Subspace Fitting Meets Regression: The Effects of Supervision and Orthonormality Constraints on Double Descent of Generalization Errors

Yehuda Dar [1] **Paul Mayer** [1] **Lorenzo Luzi** [1] **Richard G. Baraniuk** [1]

Abstract

We study the linear subspace fitting problem in the *overparameterized* setting, where the estimated subspace can perfectly *interpolate* the training examples. Our scope includes the least-squares solutions to subspace fitting tasks with varying levels of supervision in the training data (i.e., the proportion of input-output examples of the desired low-dimensional mapping) and orthonormality of the vectors defining the learned operator. This flexible family of problems connects standard, unsupervised subspace fitting that enforces strict orthonormality with a corresponding regression task that is fully supervised and does not constrain the linear operator structure. This class of problems is defined over a *supervision-orthonormality plane*, where each coordinate induces a problem instance with a unique pair of supervision level and softness of orthonormality constraints. We explore this plane and show that the generalization errors of the corresponding subspace fitting problems follow *double descent* trends as the settings become more supervised and less orthonormally constrained.

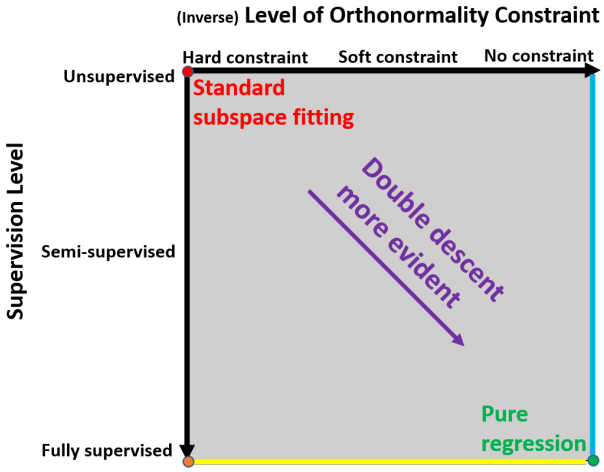


Figure 1. The supervision-orthonormality plane of subspace fitting problems.

1. Introduction

Learning processes are naturally limited by the amount of data available for making inferences according to the chosen model. In particular, the interplay between the number of training examples and the complexity of the model (the number of parameters) is fundamental to successful learning in terms of generalization ability.

The classical problem of linear regression, where one learns a linear mapping from a given set of input-output pairs, has been addressed for many years from the bias-variance tradeoff perspective. This design approach requires the number of parameters of the learned mapping to be sufficiently high, to avoid errors due to model bias, yet sufficiently low, to prevent overfitting to the training data. The established guideline (Breiman & Freedman, 1983) is that highly parameterized models, which lead to very low (or even zero) training error, are bad design choices that result in poor generalization performance.

The incredible success of highly overparameterized, deep neural networks has recently revived scientific interest in understanding the generalization errors induced by overparameterized models that are learned without explicit regularization mechanisms. One such prominent research line in (Spigler et al., 2018; Geiger et al., 2019; Belkin et al., 2019a) shows that the generalization error actually decreases as the learned model is more overparameterized, even though all such models perfectly interpolate the training data (i.e., have zero training error). This generalization-error behavior (as a function of the number of model parameters) has been termed *double descent*, due to the second decrease in the generalization error after entering the range of interpolating models. This finding has motivated an impressive series

[1]ECE Department, Rice University, Houston, TX, USA. Correspondence to: Yehuda Dar <ydar@rice.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

of mathematical studies, e.g., (Belkin et al., 2019b; Hastie et al., 2019; Xu & Hsu, 2019; Mei & Montanari, 2019) that formulate the double-descent phenomenon for the least-squares forms of various linear regression problems.

In this paper, we extend the study of overparameterized models to the realm of *dimensionality reduction*. We begin by considering the standard *subspace fitting problem*, where one estimates an underlying linear operator (in the form of a matrix with orthonormal columns) that generates a given set of noisy examples. For this unsupervised subspace fitting problem, we define and explore the meaning of interpolating solutions and their generalization errors. We show that, while overparameterization is beneficial, the generalization error follows a *single descent* trend throughout the entire range of parameterization levels, differing from the double descent shape that appears in (fully supervised) regression problems.

