Exact expressions for double descent and implicit regularization via surrogate random design

author names withheld

Editor: Under Review for COLT 2020

Abstract

Double descent refers to the phase transition that is exhibited by the generalization error of unregularized learning models when varying the ratio between the number of parameters and the number of training samples. The recent success of highly over-parameterized machine learning models such as deep neural networks has motivated a theoretical analysis of the double descent phenomenon in classical models such as linear regression which can also generalize well in the over-parameterized regime. We build on recent advances in Randomized Numerical Linear Algebra (RandNLA) to provide the first exact non-asymptotic expressions for double descent of the minimum norm linear estimator. Our approach involves constructing what we call a surrogate random design to replace the standard i.i.d. design of the training sample. This surrogate design admits exact expressions for the mean squared error of the estimator while preserving the key properties of the standard design. We also establish an exact implicit regularization result for over-parameterized training samples. In particular, we show that, for the surrogate design, the implicit bias of the unregularized minimum norm estimator precisely corresponds to solving a ridge-regularized least squares problem on the population distribution.

Keywords: Least squares, determinantal point process, double descent, implicit regularization

1. Introduction

Classical statistical learning theory asserts that to achieve generalization one must use training sample size that sufficiently exceeds the complexity of the learning model, where the latter is typically represented by the number of parameters (or some related structural parameter; see Friedman et al., 2001). In particular, this seems to suggest the conventional wisdom that one should not use models that fit the training data exactly. However, modern machine learning practice often seems to go against this intuition, using models with so many parameters that the training data can be perfectly interpolated, in which case the training error vanishes. It has been shown that models such as deep neural networks, as well as certain so-called interpolating kernels and decision trees, can generalize well in this regime. In particular, Belkin et al. (2019a) empirically demonstrated a phase transition in generalization performance of learning models which occurs at an *interpolation thershold*, i.e., a point where training error goes to zero (as one varies the ratio between the model complexity and the sample size). Moving away from this threshold in either direction tends to reduce the generalization error, leading to the so-called *double descent* curve.

To understand this surprising phenomenon, in perhaps the simplest possible setting, we study it in the context of linear or least squares regression. Consider a full rank $n \times d$ data matrix $\mathbf{X}$ and a vector $\mathbf{y}$ of responses corresponding to each of the n data points (the rows of $\mathbf{X}$), where we wish to find the best linear model $\mathbf{X}\mathbf{w} \approx \mathbf{y}$, parameterized by a d-dimensional vector $\mathbf{w}$. The simplest example of an estimator that has been shown to exhibit the double descent phenomenon (Belkin et al.,

2019c) is the Moore-Penrose estimator, $\widehat{\mathbf{w}} = \mathbf{X}^\dagger \mathbf{y}$: in the so-called over-determined regime, i.e., when $n > d$, it corresponds to the least squares solution, i.e., $\operatorname{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$; and in the under-determined regime (also known as over-parameterized or interpolating), i.e., when $n < d$, it corresponds to the minimum norm solution to the linear system $\mathbf{X}\mathbf{w} = \mathbf{y}$. Given the ubiquity of linear regression and the Moore-Penrose solution, e.g., in kernel-based machine learning, studying the performance of this estimator can shed some light on the effects of over-parameterization/interpolation in machine learning more generally. Of particular interest are results that are exact (i.e., not upper/lower bounds) and non-asymptotic (i.e., for large but still finite n and d).

We build on methods from Randomized Numerical Linear Algebra (RandNLA) in order to obtain *exact non-asymptotic expressions* for the mean squared error (MSE) of the Moore-Penrose estimator (see Theorem 1). This provides a precise characterization of the double descent phenomenon for perhaps the simplest and most ubiquitous regression problem. In obtaining these results, we are able to provide precise formulas for the *implicit regularization* induced by minimum norm solutions of under-determined training samples, relating it to classical ridge regularization (see Theorem 3). This result has been observed empirically for RandNLA methods (Mahoney, 2011), but it has also been shown in deep learning (Neyshabur, 2017) and machine learning (Mahoney, 2012) more generally. To obtain our precise results, we use a somewhat non-standard random design, which we term surrogate random design (see Section 3 for a detailed discussion), and which we expect will be of more general interest. Informally, the goal of a surrogate random design is to modify an original design to capture its main properties while being "nicer" in some useful way. In Theorem 4 and Appendix F we show, both theoretically and empirically, that our surrogate design accurately preserves the key properties of the original design when the data distribution is a multivariate Gaussian.

1.1. Main results: double descent and implicit regularization

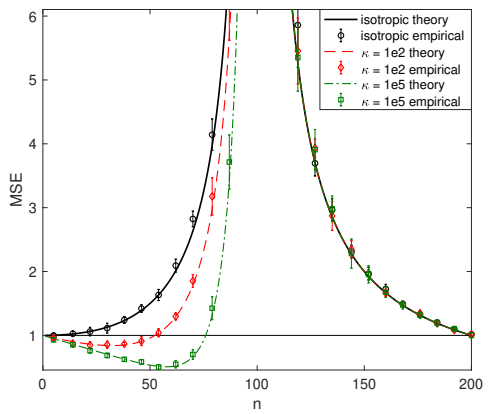
As the performance metric in our analysis, we use the *mean squared error* (MSE), defined as $\mathrm{MSE}[\widehat{\mathbf{w}}] = \mathbb{E}\big[\|\widehat{\mathbf{w}} - \mathbf{w}^*\|^2\big]$, where $\mathbf{w}^*$ is a fixed underlying linear model of the responses. In analyzing the MSE, we make the following standard assumption that the response noise is homoscedastic with variance σ^2.

Assumption 1 (Homoscedastic noise) *Responses are* $y(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}^* + \xi$ *where* $\xi \sim \mathcal{N}(0, \sigma^2)$.

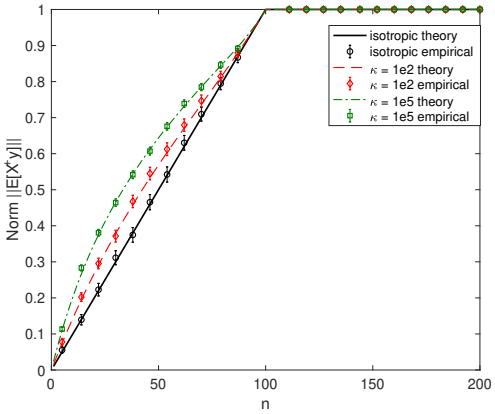
Our main result provides an exact expression for the MSE of the Moore-Penrose estimator under our surrogate design denoted $\bar{\mathbf{X}} \sim S_\mu^n$, where μ is the d-variate distribution of the row vector $\mathbf{x}^\top$ and n is the sample size (details in Section 3). This surrogate is used in place of the standard $n \times d$ random design $\mathbf{X} \sim \mu^n$, where n data points (the rows of $\mathbf{X}$) are sampled independently from μ. Unlike for the standard design, our MSE formula is fully expressible as a function of the covariance matrix $\mathbf{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$. To state our main result, we need an additional minor assumption on μ which is satisfied by most standard continuous distributions such as any multivariate Gaussian with positive definite covariance matrix.

Assumption 2 (General position) *For* $1 \leq n \leq d$, *if* $\mathbf{X} \sim \mu^n$, *then* $\operatorname{rank}(\mathbf{X}) = n$ *almost surely.*

Under Assumptions 1 and 2, we can establish our first main result, stated as the following theorem, where we use $\mathbf{X}^\dagger$ to denote the Moore-Penrose inverse of $\mathbf{X}$.



(*a*) Surrogate MSE expressions (Theorem 1) closely match numerical estimates even for non-isotropic features. Eigenvalue decay leads to a steeper descent curve in the underdetermined regime ($n < d$).



(*b*) The mean of the estimator $\mathbf{X}^\dagger\mathbf{y}$ exhibits shrinkage which closely matches the shrinkage of a ridge-regularized least squares optimum (theory lines), as characterized by Theorem 3.

Figure 1: Illustration of the main results for $d = 100$ and $\mu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is diagonal with eigenvalues decaying exponentially and scaled so that $\operatorname{tr}(\mathbf{\Sigma}^{-1}) = d$. We use our surrogate formulas to plot (a) the MSE (Theorem 1) and (b) the norm of the expectation (Theorem 3) of the Moore-Penrose estimator (*theory* lines), accompanied by the empirical estimates based on the standard i.i.d. design (error bars are three times the standard error of the mean). We consider three different condition numbers κ of $\mathbf{\Sigma}$, with *isotropic* corresponding to $\kappa = 1$, i.e., $\mathbf{\Sigma} = \mathbf{I}$. We use $\sigma^2 = 1$ and $\mathbf{w}^* = \frac{1}{\sqrt{d}}\mathbf{1}$.

Theorem 1 (Exact non-asymptotic MSE) *If the response noise is homoscedastic (Assumption 1) and μ is in general position (Assumption 2), then for $\bar{\mathbf{X}} \sim S_\mu^n$ (Definition 7) and $\bar{y}_i = y(\bar{\mathbf{x}}_i)$,*

$$\operatorname{MSE}\big[\bar{\mathbf{X}}^\dagger\bar{\mathbf{y}}\big] = \begin{cases} \sigma^2 \operatorname{tr}\big((\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big) \cdot \frac{1-\alpha_n}{d-n} \;+\; \frac{\mathbf{w}^{*\top}(\mathbf{\Sigma}_\mu+\lambda_n\mathbf{I})^{-1}\mathbf{w}^*}{\operatorname{tr}((\mathbf{\Sigma}_\mu+\lambda_n\mathbf{I})^{-1})} \cdot (d-n), & \text{for } n < d, \\ \sigma^2 \operatorname{tr}(\mathbf{\Sigma}_\mu^{-1}), & \text{for } n = d, \\ \sigma^2 \operatorname{tr}(\mathbf{\Sigma}_\mu^{-1}) \cdot \frac{1-\beta_n}{n-d}, & \text{for } n > d, \end{cases}$$

with $\lambda_n \geq 0$ defined by $n = \operatorname{tr}(\mathbf{\Sigma}_\mu(\mathbf{\Sigma}_\mu + \lambda_n\mathbf{I})^{-1})$, $\alpha_n = \det(\mathbf{\Sigma}_\mu(\mathbf{\Sigma}_\mu + \lambda_n\mathbf{I})^{-1})$ and $\beta_n = \mathrm{e}^{d-n}$.

Definition 2 *We will use $\mathcal{M}(\mathbf{\Sigma}_\mu, \mathbf{w}^*, \sigma^2, n)$ to denote the above expressions for $\operatorname{MSE}\big[\bar{\mathbf{X}}^\dagger\bar{\mathbf{y}}\big]$.*

Proof of Theorem 1 is given in Appendix C. For illustration, we plot the MSE expressions in Figure 1a, comparing them with empirical estimates of the true MSE under the i.i.d. design for a multivariate Gaussian distribution $\mu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with several different covariance matrices $\mathbf{\Sigma}$. We keep the number of features d fixed to 100 and vary the number of samples n, observing a double descent peak at $n = d$. We observe that our theory aligns well with the empirical estimates, whereas previously, no such theory was available except for special cases such as $\mathbf{\Sigma} = \mathbf{I}$ (more details in Theorem 4 and Appendices E and F). The plots show that varying the spectral decay of $\mathbf{\Sigma}$ has a

significant effect on the shape of the curve in the under-determined regime. We use the horizontal line to denote the MSE of the null estimator $\mathrm{MSE}[\mathbf{0}] = \|\mathbf{w}^*\|^2 = 1$. When the eigenvalues of $\boldsymbol{\Sigma}$ decay rapidly, then the Moore-Penrose estimator suffers less error than the null estimator for some values of $n < d$, and the curve exhibits a local optimum in this regime.

One important aspect of Theorem 1 comes from the relationship between n and the parameter λ_n, which together satisfy $n = \mathrm{tr}(\boldsymbol{\Sigma}_\mu(\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1})$. This expression is precisely the classical notion of *effective dimension* for ridge regression regularized with λ_n (Alaoui and Mahoney, 2015), and it arises here even though there is no explicit ridge regularization in the problem being considered in Theorem 1. The global solution to the ridge regression task (i.e., ℓ_2-regularized least squares) with parameter λ is defined as:

$$\operatorname*{argmin}_{\mathbf{w}} \Big\{ \mathbb{E}_{\mu,y}\big[\big(\mathbf{x}^\top \mathbf{w} - y(\mathbf{x})\big)^2\big] + \lambda \|\mathbf{w}\|^2 \Big\} \;=\; (\boldsymbol{\Sigma}_\mu + \lambda \mathbf{I})^{-1} \mathbf{v}_{\mu,y}, \quad \text{where } \mathbf{v}_{\mu,y} = \mathbb{E}_{\mu,y}[y(\mathbf{x})\,\mathbf{x}].$$

When Assumption 1 holds, then $\mathbf{v}_{\mu,y} = \boldsymbol{\Sigma}_\mu \mathbf{w}^*$, however ridge-regularized least squares is well-defined for much more general response models. Our second result makes a direct connection between the (expectation of the) unregularized minimum norm solution on the sample and the global ridge-regularized solution. While the under-determined regime (i.e., $n < d$) is of primary interest to us, for completeness we state this result for arbitrary values of n and d. Note that, just like the definition of regularized least squares, this theorem applies more generally than Theorem 1, in that it does *not* require the responses to follow any linear model as in Assumption 1 (proof in Appendix D).

