $O(nd\log n + d^4\log d)$. Similar improvements in the preprocessing cost for our algorithm may be possible. We leave this as an open question.

2. Related work

There is a large body of related work, and we describe only that which most informed our approach.

Classical experimental design. Many optimality criteria have been considered as functions $F(S)$ of a subset $S \subseteq [n]$ (Pukelsheim, 2006), assuming that $y_i = \mathbf{x}_i^\top\mathbf{w}^* + \xi_i$ where $\xi_i \sim \mathcal{N}(0,\sigma^2)$, and these typically have natural interpretations for the least squares estimator. Recent work has studied

the tractability of finding an approximately optimal subset $\widehat{S}$ of size k, i.e., such that $F(\widehat{S}) \leq (1+\epsilon)\min_{S:|S|=k} F(S)$. For example, Allen-Zhu et al. (2017) showed that polynomial time algorithms are possible for many classical optimality criteria, such as A-optimality, $F_A(S) = \mathrm{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})$, D-optimality, $F_D(S) = \det(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}$, V-optimality, $F_V(S) = \mathrm{tr}(\mathbf{X}(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}^\top)$ and others, as long as $k = \Omega(d/\epsilon^2)$; Wang et al. (2017) showed tractable approximability of A/V-optimality for $k = \Omega(d^2/\epsilon)$; and this was later improved by Nikolov et al. (2019) to $k = \Omega(d/\epsilon + (\log \epsilon^{-1})/\epsilon^2)$. Robust variants of experimental design have been considered to address more general response models. In particular, Ou and Zhou (2009) assume that the covariance matrix of the noise is known only approximately and defines a minimax-type criterion where the maximization goes over a neighborhood of that covariance; and Wiens and Li (2014) use an active learning procedure to estimate the individual noise variances before constructing the design. None of these procedures, however, are truly agnostic to the response model.

Subset selection for worst-case regression. Subset selection has been studied extensively for both statistical and worst-case regression models. Perhaps most relevant is the work of Boutsidis et al. (2013), which showed a lower bound for any deterministically chosen subset S and function $\widehat{\mathbf{w}}$, when the hidden response vector $\mathbf{y}$ is arbitrary but fixed. This implies that random sampling is necessary in this setting. In the context of *randomized numerical linear algebra* (RandNLA; see Woodruff, 2014; Drineas and Mahoney, 2016), it was shown by Drineas et al. (2006) that a random sampling algorithm based on the statistical leverage scores constructs an estimator $\widehat{\mathbf{w}}(\mathbf{y}_S)$ which, with constant probability, achieves $\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2 \leq \epsilon \cdot \|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2$ by using $k = O(d\log d + \lambda_{\max}((\mathbf{X}^\top\mathbf{X})^{-1}) \cdot d/\epsilon)$ samples. The estimator we propose in Theorem 4 achieves the same bound for $k = O(d\log d + \phi/\epsilon)$. Since $\phi = \mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^{-1}\big) \leq \lambda_{\max}\big((\mathbf{X}^\top\mathbf{X})^{-1}\big) \cdot d \leq \phi \cdot d$, our result is better by up to a factor of d.

Statistical versus algorithmic approaches. Ma et al. (2014) and Raskutti and Mahoney (2015) were the first to consider statistical guarantees (such as A-optimality) that can be obtained by sampling methods developed for RandNLA (primarily leverage score sampling), contrasting them with some common worst-case guarantees. However, these works treat those two settings separately (in particular, the statistical setting is limited to i.i.d. Gaussian noise), rather than putting them under one umbrella of minimax experimental design, as we do. Subsequently, Chen and Price (2019) showed loss bounds for worst-case regression which extend to a randomized response model that is comparable to ours. They give a randomized estimator (*not unbiased*) that with constant probability achieves the following bound on the square loss: $L(\widehat{\mathbf{w}}(\mathbf{y}_S)) \leq (1+\epsilon)L(\mathbf{w}^*)$, for sample size $k = O(d/\epsilon)$, where $L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$. In contrast we obtain an *unbiased* estimator achieving the same bound *in expectation* with only slightly larger sample size of $k = O(d\log n + d/\epsilon)$.

Constant probability versus unbiased expectations. Unlike our Theorems 4 and 5, most results in RandNLA are stated to hold with high or constant probability (Woodruff, 2014; Drineas and Mahoney, 2016, 2017) as opposed to in expectation, and they do not provide unbiased estimators, which often makes them incomparable to statistical approaches. In fact, expected bounds are often impossible for these techniques (e.g., for leverage score sampling; see Dereziński and Warmuth, 2018). Unbiased estimators were first introduced to worst-case regression by Dereziński and Warmuth (2017), who gave the first *expected* square loss bound for sample size of $k = O(d^2/\epsilon)$ via volume sampling. Subsequently, Dereziński et al. (2018) demonstrated an unbiased estimator with

a *constant probability* loss bound for sample size $k = O(d \log d + d/\epsilon)$. Our result builds on the latter by obtaining an unbiased estimator with an *expected* loss bound for $k = O(d \log n + d/\epsilon)$.

3. Rescaled volume sampling

We now discuss the sampling distribution introduced by Dereziński et al. (2018), based on earlier work by Avron and Boutsidis (2013), that allows for constructing unbiased least squares estimators.

Definition 11 *Given full rank matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a distribution $q = (q_1, \ldots, q_n)$ s.t. $q_i > 0$ for all $i \in [n]$, we define q-rescaled volume sampling of size $k \geq d$, written $\mathrm{VS}_q^k(\mathbf{X})$, as a distribution over index sequences $\pi = (\pi_1, \ldots, \pi_k) \in [n]^k$ such that:*

$$\Pr(\pi) = \frac{\det\big(\mathbf{X}^\top \mathbf{S}_\pi^\top \mathbf{S}_\pi \mathbf{X}\big)}{\frac{d!}{k^d}\binom{k}{d}\det(\mathbf{X}^\top\mathbf{X})} \prod_{i=1}^k q_{\pi_i}, \qquad \textit{where} \qquad \mathbf{S}_\pi = \begin{bmatrix} \frac{1}{\sqrt{kq_{\pi_1}}}\mathbf{e}_{\pi_1}^\top \\ \vdots \\ \frac{1}{\sqrt{kq_{\pi_k}}}\mathbf{e}_{\pi_k}^\top \end{bmatrix} \in \mathbb{R}^{k\times n}.$$

It is easy to see that for $k = d$, the matrix $\mathbf{S}_\pi\mathbf{X}$ is square and thus

$$\det(\mathbf{X}^\top \mathbf{S}_\pi^\top \mathbf{S}_\pi \mathbf{X}) = \det(\mathbf{S}_\pi\mathbf{X})^2 = \frac{\det(\mathbf{X}_\pi)^2}{d^d \prod_i q_{\pi_i}},$$

where $\mathbf{X}_\pi$ selects the rows indexed by π from $\mathbf{X}$. So the distribution $\mathrm{VS}_q^d(\mathbf{X})$ is the same for every q. For this reason, we will write it simply as $\mathrm{VS}^d(\mathbf{X})$. We mention the following recently shown results regarding rescaled volume sampling which we use later in the proofs.