Pushing further, *we bridge the tasks of subspace fitting and regression* using a flexible optimization framework that generates a family of learning problems, with member of the family aiming to recover the same underlying subspace under a different setting. Specifically, we develop a general problem structure with two adjustable aspects. The first is the *supervision level*, referring to the relative proportion between input-output and input-only examples given for learning. This essentially covers the range of problems from unsupervised, through semi-supervised, to fully supervised. The second adjustable aspect involves the structure of the learned linear operator that characterizes the fitted subspace, specifically, the *degree of orthonormality* required of the columns of the estimated matrix. This provides a continuum of optimization forms, from unconstrained to strictly constrained, including intermediate settings with soft constraints.

We interpret this entire class of problems as residing over a *supervision-orthonormality plane* (see Fig. 1), where each coordinate instantiates a distinct problem with its own coupled levels of supervision and orthonormality constraints. The two extreme, diagonal corners of the plane correspond to the standard subspace fitting and regression problems.

Since the non-standard problems on the supervision-orthonormality plane do not have closed-form solutions, we explore them by developing iterative optimization procedures based on the projected gradient descent (PGD) technique. Interestingly, the soft orthonormality constraints reduce to thresholding operations applied to the singular values of the evolving solutions. Equipped with these PGD-based solutions, we empirically explore the generalization errors induced by the subspace estimation settings throughout the supervision-orthonormality plane. *Our results clearly demonstrate that the double-descent phenomenon emerges as the problems become increasingly supervised and less orthonormally constrained.*

1.1. Related Work

As explained above, our study directly relates to the recent research line on the double descent phenomenon (Belkin et al., 2019b; Hastie et al., 2019; Xu & Hsu, 2019; Mei & Montanari, 2019). In addition, our study includes learning problems in settings that may resemble concepts available in the existing literature described next.

In Section 3 we address the linear subspace fitting problem via principal component analysis (PCA) that considers only a subset of coordinates of the given data vectors (as this design enables us to determine the number of parameters in the learned model, see Section 2.2). Interestingly, the study of dimensionality reduction mechanisms applied on a subset of the available input variables (or features) dates back to (Jolliffe, 1972; 1973), where PCA was improved and/or made more computationally efficient. The approach of variable selection was developed further into sparse variable PCA methods, e.g., the transform-based preprocessing (Johnstone & Lu, 2009) and expectation-maximization based approach (Ulfarsson & Solo, 2011). These motivated corresponding studies of PCA in overparameterized settings under asymptotic assumptions, e.g., (Paul, 2007; Johnstone & Lu, 2009; Shen et al., 2016).

To motivate our supervised and semi-supervised settings in Sections 4 and 5, we refer to (Yang et al., 2006), where dimensionality reduction is improved using a subset of high-dimensional data examples and their corresponding exact low-dimensional representations. Beyond that, dimensionality reduction applied on multi-class data can be improved by supervised examples of class-labeled input data, e.g., (Sugiyama, 2006; Zhang et al., 2007; Nie et al., 2010).

1.2. Paper Outline

This paper is organized as follows. In Section 2, we describe the subspace fitting data model and its related definitions. In Section 3, we study the standard, unsupervised subspace fitting problem; this problem lies at the red point in the supervision-orthonormality plane in Fig. 1. In Section 4, we explore a range of fully supervised learning problems that aim to recover the underlying subspace; these problems reside along the yellow axis of the supervision-orthonormality plane in Fig. 1 and include the green point of pure, standard regression. In Section 5, we define a general optimization framework that supports any level of supervision and orthonormality constraint. This enables us to explore problems residing throughout the supervision-orthonormality plane. As two representative sets of problems, we evaluate the range of unconstrained settings (marked in blue in Fig. 1) and the diagonal trajectory connecting the standard subspace fitting with pure regression (the direction of the

purple arrow in Fig. 1). We conclude with a discussion of our findings in Section 6. All proofs plus additional experimental details are provided in the Appendices included in the Supplementary Materials.

2. Basic Settings

2.1. Data Model

Consider a vector $\mathbf{x} \in \mathbb{R}^d$ that satisfies a noisy linear model in the form of

$$\mathbf{x} = \mathbf{U}_m \mathbf{z} + \epsilon \tag{1}$$

where $\mathbf{U}_m \in \mathbb{R}^{d \times m}$ is a matrix consisting of $m < d$ orthonormal column vectors $\{\mathbf{u}^{(i)}\}_{i=1}^m \in \mathbb{R}^d$ that span a rank-m linear subspace. The underlying m coefficients, organized in $\mathbf{z} \in \mathbb{R}^m$, are independent and identically distributed (i.i.d.) and standard Gaussian: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, where $\mathbf{I}_m$ is the $m \times m$ identity matrix. The random vector $\epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_d\right)$, which is independent of $\mathbf{z}$, plays the role of a Gaussian noise vector in $\mathbb{R}^d$. Thus, $\mathbf{x}$ is zero mean with covariance matrix

$$\mathbf{C}_\mathbf{x} = \mathbf{U}_m \mathbf{U}_m^T + \sigma_\epsilon^2 \mathbf{I}_d. \tag{2}$$