Theorem 3 (Implicit regularization of Moore-Penrose estimator) *For μ satisfying[1] Assumption 2 and any $y(\cdot)$ such that $\mathbf{v}_{\mu,y} = \mathbb{E}_{\mu,y}[y(\mathbf{x})\,\mathbf{x}]$ is well-defined, if $\bar{\mathbf{X}} \sim S_\mu^n$ and $\bar{y}_i = y(\bar{\mathbf{x}}_i)$, then*

$$\mathbb{E}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] = \begin{cases} (\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \mathbf{v}_{\mu,y} & \text{for } n < d, \\ \boldsymbol{\Sigma}_\mu^{-1} \mathbf{v}_{\mu,y} & \text{for } n \geq d, \end{cases}$$

where, as in Theorem 1, λ_n is such that the effective dimension $\mathrm{tr}(\boldsymbol{\Sigma}_\mu(\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1})$ equals n.

That is, when $n < d$, the Moore-Penrose estimator (which itself is not regularized), computed on the random training sample, in expectation equals the global ridge-regularized least squares solution of the underlying regression problem. Moreover, λ_n, i.e., the amount of implicit ℓ_2-regularization, is controlled by the degree of over-parameterization in such a way as to ensure that n becomes the ridge effective dimension (a.k.a. the effective degrees of freedom).

We illustrate this result in Figure 1b, plotting the norm of the expectation of the Moore-Penrose estimator. As for the MSE, our surrogate theory aligns well with the empirical estimates for i.i.d. Gaussian designs, showing that the shrinkage of the unregularized estimator in the under-determined regime matches the implicit ridge-regularization characterized by Theorem 3. While the shrinkage is a linear function of the sample size n for isotropic features (i.e., $\boldsymbol{\Sigma} = \mathbf{I}$), it exhibits a non-linear behavior for other spectral decays. Such *implicit regularization* has been studied previously (see, e.g., Mahoney and Orecchia, 2011; Mahoney, 2012); it has been observed empirically for RandNLA sampling algorithms (Ma et al., 2015); and it has also received attention more generally within the context of neural networks (Neyshabur, 2017). While our implicit regularization result is limited to the Moore-Penrose estimator, this new connection (and others, described

1. The proof of Theorem 3 can be easily extended to probability measures μ that do not satisfy Assumption 2 (such as discrete distributions). We include this assumption here to simplify the presentation.

below) between the minimum norm solution of an unregularized under-determined system and a ridge-regularized least squares solution offers a simple interpretation for the implicit regularization observed in modern machine learning architectures.

Our exact non-asymptotic expressions in Theorem 1 and our exact implicit regularization results in Theorem 3 are derived for the surrogate design, but Figure 1 suggests that they accurately describe the MSE (up to lower order terms) also under the standard i.i.d. design $\mathbf{X} \sim \mu^n$, particularly when μ is a multivariate Gaussian. As a third result, we can verify this in the cases where there exist known expressions for the MSE under the i.i.d. design (standard Gaussian for the under-determined setting, and arbitrary Gaussian for the over-determined one). We give the proof in Appendix E.

Theorem 4 (Asymptotic consistency of surrogate design) *Let $\rho = n/d \neq 1$, $\mathbf{X} \sim \mu^n$ and $y_i = y(\mathbf{x}_i)$ satisfy Assumption 1. If $d \geq c_\rho = \frac{3}{|1-\rho|}$ and*

$$\mu = \begin{cases} \mathcal{N}(\mathbf{0}, \mathbf{I}) & \text{when } n < d-1, \\ \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),\ \mathbf{\Sigma} \succ \mathbf{0} & \text{when } n > d+1, \end{cases}$$

then the absolute difference between surrogate expressions and the true MSE is bounded as follows:

$$\Big|\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] - \mathcal{M}(\mathbf{\Sigma}, \mathbf{w}^*, \sigma^2, n)\Big| \ \leq\ \frac{c_\rho}{d} \cdot \mathcal{M}(\mathbf{\Sigma}, \mathbf{w}^*, \sigma^2, n).$$

Remark 5 *For n equal to $d-1$, d or $d+1$, the true MSE under Gaussian random design can be infinite, whereas the surrogate MSE is finite and has a closed form expression.*

Empirical estimates given in Figure 1 suggest that the consistency of surrogate expressions holds much more generally than it is stated above. Based on a detailed empirical analysis described in Appendix F, we conjecture that an asymptotic consistency result of the form similar to the statement of Theorem 4 holds true in the under-determined regime without the assumption that $\mathbf{\Sigma} = \mathbf{I}$ (see Conjectures 15 and 16). In this case, no formula is known for $\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big]$, whereas the expressions for the surrogate Gaussian design naturally extend.

1.2. Key techniques: surrogate designs and determinant preserving random matrices

The standard random design model for linear regression assumes that each pair $(\mathbf{x}_i^\top, y_i)$ is drawn independently, where the row vector $\mathbf{x}_i^\top$ comes from some d-variate distribution μ and $y_i = y(\mathbf{x}_i)$ is a random response variable drawn conditionaly on $\mathbf{x}_i$. Precise theoretical analysis of under-determined regression in this setting poses significant challenges, even in such special cases as the Moore-Penrose estimator and a Gaussian data distribution μ. Rather than trying to directly analyze the usual i.i.d. random design $\mathbf{X} \sim \mu^n$ described above, we modify it slightly by introducing the notion of a *surrogate random design*, $\bar{\mathbf{X}} \sim S_\mu^n$. Informally, the goal of a surrogate random design is to modify an original design to capture the main properties of the original design, while being "nicer" for theoretical or empirical analysis. In particular, here, we will modify the distribution of matrix $\mathbf{X}$ so as to:

1. closely preserve the behavior of the Moore-Penrose estimator from the i.i.d. design; and
2. obtain exact expressions for double descent in terms of the mean squared error.

A key element in the construction of our surrogate designs involves rescaling the measure $\mathbf{X} \sim \mu^n$ by the pseudo-determinant $\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)$, i.e., a product of the non-zero eigenvalues. A similar type of determinantal design was suggested by Dereziński et al. (2019), but it was restricted there only to $n \geq d$. We broaden this definition by not only allowing the sample size to be less than d, but also allowing it to be randomized. A key property of our determinantal design matrix $\bar{\mathbf{X}}$ is that the expectation of a function $F(\bar{\mathbf{X}})$ can be expressed (up to normalization, see Definition 6) as:

$$\mathbb{E}[F(\bar{\mathbf{X}})] \;\propto\; \mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)F(\mathbf{X})],$$

where $\mathbf{X} \sim \mu^K$ and K is a random variable. Then, we define (in Definition 7) our surrogate design S_μ^n for each $n > 0$ as a determinantal design with a carefully chosen random variable K, so that the expected sample size is equal to n and so that it is possible to derive closed form expressions for the MSE. We achieve this by using the Poisson distribution to construct the variable K.

The key technical contribution that allows us to derive the MSE for determinantal designs is the concept of *determinant preserving random matrices*, a notion that we expect to be useful more generally. Specifically, in Section 4 we define a class of $d \times d$ random matrices $\mathbf{A}$ for which taking the determinant commutes with taking the expectation, for the matrix itself and any of its square submatrices (see Definition 11):

$$\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}})\big] = \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) \quad \text{for all } \mathcal{I},\mathcal{J} \subseteq [d] \text{ s.t. } |\mathcal{I}| = |\mathcal{J}|.$$

Not all random matrices satisfy this property, however many interesting and non-trivial examples can be found. Constructing these examples is facilitated by the closure properties that this class enjoys. In particular, if $\mathbf{A}$ and $\mathbf{B}$ are determinant preserving and independent, then $\mathbf{A}+\mathbf{B}$ and $\mathbf{A}\mathbf{B}$ are also determinant preserving (see Lemma 12). We use these techniques to prove a number of determinantal expectation formulas needed in obtaining Theorems 1 and 3. For example, we show that if $\mathbf{X} \sim \mu^K$, where K is a Poisson random variable, then:

$$\text{(Lemma 13)} \quad \mathbb{E}\big[\det(\mathbf{X}^\top\mathbf{X})\big] = \det\big(\mathbb{E}[\mathbf{X}^\top\mathbf{X}]\big),$$
$$\text{(Lemma 14)} \quad \mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top)\big] = \mathrm{e}^{-\mathbb{E}[K]}\det\big(\mathbf{I}+\mathbb{E}[\mathbf{X}^\top\mathbf{X}]\big).$$

2. Related work

There is a large body of related work, which for simplicity we cluster into three groups.

Double descent. The double descent phenomenon (a term introduced by Belkin et al., 2019a) corresponds to the phase transition in the generalization error that occurs when the ratio between the model complexity and the sample size crosses the so-called interpolation threshold. It has been observed empirically in a number of learning models, including neural networks (Belkin et al., 2019a; Geiger et al., 2019), kernel methods (Belkin et al., 2018a, 2019b), nearest neighbor models (Belkin et al., 2018b), and decision trees (Belkin et al., 2019a). The theoretical analysis of double descent, and more broadly the generalization properties of interpolating estimators, have primarily focused on various forms of linear regression. The most comparable to our work are Bartlett et al. (2019); Liang and Rakhlin (2019) and Hastie et al. (2019), who provide non-asymptotic upper/lower bounds and asymptotic formulas, respectively, for the generalization error of the Moore-Penrose estimator under essentially the same i.i.d. random design setting as ours. On the other hand, Muthukumar et al. (2019) provide bounds for the error of the ideal linear interpolator (instead of the minimum

norm one). Note that while we analyze the classical mean squared error, many works focus on the squared prediction error instead (some of them still refer to it as the MSE). Another line of literature deals with linear regression in the so-called *misspecified* setting, where the set of observed features does not match the feature space in which the response model is linear (Belkin et al., 2019c; Hastie et al., 2019; Mitra, 2019; Mei and Montanari, 2019), e.g., when the learner observes a random subset of d features from a larger population. This is an important distinction, because it allows varying the model complexity by changing the number of observed features while keeping the linear model fixed (see further discussion in Section 5). We believe that our results can be extended to this important setting, and we leave this as a direction for future work.

RandNLA. Randomized numerical linear algebra (Drineas and Mahoney, 2016, 2017) has traditionally focused on obtaining algorithmic improvements for tasks such as least squares and low-rank approximation via techniques that include sketching (Sarlos, 2006) and i.i.d. leverage score sampling (Drineas et al., 2006). However, there has been growing interest in understanding the statistical properties of these randomized methods (Ma et al., 2015; Raskutti and Mahoney, 2016), for example looking at the mean squared error of the least squares estimator obtained via i.i.d. subsampling under the standard linear response model. Determinantal sampling methods (a.k.a. volume sampling, or determinantal point processes), which first found their way into RandNLA in the context of low-rank approximation (Deshpande et al., 2006), have been recently shown to combine strong worst-case guarantees with elegant statistical properties. In particular, Dereziński and Warmuth (2017) showed that the least-squares estimator subsampled via the so-called size d volume sampling (loosely corresponding to the special case of our surrogate design S_μ^n where $n = d$) is an unbiased estimator that admits exact formulas for both the expected square loss (a worst-case metric) and the mean squared error (a statistical metric). These results were developed further by Dereziński et al. (2018, 2019); Dereziński et al. (2019), however they were still limited to the over-determined setting (with the exception of Dereziński and Warmuth, 2018; Dereziński et al., 2019, who gave upper bounds on the mean squared error of the ridge estimator under different determinantal samplings). Also in the over-determined setting, Dereziński et al. (2019) provided evidence for the fact that determinantal rescaling can be used to modify the original data distribution (particularly, a multivariate Gaussian) without a significant distortion to the estimator, while making certain statistical quantities expressible analytically. We take this direction further by analyzing the unregularized least squares estimator in the under-determined setting which is less well understood, partly due to the presence of implicit regularization.

Implicit regularization. The term implicit regularization typically refers to the notion that approximate computation (e.g., rather than exactly minimizing a function f, instead running an approximation algorithm to get an approximately optimal solution) can implicitly lead to statistical regularization (e.g., exactly minimizing an objective of the form $f + \lambda g$, for some well-specified λ and g). See Mahoney and Orecchia (2011); Perry and Mahoney (2011); Gleich and Mahoney (2014) and references therein for early work on the topic; and see Mahoney (2012) for an overview. More recently, often motivated by neural networks, there has been work on implicit regularization that typically considered SGD-based optimization algorithms. See, e.g., theoretical results on simplified models (Neyshabur et al., 2014; Neyshabur, 2017; Soudry et al., 2018; Gunasekar et al., 2017; Arora et al., 2019; Kubo et al., 2019) as well as extensive empirical and phenomenological results on state-of-the-art neural network models (Martin and Mahoney, 2018, 2019). The implicit regularization observed by us is different in that it is not caused by an inexact approximation algorithm (such as SGD) but rather by the selection of one out of many exact solutions (e.g., the minimum norm

solution). In this context, most relevant are the asymptotic results of LeJeune et al. (2019) (which used the asymptotic risk results for ridge regression of Dobriban and Wager, 2018) and Kobak et al. (2018). Our non-asymptotic results are also related to recent work in RandNLA on the expectation of the inverse (Dereziński and Mahoney, 2019) and generalized inverse (Mutný et al., 2019) of a subsampled matrix.

3. Surrogate random designs

In this section, we provide the definition of our surrogate random design S_μ^n, where μ is a d-variate probability measure and n is the sample size. This distribution is used in place of the standard random design μ^n consisting of n row vectors drawn independently from μ. Our surrogate design uses determinantal rescaling to alter the joint distribution of the vectors so that certain expected quantities (such as the mean squared error of the Moore-Penrose estimator) can be expressed in a closed form. We start by introducing notation.