Lemma 12 (Dereziński et al., 2018) *For any $\mathbf{X}$, q and k as in Definition 11, if $\pi \sim \mathrm{VS}_q^k(\mathbf{X})$, then*

$$\mathbb{E}\big[(\mathbf{S}_\pi\mathbf{X})^\dagger \mathbf{S}_\pi\big] = \mathbf{X}^\dagger, \tag{2}$$

$$\mathbb{E}\big[(\mathbf{X}^\top \mathbf{S}_\pi^\top \mathbf{S}_\pi \mathbf{X})^{-1}\big] \preceq \frac{k}{k-d+1}(\mathbf{X}^\top\mathbf{X})^{-1}. \tag{3}$$

The following random experimental design emerges as a natural candidate for proving Theorem 4:

$$S = \{\pi_1\} \cup \cdots \cup \{\pi_k\}, \quad \widehat{\mathbf{w}}(\mathbf{y}_S) = (\mathbf{S}_\pi\mathbf{X})^\dagger \mathbf{S}_\pi \mathbf{y}, \quad \text{where } \pi \sim \mathrm{VS}_q^k(\mathbf{X}). \tag{4}$$

Note that since sequence π may include repetitions whereas set S may not (it is not a multi-set), the function $\widehat{\mathbf{w}}(\mathbf{y}_S)$ has to depend not only on $\mathbf{y}_S$ but also on the multiplicities of each response in sequence π. Thus both set S and function $\widehat{\mathbf{w}}(\cdot)$ in this design are in fact randomized and satisfy Definition 3. Lemma 12 shows that this design is unbiased for any $\mathbf{y}$, leading to a restricted notion of minimax-optimality which provides an upper-bound on $R_k^*(\mathbf{X})$ (proof in Appendix A):

Lemma 13 *Let $\mathcal{V}_k(\mathbf{X})$ consist of all random experimental designs based on q-rescaled volume sampling as in (4), parameterized by distribution q, and let $\mathrm{Sp}(\mathbf{X})$ be the column span of $\mathbf{X}$. Then:*

$$R_k^*(\mathbf{X}) \leq \min_{(S,\widehat{\mathbf{w}})\in\mathcal{V}_k(\mathbf{X})} \max_{\mathbf{y}\in\mathbb{R}^n\setminus\mathrm{Sp}(\mathbf{X})} \frac{\mathbb{E}_{S,\widehat{\mathbf{w}}}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\|^2\big]}{\|\mathbf{X}\,\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X}) - \mathbf{y}\|^2}.$$

Even for the restricted minimax-optimality, finding the exact or even approximate optimum q^* is open. However, in Section 4 we bound the restricted minimax value by selecting a particular distribution q and utilizing the following decomposition of $q-$rescaled volume sampling in the analysis.

Lemma 14 (Dereziński et al., 2019) *For any* $\mathbf{X}$, q *and* k *as in Definition 11, let* $\pi \sim \mathrm{VS}^d(\mathbf{X})$ *and* $\tilde{\pi}_1, \dots, \tilde{\pi}_{k-d} \overset{\text{i.i.d.}}{\sim} q$. *Finally let* σ *be a permutation of* $(1, \dots, k)$ *drawn uniformly at random. Then:*

$$\sigma\big(\pi_1, \dots, \pi_d, \tilde{\pi}_1, \dots, \tilde{\pi}_{k-d}\big) \sim \mathrm{VS}_q^k(\mathbf{X}).$$

If distribution q is sufficiently close to the leverage score sampling distribution p^{lev}, then even the initial volume sample of size d can be *selected out of* an i.i.d. sample of size $O(d \log d)$ as shown in Algorithm 1. This algorithm is a new implementation of a classical method for sampling from a so-called elementary determinantal point process, due to Hough et al. (2006). To our knowledge, the best previously known runtime for this method was $O(nd^2)$ for each produced volume sample (see Li et al. (2016)), whereas the runtime of this implementation is $O(d^3 \log d)$ (in addition to a preprocessing step which involves computing distribution q and matrix $(\mathbf{X}^\top \mathbf{X})^{-1}$). Note that for some applications of volume sampling, such as determinantal point process sampling, one can often assume that $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$ (see Dereziński, 2019), in which case the preprocessing becomes much cheaper than $O(nd^2)$. We now prove Theorem 10 by establishing correctness and runtime of Algorithm 1.

Algorithm 1 (Bottom-up) volume sampling

input: $\mathbf{X} \in \mathbb{R}^{n \times d}$, q, $\mathbf{A}_1 = (\mathbf{X}^\top \mathbf{X})^{-1}$
output: $\pi \sim \mathrm{VS}^d(\mathbf{X})$
for $i = 1..d$
  repeat
    Sample $\pi_i \sim q$
    Sample $a \sim \mathrm{Bernoulli}\Big(\frac{\mathbf{x}_{\pi_i}^\top \mathbf{A}_i \mathbf{x}_{\pi_i}}{2d\, q_{\pi_i}}\Big)$
  until $a = 1$
  $\mathbf{A}_{i+1} \leftarrow \mathbf{A}_i - \frac{\mathbf{A}_i \mathbf{x}_{\pi_i} \mathbf{x}_{\pi_i}^\top \mathbf{A}_i}{\mathbf{x}_{\pi_i}^\top \mathbf{A}_i \mathbf{x}_{\pi_i}}$
end for
return $\pi_1, \dots, \pi_d$

Proof of Theorem 10 Since $\mathbf{A}_i \preceq (\mathbf{X}^\top \mathbf{X})^{-1}$, then $\mathbf{x}_{\pi_i}^\top \mathbf{A}_i \mathbf{x}_{\pi_i} / (2d\, q_{\pi_i}) \leq p^{\mathrm{lev}}_{\pi_i} / (2 q_{\pi_i}) \leq 1$ is a valid Bernoulli probability. We start with the proof of correctness, which is an adaptation of the one given by Hough et al. (2006). For any i, j define $\mathbf{u}_j^{(i)} = \mathbf{A}_i^{1/2} \mathbf{x}_j$. The marginal probability of sampling π_{i+1} conditioned on previous steps is proportional to $\|\mathbf{u}_{\pi_{i+1}}^{(i+1)}\|^2$, which can be written as:

$$\begin{aligned}
\|\mathbf{u}_{\pi_{i+1}}^{(i+1)}\|^2 &= \mathbf{x}_{\pi_{i+1}}^\top \mathbf{A}_{i+1} \mathbf{x}_{\pi_{i+1}} = \mathbf{x}_{\pi_{i+1}}^\top \mathbf{A}_i^{1/2} \left(\mathbf{I} - \frac{\mathbf{A}_i^{1/2} \mathbf{x}_{\pi_i} \mathbf{x}_{\pi_i}^\top \mathbf{A}_i^{1/2}}{\mathbf{x}_{\pi_i}^\top \mathbf{A}_i \mathbf{x}_{\pi_i}} \right) \mathbf{A}_i^{1/2} \mathbf{x}_{\pi_{i+1}} \\
&= \mathbf{u}_{\pi_{i+1}}^{(i)\top} \left(\mathbf{I} - \frac{\mathbf{u}_{\pi_i}^{(i)} \mathbf{u}_{\pi_i}^{(i)\top}}{\|\mathbf{u}_{\pi_i}^{(i)}\|^2} \right) \mathbf{u}_{\pi_{i+1}}^{(i)} = \big\|\mathbf{P}_i \mathbf{u}_{\pi_{i+1}}^{(i)}\big\|^2, \quad \text{where } \mathbf{P}_i = \mathbf{I} - \frac{\mathbf{u}_{\pi_i}^{(i)} \mathbf{u}_{\pi_i}^{(i)\top}}{\|\mathbf{u}_{\pi_i}^{(i)}\|^2}
\end{aligned}$$

is a projection onto the $(d-1)$-dimensional subspace of $\mathbb{R}^d$ orthogonal to $\mathbf{u}_{\pi_i}^{(i)}$. We conclude that vectors $\mathbf{u}_j^{(i)}$ are obtained from $\mathbf{u}_j^{(1)} = (\mathbf{X}^\top \mathbf{X})^{-1/2} \mathbf{x}_j$ by repeatedly *projecting away* the points that were already sampled. This means that since $\mathbf{U} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1/2}$ satisfies $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, we have:

$$\sum_{j=1}^n \|\mathbf{u}_j^{(i+1)}\|^2 = \mathrm{tr}\left(\sum_{j=1}^n \mathbf{u}_j^{(1)} \mathbf{u}_j^{(1)\top} \cdot \prod_{t=1}^i \mathbf{P}_t \right) = \mathrm{tr}\left(\mathbf{U}^\top \mathbf{U} \cdot \prod_{t=1}^i \mathbf{P}_t \right) = d - i.$$