The problems in this paper are defined for learning settings where the data model (1) is unknown, and only a dataset $\mathcal{D} \triangleq \left\{\mathbf{x}^{(\ell)}\right\}_{\ell=1}^n \in \mathbb{R}^d$ of n i.i.d. examples of data vectors satisfying (1) is available. The vectors in $\mathcal{D}$ are centered with respect to their sample mean. Note that $\mathcal{D}$, as defined here, enables unsupervised learning. In a later stage in the paper, where we discuss supervised and semi-supervised learning problems, the formulation of $\mathcal{D}$ will be extended.

2.2. Learning Mappings with Desired Parameterization Levels

Our interest is in learning tasks that infer mappings $f : \mathbb{R}^d \to \mathbb{R}^k$ to be applied on d-dimensional data and provide k-dimensional results, where $k < d$. The learned mapping is used for computing $f(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$ realizing (1) beyond the examples in $\mathcal{D}$. The common case of a very high dimension d induces complex and highly parameterized instances of f, which are usually more difficult to learn. This challenge can be addressed by simplifying the learned mapping using the following design. A single set $\mathcal{S}$ of p out of d coordinates is determined arbitrarily (i.e., without any adaptation to the data). Specifically, $\mathcal{S} = \{s_1, ..., s_p\}$, where $1 \le s_1 < s_2 < ... < s_p \le d$. The p-dimensional *feature vector* of $\mathbf{x} \in \mathbb{R}^d$ is defined as $\mathbf{x}_\mathcal{S} \triangleq \left[x_{s_1}, x_{s_2}, ..., x_{s_p}\right]^T$, where x_{s_j} is the s_j-th component of $\mathbf{x}$. Then, the overall mapping is defined as $f(\mathbf{x}) = f_\mathcal{S}(\mathbf{x}_\mathcal{S})$, where $f_\mathcal{S} : \mathbb{R}^p \to \mathbb{R}^k$ is a learned mapping that requires fewer parameters (than f) due to the lower dimension of its inputs. This simple approach lets us determine the actual number of parameters in the learned mappings by choosing the size of $\mathcal{S}$ (i.e., p).

We consider procedures that learn $f_\mathcal{S}$ using only p-dimensional subvectors, specified by $\mathcal{S}$, of the vectors in $\mathcal{D}$ (a similar approach was used in (Belkin et al., 2019b) for non-asymptotic analysis of linear regression). Accordingly, the dataset of the p-dimensional feature vectors used for the learning process is denoted by $\mathcal{D}_\mathcal{S} \triangleq \left\{\mathbf{x}_\mathcal{S}^{(\ell)}\right\}_{\ell=1}^n$.

3. Linear Subspace Fitting: The Standard, Unsupervised Setting

3.1. Problem Definition

The goal is to find the linear subspace of rank k that provides the best approximation ability, in the squared-error sense, of the data satisfying (1). Recall that m, the true rank of the underlying linear part in (1), is unknown and, hence, k is not necessarily equal to m. The subspace estimate is formed based on the dataset $\mathcal{D}_\mathcal{S}$ of p-dimensional feature vectors. Nevertheless, the desired representation ability is for d-dimensional vectors beyond the dataset $\mathcal{D}$, namely, the out-of-sample squared error with respect to the data model in (1) is the performance criterion of interest.