Preliminaries. The set $\{1, ..., n\}$ will be denoted by $[n]$. For an $n \times n$ matrix $\mathbf{A}$, we use $\mathrm{pdet}(\mathbf{A})$ to denote the pseudo-determinant of $\mathbf{A}$, which is the product of non-zero eigenvalues. For index subsets $\mathcal{I}$ and $\mathcal{J}$, we use $\mathbf{A}_{\mathcal{I},\mathcal{J}}$ to denote the submatrix of $\mathbf{A}$ with rows indexed by $\mathcal{I}$ and columns indexed by $\mathcal{J}$. We may write $\mathbf{A}_{\mathcal{I},*}$ to indicate that we take a subset of rows. We use $\mathrm{adj}(\mathbf{A})$ to denote the adjugate of $\mathbf{A}$, defined as follows: the (i,j)th entry of $\mathrm{adj}(\mathbf{A})$ is $(-1)^{i+j}\det(\mathbf{A}_{[n]\setminus\{j\},[n]\setminus\{i\}})$. For a probability measure μ over $\mathbb{R}^d$, we use $\mathbf{x}^\top \sim \mu$ to denote a random row vector $\mathbf{x}^\top$ sampled according to this distibution. We let $\mathbf{X} \sim \mu^k$ denote a $k \times d$ random matrix with rows drawn i.i.d. according to μ, and the ith row is denoted as $\mathbf{x}_i^\top$. We also let $\mathbf{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$, where $\mathbb{E}_\mu$ refers to the expectation with respect to $\mathbf{x}^\top \sim \mu$, assuming throughout that $\mathbf{\Sigma}_\mu$ is well-defined and positive definite. We use $\mathrm{Poisson}(\gamma)_{\leq a}$ as the Poisson distribution restricted to $[0, a]$, whereas $\mathrm{Poisson}(\gamma)_{\geq a}$ is restricted to $[a, \infty)$. We also let $\#(\mathbf{X})$ denote the number of rows of $\mathbf{X}$.

We now define a family of determinantal distributions over random matrices $\bar{\mathbf{X}}$, where not only the entries but also the number of rows is randomized. This randomized sample size is a crucial property of our designs that enables our analysis.

Definition 6 *Let μ satisfy Assumption 2 and let K be a random variable over non-negative integers. A determinantal design $\bar{\mathbf{X}} \sim \mathrm{Det}(\mu, K)$ is a distribution with the same domain as $\mathbf{X} \sim \mu^K$ such that for any event E measurable w.r.t. $\mathbf{X}$, we have*

$$\Pr\{\bar{\mathbf{X}} \in E\} = \frac{\mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)\mathbf{1}_{[\mathbf{X}\in E]}]}{\mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)]}.$$

The above definition can be interpreted as rescaling the density function of μ^K by the pseudo-determinant, and then renormalizing it. We now construct our surrogate design S_μ^n by appropriately selecting the random variable K. The obvious choice of $K = n$ does *not* result in simple closed form expressions for the MSE in the under-determined regime (i.e., $n < d$), which is the regime of primary interest to us. Instead, we derive our random variables K from the Poisson distribution.

Definition 7 *For μ satisfying Assumption 2, define surrogate design S_μ^n as $\mathrm{Det}(\mu, K)$ where:*

1. *if $n < d$, then $K \sim \mathrm{Poisson}(\gamma_n)_{\leq d}$ with γ_n being the solution of $n = \mathrm{tr}(\mathbf{\Sigma}_\mu(\mathbf{\Sigma}_\mu + \frac{1}{\gamma_n}\mathbf{I})^{-1})$,*
2. *if $n = d$, then we simply let $K = d$,*
3. *if $n > d$, then $K \sim \mathrm{Poisson}(\gamma_n)_{\geq d}$ with $\gamma_n = n - d$.*

Note that the under-determined case, i.e., $n < d$, is restricted to $K \leq d$ so that, under Assumption 2, $\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top) = \det(\mathbf{X}\mathbf{X}^\top)$ with probability 1. On the other hand in the over-determined case, i.e., $n > d$, we have $K \geq d$ so that $\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top) = \det(\mathbf{X}^\top\mathbf{X})$. In the special case of $n = d = K$ both of these equations are satisfied: $\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top) = \det(\mathbf{X}^\top\mathbf{X}) = \det(\mathbf{X}\mathbf{X}^\top) = \det(\mathbf{X})^2$.

The first non-trivial property of the surrogate design S_μ^n is that the expected sample size is in fact always equal to n, which we prove in Appendix A.

Lemma 8 *Let $\bar{\mathbf{X}} \sim S_\mu^n$ for any $n > 0$. Then, we have $\mathbb{E}[\#(\bar{\mathbf{X}})] = n$.*

Our general template for computing expectations under a surrogate design $\bar{\mathbf{X}} \sim \mathbf{S}_\mu^n$ is to use the following expressions based on the i.i.d. random design $\mathbf{X} \sim \mu^K$:

$$\mathbb{E}[F(\bar{\mathbf{X}})] = \begin{cases} \frac{\mathbb{E}[\det(\mathbf{X}\mathbf{X}^\top)F(\mathbf{X})]}{\mathbb{E}[\det(\mathbf{X}\mathbf{X}^\top)]} & K \sim \mathrm{Poisson}(\gamma_n) \quad \text{for } n < d, \\ \frac{\mathbb{E}[\det(\mathbf{X})^2F(\mathbf{X})]}{\mathbb{E}[\det(\mathbf{X})^2]} & K = d \quad \text{for } n = d, \\ \frac{\mathbb{E}[\det(\mathbf{X}^\top\mathbf{X})F(\mathbf{X})]}{\mathbb{E}[\det(\mathbf{X}^\top\mathbf{X})]} & K \sim \mathrm{Poisson}(\gamma_n) \quad \text{for } n > d. \end{cases} \tag{1}$$

These formulas follow from Definitions 6 and 7 because the determinants $\det(\mathbf{X}\mathbf{X}^\top)$ and $\det(\mathbf{X}^\top\mathbf{X})$ are non-zero precisely in the regimes $n \leq d$ and $n \geq d$, respectively, which is why we can drop the restrictions on the range of the Poisson distribution. The normalization constants for computing the expectations can be obtained using the following formulas: if $\mathbf{X} \sim \mu^K$ then

$$\begin{aligned} &\text{(Lemma 14)} && \mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top)\big] = \mathrm{e}^{-\gamma_n}\det(\mathbf{I} + \gamma_n\boldsymbol{\Sigma}_\mu) && \text{for } K \sim \mathrm{Poisson}(\gamma_n),\ n < d, \\ &\text{(Lemma 18)} && \mathbb{E}\big[\det(\mathbf{X})^2\big] = d!\det(\boldsymbol{\Sigma}_\mu), && \text{for } K = n = d, \\ &\text{(Lemma 13)} && \mathbb{E}\big[\det(\mathbf{X}^\top\mathbf{X})\big] = \det(\gamma_n\boldsymbol{\Sigma}_\mu), && \text{for } K \sim \mathrm{Poisson}(\gamma_n),\ n > d. \end{aligned}$$

We prove Lemmas 13 and 14 in Section 4 by introducing the concept of determinant preserving random matrices. On the other hand, Lemma 18 and the design S_μ^d can be found in the literature (van der Vaart, 1965; Dereziński et al., 2019), and we will rely on those known results in this case.

We now highlight some key expectation formulas for surrogate designs, which are used to derive the MSE expressions from Theorem 1. A standard decomposition of the MSE yields:

$$\mathrm{MSE}\big[\bar{\mathbf{X}}^\dagger\bar{\mathbf{y}}\big] = \mathbb{E}\big[\|\bar{\mathbf{X}}^\dagger(\bar{\mathbf{X}}\mathbf{w}^* + \boldsymbol{\xi}) - \mathbf{w}^*\|^2\big] = \sigma^2\mathbb{E}\big[\mathrm{tr}\big((\bar{\mathbf{X}}^\top\bar{\mathbf{X}})^\dagger\big)\big] + \mathbf{w}^{*\top}\mathbb{E}\big[\mathbf{I} - \bar{\mathbf{X}}^\dagger\bar{\mathbf{X}}\big]\mathbf{w}^*. \tag{2}$$

Thus, our task is to find closed form expressions for the two expectations above. If $n \geq d$, then the latter goes away because $\bar{\mathbf{X}}^\dagger\bar{\mathbf{X}}$ is the projection onto the row-span of $\bar{\mathbf{X}}$ so when $\bar{\mathbf{X}}$ has rank d then $\mathbf{I} - \bar{\mathbf{X}}^\dagger\bar{\mathbf{X}} = \mathbf{0}$. When $n < d$, it is given in the following result (proof in Appendix D).

Lemma 9 *If $\bar{\mathbf{X}} \sim S_\mu^n$ and $n < d$, then we have: $\mathbb{E}\big[\mathbf{I} - \bar{\mathbf{X}}^\dagger\bar{\mathbf{X}}\big] = (\gamma_n\boldsymbol{\Sigma}_\mu + \mathbf{I})^{-1}$.*

No such expectation formula is known for i.i.d. designs, except when μ is an isotropic Gaussian. In Appendix D, we also prove a generalization of Lemma 9 which is then used to establish our implicit regularization result (Theorem 3). We next give an expectation formula for the trace of the Moore-Penrose inverse of the covariance matrix for a surrogate design (proof in Appendix C).

Lemma 10 *If $\bar{\mathbf{X}} \sim S_\mu^n$, then:*

$$\mathbb{E}\big[\mathrm{tr}\big((\bar{\mathbf{X}}^\top\bar{\mathbf{X}})^\dagger\big)\big] = \begin{cases} \gamma_n\big(1 - \det\big((\frac{1}{\gamma_n}\mathbf{I} + \boldsymbol{\Sigma}_\mu)^{-1}\boldsymbol{\Sigma}_\mu\big)\big), & \textit{for } n < d, \\ \mathrm{tr}(\boldsymbol{\Sigma}_\mu^{-1}), & \textit{for } n = d, \\ \mathrm{tr}(\boldsymbol{\Sigma}_\mu^{-1})\,\frac{1-\mathrm{e}^{-\gamma_n}}{\gamma_n}, & \textit{for } n > d. \end{cases}$$

Our key contribution is the under-determined regime (i.e., $n < d$), where we observe implicit regularization given by $\lambda_n = \frac{1}{\gamma_n}$. Since $n = \mathrm{tr}(\boldsymbol{\Sigma}_\mu(\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}) = d - \lambda_n \mathrm{tr}((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1})$, it follows that $\lambda_n = (d-n)/\mathrm{tr}((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1})$. Combining this with Lemmas 9 and 10, we recover the surrogate MSE expression in Theorem 1 (see details in Appendix C).

4. Determinant preserving random matrices

In this section, we introduce the key tool for computing expectation formulas of matrix determinants. It is used in our analysis of the surrogate design, and it should be of independent interest.

The key question motivating the following definition is: When does taking expectation commute with computing a determinant for a square random matrix?

Definition 11 *A random $d \times d$ matrix $\mathbf{A}$ is called determinant preserving (d.p.), if*

$$\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}})\big] = \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) \quad \textit{for all } \mathcal{I}, \mathcal{J} \subseteq [d] \textit{ s.t. } |\mathcal{I}| = |\mathcal{J}|.$$

Note that from the definition of an adjugate matrix (see preliminaries in Section 3) it immediately follows that if $\mathbf{A}$ is determinant preserving then adjugate commutes with expectation for this matrix:

$$\begin{aligned}\mathbb{E}\big[\big(\mathrm{adj}(\mathbf{A})\big)_{i,j}\big] &= \mathbb{E}\big[(-1)^{i+j}\det(\mathbf{A}_{[d]\setminus\{j\},[d]\setminus\{i\}})\big] \\ &= (-1)^{i+j}\det\big(\mathbb{E}[\mathbf{A}_{[d]\setminus\{j\},[d]\setminus\{i\}}]\big) = \big(\mathrm{adj}(\mathbb{E}[\mathbf{A}])\big)_{i,j}. \qquad (3)\end{aligned}$$

The adjugate is useful in our analysis because it connects the determinant and the inverse via the formula $\mathrm{adj}(\mathbf{A}) = \det(\mathbf{A})\mathbf{A}^{-1}$, which holds for any invertible $\mathbf{A}$. We next give a few simple examples to provide some intuition. First, note that every 1×1 random matrix is determinant preserving simply because taking a determinant is an identity transfomation in one dimension. Similarly, every fixed matrix is determinant preserving because in this case taking the expectation is an identity transformation. In all other cases, however, Definition 11 has to be verified more carefully. Further examples (positive and negative) follow.

Example 1 *If $\mathbf{A}$ has i.i.d. Gaussian entries $a_{ij} \sim \mathcal{N}(0,1)$, then $\mathbf{A}$ is d.p. because $\mathbb{E}[\det(\mathbf{A})] = 0$.*

In fact, it can be shown that all random matrices with independent entries are determinant preserving. However, this is not a necessary condition.

Example 2 *Let $\mathbf{A} = s\,\mathbf{Z}$, where $\mathbf{Z}$ is fixed with $\mathrm{rank}(\mathbf{Z}) = r$, and s is a scalar random variable. Then for $|\mathcal{I}| = |\mathcal{J}| = r$ we have*

$$\mathbb{E}\big[\det(s\,\mathbf{Z}_{\mathcal{I},\mathcal{J}})\big] = \mathbb{E}[s^r]\det(\mathbf{Z}_{\mathcal{I},\mathcal{J}}) = \det\Big(\big(\mathbb{E}[s^r]\big)^{\frac{1}{r}}\mathbf{Z}_{\mathcal{I},\mathcal{J}}\Big),$$

so if $r = 1$ then $\mathbf{A}$ is determinant preserving, whereas if $r > 1$ and $\mathrm{Var}[s] > 0$ then it is not.