We can now write the probability of sampling a sequence $\pi_1, \dots, \pi_d$ as:

$$\Pr(\pi) = \prod_{i=1}^d \frac{\|\mathbf{u}_{\pi_i}^{(i)}\|^2}{d-i+1} = \frac{\det(\mathbf{U}_\pi)^2}{d!} = \frac{\det(\mathbf{X}_\pi)^2}{d! \det(\mathbf{X}^\top \mathbf{X})},$$

which follows because $\det(\mathbf{U}_\pi)^2$ is the squared volume spanned by the vectors $\mathbf{u}_{\pi_1}^{(1)}, \ldots, \mathbf{u}_{\pi_d}^{(1)}$ and it is obtained as a series of applications of the "base $\times$ height" formula. To bound the runtime, we note that the expected acceptance probability in the ith step of Algorithm 1 is:

$$\sum_{j=1}^{n} q_j \cdot \frac{\mathbf{x}_j^\top \mathbf{A}_i \mathbf{x}_j}{2d\, q_j} = \frac{1}{2d} \sum_{j=1}^{n} \|\mathbf{u}_j^{(i)}\|^2 = \frac{d-i+1}{2d}.$$

Thus, the expected total number of trials of rejection sampling throughout the algorithm is:

$$\sum_{i=1}^{d} \frac{2d}{d-i+1} = 2d \sum_{i=1}^{d} \frac{1}{i} \leq 2d\big(\ln(d)+1\big).$$

Standard tail bounds for a sum of geometric random variables show that with probability at least $1-\delta$ the number of rejection sampling trials is $O(d \log d \log \frac{1}{\delta})$. Each trial costs $O(d^2)$, as does updating the matrix $\mathbf{A}_i$, which concludes the proof. ■

4. Proof of Theorem 4

In this section we use $\widehat{\mathbf{w}}$ and $\mathbf{w}_{\mathrm{LS}}$ as shorthands for $\widehat{\mathbf{w}}(\mathbf{y}_S)$ and $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})$. To prove the error bound in Theorem 4 we will invoke Lemma 13, thereby restricting ourselves to a fixed response vector $\mathbf{y} \in \mathbb{R}^n$, in which case $\mathbf{w}^* = \mathbf{w}_{\mathrm{LS}}$, and a volume sampled random design as discussed in the previous section. The construction in our proof uses leverage scores and inverse scores, as discussed in Subsection 1.3.

Definition 15 *Given full rank matrix $\mathbf{X}$, its ith* ***leverage score*** *is defined as $l_i(\mathbf{X}) \overset{def}{=} \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$, and its ith* ***inverse score*** *as $v_i(\mathbf{X}) \overset{def}{=} \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-2} \mathbf{x}_i$.*

The key challenge in obtaining the result is that standard techniques developed for i.i.d. sampling (see, e.g., Drineas et al., 2006) only show the least squares error bounds with constant probability. Such bounds do not suffice to show an expected bound because we do not have control over what happens in the *failure event* (where the expectation may be unbounded). In fact, an expected bound of this type is not possible for any i.i.d. sampling (see Proposition 11 in Dereziński and Warmuth, 2018). Our key contribution is to define an event A s.t.:

1. if A occurs, then we can show a strong expected bound relying on i.i.d. sampling techniques,
2. if A fails to occur, a weaker bound still holds because of the jointness of volume sampling.

Crucially, the probability of failure will be exponentially small, thus allowing us to obtain the desired result. This technique is described in the proof of the following key lemma.

Lemma 16 *There is $C > 0$ s.t. for any full rank matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, if $\pi \sim \mathrm{VS}^k_{q(\alpha)}(\mathbf{X})$ where*

$$q(\alpha) = \alpha\left(0.5 \cdot p^{\mathrm{uni}} + 0.5 \cdot p^{\mathrm{inv}}\right) + (1-\alpha)\, p^{\mathrm{lev}} \quad \text{for} \quad \alpha \in \big[0.5, 0.75\big],$$
$$\text{with} \quad p_i^{\mathrm{uni}} = 1/n, \quad p_i^{\mathrm{inv}} = v_i(\mathbf{X})/\phi, \quad p_i^{\mathrm{lev}} = l_i(\mathbf{X})/d, \quad \text{and} \quad \phi = \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big),$$

then for any $k \geq d + C \max\{d \log n, \phi/\epsilon\}$ and an arbitrary vector $\boldsymbol{\xi} \in \mathbb{R}^n$ we have

$$\mathbb{E}\big[\|(\mathbf{S}_\pi \mathbf{X})^\dagger \mathbf{S}_\pi \boldsymbol{\xi}\|^2\big] \leq \frac{\epsilon}{8}\|\boldsymbol{\xi}\|^2 + 4\left\|\mathbf{X}^\dagger \mathbb{E}[\mathbf{S}_\pi^\top \mathbf{S}_\pi]\boldsymbol{\xi}\right\|^2.$$

Proof Observe that we chose $q(\alpha)$ as a mixture of three distributions in such a way that each of them has at least 0.25 weight in the mixture. Lemma 14 allows us to decompose sample π into the volume part, $\pi_{[d]} = (\pi_1, \dots, \pi_d) \sim \mathrm{VS}^d(\mathbf{X})$, and the i.i.d. part, $\tilde{\pi} = (\pi_{d+1}, \dots, \pi_k) \sim q$ (technically, this requires reordering the sequence π). We now define an event A as a variant of the so-called *subspace embedding* condition:

$$\text{event } A \text{ holds iff} \quad \frac{1}{k}\sum_{i=d+1}^{k} \frac{1}{q_{\pi_i}} \mathbf{x}_{\pi_i}\mathbf{x}_{\pi_i}^\top \succeq \frac{1}{2}\mathbf{X}^\top\mathbf{X}.$$

Note that A is defined only over the i.i.d. samples $\tilde{\pi}$ and is therefore completely independent of the volume sample $\pi_{[d]}$. We start by decomposing the expectation into two terms:

$$\mathbb{E}\big[\|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\boldsymbol{\xi}\|^2\big] = \Pr(A)\,\mathbb{E}\big[\|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\boldsymbol{\xi}\|^2 \mid A\big] + \Pr(\neg A)\,\mathbb{E}\big[\|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\boldsymbol{\xi}\|^2 \mid \neg A\big]. \quad (5)$$

To bound the first term we decompose the squared norm into two factors

$$\begin{aligned}\|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\boldsymbol{\xi}\|^2 &= \big\|(\mathbf{X}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\boldsymbol{\xi}\big\|^2 \\ &\le \big\|(\mathbf{X}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\big\|^2 \cdot \big\|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\boldsymbol{\xi}\big\|^2.\end{aligned}$$

When event A occurs, then the first factor can be easily bounded by 4, because

$$(\mathbf{X}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} \preceq \left(\frac{1}{k}\sum_{i=d+1}^{k}\frac{1}{q_{\pi_i}}\mathbf{x}_{\pi_i}\mathbf{x}_{\pi_i}^\top\right)^{-1}\mathbf{X}^\top\mathbf{X} \preceq 2\,\mathbf{I}.$$

Thus, it remains to bound the second factor in expectation, i.e. $\mathbb{E}\big[\|\mathbf{X}^\dagger\mathbf{S}_\pi^\top\mathbf{S}_\pi\boldsymbol{\xi}\|^2 \,|A\big]$. For this, we need an extension of a result by Dereziński et al. (2018) (see proof in Appendix B).