A simple approach to address the subspace fitting problem, for $k \le p$, in a way conforming to the guidelines given in Section 2.2, is as follows. The first stage is to define the linear subspace $\widehat{\mathcal{U}}_{k,\mathcal{S}}$ that has rank k and resides in $\mathbb{R}^p$, that minimizes the Euclidean distance between the data vectors in $\mathcal{D}_\mathcal{S}$ and their corresponding orthogonal projections onto the estimated subspace. Denote the k orthonormal vectors spanning $\widehat{\mathcal{U}}_{k,\mathcal{S}}$ by $\widehat{\mathbf{u}}_\mathcal{S}^{(1)}, \ldots, \widehat{\mathbf{u}}_\mathcal{S}^{(k)} \in \mathbb{R}^p$; organize them into the columns of a $p \times k$ matrix $\widehat{\mathbf{U}}_{k,\mathcal{S}} \triangleq \left[\widehat{\mathbf{u}}_\mathcal{S}^{(1)}, \ldots, \widehat{\mathbf{u}}_\mathcal{S}^{(k)}\right]$. Note that, for $k < p$, $\widehat{\mathbf{U}}_{k,\mathcal{S}}^T \widehat{\mathbf{U}}_{k,\mathcal{S}} = \mathbf{I}_k$, whereas $\widehat{\mathbf{U}}_{k,\mathcal{S}} \widehat{\mathbf{U}}_{k,\mathcal{S}}^T \neq \mathbf{I}_p$. Then, the closest point in $\widehat{\mathcal{U}}_{k,\mathcal{S}}$ to an arbitrary vector $\mathbf{v} \in \mathbb{R}^p$ is $\widehat{\mathbf{v}} = \widehat{\mathbf{U}}_{k,\mathcal{S}} \widehat{\mathbf{U}}_{k,\mathcal{S}}^T \mathbf{v}$. This produces the standard form of the subspace fitting problem, namely,

$$\begin{aligned}\widehat{\mathbf{U}}_{k,\mathcal{S}} &= \underset{\mathbf{W} \in \mathbb{R}^{p \times k}:\ \mathbf{W}^T\mathbf{W} = \mathbf{I}_k}{\arg\min} \frac{1}{n} \sum_{\ell=1}^n \left\| \left(\mathbf{I}_p - \mathbf{W}\mathbf{W}^T\right) \mathbf{x}_\mathcal{S}^{(\ell)} \right\|_2^2 \\ &= \underset{\mathbf{W} \in \mathbb{R}^{p \times k}:\ \mathbf{W}^T\mathbf{W} = \mathbf{I}_k}{\arg\min} \frac{1}{n} \left\| \left(\mathbf{I}_p - \mathbf{W}\mathbf{W}^T\right) \mathbf{X}_\mathcal{S} \right\|_F^2\end{aligned}$$

where $\mathbf{X}_\mathcal{S} \triangleq \left[\mathbf{x}_\mathcal{S}^{(1)}, \ldots, \mathbf{x}_\mathcal{S}^{(n)}\right] \in \mathbb{R}^{p \times n}$ is the data matrix having the examples in $\mathcal{D}_\mathcal{S}$ as its columns. As is commonly known, the last optimization form is equivalent to

$$\widehat{\mathbf{U}}_{k,\mathcal{S}} = \underset{\mathbf{W} \in \mathbb{R}^{p \times k}:\ \mathbf{W}^T\mathbf{W} = \mathbf{I}_k}{\arg\max} \operatorname{Tr}\left\{\mathbf{W}^T \mathbf{X}_\mathcal{S} \mathbf{X}_\mathcal{S}^T \mathbf{W}\right\}, \tag{3}$$

which can be solved via a principal component analysis (PCA) procedure. Specifically, the orthonormal columns of

$\widehat{\mathbf{U}}_{k,\mathcal{S}}$ are the eigenvectors corresponding to the first k principal components of the sample covariance matrix induced by $\mathcal{D}_{\mathcal{S}}$.

The learned rank-k subspace $\widehat{\mathcal{U}}_{k,\mathcal{S}} \subset \mathbb{R}^p$ is extended to a rank-k subspace $\widehat{\mathcal{U}}_k$ that resides in $\mathbb{R}^d$ and is spanned by k orthonormal vectors, denoted as $\widehat{\mathbf{u}}^{(1)}, \ldots, \widehat{\mathbf{u}}^{(k)} \in \mathbb{R}^d$. The suggested construction defines $\widehat{\mathbf{u}}^{(j)}$ (for $j = 1, ..., k$) such that its subvector corresponding to its coordinates in $\mathcal{S}$ is the learned $\widehat{\mathbf{u}}^{(j)}_{\mathcal{S}}$, and the rest of its $d - p$ components are zeros. Organizing these orthonormal vectors as the columns of a $d \times k$ matrix $\widehat{\mathbf{U}}_k \triangleq \left[\widehat{\mathbf{u}}^{(1)}, \ldots, \widehat{\mathbf{u}}^{(k)}\right]$ provides a linear operator that, as required, creates k-dimensional representations for d-dimensional inputs. Namely,

$$\widehat{\mathbf{v}} = \widehat{\mathbf{U}}_k^T \mathbf{x} \tag{4}$$

for $\mathbf{x} \in \mathbb{R}^d$ satisfying the data model (1).