To construct more complex examples, we show that determinant preserving random matrices are closed under addition and multiplication. The proof of this result is an extension of an existing argument, given by Dereziński and Mahoney (2019) in the proof of Lemma 7, for computing the expected determinant of the sum of rank-1 random matrices (proof in Appendix B).

Lemma 12 *If $\mathbf{A}, \mathbf{B}$ are independent and d.p. then $\mathbf{A}+\mathbf{B}$ and $\mathbf{AB}$ are also determinant preserving.*

Next, we introduce another important class of d.p. matrices: a sum of i.i.d. rank-1 random matrices with the number of i.i.d. samples being a Poisson random variable. Our use of the Poisson distribution is crucial for the below result to hold. It is an extension of an expectation formula given by Dereziński (2019) for sampling from discrete distributions (proof in Appendix B).

Lemma 13 *If K is a Poisson random variable and $\mathbf{A}, \mathbf{B}$ are random $K \times d$ matrices whose rows are sampled as an i.i.d. sequence of joint pairs of random vectors, then $\mathbf{A}^\top\mathbf{B}$ is d.p., and so:*

$$\mathbb{E}\big[\det(\mathbf{A}^\top\mathbf{B})\big] = \det\big(\mathbb{E}[\mathbf{A}^\top\mathbf{B}]\big).$$

Finally, we show the expectation formula needed for obtaining the normalization constant of the under-determined surrogate design, given in (1). The below result is more general than the normalization constant requires, because it allows the matrices $\mathbf{A}$ and $\mathbf{B}$ to be different (the constant is obtained by setting $\mathbf{A} = \mathbf{B} = \mathbf{X} \sim \mu^K$). In fact, we use this more general statement to show Theorems 1 and 3. The proof uses Lemmas 12 and 13 (see Appendix B).

Lemma 14 *If K is a Poisson random variable and $\mathbf{A}$, $\mathbf{B}$ are random $K \times d$ matrices whose rows are sampled as an i.i.d. sequence of joint pairs of random vectors, then*

$$\mathbb{E}\big[\det(\mathbf{A}\mathbf{B}^\top)\big] = \mathrm{e}^{-\mathbb{E}[K]}\det\big(\mathbf{I} + \mathbb{E}[\mathbf{B}^\top\mathbf{A}]\big).$$

5. Conclusions and open problems

We derived exact non-asymptotic expressions for the MSE of the Moore-Penrose estimator in the standard regression task, reproducing the double descent phenomenon as the sample size crosses between the under- and over-determined regime. To achieve this, we modified the standard i.i.d. random design distribution using a determinantal rescaling to obtain a surrogate design which admits exact MSE expressions, while capturing the key properties of the i.i.d. design. We also provided a result that relates the expected value of the Moore-Penrose estimator of a training sample in the under-determined regime (i.e., the minimum norm solution) to the ridge-regularized least squares solution for the population distribution, thereby providing an interpretation for the implicit regularization resulting from over-parameterization.

An important technical issue is that, in this work, we focus on the classical *well-specified* linear regression task, where the underlying response model is linear with respect to the observed feature space. A significant effort in the related literature (see Section 2) has been directed towards a number of *misspecified* linear regression tasks, where the set of d observed features is different than the set of D features which define the linear model (typically, $d \ll D$). Crucially, unlike in the well-specified task, here it is possible to vary the number of observed features without changing the underlying linear model. Recent work (Hastie et al., 2019) has compared how varying the feature dimension affects the (asymptotic) generalization error for both well-specified and misspecified

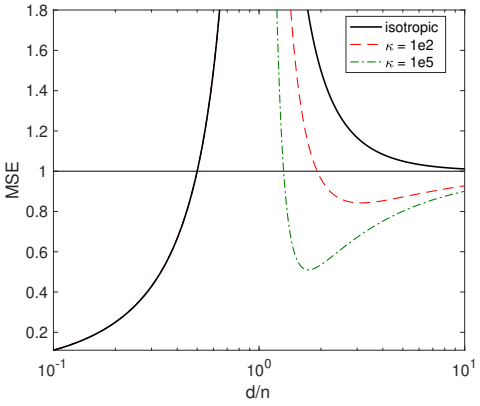


Figure 2: Surrogate MSE as a function of d/n, with n fixed to 100 and varying d.

tasks, however their analysis was limited to certain special settings such as an isotropic data distribution. As an additional point of comparison, in Figure 2 we plot the MSE expressions of Theorem 1 for our well-specified setting when varying the feature dimension d. The model is chosen just like in Figure 1, where the covariances $\boldsymbol{\Sigma}_\mu$ are diagonal with condition number κ and exponentially decaying spectrum scaled so that $\mathrm{tr}(\boldsymbol{\Sigma}_\mu^{-1}) = d$. We also use $\sigma^2 = 1$ and $\mathbf{w}^* = \frac{1}{\sqrt{d}}\mathbf{1}$. Qualitatively, our plots follow the trends outlined by Hastie et al. (2019) for the isotropic case (i.e., $\kappa = 1$), but the spectral decay of the covariance matrix (captured by our new MSE expressions) does have a significant effect on the descent curve in the under-determined regime. Note that the plots achieve their minimum as d goes to zero because in the well-specified task as the complexity of the prediction model decreases, so does the complexity of the true response model. Nevertheless, we observe generalization in the under-determined regime, as seen by the fact that the MSE curve goes below the error of the null estimator, $\mathrm{MSE}[\mathbf{0}] = 1$.

Our work opens up a number of new directions for future research. This includes extending our surrogate analysis to the misspecified linear regression discussed above. Also, it remains open whether the analysis we provided for the mean squared error can be reproduced in the context of mean squared *prediction* error, which is relevant in many machine learning tasks. Finally, while Theorem 4 states that our surrogate expressions for the MSE under certain Gaussian designs are asymptotically consistent with the multiplicative error rate of $O(1/d)$, we believe that this fact extends to the setting not covered by the theorem: under-determined regime (i.e., $n < d$) with a non-isotropic Gaussian distribution (i.e., $\boldsymbol{\Sigma} \neq \mathbf{I}$). We break down our analysis into verifying two conjectures which are of independent interest to multivariate Gaussian analysis. The first conjecture addresses the first term in the MSE derivation (2) and postulates an asymptotically consistent formula for the expected Moore-Penrose inverse of the pseudo-Wishart distribution. Matrix $\mathbf{W} \sim \mathcal{PW}(\boldsymbol{\Sigma}, n)$ is distributed according to the pseudo-Wishart distribution with $n < d$ degrees of freedom (Srivastava and Khatri, 1979) if it can be written as $\mathbf{W} = \mathbf{X}^\top\mathbf{X}$, where $\mathbf{X} \sim \mathcal{N}_{n,d}(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$ is the matrix-variate Normal (Gupta and Nagar, 2018), i.e., an i.i.d. Gaussian design.

Conjecture 15 (Moore-Penrose inverse of pseudo-Wishart) *Fix $n/d < 1$ and let $\mathbf{W} \sim \mathcal{PW}(\boldsymbol{\Sigma}, n)$, where $\boldsymbol{\Sigma}$ is $d \times d$ positive definite with condition number bounded by a constant. Then:*

$$\left|\frac{\mathbb{E}\big[\mathrm{tr}(\mathbf{W}^\dagger)\big]}{\mathcal{V}(\boldsymbol{\Sigma}, n)} - 1\right| = O(1/d) \qquad \text{for} \quad \mathcal{V}(\boldsymbol{\Sigma}, n) = \frac{1-\alpha_n}{\lambda_n}, \tag{4}$$

where $\lambda_n \geq 0$ satisfies $n = \mathrm{tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_n\mathbf{I})^{-1})$ and $\alpha_n = \det(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_n\mathbf{I})^{-1})$.

Our second conjecture involves the projection onto the orthogonal complement of a Gaussian sample $\mathbf{X}$, i.e., the matrix $\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}$, and addresses the second term in the MSE derivation (2).

Conjecture 16 (Gaussian orthogonal projection) *Fix $n/d < 1$ and let $\mathbf{X} \sim \mathcal{N}_{n,d}(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is $d \times d$ positive definite with condition number bounded by a constant. Then:*

$$\sup_{\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \left|\frac{\mathbf{w}^\top\mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]\mathbf{w}}{\mathbf{w}^\top\mathcal{B}(\boldsymbol{\Sigma}, n)\mathbf{w}} - 1\right| = O(1/d) \qquad \text{for} \quad \mathcal{B}(\boldsymbol{\Sigma}, n) = \lambda_n(\boldsymbol{\Sigma} + \lambda_n\mathbf{I})^{-1}. \tag{5}$$

Recall that $\lambda_n = \frac{d-n}{\mathrm{tr}((\boldsymbol{\Sigma}+\lambda_n\mathbf{I})^{-1})}$, so our surrogate MSE is recovered as $\sigma^2\mathcal{V}(\boldsymbol{\Sigma}, n) + \mathbf{w}^{*\top}\mathcal{B}(\boldsymbol{\Sigma}, n)\mathbf{w}^*$. Conjectures 15 and 16 provide new insights into classical matrix-variate distributions with extensive literature dedicated to them (see, e.g., Chikuse, 1990; Cook and Forzani, 2011). We discuss this further in Appendix F, where we also provide detailed empirical evidence supporting these claims.

References

A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. In *Annual Advances in Neural Information Processing Systems 28: Proceedings of the 2015 Conference*, pages 775–783, 2015.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7411–7422. Curran Associates, Inc., 2019.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. Technical Report Preprint: arXiv:1906.11300, 2019.

M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35st International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, Stockholm, Sweden, 2018a. PMLR.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. USA*, 116:15849–15854, 2019a.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, Naha, Okinawa, Japan, 2019b. PMLR.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2300–2311. Curran Associates, Inc., 2018b.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019c.

Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, second edition, 2011.

Yasuko Chikuse. The matrix angular central gaussian distribution. *Journal of Multivariate Analysis*, 33(2):265–274, 1990.

Yasuko Chikuse. High dimensional limit theorems and matrix decompositions on the stiefel manifold. *Journal of Multivariate Analysis*, 36(2):145 – 162, 1991.

Yasuko Chikuse. Density estimation on the stiefel manifold. *Journal of Multivariate Analysis*, 66 (2):188 – 206, 1998.

R. Dennis Cook and Liliana Forzani. On the mean and variance of the generalized inverse of a singular wishart matrix. *Electron. J. Statist.*, 5:146–158, 2011.

Michał Dereziński. Fast determinantal point processes via distortion-free intermediate sampling. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on*

Learning Theory, volume 99 of *Proceedings of Machine Learning Research*, pages 1029–1049, Phoenix, USA, 25–28 Jun 2019.

Michał Dereziński and Michael W Mahoney. Distributed estimation of the inverse hessian by determinantal averaging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11401–11411. Curran Associates, Inc., 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, 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.

Michał Dereziński, Kenneth L. Clarkson, Michael W. Mahoney, and Manfred K. Warmuth. Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1050–1069, Phoenix, USA, 25–28 Jun 2019.

Michał Dereziński, Feynman Liang, and Michael W. Mahoney. Bayesian experimental design using regularized determinantal point processes. *arXiv e-prints*, art. arXiv:1906.04133, Jun 2019.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Correcting the bias in least squares regression with volume-rescaled sampling. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 944–953. PMLR, 16–18 Apr 2019.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Unbiased estimators for random design regression. *arXiv e-prints*, art. arXiv:1907.03411, Jul 2019.

Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pages 1117–1126, Miami, FL, USA, January 2006.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

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.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart. Scaling description of generalization with number of parameters in deep learning. Technical Report Preprint: arXiv:1901.01608, 2019.

D. F. Gleich and M. W. Mahoney. Anti-differentiating approximation algorithms: A case study with min-cuts, spectral, and flow. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1018–1025, 2014.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6151–6159. Curran Associates, Inc., 2017.

A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Monographs and Surveys in Pure and Applied Mathematics. CRC Press, 2018. ISBN 9781351433006.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. Technical Report Preprint: arXiv:1903.08560, 2019.

D. Kobak, J. Lomond, and B. Sanchez. Optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. Technical report, 2018. Preprint: arXiv:1805.10939.

M. Kubo, R. Banno, H. Manabe, and M. Minoji. Implicit regularization in over-parameterized neural networks. Technical Report Preprint: arXiv:1903.01997, 2019.

Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012.

D. LeJeune, H. Javadi, and R. G. Baraniuk. The implicit regularization of ordinary least squares ensembles. Technical report, 2019. Preprint: arXiv:1910.04743.

T. Liang and A. Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics, to appear*, 2019.

Miles E Lopes, N Benjamin Erichson, and Michael W Mahoney. Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching. *arXiv preprint arXiv:1909.06120*, 2019.

P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16:861–911, 2015.

M. W. Mahoney. Approximate computation and implicit regularization for very large-scale data analysis. In *Proceedings of the 31st ACM Symposium on Principles of Database Systems*, pages 143–154, 2012.

M. W. Mahoney and L. Orecchia. Implementing regularization implicitly via approximate eigenvector computation. In *Proceedings of the 28th International Conference on Machine Learning*, pages 121–128, 2011.

Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011. Also available at: arXiv:1104.5557.

C. H. Martin and M. W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. Technical Report Preprint: arXiv:1810.01075, 2018.

C. H. Martin and M. W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4284–4293, 2019.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. Technical Report Preprint: arXiv:1908.05355, 2019.

P. P. Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for l2 and l1 penalized interpolation. Technical Report Preprint: arXiv:1906.03667, 2019.