Lemma 17 *Given $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$ and $\beta > 0$, let $\pi \sim \mathrm{VS}_q^k(\mathbf{A})$ where $q_i \ge \beta \max\big\{\frac{\|\mathbf{a}_i\|^2}{\|\mathbf{A}\|_F^2}, \frac{l_i(\mathbf{A})}{d}\big\}$ for all $i \in [n]$ and $k \ge d$. Then $\mathbf{v}_\pi = \mathbf{A}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\mathbf{b}$ satisfies*

$$\mathbb{E}\big[\|\mathbf{v}_\pi - \mathbb{E}[\mathbf{v}_\pi]\|^2\big] \le \frac{1}{\beta^2 k}\|\mathbf{A}\|_F^2\|\mathbf{b}\|^2.$$

We use Lemma 17 with $\mathbf{A} = \mathbf{X}^{\dagger\top}$ and $\mathbf{b} = \boldsymbol{\xi}$. In this case $\|\mathbf{a}_i\|^2 = v_i(\mathbf{X})$ and $l_i(\mathbf{A}) = l_i(\mathbf{X})$. Also let $\beta = 0.25$ and observe that $\phi = \|\mathbf{X}^\dagger\|_F^2$. Now for $k \ge C\cdot\phi/\epsilon$ and $C \ge 16\cdot 4/\beta^2$ we have:

$$\begin{aligned}\Pr(A)\,\mathbb{E}\big[\|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\boldsymbol{\xi}\|^2 \mid A\big] &\le 4\Pr(A)\,\mathbb{E}\Big[\big\|\mathbf{X}^\dagger\mathbf{S}_\pi^\top\mathbf{S}_\pi\boldsymbol{\xi}\big\|^2 \mid A\Big] \\ &\le 4\,\mathbb{E}\big[\|\mathbf{v}_\pi\|^2\big] \;=\; 4\,\mathbb{E}\Big[\big\|\mathbf{v}_\pi - \mathbb{E}[\mathbf{v}_\pi]\big\|^2\Big] + 4\,\big\|\mathbb{E}[\mathbf{v}_\pi]\big\|^2 \\ &\le \frac{4\epsilon}{C\beta^2\phi}\|\mathbf{X}^\dagger\|_F^2\|\boldsymbol{\xi}\|^2 + 4\,\big\|\mathbf{X}^\dagger\mathbb{E}[\mathbf{S}_\pi^\top\mathbf{S}_\pi]\boldsymbol{\xi}\big\|^2 \\ &\le \frac{\epsilon}{16}\|\boldsymbol{\xi}\|^2 + 4\,\big\|\mathbf{X}^\dagger\mathbb{E}[\mathbf{S}_\pi^\top\mathbf{S}_\pi]\boldsymbol{\xi}\big\|^2. \quad (6)\end{aligned}$$

Next, we bound the second term in (5) by using a different decomposition of $\|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\boldsymbol{\xi}\|^2$. We will use the fact that since $q_i(\alpha) \ge \frac{1}{4}p_i^{\mathrm{uni}} = \frac{1}{4n}$ for all $i \in [n]$, then $\mathbf{S}_\pi^\top\mathbf{S}_\pi \preceq 4n\,\mathbf{I}$ for all $\pi \in [n]^k$. It is only here that we use the p^{uni} term in $q(\alpha)$. It follows that

$$\begin{aligned}\big\|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\big\|^2 \le \|(\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\|_F^2 &\le \|(\mathbf{S}_\pi\mathbf{X})^\dagger\|_F^2\|\mathbf{S}_\pi\|^2 \le \|(\mathbf{S}_\pi\mathbf{X})^\dagger\|_F^2 4n \\ &= 4n\,\mathrm{tr}\big((\mathbf{X}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\mathbf{X})^{-1}\big) \le 4n\,\mathrm{tr}\big((\mathbf{X}^\top\mathbf{S}_\pi^\top\mathbf{I}_{[d]}\mathbf{S}_\pi\mathbf{X})^{-1}\big),\end{aligned}$$

where $\mathbf{I}_{[d]} = \sum_{i=1}^{d} \mathbf{e}_i \mathbf{e}_i^\top$ selects the first d rows from $\mathbf{S}_\pi$. Since $\pi_{[d]} \sim \mathrm{VS}^d(\mathbf{X})$ and it is independent of the event A, from inequality (3) in Lemma 12 it follows that:

$$\mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^\top \mathbf{S}_\pi^\top \mathbf{I}_{[d]} \mathbf{S}_\pi \mathbf{X})^{-1}\big) \,|\, \neg A\big] = \frac{k}{d} \cdot \mathrm{tr}\big(\mathbb{E}\big[(\mathbf{X}^\top \mathbf{S}_{\pi_{[d]}}^\top \mathbf{S}_{\pi_{[d]}} \mathbf{X})^{-1}\big]\big) \leq \frac{k}{d} \cdot d\,\mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big) = k\phi.$$

Here the k/d term arises from the different multipliers used for $\mathbf{S}_\pi$ and $\mathbf{S}_{\pi_{[d]}}$. Thus, we obtain that $\mathbb{E}\big[\|(\mathbf{S}_\pi \mathbf{X})^\dagger \mathbf{S}_\pi \boldsymbol{\xi}\|^2 \,|\, \neg A\big] \leq 4nk\phi\, \|\boldsymbol{\xi}\|^2$. It remains to show that $\Pr(\neg A)$ is sufficiently small to obtain the desired bound. For this, we refer to a standard matrix concentration result for obtaining subspace embeddings, which follows from Tropp (2012).

Lemma 18 *Given a full rank $\mathbf{X} \in \mathbb{R}^{n\times d}$, if distribution q is such that $q_i \geq \beta \frac{l_i(\mathbf{X})}{d}$ for all $i \in [n]$, then an i.i.d. sample $\pi_1, \ldots, \pi_s \sim q$ for $s \geq C' \frac{d}{\beta} \log \frac{d}{\delta}$ with probability at least $1-\delta$ satisfies*

$$\frac{1}{2}\mathbf{X}^\top \mathbf{X} \preceq \frac{1}{s}\sum_{i=1}^{s} \frac{1}{q_{\pi_i}} \mathbf{x}_{\pi_i} \mathbf{x}_{\pi_i}^\top \preceq \frac{3}{2}\mathbf{X}^\top \mathbf{X}.$$

Setting $\beta = \frac{1}{4}$, $\delta = \frac{1}{6nk^2}$, $s = k - d$ and $C \geq 16C'$, we conclude that since $k - d \geq 16C' d \log n \geq C' \frac{d}{\beta} \log \frac{d}{\delta}$, we have $\Pr(\neg A)\mathbb{E}\big[\|(\mathbf{S}_\pi \mathbf{X})^\dagger \mathbf{S}_\pi \boldsymbol{\xi}\|^2 \,|\, \neg A\big] \leq \frac{\phi}{k}\|\boldsymbol{\xi}\|^2 \leq \frac{\epsilon}{16}\|\boldsymbol{\xi}\|^2$. Combining this with (5) and (6), we complete the proof of the bound of Lemma 16. ■

The last term in the bound of Lemma 16 can be controlled when the vector $\boldsymbol{\xi}$ is orthogonal to the columns of $\mathbf{X}$. The following bound is shown in Appendix B.

Lemma 19 *For $\mathbf{X}$ and $\boldsymbol{\xi}$ s.t. $\mathbf{X}^\top \boldsymbol{\xi} = \mathbf{0}$, if $\pi \sim \mathrm{VS}^k_{q(\alpha)}(\mathbf{X})$ (as in Lemma 16) with $\alpha = 0.5$, then*

$$\big\|\mathbf{X}^\dagger \mathbb{E}[\mathbf{S}_\pi^\top \mathbf{S}_\pi]\boldsymbol{\xi}\big\|^2 \leq \frac{\epsilon}{8}\|\boldsymbol{\xi}\|^2.$$

We put the two lemmas together to complete the proof of our main result.