Consider the case of $k = m$ and note that the unsupervised learning is defined to minimize d-dimensional reconstruction errors and, therefore, the columns of $\widehat{\mathbf{U}}_m$ do not necessarily match in their indices to their closest columns of the true matrix $\mathbf{U}_m$. Hence, the vector $\widehat{\mathbf{v}}$ is not a straightforward estimate of the underlying $\mathbf{z} \in \mathbb{R}^m$ that generates the given $\mathbf{x}$. This leads to the test error evaluation metric that is described next.

While $\widehat{\mathbf{U}}_k$ is optimized to approximate the given sample $\mathcal{D}_{\mathcal{S}}$, the real interest is in representing arbitrary realizations of the model in (1). Hence, the quality of $\widehat{\mathbf{U}}_k$ should be evaluated for test data, $\mathbf{x}_{\text{test}} \in \mathbb{R}^d$, randomly drawn from the probability distribution $P_{\mathbf{x}}$ induced by (1). This provides the out-of-sample error of interest

$$\begin{aligned} \mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right) &\triangleq \mathbb{E}\left\|\left(\mathbf{I}_d - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\right)\mathbf{x}_{\text{test}}\right\|_2^2 \\ &= \text{Tr}\left\{\left(\mathbf{I}_d - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\right)\mathbf{C}_{\mathbf{x}}\left(\mathbf{I}_d - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\right)^T\right\} \end{aligned} \tag{5}$$

where the expectation is for $\mathbf{x}_{\text{test}} \sim P_{\mathbf{x}}$, and $\mathbf{C}_{\mathbf{x}}$ is the covariance matrix from (2). Naturally, the formula for $\mathcal{E}^{\text{unsup}}_{\text{out}}$ has an empirical counterpart defined for a set of test data vectors.

Another metric useful for studying properties of learned subpaces is the in-sample approximation error of $\mathcal{D}$

$$\begin{aligned} &\mathcal{E}^{\text{unsup}}_{\text{in}}\left(\widehat{\mathbf{U}}_k\right) \triangleq \\ &\text{Tr}\left\{\left(\mathbf{I}_d - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\right)\widehat{\mathbf{C}}^{(n)}_{\mathbf{x}}\left(\mathbf{I}_d - \widehat{\mathbf{U}}_k\widehat{\mathbf{U}}_k^T\right)^T\right\}. \end{aligned} \tag{6}$$

Here $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x}} \triangleq \frac{1}{n}\mathbf{X}\mathbf{X}^T$ is the $d \times d$ sample covariance matrix corresponding to the n examples provided in $\mathcal{D}$ (recall that the data is centered).

Since the actual learning in the proposed construction of $\widehat{\mathbf{U}}_k$ involves an actual learning only with respect to $\mathcal{D}_{\mathcal{S}}$, we define an additional in-sample approximation error as

$$\begin{aligned} &\mathcal{E}^{\text{unsup}}_{\text{in},\mathcal{S}}\left(\widehat{\mathbf{U}}_{k,\mathcal{S}}\right) \triangleq \\ &\text{Tr}\left\{\left(\mathbf{I}_p - \widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}_{k,\mathcal{S}}^T\right)\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}\left(\mathbf{I}_p - \widehat{\mathbf{U}}_{k,\mathcal{S}}\widehat{\mathbf{U}}_{k,\mathcal{S}}^T\right)^T\right\} \end{aligned} \tag{7}$$

where $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}} \triangleq \frac{1}{n}\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^T$ is a $p \times p$ sample-covariance matrix corresponding to $\mathcal{D}_{\mathcal{S}}$. Note that

$$\mathcal{E}^{\text{unsup}}_{\text{in}}\left(\widehat{\mathbf{U}}_k\right) = \mathcal{E}^{\text{unsup}}_{\text{in},\mathcal{S}}\left(\widehat{\mathbf{U}}_{k,\mathcal{S}}\right) + \frac{1}{n}\left\|\mathbf{X}_{\mathcal{S}_c}\right\|_F^2 \tag{8}$$

where $\mathcal{S}_c \triangleq \{1, ..., d\} \setminus \mathcal{S}$ is the subset of coordinates excluded from the actual learning process, and $\mathbf{X}_{\mathcal{S}_c} \triangleq \left[\mathbf{x}^{(1)}_{\mathcal{S}_c}, \ldots, \mathbf{x}^{(n)}_{\mathcal{S}_c}\right] \in \mathbb{R}^{(d-p)\times n}$ includes the corresponding subvectors from the dataset as its columns. Accordingly, the term $\|\mathbf{X}_{\mathcal{S}_c}\|_F^2$ in (8) is a quantity stemming from $\mathcal{S}$ and the number of parameters p, but independent of the specific subspace estimate.