V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. Technical Report Preprint: arXiv:1903.09139, 2019.

M. Mutný, M. Dereziński, and A. Krause. Convergence analysis of the randomized Newton method with determinantal sampling. Technical report, 2019. Preprint: arXiv:1910.11561.

B. Neyshabur. Implicit regularization in deep learning. Technical report, 2017. Preprint: arXiv:1709.01953.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: on the role of implicit regularization in deep learning. Technical Report Preprint: arXiv:1412.6614, 2014.

P. O. Perry and M. W. Mahoney. Regularized Laplacian estimation and fast eigenvector approximation. In *Annual Advances in Neural Information Processing Systems 24: Proceedings of the 2011 Conference*, 2011.

G. Raskutti and M. W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(214):1–31, 2016.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 143–152, Washington, DC, USA, 2006. IEEE Computer Society.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

M.S. Srivastava. Singular wishart and multivariate beta distributions. *Ann. Statist.*, 31(5):1537–1560, 10 2003.

M.S. Srivastava and C.G. Khatri. *An introduction to multivariate statistics*. North-Holland/New York, 1979. ISBN 9780444003027.

H. Robert van der Vaart. A note on Wilks' internal scatter. *Ann. Math. Statist.*, 36(4):1308–1312, 08 1965.

Appendix A. Proof of Lemma 8

We first record an important property of the design S_μ^d which can be used to construct an over-determined design for any $n > d$. A similar version of this result was also previously shown by Dereziński et al. (2019) for a different determinantal design.

Lemma 17 *Let $\bar{\mathbf{X}} \sim S_\mu^d$ and $\mathbf{X} \sim \mu^K$, where $K \sim \mathrm{Poisson}(\gamma)$. Then the matrix composed of a random permutation of the rows from $\bar{\mathbf{X}}$ and $\mathbf{X}$ is distributed according to $S_\mu^{d+\gamma}$.*

Proof Let $\widetilde{\mathbf{X}}$ denote the matrix constructed from the permuted rows of $\bar{\mathbf{X}}$ and $\mathbf{X}$. Letting $\mathbf{Z} \sim \mu^{K+d}$, we derive the probability $\Pr\{\widetilde{\mathbf{X}} \in E\}$ by summing over the possible index subsets $S \subseteq [K+d]$ that correspond to the rows coming from $\bar{\mathbf{X}}$:

$$\begin{aligned}
\Pr\{\widetilde{\mathbf{X}} \in E\} &= \mathbb{E}\bigg[\frac{1}{\binom{K+d}{d}} \sum_{S:\,|S|=d} \frac{\mathbb{E}[\det(\mathbf{Z}_{S,*})^2 \mathbf{1}_{[\mathbf{Z}\in E]} \mid K]}{d!\det(\boldsymbol{\Sigma}_\mu)}\bigg] \\
&= \sum_{k=0}^{\infty} \frac{\gamma^k \mathrm{e}^{-\gamma}}{k!} \frac{\gamma^d k!}{(k+d)!} \frac{\mathbb{E}\big[\sum_{S:\,|S|=d} \det(\mathbf{Z}_{S,*})^2 \mathbf{1}_{[\mathbf{Z}\in E]} \mid K=k\big]}{\det(\gamma\boldsymbol{\Sigma}_\mu)} \\
&\overset{(*)}{=} \sum_{k=0}^{\infty} \frac{\gamma^{k+d} \mathrm{e}^{-\gamma}}{(k+d)!} \frac{\mathbb{E}[\det(\mathbf{Z}^\top \mathbf{Z}) \mathbf{1}_{[\mathbf{Z}\in E]} \mid K=k]}{\det(\gamma\boldsymbol{\Sigma}_\mu)},
\end{aligned}$$

where $(*)$ uses the Cauchy-Binet formula to sum over all subsets S of size d. Finally, since the sum shifts from k to $k+d$, the last expression can be rewritten as $\mathbb{E}[\det(\mathbf{X}^\top\mathbf{X})\mathbf{1}_{[\mathbf{X}\in E]}]/\det(\gamma\boldsymbol{\Sigma}_\mu)$, where recall that $\mathbf{X} \sim \mu^K$ and $K \sim \mathrm{Poisson}(\gamma)$, matching the definition of $S_\mu^{d+\gamma}$. ■

We now proceed with the proof of Lemma 8, where we establish that the expected sample size of S_μ^n is indeed n.

Proof of Lemma 8 The result is obvious when $n = d$, whereas for $n > d$ it is an immediate consequence of Lemma 17. Finally, for $n < d$ the expected sample size follows as a corollary of Lemma 9, which states that

$$\text{(Lemma 9)} \qquad \mathbb{E}\big[\mathbf{I} - \bar{\mathbf{X}}^\dagger \bar{\mathbf{X}}\big] = (\gamma_n \boldsymbol{\Sigma}_\mu + \mathbf{I})^{-1},$$

where $\bar{\mathbf{X}}^\dagger\bar{\mathbf{X}}$ is the orthogonal projection onto the subspace spanned by the rows of $\bar{\mathbf{X}}$. Since the rank of this subspace is equal to the number of the rows, we have $\#(\bar{\mathbf{X}}) = \operatorname{tr}(\bar{\mathbf{X}}^\dagger\bar{\mathbf{X}})$, so

$$\mathbb{E}\big[\#(\bar{\mathbf{X}})\big] = d - \operatorname{tr}\big((\gamma_n\boldsymbol{\Sigma}_\mu + \mathbf{I})^{-1}\big) = \operatorname{tr}\big(\gamma_n\boldsymbol{\Sigma}_\mu(\gamma_n\boldsymbol{\Sigma}_\mu + \mathbf{I})^{-1}\big) = n,$$

which completes the proof. ■

Appendix B. Proofs for Section 4

Proof of Lemma 12 First, we show that $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ is d.p. for any fixed $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Below, we use a standard identity for the rank one update of a determinant: $\det(\mathbf{A} + \mathbf{u}\mathbf{v}^\top) = \det(\mathbf{A}) + \mathbf{v}^\top\operatorname{adj}(\mathbf{A})\mathbf{u}$. It follows that for any $\mathcal{I}$ and $\mathcal{J}$ of the same size,

$$\begin{aligned}
\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}} + \mathbf{u}_\mathcal{I}\mathbf{v}_\mathcal{J}^\top)\big] &= \mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}}) + \mathbf{v}_\mathcal{J}^\top\operatorname{adj}(\mathbf{A}_{\mathcal{I},\mathcal{J}})\mathbf{u}_\mathcal{I}\big] \\
&\overset{(*)}{=} \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) + \mathbf{v}_\mathcal{J}^\top\operatorname{adj}\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big)\mathbf{u}_\mathcal{I} \\
&= \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}} + \mathbf{u}_\mathcal{I}\mathbf{v}_\mathcal{J}^\top]\big),
\end{aligned}$$

where $(*)$ used (3), i.e., the fact that for d.p. matrices, adjugate commutes with expectation. Crucially, through the definition of an adjugate this step implicitly relies on the assumption that all the square submatrices of $\mathbf{A}_{\mathcal{I},\mathcal{J}}$ are also determinant preserving. Iterating this, we get that $\mathbf{A} + \mathbf{Z}$ is d.p. for any fixed $\mathbf{Z}$. We now show the same for $\mathbf{A} + \mathbf{B}$:

$$\begin{aligned}
\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}} + \mathbf{B}_{\mathcal{I},\mathcal{J}})\big] &= \mathbb{E}\Big[\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}} + \mathbf{B}_{\mathcal{I},\mathcal{J}}) \mid \mathbf{B}\big]\Big] \\
&\overset{(*)}{=} \mathbb{E}\Big[\det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}] + \mathbf{B}_{\mathcal{I},\mathcal{J}}\big)\Big] \\
&= \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}} + \mathbf{B}_{\mathcal{I},\mathcal{J}}]\big),
\end{aligned}$$

where $(*)$ uses the fact that after conditioning on $\mathbf{B}$ we can treat it as a fixed matrix. Next, we show that $\mathbf{AB}$ is determinant preserving via the Cauchy-Binet formula:

$$\begin{aligned}
\mathbb{E}\big[\det\big((\mathbf{AB})_{\mathcal{I},\mathcal{J}}\big)\big] &= \mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},*}\mathbf{B}_{*,\mathcal{J}})\big] \\
&= \mathbb{E}\Big[\sum_{S:\,|S|=|\mathcal{I}|}\det\big(\mathbf{A}_{\mathcal{I},S}\big)\det\big(\mathbf{B}_{S,\mathcal{J}}\big)\Big] \\
&= \sum_{S:\,|S|=|\mathcal{I}|}\det\big(\mathbb{E}[\mathbf{A}]_{\mathcal{I},S}\big)\det\big(\mathbb{E}[\mathbf{B}]_{S,\mathcal{J}}\big) \\
&= \det\big(\mathbb{E}[\mathbf{A}]_{\mathcal{I},*}\,\mathbb{E}[\mathbf{B}]_{*,\mathcal{J}}\big) \\
&= \det\big(\mathbb{E}[\mathbf{AB}]_{\mathcal{I},\mathcal{J}}\big),
\end{aligned}$$

where recall that $\mathbf{A}_{\mathcal{I},*}$ denotes the submatrix of $\mathbf{A}$ consisting of its (entire) rows indexed by $\mathcal{I}$. ■

To prove Lemma 13, we will use the following lemma, many variants of which appeared in the literature (e.g., van der Vaart, 1965). We use the one given by Dereziński et al. (2019).

Lemma 18 (Dereziński et al. (2019)) *If the rows of random $k \times d$ matrices $\mathbf{A}, \mathbf{B}$ are sampled as an i.i.d. sequence of $k \geq d$ pairs of joint random vectors, then*

$$k^d\,\mathbb{E}\big[\det(\mathbf{A}^\top\mathbf{B})\big] = k^{\underline{d}}\det\big(\mathbb{E}[\mathbf{A}^\top\mathbf{B}]\big). \tag{6}$$

Here, we use the following standard shorthand: $k^{\underline{d}} = \frac{k!}{(k-d)!} = k\,(k-1)\cdots(k-d+1)$. Note that the above result almost looks like we are claiming that the matrix $\mathbf{A}^\top\mathbf{B}$ is d.p., but in fact it is not because $k^d \neq k^{\underline{d}}$. The difference in those factors is precisely what we are going to correct with the Poisson random variable. We now present the proof of Lemma 13.

Proof of Lemma 13 Without loss of generality, it suffices to check Definition 11 with both $\mathcal{I}$ and $\mathcal{J}$ equal $[d]$. We first expand the expectation by conditioning on the value of K and letting $\gamma = \mathbb{E}[K]$:

$$\begin{aligned}
\mathbb{E}\big[\det(\mathbf{A}^\top\mathbf{B})\big] &= \sum_{k=0}^{\infty} \mathbb{E}\big[\det(\mathbf{A}^\top\mathbf{B}) \mid K=k\big]\,\Pr(K=k) \\
\text{(Lemma 18)} \quad &= \sum_{k=d}^{\infty} \frac{k!k^{-d}}{(k-d)!}\det\big(\mathbb{E}[\mathbf{A}^\top\mathbf{B} \mid K=k]\big)\frac{\gamma^k \mathrm{e}^{-\gamma}}{k!} \\
&= \sum_{k=d}^{\infty} \Big(\frac{\gamma}{k}\Big)^d \det\big(\mathbb{E}[\mathbf{A}^\top\mathbf{B} \mid K=k]\big)\frac{\gamma^{k-d}\mathrm{e}^{-\gamma}}{(k-d)!}.
\end{aligned}$$

Note that $\frac{\gamma}{k}\,\mathbb{E}[\mathbf{A}^\top\mathbf{B} \mid K=k] = \mathbb{E}[\mathbf{A}^\top\mathbf{B}]$, which is independent of k. Thus we can rewrite the above expression as:

$$\det\big(\mathbb{E}[\mathbf{A}^\top\mathbf{B}]\big)\sum_{k=d}^{\infty}\frac{\gamma^{k-d}\mathrm{e}^{-\gamma}}{(k-d)!} = \det\big(\mathbb{E}[\mathbf{A}^\top\mathbf{B}]\big)\sum_{k=0}^{\infty}\frac{\gamma^{k}\mathrm{e}^{-\gamma}}{k!} = \det\big(\mathbb{E}[\mathbf{A}^\top\mathbf{B}]\big),$$

which concludes the proof. ■

To prove Lemma 14, we use the following standard determinantal formula which is used to derive the normalization constant of a discrete determinantal point process.