Proof of Theorem 4 Setting $\widehat{\mathbf{w}} = (\mathbf{S}_\pi \mathbf{X})^\dagger \mathbf{S}_\pi \mathbf{y}$ for a fixed vector $\mathbf{y} \in \mathbb{R}^n$ with π as in Lemma 16 and $\alpha = 0.5$, using Lemmas 16 and 19 with $\boldsymbol{\xi} = \mathbf{y} - \mathbf{X}\mathbf{w}_{\mathrm{LS}}$ we obtain that:

$$\mathbb{E}\big[\|\widehat{\mathbf{w}} - \mathbf{w}_{\mathrm{LS}}\|^2\big] \overset{(*)}{=} \mathbb{E}\big[\|(\mathbf{S}_\pi \mathbf{X})^\dagger \mathbf{S}_\pi (\mathbf{y} - \mathbf{X}\mathbf{w}_{\mathrm{LS}})\|^2\big] \leq \frac{\epsilon}{8}\|\boldsymbol{\xi}\|^2 + 4 \cdot \frac{\epsilon}{8}\|\boldsymbol{\xi}\|^2 \leq \epsilon \cdot \|\boldsymbol{\xi}\|^2,$$

where $(*)$ follows because volume sampling ensures that $\mathrm{rank}(\mathbf{S}_\pi \mathbf{X}) = d$, so $(\mathbf{S}_\pi \mathbf{X})^\dagger \mathbf{S}_\pi \mathbf{X} = \mathbf{I}$. ■

In Appendix C we prove Theorem 5 as a reduction from Theorem 4 by transforming the matrix $\mathbf{X}$ into matrix $\mathbf{U} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1/2}$. This transformation makes MSE equal to MSPE, while preserving most key properties of least squares estimators. The random experimental design obtained via this reduction is different than the one used to bound the mean squared error. While the leverage scores are preserved during the transformation from $\mathbf{X}$ to $\mathbf{U}$, the inverse scores change, and in fact they become equal to the leverage scores, i.e., $v_i(\mathbf{U}) = l_i(\mathbf{U})$, so distribution q is somewhat simpler in this case. However, it can be shown that the exact experimental design used for Theorem 4 also satisfies the guarantee from Theorem 5, albeit with slightly different constants.

The logarithmic dependence on n in the sample size k for Theorems 4 and 5 comes from our analysis of the expected error in the tail of the sampling distribution. It is possible that the dependence on n can be eliminated altogether, even when using the same distribution. We leave this as an open question for future work.

Acknowledgments

MWM would like to acknowledge ARO, DARPA, NSF and ONR for providing partial support of this work. Also, MWM and MD thank the NSF for funding via the NSF TRIPODS program. Part of this work was done while MD, KLC and MWM were visiting the Simons Institute for the Theory of Computing and while MKW was at UC Santa Cruz, supported by NSF grant IIS-1619271.

References

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 126–135, Sydney, Australia, August 2017. URL http://proceedings.mlr.press/v70/allen-zhu17e.html.

Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE Trans. Information Theory*, 59(10):6880–6892, 2013.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statist. Sci.*, 10 (3):273–304, 08 1995. doi: 10.1214/ss/1177009939. URL https://doi.org/10.1214/ss/1177009939.

Xue Chen and Eric Price. Active regression via linear-sample sparsification. In *Proceedings of the 32nd Conference on Learning Theory*, 2019.

Michał Dereziński. Fast determinantal point processes via distortion-free intermediate sampling. In *Proceedings of the 32nd Conference on Learning Theory*, 2019.

Michał Dereziński and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems 30*, pages 3087–3096, Long Beach, CA, USA, December 2017.

Michał Dereziński and Manfred K. Warmuth. Subsampling for ridge regression via regularized volume sampling. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 716–725, Playa Blanca, Lanzarote, Canary Islands, April 2018.

Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *Journal of Machine Learning Research*, 19(23):1–39, 2018. URL http://jmlr.org/papers/v19/17-781.html.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Leveraged volume sampling for linear regression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2510–2519. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7517-leveraged-volume-sampling-for-linear-regression.pdf.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Correcting the bias in least squares regression with volume-rescaled sampling. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Petros Drineas and Michael W. Mahoney. RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59:80–90, 2016.

Petros Drineas and Michael W. Mahoney. Lectures on randomized numerical linear algebra. Technical report, 2017. Preprint: arXiv:1712.08880; To appear in: *Lectures of the 2016 PCMI Summer School on Mathematics of Data*.

Petros Drineas, Michael W Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13: 3475–3506, 2012.

Valerii V. Fedorov. *Theory of optimal experiments*. Probability and mathematical statistics. Academic Press, New York, NY, USA, 1972.

Robert M. Gray and Lee D. Davisson. *An Introduction to Statistical Signal Processing*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. ISBN 0521131820, 9780521131827.

J. Ben Hough, Manjunath Krishnapur, Yuval Peres, Bálint Virág, et al. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Efficient sampling for k-determinantal point processes. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1328–1337, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/li16f.html.

Ping Ma, Michael Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 91–99, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/ma14.html.

Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for a -optimal design. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1369–1386, January 2019.

Beiyan Ou and Julie Zhou. Minimax robust designs for field experiments. *Metrika*, 69(1):45–54, Jan 2009.

Friedrich Pukelsheim. *Optimal Design of Experiments (Classics in Applied Mathematics) (Classics in Applied Mathematics, 50)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006. ISBN 0898716047.

Garvesh Raskutti and Michael Mahoney. Statistical and algorithmic perspectives on randomized sketching for ordinary least-squares. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 617–625, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/raskutti15.html.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2012.

Yining Wang, Adams W. Yu, and Aarti Singh. On computationally tractable selection of experiments in measurement-constrained regression models. *J. Mach. Learn. Res.*, 18(1):5238–5278, January 2017. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=3122009.3208024.

Douglas P. Wiens and Pengfei Li. V-optimal designs for heteroscedastic regression. *Journal of Statistical Planning and Inference*, 145:125 – 138, 2014. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2013.09.007. URL http://www.sciencedirect.com/science/article/pii/S0378375813002310.

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

Appendix A. Basic properties of the minimax-optimal experimental design

We start by formally showing that the least squares estimator $\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X}) = \mathbf{X}^{\dagger}\mathbf{y}$ is the minimum variance unbiased estimator (MVUE) for $\mathbf{w}^* = \mathbf{X}^{\dagger}\mathbb{E}[\mathbf{y}]$ w.r.t. the family $\mathcal{F}$ consisting of all random response vectors $\mathbf{y}$ with finite second moment.

Proof of Proposition 7 Since all fixed vectors $\mathbf{y} \in \mathbb{R}^n$ belong to $\mathcal{F}$, it follows that $\mathbb{E}\big[\widehat{\mathbf{w}}(\mathbf{y}) \,|\, \mathbf{y}\big] = \mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})$ and using shorthands $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}(\mathbf{y})$, $\mathbf{w}_{\mathrm{LS}} = \mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})$ and $\mathbf{w}^* = \mathbf{X}^{\dagger}\mathbb{E}[\mathbf{y}]$ we have

$$\begin{aligned}
\mathrm{Var}\big[\widehat{\mathbf{w}}\big] &= \mathbb{E}\big[\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\top}\big] - \mathbf{w}^*\mathbf{w}^{*\top} \\
&= \mathbb{E}_{\mathbf{y}}\Big[\mathbb{E}_{\widehat{\mathbf{w}}}\big[\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\top} - \mathbf{w}_{\mathrm{LS}}\mathbf{w}_{\mathrm{LS}}^{\top} \,|\, \mathbf{y}\big] + \mathbf{w}_{\mathrm{LS}}\mathbf{w}_{\mathrm{LS}}^{\top}\Big] - \mathbf{w}^*\mathbf{w}^{*\top} \\
&= \mathbb{E}\big[(\widehat{\mathbf{w}} - \mathbf{w}_{\mathrm{LS}})(\widehat{\mathbf{w}} - \mathbf{w}_{\mathrm{LS}})^{\top}\big] + \mathbb{E}\big[\mathbf{w}_{\mathrm{LS}}\mathbf{w}_{\mathrm{LS}}^{\top}\big] - \mathbf{w}^*\mathbf{w}^{*\top} \succeq \mathrm{Var}\big[\mathbf{w}_{\mathrm{LS}}\big],
\end{aligned}$$

because the first term is a positive semi-definite matrix. ■ In the next lemma, we observe that it suffices to consider fixed vectors $\mathbf{y} \in \mathbb{R}^n$ to bound $R_k^*(\mathbf{X})$.