3.2. Interpolating Subspaces

We now turn to define two central concepts in our analysis.

Definition 3.1. A subspace estimate $\widehat{\mathcal{U}}_k$, constructed based on the learning of $\widehat{\mathcal{U}}_{k,\mathcal{S}}$, is *$\mathcal{S}$-interpolating* if $\mathcal{E}^{\text{unsup}}_{\text{in},\mathcal{S}}\left(\widehat{\mathbf{U}}_{k,\mathcal{S}}\right) = 0$.

That is, an $\mathcal{S}$-interpolating subspace is able to perfectly represent the information embodied in $\mathcal{D}_{\mathcal{S}}$.

Definition 3.2. A subspace estimate $\widehat{\mathcal{U}}_k$, constructed based on learning $\widehat{\mathcal{U}}_{k,\mathcal{S}}$, is *overparameterized* if $p \in \{n+1, ..., d\}$ and *rank-overparameterized* if $p \in \{n+1, ..., d\}$ and $k \in \{n, ..., p\}$.

Remark 3.1. *A rank-overparameterized estimate of a subspace is also overparameterized.*

Recall that $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$ is a $p \times p$ matrix constructed from n centered samples.

Corollary 3.1. *An overparameterized subspace estimate $\widehat{\mathcal{U}}_k$ is formed based on a rank-deficient sample covariance matrix $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$ of rank $\rho \triangleq rank\left\{\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}\right\} \leq n - 1$. If the subspace estimate $\widehat{\mathcal{U}}_k$ is also rank-overparameterized, then the rank-deficiency of $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$ affects $\widehat{\mathcal{U}}_k$.*

Corollary 3.2. *A rank-overparameterized subspace estimate (of rank k) is spanned by the ρ eigenvectors of $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$ corresponding to all the nonzero eigenvalues. The additional $k - \rho$ orthonormal vectors can be arbitrarily chosen from the $p - \rho$ eigenvectors of $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$ that match to its zero eigenvalues.*

Remark 3.2. *Corollary 3.2 provides a suggested construction for a rank-overparameterized subspace estimate. In*

general, the additional $k-\rho$ orthonormal vectors defined above can be any set spanning a rank-$(k-\rho)$ subspace of the null space of the sample covariance $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$.

This means that the PCA procedure required for solving (3) reduces to a significantly simpler task. The following is proved in Appendix A.

Proposition 3.1. *A rank-overparameterized subspace estimate is also an $\mathcal{S}$-interpolating subspace.*

3.3. Generalization Error vs. Parameterization Level

We now turn to characterize the benefits of overparameterized solutions to the unsupervised subspace fitting problem.

Proposition 3.2. *The out-of-sample error (5) can be expressed as*

$$\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)=\sum_{i=1}^{d}\lambda^{(i)}-\sum_{i\in\widehat{\mathcal{S}}^{(k)}_{\max}}\sum_{j=1}^{p}\lambda^{(j)}_{\mathcal{S}}\left|\left\langle\boldsymbol{\psi}^{(j)}_{\mathcal{S}},\widehat{\boldsymbol{\psi}}^{(i)}_{\mathcal{S}}\right\rangle\right|^2 \tag{9}$$

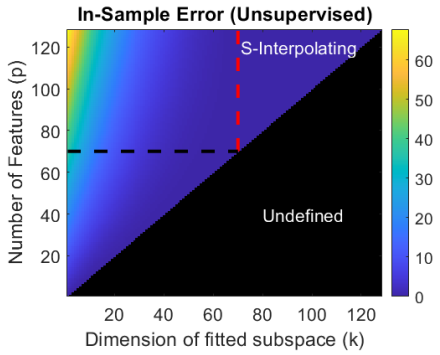
where $\lambda^{(i)}$ is the i^{th} eigenvalue of $\mathbf{C}_{\mathbf{x}}$, the eigenvalues $\left\{\lambda^{(j)}_{\mathcal{S}}\right\}_{j=1}^{p}$ and eigenvectors $\left\{\boldsymbol{\psi}^{(j)}_{\mathcal{S}}\right\}_{j=1}^{p}$ correspond to the true covariance matrix of the p-dimensional feature vectors $\mathbf{C}_{\mathbf{x},\mathcal{S}}$, and $\widehat{\boldsymbol{\psi}}^{(j)}_{\mathcal{S}}$ is the j^{th} eigenvector of the sample covariance $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$. Also, $\widehat{\mathcal{S}}^{(k)}_{\max}$ is the set of indices corresponding to the k maximal eigenvalues of $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$.