Lemma 19 (Kulesza and Taskar (2012)) *For any $k \times d$ matrices $\mathbf{A}, \mathbf{B}$ we have*

$$\det(\mathbf{I} + \mathbf{A}\mathbf{B}^\top) = \sum_{S\subseteq[k]} \det(\mathbf{A}_{S,*}\mathbf{B}_{S,*}^\top).$$

Proof of Lemma 14 By Lemma 13, the matrix $\mathbf{B}^\top\mathbf{A}$ is determinant preserving. Applying Lemma 12 we conclude that $\mathbf{I} + \mathbf{B}^\top\mathbf{A}$ is also d.p., so

$$\det\big(\mathbf{I} + \mathbb{E}[\mathbf{B}^\top\mathbf{A}]\big) = \mathbb{E}\big[\det(\mathbf{I} + \mathbf{B}^\top\mathbf{A})\big] = \mathbb{E}\big[\det(\mathbf{I} + \mathbf{A}\mathbf{B}^\top)\big],$$

where the second equality is known as Sylvester's Theorem. We rewrite the expectation of $\det(\mathbf{I} + \mathbf{A}\mathbf{B}^\top)$ by applying Lemma 19. Letting $\gamma = \mathbb{E}[K]$, we obtain:

$$\begin{aligned}
\mathbb{E}\big[\det(\mathbf{I} + \mathbf{A}\mathbf{B}^\top)\big] &= \mathbb{E}\Big[\sum_{S\subseteq[K]} \mathbb{E}\big[\det(\mathbf{A}_{S,*}\mathbf{B}_{S,*}^\top) \mid K\big]\Big] \\
&\overset{(*)}{=} \sum_{k=0}^{\infty}\frac{\gamma^k\mathrm{e}^{-\gamma}}{k!}\sum_{i=0}^{k}\binom{k}{i}\mathbb{E}\big[\det(\mathbf{A}\mathbf{B}^\top) \mid K=i\big] \\
&= \sum_{i=0}^{\infty}\mathbb{E}\big[\det(\mathbf{A}\mathbf{B}^\top) \mid K=i\big]\sum_{k\geq i}^{\infty}\binom{k}{i}\frac{\gamma^k\mathrm{e}^{-\gamma}}{k!} \\
&= \sum_{i=0}^{\infty}\frac{\gamma^i\mathrm{e}^{-\gamma}}{i!}\mathbb{E}\big[\det(\mathbf{A}\mathbf{B}^\top) \mid K=i\big]\sum_{k\geq i}^{\infty}\frac{\gamma^{k-i}}{(k-i)!} = \mathbb{E}\big[\det(\mathbf{A}\mathbf{B}^\top)\big]\cdot\mathrm{e}^{\gamma},
\end{aligned}$$

where $(*)$ follows from the exchangeability of the rows of $\mathbf{A}$ and $\mathbf{B}$, which implies that the distribution of $\mathbf{A}_{S,*}\mathbf{B}_{S,*}^\top$ is the same for all subsets S of a fixed size k. ■

Appendix C. Proof of Theorem 1

In this section we use Z_μ^n to denote the normalization constant that appears in (1) when computing an expectation for surrogate design S_μ^n. We first prove Lemma 10 by splitting the under- and over-determined cases and start with proving the former.

Lemma 20 *If $\bar{\mathbf{X}} \sim S_\mu^n$ for $n < d$, then we have*

$$\mathbb{E}\big[\mathrm{tr}\big((\bar{\mathbf{X}}^\top\bar{\mathbf{X}})^\dagger\big)\big] = \gamma_n\big(1 - \det\big((\tfrac{1}{\gamma_n}\mathbf{I} + \mathbf{\Sigma}_\mu)^{-1}\mathbf{\Sigma}_\mu\big)\big).$$

Proof Let $\mathbf{X} \sim \mu^K$ for $K \sim \mathrm{Poisson}(\gamma_n)$. Note that if $\det(\mathbf{X}\mathbf{X}^\top) > 0$ then using the fact that $\det(\mathbf{A})\mathbf{A}^{-1} = \mathrm{adj}(\mathbf{A})$ for any invertible matrix $\mathbf{A}$, we can write:

$$\begin{aligned}
\det(\mathbf{X}\mathbf{X}^\top)\mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^\dagger\big) &= \det(\mathbf{X}\mathbf{X}^\top)\mathrm{tr}\big((\mathbf{X}\mathbf{X}^\top)^{-1}\big)\\
&= \mathrm{tr}(\mathrm{adj}(\mathbf{X}\mathbf{X}^\top))\\
&= \sum_{i=1}^K \det(\mathbf{X}_{-i}\mathbf{X}_{-i}^\top),
\end{aligned}$$

where $\mathbf{X}_{-i}$ is a shorthand for $\mathbf{X}_{[K]\setminus\{i\},*}$. Assumption 2 ensures that $\Pr\big\{\det(\mathbf{X}\mathbf{X}^\top) > 0\big\} = 1$, which allows us to write:

$$\begin{aligned}
Z_\mu^n \cdot \mathbb{E}\big[\mathrm{tr}\big((\bar{\mathbf{X}}^\top\bar{\mathbf{X}})^\dagger\big)\big] &= \mathbb{E}\Big[\sum_{i=1}^K \det(\mathbf{X}_{-i}\mathbf{X}_{-i}^\top) \;\Big|\; \det(\mathbf{X}\mathbf{X}^\top) > 0\Big] \cdot \overbrace{\Pr\big\{\det(\mathbf{X}\mathbf{X}^\top) > 0\big\}}^{1}\\
&= \sum_{k=0}^d \frac{\gamma_n^k \mathrm{e}^{-\gamma_n}}{k!}\mathbb{E}\Big[\sum_{i=1}^k \det(\mathbf{X}_{-i}\mathbf{X}_{-i}^\top) \;\Big|\; K = k\Big]\\
&= \sum_{k=0}^d \frac{\gamma_n^k \mathrm{e}^{-\gamma_n}}{k!}\, k\, \mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top) \mid K = k-1\big]\\
&= \gamma_n \sum_{k=0}^{d-1} \frac{\gamma_n^k \mathrm{e}^{-\gamma_n}}{k!}\mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top) \mid K = k\big]\\
&= \gamma_n\Big(\mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top)\big] \;-\; \frac{\gamma_n^d \mathrm{e}^{-\gamma_n}}{d!}\mathbb{E}\big[\det(\mathbf{X})^2 \mid K = d\big]\Big)\\
&\overset{(*)}{=} \gamma_n\big(\mathrm{e}^{-\gamma_n}\det(\mathbf{I} + \gamma_n\mathbf{\Sigma}_\mu) - \mathrm{e}^{-\gamma_n}\det(\gamma_n\mathbf{\Sigma}_\mu)\big),
\end{aligned}$$

where $(*)$ uses Lemma 14 for the first term and Lemma 18 for the second term. We obtain the desired result by dividing both sides by $Z_\mu^n = \mathrm{e}^{-\gamma_n}\det(\mathbf{I} + \gamma_n\mathbf{\Sigma}_\mu)$. ■

In the over-determined regime, a more general matrix expectation formula can be shown (omitting the trace). The following result is related to an expectation formula derived by Dereziński et al. (2019), however they use a slightly different determinantal design so the results are incomparable.

Lemma 21 *If $\bar{\mathbf{X}} \sim S_\mu^n$ and $n > d$, then we have*

$$\mathbb{E}\big[(\bar{\mathbf{X}}^\top \bar{\mathbf{X}})^\dagger\big] = \boldsymbol{\Sigma}_\mu^{-1} \cdot \frac{1 - \mathrm{e}^{-\gamma_n}}{\gamma_n}.$$

Proof Let $\mathbf{X} \sim \mu^K$ for $K \sim \mathrm{Poisson}(\gamma_n)$. Assumption 2 implies that for $K \neq d-1$ we have

$$\det(\mathbf{X}^\top \mathbf{X})(\mathbf{X}^\top \mathbf{X})^\dagger = \mathrm{adj}(\mathbf{X}^\top \mathbf{X}), \tag{7}$$

however when $k = d-1$ then (7) does not hold because $\det(\mathbf{X}^\top \mathbf{X}) = 0$ while $\mathrm{adj}(\mathbf{X}^\top \mathbf{X})$ may be non-zero. It follows that:

$$\begin{aligned}
Z_\mu^n \cdot \mathbb{E}\big[(\bar{\mathbf{X}}^\top \bar{\mathbf{X}})^\dagger\big] &= \mathbb{E}\big[\det(\mathbf{X}^\top \mathbf{X})(\mathbf{X}^\top \mathbf{X})^\dagger\big] \\
&= \mathbb{E}\big[\mathrm{adj}(\mathbf{X}^\top \mathbf{X})\big] - \frac{\gamma_n^{d-1}\mathrm{e}^{-\gamma_n}}{(d-1)!}\mathbb{E}\big[\mathrm{adj}(\mathbf{X}^\top \mathbf{X}) \mid K = d-1\big] \\
&\overset{(*)}{=} \mathrm{adj}\big(\mathbb{E}[\mathbf{X}^\top \mathbf{X}]\big) - \frac{\gamma_n^{d-1}\mathrm{e}^{-\gamma_n}}{(d-1)^{d-1}}\,\mathrm{adj}\big(\mathbb{E}[\mathbf{X}^\top \mathbf{X} \mid K = d-1]\big) \\
&= \mathrm{adj}(\gamma_n \boldsymbol{\Sigma}_\mu) - \mathrm{e}^{-\gamma_n}\,\mathrm{adj}(\gamma_n \boldsymbol{\Sigma}_\mu) \\
&= \det(\gamma_n \boldsymbol{\Sigma}_\mu)\,(\gamma_n \boldsymbol{\Sigma}_\mu)^{-1}(1 - \mathrm{e}^{-\gamma_n}) \\
&= \det(\gamma_n \boldsymbol{\Sigma}_\mu)\,\boldsymbol{\Sigma}_\mu^{-1} \cdot \frac{1 - \mathrm{e}^{-\gamma_n}}{\gamma_n},
\end{aligned}$$

where the first term in $(*)$ follows from Lemma 14 and (3), whereas the second term comes from Lemma 2.3 of Dereziński et al. (2019). Dividing both sides by $Z_\mu^n = \det(\gamma_n \boldsymbol{\Sigma}_\mu)$ completes the proof. ■

Applying the closed form expressions from Lemmas 9 and 10, we derive the formula for the MSE and prove Theorem 1 (we defer the proof of Lemma 9 to Appendix D).

Proof of Theorem 1 First, assume that $n < d$, in which case we have $\gamma_n = \frac{1}{\lambda_n}$ and moreover

$$\begin{aligned}
n &= \mathrm{tr}\big(\boldsymbol{\Sigma}_\mu(\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big) \\
&= \mathrm{tr}\big((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I} - \lambda_n \mathbf{I})(\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big) \\
&= d - \lambda_n \mathrm{tr}\big((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big),
\end{aligned}$$

so we can write λ_n as $(d-n)/\mathrm{tr}((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1})$. From this and Lemmas 9 and 20, we obtain the desired expression, where recall that $\alpha_n = \det\big(\boldsymbol{\Sigma}_\mu(\boldsymbol{\Sigma}_\mu + \frac{1}{\gamma_n})^{-1}\big)$:

$$\begin{aligned}
\mathrm{MSE}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] &= \sigma^2\,\gamma_n(1 - \alpha_n) + \tfrac{1}{\gamma_n}\,\mathbf{w}^{*\top}(\boldsymbol{\Sigma}_\mu + \tfrac{1}{\gamma_n}\mathbf{I})^{-1}\mathbf{w}^* \\
&\overset{(a)}{=} \sigma^2\,\frac{1 - \alpha_n}{\lambda_n} + \lambda_n\,\mathbf{w}^{*\top}(\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\mathbf{w}^* \\
&\overset{(b)}{=} \sigma^2 \mathrm{tr}\big((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big)\frac{1 - \alpha_n}{d - n} + (d-n)\frac{\mathbf{w}^{*\top}(\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\mathbf{w}^*}{\mathrm{tr}\big((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big)}.
\end{aligned}$$

While the expression given after (a) is simpler than the one after (b), the latter better illustrates how the MSE depends on the sample size n and the dimension d. Now, assume that $n > d$. In this case, we have $\gamma_n = n - d$ and apply Lemma 21:

$$\mathrm{MSE}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] = \sigma^2 \operatorname{tr}(\boldsymbol{\Sigma}_\mu^{-1}) \frac{1 - \mathrm{e}^{-\gamma_n}}{\gamma_n} = \sigma^2 \operatorname{tr}(\boldsymbol{\Sigma}_\mu^{-1}) \frac{1 - \beta_n}{n - d}.$$

The case of $n = d$ was shown in Theorem 2.12 of Dereziński et al. (2019). This concludes the proof. ■

Appendix D. Proof of Theorem 3

As in the previous section, we use Z_μ^n to denote the normalization constant that appears in (1) when computing an expectation for surrogate design S_μ^n. Recall that our goal is to compute the expected value of $\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}$ under the surrogate design S_μ^n. Similarly as for Theorem 1, the case of $n = d$ was shown in Theorem 2.10 of Dereziński et al. (2019). We break the rest down into the under-determined case $(n < d)$ and the over-determined case $(n > d)$, starting with the former. Recall that we do *not* require any modeling assumptions on the responses.