Lemma 20 *Given a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, suppose that a random design $(S, \widehat{\mathbf{w}})$ for all fixed response vectors $\mathbf{y} \in \mathbb{R}^n$ satisfies*

$$\mathbb{E}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \mathbf{w}_{\mathrm{LS}} \quad \textit{and} \quad \mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}_{\mathrm{LS}}\|^2\big] \leq \epsilon \cdot \|\mathbf{X}\mathbf{w}_{\mathrm{LS}} - \mathbf{y}\|^2,$$

where $\mathbf{w}_{\mathrm{LS}} = \mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})$. Then, for all random response vectors $\mathbf{y} \in \mathcal{F}$, we have

$$\mathbb{E}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] = \mathbf{w}^* \quad \textit{and} \quad \mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2\big] \leq \mathbb{E}\big[\|\mathbf{w}_{\mathrm{LS}} - \mathbf{w}^*\|^2\big] + \epsilon \cdot \mathbb{E}\big[\|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2\big].$$

Proof We first decompose the mean squared error using the unbiasedness of $\widehat{\mathbf{w}}(\mathbf{y}_S)$ as follows:

$$\begin{aligned}
\mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}^*\|^2\big] &= \mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}_{\mathrm{LS}}\|^2\big] + \mathbb{E}\big[\|\mathbf{w}_{\mathrm{LS}} - \mathbf{w}^*\|^2\big] \\
&\leq \epsilon \cdot \mathbb{E}\big[\|\mathbf{X}\mathbf{w}_{\mathrm{LS}} - \mathbf{y}\|^2\big] + \mathbb{E}\big[\|\mathbf{w}_{\mathrm{LS}} - \mathbf{w}^*\|^2\big].
\end{aligned}$$

Also note that $\mathbb{E}[\widehat{\mathbf{w}}(\mathbf{y}_S)] = \mathbb{E}\big[\mathbb{E}[\widehat{\mathbf{w}}(\mathbf{y}_S) \,|\, \mathbf{y}]\big] = \mathbb{E}[\mathbf{w}_{\mathrm{LS}}] = \mathbf{w}^*$. It remains to bound $\mathbb{E}\big[\|\mathbf{X}\mathbf{w}_{\mathrm{LS}} - \mathbf{y}\|^2\big]$. Note that $\mathbf{w}_{\mathrm{LS}} = \mathrm{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, so that in particular $\|\mathbf{X}\mathbf{w}_{\mathrm{LS}} - \mathbf{y}\|^2 \leq \|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2$, and therefore $\mathbb{E}\big[\|\mathbf{X}\mathbf{w}_{\mathrm{LS}} - \mathbf{y}\|^2\big] \leq \mathbb{E}\big[\|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2\big]$, which concludes the proof. ■

Proof of Lemma 13 Follows immediately from Lemma 20 and the fact that $\mathcal{V}_k(\mathbf{X}) \subseteq \mathcal{W}_k(\mathbf{X})$. ■ We next prove Proposition 9, showing that the minimax-optimal value $R_k^*(\mathbf{X})$ given in Definition 8 is well-defined for all $d \leq k \leq n$ and for most k it has matching upper and lower bounds of $\Theta(\phi/\epsilon)$.

Proof of Proposition 9 Since $\mathrm{MSE}[\widehat{\mathbf{w}}] = \mathrm{tr}\big(\mathrm{Var}[\widehat{\mathbf{w}}]\big)$ for any unbiased estimator, Proposition 7 immediately implies that $R_k^*(\mathbf{X}) \geq 0$. Next, let $\pi \sim \mathrm{VS}_q^k(\mathbf{X})$ with q chosen as in the proof of Theorem 4 and consider the volume sampled estimator defined as in (4), i.e. $\widehat{\mathbf{w}} = (\mathbf{S}_\pi\mathbf{X})^{\dagger}\mathbf{S}_\pi\mathbf{y}$. First, we can use Lemma 13 to assume w.l.o.g. that $\mathbf{y}$ is a fixed vector in $\mathbb{R}^n$ so that $\mathbf{w}^* = \mathbf{X}^{\dagger}\mathbf{y}$. Then, as in the proof of Theorem 4 we use equation (3) in Lemma 12:

$$\begin{aligned}
\mathbb{E}\big[\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2\big] &\leq 4n\|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2\,\mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^{\top}\mathbf{S}_\pi^{\top}\mathbf{S}_\pi\mathbf{X})^{-1}\big)\big] \\
&\leq 4n\|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2\frac{k}{k-d+1}\mathrm{tr}\big((\mathbf{X}^{\top}\mathbf{X})^{-1}\big).
\end{aligned}$$

This implies that $R_k^*(\mathbf{X}) \leq 4n\frac{k}{k-d+1}\,\phi < \infty$, where unlike in the proof of Theorem 4 we did not need to assume any lower bound on k other than that $k \geq d$.

Part 1. Theorem 4 implies that there is a constant $C > 0$ such that for any $\mathbf{X}$, $\epsilon > 0$ and $k \geq C \cdot (d \log n + \phi/\epsilon)$ there is a random experimental design $(S, \widehat{\mathbf{w}})$ of size at most k which demonstrates an upper bound on the minimax-optimal value, i.e., that $R_k^*(\mathbf{X}) \leq \epsilon$. Now, suppose that $k \geq 2C \cdot d \log n$ and let $\epsilon = 2C \cdot \phi/k$. Then, $k \geq 2C \cdot \max\{d \log n,\ \phi/\epsilon\} \geq C \cdot (d \log n + \phi/\epsilon)$, which means that for any $\mathbf{X}$ we have $R_k^*(\mathbf{X}) \leq \epsilon = 2C \cdot \phi/k$.

Part 2. This result is based on the following lower bound for classical A-optimal design.

Theorem 21 (Avron and Boutsidis, 2013, Theorem 4.5) *For any $\alpha > 0$, n, d such that $n > 2d$ and $\mathrm{mod}(n, d) = 0$ there is a full rank $n \times d$ matrix $\mathbf{X}$ such that for any subset $S \subseteq [n]$ with $\mathrm{rank}(\mathbf{X}_S) = d$ and $|S| = k$, we have*

$$\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\big) \geq \left(\frac{n-k}{k+\alpha^2} + 1 - \frac{k}{d}\right) \cdot \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^{-1}\big).$$

We remark that $\mathbf{X}$ in the theorem depends on α. Note that the condition $\mathrm{mod}(n, d) = 0$ can be eliminated by padding matrix $\mathbf{X}$ with appropriate number of $\mathbf{0}$ rows and replacing n in the bound with $n - d$. Let $\mathbf{X}$ be the matrix from Theorem 21 (padded if necessary, with α chosen later) and let $\mathcal{F}_N(\mathbf{X})$ be the family of random response vectors $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}$ such that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{w}^* \in \mathbb{R}^d$. Also let $\mathcal{F}_B(\mathbf{X})$ be the Bayesian counterpart, where $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi}$ for independent Gaussian random vectors $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathbf{w}^2 \mathbf{I})$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, with any $\sigma_\mathbf{w}^2 > 0$. Given any $\mathbf{y} \in \mathcal{F}_B(\mathbf{X})$ with prior variance $\sigma_\mathbf{w}^2$, the following bound is known for any (possibly biased) estimator $\widehat{\mathbf{w}}(\mathbf{y})$ of $\mathbf{w}$, called the minimum mean squared error bound (MMSE; see Gray and Davisson, 2010):

$$\mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}) - \mathbf{w}\|^2\big] \geq \mathrm{tr}\big((\mathbf{X}^\top \mathbf{X} + (1/\sigma_\mathbf{w}^2)\mathbf{I})^{-1}\big).$$