Remark 3.3. *In case the subspace estimate is rank-overparameterized, then the definition of $\widehat{\mathcal{S}}^{(k)}_{\max}$ in Proposition 3.2 assumes the construction suggested in Corollary 3.2. This means that when $k>\rho$, the set $\hat{S}^{(k)}_{\max}$ includes $k-\rho$ indices that correspond to $k-\rho$ out of the $p-\rho$ zero eigenvalues of the sample covariance matrix $\widehat{\mathbf{C}}^{(n)}_{\mathbf{x},\mathcal{S}}$.*

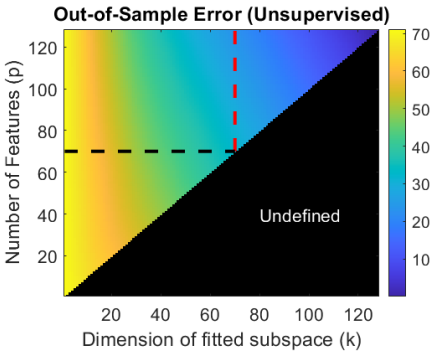
There are two axes along which to study how $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$ decays: along k and along p. For k, we can state the following (see the proof in Appendix A).

Proposition 3.3. *A subspace estimate induces an out-of-sample error $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$ that monotonically decreases as $k\in\{1,...,p\}$ increases and $\widehat{\mathbf{U}}_k$ is gradually extended.*

For p, the situation is more delicate. A rigorous proof has so far eluded us, possibly due to our non-asymptotic setting that hinders the important characterization of the sample covariance eigenvectors (e.g., as provided in the asymptotic frameworks in (Paul, 2007; Shen et al., 2016)). Yet, the results of extensive simulations indicate that, on average with respect to $\mathcal{S}$ that is uniformly chosen at random, $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$ decays monotonically in p as well (see Fig. 2b and the additional results provided in Appendix A).



(a)



(b)

Figure 2. Evaluation of unsupervised learning at various parameterization settings. (a) The in-sample errors, $\mathcal{E}^{\text{unsup}}_{\text{in}}\left(\widehat{\mathbf{U}}_k\right)$. (b) The out-of-sample errors, $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$. The border lines of the overparamaeterization and rank-overparameterization regions are marked with black and red dashed lines, respectively.

To summarize what we have learned so far, increased overparameterization and/or rank-overparameterization of unsupervised subspace estimates provide lower generalization errors. Moreover, the overall trend induced by increasing the number of features, p, significantly differs from the double-descent behavior arising in regression problems (see, e.g., (Belkin et al., 2019b)).

3.4. Empirical Demonstrations

We now present results for unsupervised learning settings, where $d=128$, $n=70$, and the $m=40$ columns of $\mathbf{U}_m$ are set as the first 40 normalized columns of the Hadamard matrix of order 128. Figure 2a shows the in-sample error, $\mathcal{E}^{\text{unsup}}_{\text{in},\mathcal{S}}\left(\widehat{\mathbf{U}}_{k,\mathcal{S}}\right)$, obtained for the various parameterization combinations of p and k (recall that $k\leq p$, and this is the reason for the undefined regions in Figs. 2a–2b). Figure 2b demonstrates the out-of-sample errors, $\mathcal{E}^{\text{unsup}}_{\text{out}}\left(\widehat{\mathbf{U}}_k\right)$, that are empirically evaluated using a test set of 1000 out-of-sample realizations of data vectors $\mathbf{x}$ satisfying (1). The border lines of the overparamaeterization and rank-

overparameterization regions are marked with black and red dashed lines, respectively. The monotonic decrease of the out-of-sample error with the increase in p and/or k is evident (see Fig. 2b). The fact that rank-overparameterization induces $\mathcal{S}$-interpolating subspace estimates is also visible in Fig. 2a.

4. Supervised Subspace Fitting

The previous section demonstrated the behavior of the generalization error with respect to the number of features p for the unsupervised subspace fitting setting. We now turn to define *fully supervised* forms that are related to the above defined problem (and reside along the bottom, yellow-colored border line of the supervision-orthonormality plane in Fig. 1). Our main goal is to study how the aspects of supervision and constraints affect the trends of generalization errors observed for the unsupervised setting.