Lemma 22 *If $\bar{\mathbf{X}} \sim S_\mu^n$ and $n < d$, then for any $y(\cdot)$ such that $\mathbb{E}_{\mu,y}[y(\mathbf{x})\,\mathbf{x}]$ is well-defined, denoting $\bar{y}_i$ as $y(\bar{\mathbf{x}}_i)$, we have*

$$\mathbb{E}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] = \big(\boldsymbol{\Sigma}_\mu + \tfrac{1}{\gamma_n}\mathbf{I}\big)^{-1} \mathbb{E}_{\mu,y}[y(\mathbf{x})\,\mathbf{x}].$$

Proof Let $\mathbf{X} \sim \mu^K$ for $K \sim \mathrm{Poisson}(\gamma_n)$ and denote $y(\mathbf{x}_i)$ as y_i. Note that when $\det(\mathbf{X}\mathbf{X}^\top) > 0$, then the jth entry of $\mathbf{X}^\dagger \mathbf{y}$ equals $\mathbf{f}_j^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}$, where $\mathbf{f}_j$ is the jth column of $\mathbf{X}$, so:

$$\begin{aligned}
\det(\mathbf{X}\mathbf{X}^\top)\,(\mathbf{X}^\dagger \mathbf{y})_j &= \det(\mathbf{X}\mathbf{X}^\top)\,\mathbf{f}_j^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} \\
&= \det(\mathbf{X}\mathbf{X}^\top + \mathbf{y}\mathbf{f}_j^\top) - \det(\mathbf{X}\mathbf{X}^\top).
\end{aligned}$$

If $\det(\mathbf{X}\mathbf{X}^\top) = 0$, then also $\det(\mathbf{X}\mathbf{X}^\top + \mathbf{y}\mathbf{f}_j^\top) = 0$, so we can write:

$$\begin{aligned}
Z_\mu^n \cdot \mathbb{E}\big[(\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}})_j\big] &= \mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top)(\mathbf{X}^\dagger \mathbf{y})_j\big] \\
&= \mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top + \mathbf{y}\mathbf{f}_j^\top) - \det(\mathbf{X}\mathbf{X}^\top)\big] \\
&= \mathbb{E}\big[\det\big([\mathbf{X},\mathbf{y}][\mathbf{X},\mathbf{f}_j]^\top\big)\big] - \mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top)\big] \\
&\overset{(a)}{=} \mathrm{e}^{-\gamma_n} \det\left(\mathbf{I} + \gamma_n\, \mathbb{E}_{\mu,y}\left[\begin{pmatrix} \mathbf{x}\mathbf{x}^\top & \mathbf{x}\, y(\mathbf{x}) \\ x_j\, \mathbf{x}^\top & x_j\, y(\mathbf{x}) \end{pmatrix}\right]\right) - \mathrm{e}^{-\gamma_n} \det(\mathbf{I} + \gamma_n \boldsymbol{\Sigma}_\mu) \\
&\overset{(b)}{=} \mathrm{e}^{-\gamma_n} \det(\mathbf{I} + \gamma_n \boldsymbol{\Sigma}_\mu) \\
&\qquad \times \Big(\mathbb{E}_{\mu,y}\big[\gamma_n x_j\, y(\mathbf{x})\big] - \mathbb{E}_\mu\big[\gamma_n x_j\, \mathbf{x}^\top\big](\mathbf{I} + \gamma_n \boldsymbol{\Sigma}_\mu)^{-1} \mathbb{E}_{\mu,y}\big[\gamma_n \mathbf{x}\, y(\mathbf{x})\big]\Big),
\end{aligned}$$

where (a) uses Lemma 14 twice, with the first application involving two different matrices $\mathbf{A} = [\mathbf{X},\mathbf{y}]$ and $\mathbf{B} = [\mathbf{X},\mathbf{f}_j]$, whereas (b) is a standard determinantal identity (see Fact 2.14.2 in Bernstein, 2011). Dividing both sides by Z_μ^n and letting $\mathbf{v}_{\mu,y} = \mathbb{E}_{\mu,y}[y(\mathbf{x})\,\mathbf{x}]$, we obtain that:

$$\begin{aligned}
\mathbb{E}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] &= \gamma_n \mathbf{v}_{\mu,y} - \gamma_n^2 \boldsymbol{\Sigma}_\mu (\mathbf{I} + \gamma_n \boldsymbol{\Sigma}_\mu)^{-1} \mathbf{v}_{\mu,y} \\
&= \gamma_n \big(\mathbf{I} - \gamma_n \boldsymbol{\Sigma}_\mu (\mathbf{I} + \gamma_n \boldsymbol{\Sigma}_\mu)^{-1}\big)\mathbf{v}_{\mu,y} = \gamma_n (\mathbf{I} + \gamma_n \boldsymbol{\Sigma}_\mu)^{-1} \mathbf{v}_{\mu,y},
\end{aligned}$$

which completes the proof. ■

We return to Lemma 9, regarding the expected orthogonal projection onto the complement of the row-span of $\bar{\mathbf{X}}$, i.e., $\mathbb{E}[\mathbf{I} - \bar{\mathbf{X}}^\dagger\bar{\mathbf{X}}]$, which follows as a corollary of Lemma 22.

Proof of Lemma 9 We let $y(\mathbf{x}) = x_j$ where $j \in [d]$ and apply Lemma 22 for each j, obtaining:

$$\mathbf{I} - \mathbb{E}\big[\bar{\mathbf{X}}^\dagger\bar{\mathbf{X}}\big] = \mathbf{I} - (\boldsymbol{\Sigma}_\mu + \tfrac{1}{\gamma_n}\mathbf{I})^{-1}\boldsymbol{\Sigma}_\mu,$$

from which the result follows by simple algebraic manipulation. ■

We move on to the over-determined case, where the ridge regularization of adding the identity to $\boldsymbol{\Sigma}_\mu$ vanishes. Recall that we assume throughout the paper that $\boldsymbol{\Sigma}_\mu$ is invertible.

Lemma 23 *If $\bar{\mathbf{X}} \sim S_\mu^n$ and $n > d$, then for any real-valued random function $y(\cdot)$ such that $\mathbb{E}_{\mu,y}[y(\mathbf{x})\,\mathbf{x}]$ is well-defined, denoting $\bar{y}_i$ as $y(\bar{\mathbf{x}}_i)$, we have*

$$\mathbb{E}\big[\bar{\mathbf{X}}^\dagger\bar{\mathbf{y}}\big] = \boldsymbol{\Sigma}_\mu^{-1}\mathbb{E}_{\mu,y}\big[y(\mathbf{x})\,\mathbf{x}\big].$$

Proof Let $\mathbf{X} \sim \mu^K$ for $K \sim \mathrm{Poisson}(\gamma_n)$ and denote $y_i = y(\mathbf{x}_i)$. Similarly as in the proof of Lemma 22, we note that when $\det(\mathbf{X}^\top\mathbf{X}) > 0$, then the jth entry of $\mathbf{X}^\dagger\mathbf{y}$ equals $\mathbf{e}_j^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, where $\mathbf{e}_j$ is the jth standard basis vector, so:

$$\det(\mathbf{X}^\top\mathbf{X})\,(\mathbf{X}^\dagger\mathbf{y})_j = \det(\mathbf{X}^\top\mathbf{X})\,\mathbf{e}_j^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \det(\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\mathbf{y}\mathbf{e}_j^\top) - \det(\mathbf{X}^\top\mathbf{X}).$$

If $\det(\mathbf{X}^\top\mathbf{X}) = 0$, then also $\det(\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\mathbf{y}\mathbf{e}_j^\top) = 0$. We proceed to compute the expectation:

$$\begin{aligned}
Z_\mu^n \cdot \mathbb{E}\big[(\bar{\mathbf{X}}^\dagger\bar{\mathbf{y}})_j\big] &= \mathbb{E}\big[\det(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\dagger\mathbf{y})_j\big] \\
&= \mathbb{E}\big[\det(\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\mathbf{y}\mathbf{e}_j^\top) - \det(\mathbf{X}^\top\mathbf{X})\big] \\
&= \mathbb{E}\big[\det\big(\mathbf{X}^\top(\mathbf{X} + \mathbf{y}\mathbf{e}_j^\top)\big)\big] - \mathbb{E}\big[\det(\mathbf{X}^\top\mathbf{X})\big] \\
&\overset{(*)}{=} \det\Big(\gamma_n\,\mathbb{E}_{\mu,y}\big[\mathbf{x}(\mathbf{x} + y(\mathbf{x})\mathbf{e}_j)^\top\big]\Big) - \det(\gamma_n\boldsymbol{\Sigma}_\mu) \\
&= \det\big(\gamma_n\boldsymbol{\Sigma}_\mu + \gamma_n\mathbb{E}_{\mu,y}[\mathbf{x}\,y(\mathbf{x})]\mathbf{e}_j^\top\big) - \det(\gamma_n\boldsymbol{\Sigma}_\mu) \\
&= \det(\gamma_n\boldsymbol{\Sigma}_\mu)\cdot\gamma_n\mathbf{e}_j^\top(\gamma_n\boldsymbol{\Sigma}_\mu)^{-1}\mathbb{E}_{\mu,y}\big[y(\mathbf{x})\,\mathbf{x}\big],
\end{aligned}$$

where $(*)$ uses Lemma 13 twice (the first time, with $\mathbf{A} = \mathbf{X}$ and $\mathbf{B} = \mathbf{X} + \mathbf{y}\mathbf{e}_j^\top$). Dividing both sides by $Z_\mu^n = \det(\gamma_n\boldsymbol{\Sigma}_\mu)$ concludes the proof. ■

We combine Lemmas 22 and 23 to obtain the proof of Theorem 3.

Proof of Theorem 3 The case of $n = d$ follows directly from Theorem 2.10 of Dereziński et al. (2019). Assume that $n < d$. Then we have $\gamma_n = \frac{1}{\lambda_n}$, so the result follows from Lemma 22. If $n > d$, then the result follows from Lemma 23. ■

Appendix E. Proof of Theorem 4

Once again, we break down the proof into under- and over-determined cases, starting with the former. Note that in this case we require that the covariance be equal to identity.

Lemma 24 *Let $\rho = n/d$, $\mathbf{X} \sim \mathcal{N}_{n,d}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{I}_d)$ and $y_i = y(\mathbf{x}_i)$ under Assumption 1. If $d > n+1$ then*

$$0 \;\le\; \frac{\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] - \mathcal{M}(\mathbf{I}, \mathbf{w}^*, \sigma^2, n)}{\mathcal{M}(\mathbf{I}, \mathbf{w}^*, \sigma^2, n)} \;\le\; \frac{1}{d} \cdot \frac{1}{1 - \rho - \frac{1}{d}} + 3\rho^d.$$

Proof We first recall the standard decomposition of $\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big]$:

$$\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] = \sigma^2 \mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^\top \mathbf{X})^\dagger\big)\big] + \mathbf{w}^{*\top}\big(\mathbf{I} - \mathbb{E}[\mathbf{X}^\dagger \mathbf{X}]\big)\mathbf{w}^*.$$

Since the rows of $\mathbf{X}$ are standard normal random variables, $\mathbf{X}\mathbf{X}^\top$ is an $n \times n$ Wishart random matrix with $d > n+1$ degrees of freedom. Using the formula for the mean of the Inverse-Wishart distribution, it follows that

$$\mathbb{E}\big[\mathrm{tr}((\mathbf{X}^\top \mathbf{X})^\dagger)\big] = \mathbb{E}\big[\mathrm{tr}((\mathbf{X}\mathbf{X}^\top)^{-1})\big] = \frac{n}{d - n - 1}.$$

Note that $\mathbf{X}^\dagger \mathbf{X}$ is a uniformly random projection matrix so by rotational symmetry it follows that

$$\mathbf{w}^{*\top} \mathbb{E}\big[\mathbf{X}^\dagger \mathbf{X}\big] \mathbf{w}^* = \mathbb{E}\big[\|\mathbf{X}^\dagger \mathbf{X} \mathbf{w}^*\|^2\big] = \frac{n}{d} \|\mathbf{w}^*\|^2.$$

Putting this together we obtain that

$$\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] = \frac{\sigma^2 n}{d - n - 1} + \|\mathbf{w}^*\|^2 \frac{d - n}{d}.$$

On the other hand, the surrogate MSE expression can be derived by observing that for $\mathbf{\Sigma} = \mathbf{I}$ we have $\mathrm{tr}((\mathbf{\Sigma} + \lambda_n \mathbf{I})^{-1}) = d/(1 + \lambda_n) = n$ (see definition of λ_n in Theorem 1):

$$\mathcal{M}(\mathbf{I}, \mathbf{w}^*, \sigma^2, n) = \sigma^2 n \cdot \frac{1 - \alpha_n}{d - n} + \|\mathbf{w}^*\|^2 \frac{d - n}{d}.$$

Note that the second term is the same in both cases, even though this may not be true for non-isotropic Gaussians. We now compute the normalized difference between the expressions,

$$\begin{aligned}
\frac{\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] - \mathcal{M}(\mathbf{I}, \mathbf{w}^*, \sigma^2, n)}{\mathcal{M}(\mathbf{I}, \mathbf{w}^*, \sigma^2, n)} &= \frac{\sigma^2 n \cdot \big(\frac{1}{d-n-1} - \frac{1-\alpha_n}{d-n}\big)}{\sigma^2 n \cdot \frac{1-\alpha_n}{d-n} + \|\mathbf{w}^*\|^2 \frac{d-n}{d}} \\
&\le \frac{d-n}{1-\alpha_n}\Big(\frac{1}{d-n-1} - \frac{1-\alpha_n}{d-n}\Big) \\
&= \frac{d-n}{d-n-1}\frac{1}{1-\alpha_n} - 1 \\
&= \frac{1}{d-n-1} + \frac{\alpha_n}{1-\alpha_n}\Big(1 + \frac{1}{d-n-1}\Big).
\end{aligned}$$

Let $\rho = n/d$. Then $\frac{1}{d-n-1} = \frac{1}{d} \cdot \frac{1}{1-\rho-\frac{1}{d}}$ and moreover $\alpha_n = (\frac{1}{1+\lambda_n})^d = (\frac{n}{d})^d = \rho^d$. From the assumption that $d > n+1$, we conclude that $\alpha_n \le (\frac{d-2}{d})^d \le \mathrm{e}^{-2}$ so that $\frac{\alpha_n}{1-\alpha_n}(1 + \frac{1}{d-n-1}) \le \frac{2\alpha_n}{1-\mathrm{e}^{-2}} \le 3\rho^d$. This shows the right-hand-side inequality of the theorem. That fact that $\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] \ge \mathcal{M}(\mathbf{I}, \mathbf{w}^*, \sigma^2, n)$ follows easily. ∎

Lemma 25 *Let $\rho = n/d$, $\mathbf{X} \sim \mathcal{N}_{n,d}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma})$ and $y_i = y(\mathbf{x}_i)$ under Assumption 1. If $n > d+1$ then*

$$0 \;\le\; \frac{\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] - \mathcal{M}(\mathbf{\Sigma}, \mathbf{w}^*, \sigma^2, n)}{\mathcal{M}(\mathbf{\Sigma}, \mathbf{w}^*, \sigma^2, n)} \;\le\; \frac{1}{d} \cdot \frac{1}{\rho - 1 - \frac{1}{d}} + 3(\mathrm{e}^{1-\rho})^d.$$

Proof The MSE for the over-determined Gaussian design can be derived by using the formula for the mean of the Inverse-Wishart distribution:

$$\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] = \sigma^2 \mathrm{tr}\big(\mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}]\big) = \frac{\sigma^2 \mathrm{tr}(\mathbf{\Sigma}^{-1})}{n - d - 1}.$$

To compute the normalized difference we follow similar derivations as in the proof of Lemma 24:

$$\begin{aligned}\frac{\mathrm{MSE}\big[\mathbf{X}^\dagger \mathbf{y}\big] - \mathcal{M}(\mathbf{\Sigma}, \mathbf{w}^*, \sigma^2, n)}{\mathcal{M}(\mathbf{\Sigma}, \mathbf{w}^*, \sigma^2, n)} &= \frac{n-d}{1-\beta_n}\Big(\frac{1}{n-d-1} - \frac{1-\beta_n}{n-d}\Big) \\ &\le \frac{1}{d} \cdot \frac{1}{\rho - 1 - \frac{1}{d}} + \frac{2\beta_n}{1-\beta_n}.\end{aligned}$$

Recall that $\beta_n = \mathrm{e}^{d-n} = (\mathrm{e}^{1-\rho})^d$ and for $n - d \ge 2$ we have $\frac{2}{1-\beta_n} \le 3$. The desired inequalities follow immediately. ∎

Theorem 4 now follows as a consequence of Lemmas 24 and 25.