An analogous lower bound holds when applied to the same regression model restricted to a fixed subset $S_\circ \subseteq [n]$, i.e. such that $\mathbf{y}_{S_\circ} = \mathbf{X}_{S_\circ}\mathbf{w} + \boldsymbol{\xi}_{S_\circ}$, and any estimator $\widehat{\mathbf{w}}(\mathbf{y}_{S_\circ})$:

$$\mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_{S_\circ}) - \mathbf{w}\|^2\big] \geq \mathrm{tr}\big((\mathbf{X}_{S_\circ}^\top \mathbf{X}_{S_\circ} + (1/\sigma_\mathbf{w}^2)\mathbf{I})^{-1}\big). \tag{7}$$

We use this to give a lower bound for the MSE of any random experimental design $(S, \widehat{\mathbf{w}}) \in \mathcal{W}_k(\mathbf{X})$ following Definition 3 for the (non-Bayesian) response model $\mathcal{F}_N(\mathbf{X})$:

$$\begin{aligned}
\max_{\mathbf{y} \in \mathcal{F}_N(\mathbf{X})} \mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] &= \max_{\substack{\mathbf{y} \in \mathcal{F}_B(\mathbf{X}) \\ \mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi}}} \max_{\mathbf{w}^* \in \mathbb{R}^d} \mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}\|^2 \,|\, \mathbf{w} = \mathbf{w}^*\big] \\
&\overset{(a)}{\geq} \max_{\substack{\mathbf{y} \in \mathcal{F}_B(\mathbf{X}) \\ \mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi}}} \mathbb{E}_\mathbf{w}\Big[\mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}\|^2 \,|\, \mathbf{w}\big]\Big] \\
&= \max_{\substack{\mathbf{y} \in \mathcal{F}_B(\mathbf{X}) \\ \mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi}}} \mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}\|^2\big] \\
&= \max_{\substack{\mathbf{y} \in \mathcal{F}_B(\mathbf{X}) \\ \mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi}}} \mathbb{E}_S\Big[\mathbb{E}\big[\|\widehat{\mathbf{w}}(\mathbf{y}_S) - \mathbf{w}\|^2 \,|\, S\big]\Big] \\
&\overset{(b)}{\geq} \lim_{\sigma_\mathbf{w} \to \infty} \mathbb{E}_S\Big[\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S + (1/\sigma_\mathbf{w}^2)\mathbf{I})^{-1}\big)\Big] \\
&\geq \min_{\substack{S : |S| \leq k \\ \mathrm{rank}(\mathbf{X}_S) = d}} \mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\big),
\end{aligned}$$

where (a) follows because expectation $\mathbb{E}_{\mathbf{w}}$ is always upper bounded by the maximum over all possible values of $\mathbf{w}$, and (b) follows from (7) applied to the conditional expectation for a fixed S and the fact that the trace is monotonically increasing with $\sigma_{\mathbf{w}}$. We now lower bound $R_k^*(\mathbf{X})$ by the minimax value based on the response family $\mathcal{F}_N(\mathbf{X})$:

$$\begin{aligned}
R_k^*(\mathbf{X}) &\geq \min_{(S,\widehat{\mathbf{w}})\in\mathcal{W}_k(\mathbf{X})} \max_{\mathbf{y}\in\mathcal{F}_N(\mathbf{X})} \frac{\mathrm{MSE}\big[\widehat{\mathbf{w}}(\mathbf{y}_S)\big] - \mathrm{MSE}\big[\mathbf{w}_{\mathrm{LS}}(\mathbf{y}|\mathbf{X})\big]}{\mathbb{E}_{\mathbf{y}}\big[\|\boldsymbol{\xi}_{\mathbf{y}|\mathbf{X}}\|^2\big]} \\
&\geq \min_{\substack{S:|S|\leq k \\ \mathrm{rank}(\mathbf{X}_S)=d}} \frac{1}{n}\Big(\mathrm{tr}\big((\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\big) - \mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^{-1}\big)\Big) \\
\text{(Theorem 21)} \quad &\geq \frac{1}{n}\left(\frac{n-d-k}{k+\alpha^2} - \frac{k}{d}\right)\cdot \mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^{-1}\big) \\
&\geq \left(\frac{k}{k+\alpha^2} - \frac{3k^2}{nd}\right)\cdot\frac{\phi}{k} \;\geq\; (1-\epsilon)\cdot\frac{\phi}{k},
\end{aligned}$$

because $k^2 < \epsilon nd/3$ and we can choose α small enough so that $\frac{\alpha^2}{k+\alpha^2} + \frac{3k^2}{nd} \leq \epsilon$. ■

Appendix B. Omitted proofs from Section 4

Proof of Lemma 17 Recall that $\mathbf{v}_\pi = \mathbf{A}^\top\mathbf{S}_\pi^\top\mathbf{S}_\pi\mathbf{b}$. We rewrite the expectation as follows:

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{v}_\pi - \mathbb{E}[\mathbf{v}_\pi]\|^2\big] &= \mathbf{b}^\top\mathbb{E}\Big[\big(\mathbf{S}_\pi^\top\mathbf{S}_\pi - \mathbb{E}[\mathbf{S}_\pi^\top\mathbf{S}_\pi]\big)\mathbf{A}\mathbf{A}^\top\big(\mathbf{S}_\pi^\top\mathbf{S}_\pi - \mathbb{E}[\mathbf{S}_\pi^\top\mathbf{S}_\pi]\big)\Big]\,\mathbf{b} \\
&\leq \Big\|\underbrace{\Big[\mathrm{cov}\big[\tfrac{s_i}{kq_i},\tfrac{s_j}{kq_j}\big]\mathbf{a}_i^\top\mathbf{a}_j\Big]_{n\times n}}_{\mathbf{M}}\Big\|\cdot\|\mathbf{b}\|^2,
\end{aligned}$$

where $s_i = |\{t : \pi_t = i\}|$. Note that $\mathbf{M}$ is the Hadamard product of two PSD matrices, and therefore also PSD by the Schur product theorem. Next, we use two formulas shown by Dereziński et al. (2018) for rescaled volume sampling:

$$\begin{aligned}
\mathbb{E}[s_i] &= (k-d)\,q_i + l_i, \\
\mathrm{cov}(s_i,s_j) &= \mathbf{1}_{i=j}\mathbb{E}[s_i] - (k-d)q_iq_j - l_{ij}^2,
\end{aligned} \tag{8}$$

where $l_i = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{a}_i$ is the ith leverage score of $\mathbf{A}$ and $l_{ij} = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{a}_j$ is the (i,j)th cross-leverage score. Using the above we get:

$$\mathbf{M} = \mathrm{diag}\left(\left[\tfrac{\|\mathbf{a}_i\|^2}{kq_i}\,\tfrac{(k-d)q_i+l_i}{kq_i}\right]_{n\times 1}\right) - \frac{k-d}{k^2}\mathbf{A}\mathbf{A}^\top - \left[\tfrac{l_{ij}^2}{k^2q_iq_j}\mathbf{a}_i^\top\mathbf{a}_j\right]_{n\times n}.$$