Proof of Theorem 4 Since $\frac{1}{d} \le \frac{1}{3}|1-\rho|$ and $\mathrm{e}^{-|1-\rho|d} \le \frac{1}{2d|1-\rho|}$, it follows that

$$\frac{1}{d} \cdot \frac{1}{|1-\rho| - \frac{1}{d}} + 3(\mathrm{e}^{-|1-\rho|})^d \le \frac{1}{d} \cdot \frac{1}{|1-\rho| - \frac{1}{3}|1-\rho|} + \frac{1}{d} \cdot \frac{3}{2|1-\rho|} = \frac{c_\rho}{d}.$$

The case of $n > d+1$ now follows from Lemma 25. Also, since $\rho \le \mathrm{e}^{\rho-1}$, for $n < d-1$ we have $3\rho^d \le 3(\mathrm{e}^{-|1-\rho|})^d$, so the same bound follows from Lemma 24. ∎

Appendix F. Empirical analysis of the asymptotic consistency conjectures

Conjectures 15 and 16 are related to open problems which have been extensively studied in the literature. With respect to Conjecture 15, Srivastava (2003) first derived the probability density function of the pseudo-Wishart distribution (also called the singular Wishart), and Cook and Forzani (2011) computed the first and second moments of generalized inverses for the distribution. However, for the Moore-Penrose inverse and arbitrary covariance $\mathbf{\Sigma}$, Cook and Forzani (2011) claims that the quantities required to express the mean "do not have tractable closed-form representation." Conjecture 16 has connections to directional statistics. Using the SVD, we have the equivalent representation $\mathbf{X}^\dagger \mathbf{X} = \mathbf{V}\mathbf{V}^\top$ where $\mathbf{V}$ is an element of the Stiefel manifold $V_{n,d}$ (i.e., orthonormal n-frames in $\mathbb{R}^d$). The distribution of $\mathbf{V}$ is known as the matrix angular central Gaussian (MACG) distribution (Chikuse, 1990). While prior work has considered high dimensional limit theorems (Chikuse, 1991) as well as density estimation and hypothesis testing (Chikuse, 1998) on $V_{n,d}$, they only analyzed the invariant measure (which corresponds in our setting to $\mathbf{\Sigma} = \mathbf{I}$), and to our knowledge a closed form expression of $\mathbb{E}[\mathbf{V}\mathbf{V}^\top]$ where $\mathbf{V}$ is distributed according to MACG with arbitrary $\mathbf{\Sigma}$ remains an open question.

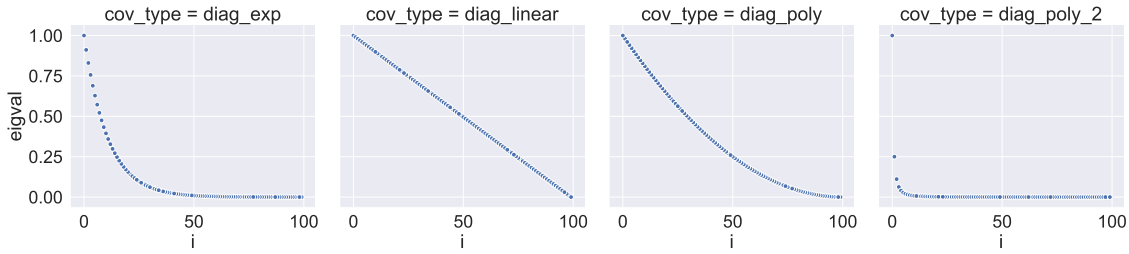


Figure 3: Scree-plots of $\mathbf{\Sigma}$ for the eigenvalue decays examined in our empirical valuations. Here $d = 100$ for visualization, whereas our experiments increase d while preserving the ratio n/d and the decay profile, with $\lambda_{\max}(\mathbf{\Sigma}) = 1$ to $\lambda_{\min}(\mathbf{\Sigma}) = 10^{-4}$.

For verifying these two conjectures, it suffices to only consider diagonal covariance matrices $\mathbf{\Sigma}$. This is because if $\mathbf{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\top$ is its eigendecomposition and $\mathbf{X} \sim \mathcal{N}_{n,d}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{Q}\mathbf{D}\mathbf{Q}^\top)$, then we have for $\mathbf{W} \sim \mathcal{PW}(\mathbf{\Sigma}, n)$ that $\mathbf{W} \overset{d}{=} \mathbf{X}^\top\mathbf{X}$ and hence, defining $\widetilde{\mathbf{X}} \sim \mathcal{N}_{n,d}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{D})$, by linearity and unitary invariance of trace,

$$\mathbb{E}[\mathrm{tr}(\mathbf{W}^\dagger)] = \mathrm{tr}\big(\mathbb{E}[(\mathbf{X}^\top\mathbf{X})^\dagger]\big) = \mathrm{tr}\Big(\mathbf{Q}\mathbb{E}\big[(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^\dagger\big]\mathbf{Q}^\top\Big) = \mathrm{tr}\Big(\mathbb{E}\big[(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^\dagger\big]\Big) = \mathbb{E}\left[\mathrm{tr}\big((\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}})^\dagger\big)\right].$$

Similarly, we have that $\mathbb{E}[\mathbf{X}^\dagger\mathbf{X}] = \mathbf{Q}\mathbb{E}\big[\widetilde{\mathbf{X}}^\dagger\widetilde{\mathbf{X}}\big]\mathbf{Q}^\top$, and a simple calculation shows that the expression in Conjecture 16 is also independent of the choice of matrix $\mathbf{Q}$.

Letting $\lambda_i(\mathbf{\Sigma})$ be the ith largest eigenvalue of $\mathbf{\Sigma}$, we consider the following eigenvalue profiles:

- `diag_linear`: linear decay, $\lambda_i(\mathbf{\Sigma}) = b - ai$;
- `diag_exp`: exponential decay, $\lambda_i(\mathbf{\Sigma}) = b\,10^{-ai}$;
- `diag_poly`: fixed-degree polynomial decay, $\lambda_i(\mathbf{\Sigma}) = (b - ai)^2$;
- `diag_poly_2`: variable-degree polynomial decay, $\lambda_i(\mathbf{\Sigma}) = bi^{-a}$.

The constants a and b are chosen to ensure $\lambda_{\max}(\mathbf{\Sigma}) = 1$ and $\lambda_{\min}(\mathbf{\Sigma}) = 10^{-4}$ (i.e., the condition number $\kappa(\mathbf{\Sigma}) = 10^4$ remains constant).

We verify our conjectures by increasing d while keeping the aspect ratio n/d fixed and examining the rate of decay of the quantities asserted in the conjectures. As no closed form expressions are available for the expectations in the conjectures, we estimate $\mathbb{E}\big[\mathrm{tr}(\mathbf{W}^\dagger)\big]$ (for Conjecture 15) and $\mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]$ (for Conjecture 16) through Monte Carlo sampling. To ensure that estimation noise is sufficiently small, we continually increase the number of Monte Carlo samples until the bootstrap confidence intervals are within $\pm 12.5\%$ of the quantities in (4) and (5). We found that while Conjecture 15 required a relatively small number of trials (up to one thousand), estimation noise was much larger in Conjecture 16 and necessitated over two million trials to obtain good estimates near $d = 100$.

The results of empirically validating Conjecture 15 are illustrated in Figure 4 (top), where we performed Monte Carlo estimation of $\mathbb{E}\big[\mathrm{tr}(\mathbf{W}^\dagger)\big]$ and plot $\big|\mathbb{E}[\mathrm{tr}(\mathbf{W}^\dagger)]\,\mathcal{V}(\mathbf{\Sigma}, n)^{-1} - 1\big|$ as d increases from 10 to 1000, across a range of aspect ratios n/d and eigenvalue decay profiles for $\mathbf{\Sigma}$.

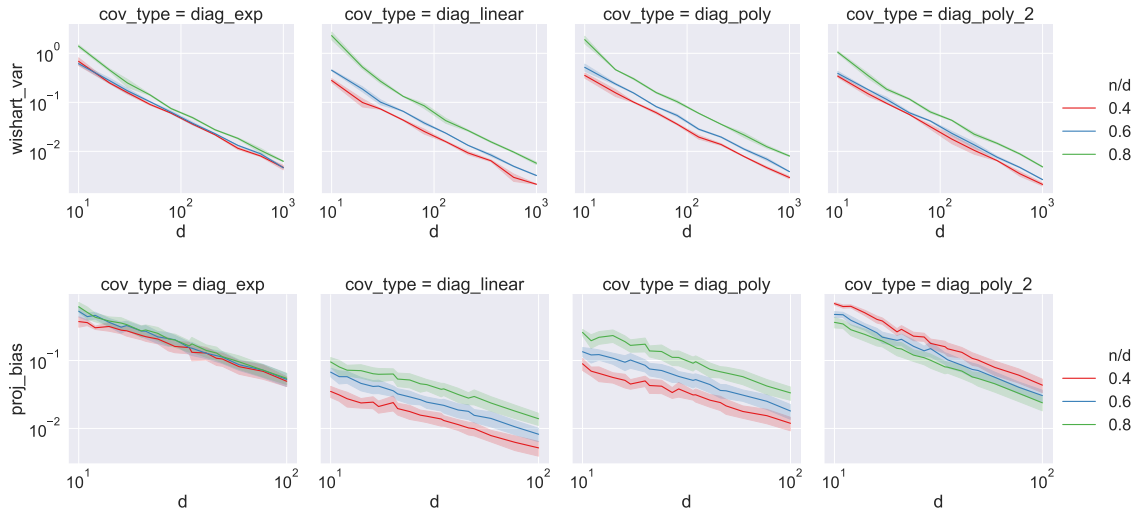


Figure 4: Empirical verification of Conjectures 15 (top) and 16 (bottom). We show the quantities $|\mathbb{E}[\mathrm{tr}(\mathbf{W}^\dagger)]\,\mathcal{V}(\boldsymbol{\Sigma}, n)^{-1} - 1|$ and $\|\mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]\mathcal{B}(\boldsymbol{\Sigma}, n)^{-1} - \mathbf{I}\|$, respectively, as d increases for various aspect ratios n/d. Consistent with our conjectures, a $O(1/d)$ decay (linear with slope -1 on a log-log plot) is exhibited across all eigenvalue decay profiles and aspect ratios investigated.

Confidence intervals are estimated by bootstrapping. We observe that on log-log axes all of the plots are decreasing with a linear -1 slope, consistent with the $O(1/d)$ rate predicted by Conjecture 15.

Conjecture 16 is handled similarly, by sampling $\mathbf{X} \sim \mu^n$ where $\mu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ to obtain a Monte Carlo estimate of $\mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]$. To handle the supremum over $\mathbf{w}$, notice that we can rewrite (5) as a spectral norm

$$\sup_{\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \left| \frac{\mathbf{w}^\top \mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]\mathbf{w}}{\mathbf{w}^\top \mathcal{B}(\boldsymbol{\Sigma}, n)\mathbf{w}} - 1 \right| \;=\; \left\| \mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]\mathcal{B}(\boldsymbol{\Sigma}, n)^{-1} - \mathbf{I} \right\|. \tag{8}$$

Confidence intervals can now be constructed using existing methods for constructing operator norm confidence intervals, and our results use the bootstrapping method described by Lopes et al. (2019).

Figure 4 (bottom) shows how $\left\|\mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]\mathcal{B}(\boldsymbol{\Sigma}, n)^{-1} - \mathbf{I}\right\|$ decays as we hold the aspect ratio n/d fixed and increase d between 10 and 100 across the listed eigenvalue decay profiles and aspect ratios. Again, we observe on log-log axes a linear decay with slope -1 consistent with the $O(1/d)$ rate posed by Conjecture 16. Note that the range of d is smaller than in Figure 4 (top) because the large number of Monte Carlo samples (up to two million) required for this experiment made the computations much more expensive.