Note that the second term is a PSD matrix being subtracted from $\mathbf{M}$. Similarly, the last term is also subtracting a PSD matrix. To see this, note that the matrix formed by the cross-leverage scores is $[l_{ij}]_{n\times n} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$, so it is PSD. Therefore, we can write the last term as a Hadamard product of PSD matrices and apply Schur product theorem. Thus it remains to bound the diagonal term:

$$\frac{\|\mathbf{a}_i\|^2}{kq_i}\,\frac{(k-d)q_i+l_i}{kq_i} \leq \frac{\|\mathbf{A}\|_F^2}{\beta k}\left(\frac{k-d}{k} + \frac{d}{k\beta}\right) \leq \frac{\|\mathbf{A}\|_F^2}{\beta^2 k}.$$

Since the spectral norm of the diagonal term is bounded by $\frac{\|\mathbf{A}\|_F^2}{\beta^2 k}$, and subtracting the two PSD terms leaves $\mathbf{M}$ a PSD matrix, we have $\|\mathbf{M}\| \leq \frac{\|\mathbf{A}\|_F^2}{\beta^2 k}$, and the result follows. ■

Proof of Lemma 19 Using the marginal expectation formula (8) for volume sampling we have:

$$\mathbb{E}[\mathbf{S}_\pi^\top \mathbf{S}_\pi] = \operatorname{diag}\left(\left[\tfrac{(k-d)q_i + l_i(\mathbf{X})}{k q_i}\right]_{n\times 1}\right) = \operatorname{diag}\left(\left[\tfrac{q_i(\frac{k-d}{k}\alpha)}{q_i(\alpha)}\right]_{n\times 1}\right) \overset{def}{=} \mathbf{D}_\alpha,$$

where $q(\alpha) = \alpha\big(0.5\cdot p^{\text{uni}} + 0.5\cdot p^{\text{inv}}\big) + (1-\alpha)p^{\text{lev}}$. Now, let $\beta = \alpha\cdot\frac{k}{k-d}$ and $\pi' \sim \mathrm{VS}^k_{q(\beta)}(\mathbf{X})$. Using expectation formula (2) of Lemma 12 and the fact that $\mathbf{X}^\dagger\boldsymbol{\xi} = \mathbf{0}$, we have:

$$\begin{aligned}
\mathbb{E}\big[(\mathbf{S}_{\pi'}\mathbf{X})^\dagger \mathbf{S}_{\pi'}\,\mathbf{D}_\beta^{-1}\boldsymbol{\xi}\big] &= \mathbf{X}^\dagger\mathbf{D}_\beta^{-1}\boldsymbol{\xi}\\
&= \mathbf{X}^\dagger(\mathbf{D}_\beta^{-1} - \mathbf{I})\boldsymbol{\xi}\\
&= \mathbf{X}^\dagger \operatorname{diag}\left(\frac{q_i(\beta)-q_i(\alpha)}{q_i(\alpha)}\right)\boldsymbol{\xi}\\
&= -\frac{k}{k-d}\mathbf{X}^\dagger(\mathbf{D}_\alpha - \mathbf{I})\boldsymbol{\xi}\\
&= -\frac{k}{k-d}\mathbf{X}^\dagger\mathbf{D}_\alpha\boldsymbol{\xi}.
\end{aligned}$$

We can assume $k \geq 3d$, adjusting C of Lemma 16. With $\alpha = 0.5$, this implies $\beta \in [0.5, 0.75]$, so applying Lemma 16 to sequence π' and vector $\mathbf{D}_\beta^{-1}\boldsymbol{\xi}$ combined with Jensen's inequality we obtain:

$$\begin{aligned}
\|\mathbf{X}^\dagger\mathbf{D}_\alpha\boldsymbol{\xi}\|^2 &= \left(\frac{k-d}{k}\right)^2 \cdot \big\|\mathbb{E}[(\mathbf{S}_{\pi'}\mathbf{X})^\dagger\mathbf{S}_{\pi'}\mathbf{D}_\beta^{-1}\boldsymbol{\xi}]\big\|^2\\
\text{(Jensen's inequality)}\quad &\leq \left(\frac{k-d}{k}\right)^2 \cdot \mathbb{E}\big[\|(\mathbf{S}_{\pi'}\mathbf{X})^\dagger\mathbf{S}_{\pi'}\mathbf{D}_\beta^{-1}\boldsymbol{\xi}\|^2\big]\\
\text{(Lemma 16)}\quad &\leq \left(\frac{k-d}{k}\right)^2 \cdot \Big(\frac{\epsilon}{8}\|\mathbf{D}_\beta^{-1}\boldsymbol{\xi}\|^2 + 4\underbrace{\|\mathbf{X}^\dagger\mathbf{D}_\beta\mathbf{D}_\beta^{-1}\boldsymbol{\xi}\|^2}_{0}\Big)\\
&\leq \frac{\epsilon}{8}\cdot\|\boldsymbol{\xi}\|^2,
\end{aligned}$$

because $\|\mathbf{D}_\beta^{-1}\| \leq \frac{k}{k-d}$, which concludes the proof. ■

Appendix C. Proof of Theorem 5

The key idea in the proof is a standard transformation of the data matrix $\mathbf{X}$ which has the property that it preserves the predictions of the least squares estimator, while transforming the actual estimator in such a way that the mean squared error becomes equal to the mean squared prediction error. Specifically, consider matrix $\mathbf{U} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}$. This matrix has the property that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ so $\operatorname{tr}((\mathbf{U}^\top\mathbf{U})^{-1}) = d$ and $\mathbf{U}\mathbf{U}^\dagger = \mathbf{U}\mathbf{U}^\top = \mathbf{X}\mathbf{X}^\dagger$. Replacing all least squares estimators for this new matrix we have $\mathbf{v}^* = \mathbf{U}^\top\mathbb{E}[\mathbf{y}]$, $\mathbf{v}_{\text{LS}} = \mathbf{U}^\top\mathbf{y}$ and $\widehat{\mathbf{v}} = (\mathbf{S}_\pi\mathbf{U})^\dagger\mathbf{S}_\pi\mathbf{y}$. Note that we have $\mathbf{U}\mathbf{v}^* = \mathbf{X}\mathbf{X}^\dagger\mathbb{E}[\mathbf{y}] = \mathbf{X}\mathbf{w}^*$ and similarly $\mathbf{U}\mathbf{v}_{\text{LS}} = \mathbf{X}\mathbf{w}_{\text{LS}}$. A simple calculation also reveals that $\mathbf{U}\widehat{\mathbf{v}} = \mathbf{X}\widehat{\mathbf{w}}$ for $\widehat{\mathbf{w}} = (\mathbf{S}_\pi\mathbf{X})^\dagger\mathbf{S}_\pi\mathbf{y}$. Suppose that $\pi \sim \mathrm{VS}^k_q(\mathbf{U})$ is produced as in the proof of

Theorem 4 when applied to matrix $\mathbf{U}$. Then:

$$\begin{aligned}
\mathbb{E}\big[\|\mathbf{X}(\widehat{\mathbf{w}} - \mathbf{w}^*)\|^2\big] &= \mathbb{E}\big[\|\mathbf{U}(\widehat{\mathbf{v}} - \mathbf{v}^*)\|^2\big] \\
&= \mathbb{E}\big[\|\widehat{\mathbf{v}} - \mathbf{v}^*\|^2\big] \\
\text{(Theorem 4)} \quad &\leq \mathbb{E}\big[\|\mathbf{v}_{\mathrm{LS}} - \mathbf{v}^*\|^2\big] + \epsilon \cdot \mathbb{E}\big[\|\mathbf{U}\mathbf{v}^* - \mathbf{y}\|^2\big] \\
&= \mathbb{E}\big[\|\mathbf{X}(\mathbf{w}_{\mathrm{LS}} - \mathbf{w}^*)\|^2\big] + \epsilon \cdot \mathbb{E}\big[\|\mathbf{X}\mathbf{w}^* - \mathbf{y}\|^2\big].
\end{aligned}$$