The data model remains the same as in Section 2.1. The only exception, here, is that the provided dataset is $\mathcal{D}^{\text{sup}} \triangleq \left\{\left(\mathbf{x}^{(\ell)}, \mathbf{z}^{(\ell)}\right)\right\}_{\ell=1}^{n} \in \mathbb{R}^d \times \mathbb{R}^m$ of n i.i.d. samples of $(\mathbf{x}, \mathbf{z})$ pairs satisfying (1). Note that the examples given for the low-dimensional representations $\mathbf{z}$ reflect the true dimension of the linear subspace underlying the noisy data. Hence, the learning is to be defined for establishing a mapping that provides m-dimensional representations. This contrasts the unsupervised case, where m is unknown and, thus, the assumed low-dimension k is possibly incorrect.

4.1. Supervised Learning with Strict Orthonormality Constraints

This subsection examines the problem induced at the lower-left corner of the supervision-orthonormality plane (see orange-colored coordinate in Fig. 1). We employ the approach described in Section 2.2 for setting a parameterization level of interest. Again, the subset of p coordinates specified in $\mathcal{S}$ is used to subsample the $\mathbf{x}$ vectors, corresponding to the data elements that the learned mapping should be applied on. Note that the $\mathbf{z}$ vectors remain in their full forms. Accordingly, the dataset used for the supervised learning is $\mathcal{D}_{\mathcal{S}}^{\text{sup}} \triangleq \left\{\left(\mathbf{x}_{\mathcal{S}}^{(\ell)}, \mathbf{z}^{(\ell)}\right)\right\}_{\ell=1}^{n} \in \mathbb{R}^p \times \mathbb{R}^m$, where $p \geq m$. The optimization problem for establishing the orthonormal set of m vectors spanning the subspace is

$$\begin{aligned}\widehat{\mathbf{U}}_{m,\mathcal{S}} &= \underset{\mathbf{W}\in\mathbb{R}^{p\times m}:\ \mathbf{W}^T\mathbf{W}=\mathbf{I}_m}{\arg\min} \frac{1}{n}\left\|\mathbf{W}\mathbf{Z} - \mathbf{W}\mathbf{W}^T\mathbf{X}_{\mathcal{S}}\right\|_F^2 \\ &= \underset{\mathbf{W}\in\mathbb{R}^{p\times m}:\ \mathbf{W}^T\mathbf{W}=\mathbf{I}_m}{\arg\min} \frac{1}{n}\left\|\mathbf{Z} - \mathbf{W}^T\mathbf{X}_{\mathcal{S}}\right\|_F^2 \quad (10)\end{aligned}$$

where $\mathbf{X}_{\mathcal{S}} \triangleq \left[\mathbf{x}_{\mathcal{S}}^{(1)}, \ldots, \mathbf{x}_{\mathcal{S}}^{(n)}\right] \in \mathbb{R}^{p\times n}$ and $\mathbf{Z} \triangleq \left[\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)}\right] \in \mathbb{R}^{m\times n}$.

The optimization problem in (10) is related to the orthonormal Procrustes problem (Gower et al., 2004). However, here the optimization variable is a rectangular, instead of a square, matrix and therefore we do not have a closed-form solution. This motivates us to address (10) by a projected gradient descent approach (see Algorithm 1, where t is the iteration index, μ is the gradient step size, and T_{hard} is defined next).

Algorithm 1 Supervised Subspace Fitting via Projected Gradient Descent: Strict Orthonormality Constraints

Input: dataset $\mathcal{D}_{\mathcal{S}}^{\text{sup}} = \left\{\left(\mathbf{x}_{\mathcal{S}}^{(\ell)}, \mathbf{z}^{(\ell)}\right)\right\}_{\ell=1}^{n}$ and a coordinate subset $\mathcal{S}$
Initialize $\mathbf{W}^{(t=0)} = T_{\text{hard}}\left(\left(\mathbf{Z}\mathbf{X}_{\mathcal{S}}^{+}\right)^T\right)$, $t = 0$
repeat
  $t \leftarrow t + 1$
  $\mathbf{Y}^{(t)} = \mathbf{W}^{(t-1)} - \mu\mathbf{X}_{\mathcal{S}}\left(\left(\mathbf{W}^{(t-1)}\right)^T\mathbf{X}_{\mathcal{S}} - \mathbf{Z}\right)^T$
  $\mathbf{W}^{(t)} = T_{\text{hard}}\left(\mathbf{Y}^{(t)}\right)$
until stopping criterion is satisfied
Set $\widehat{\mathbf{U}}_{m,\mathcal{S}} = \mathbf{W}^{(t)}$
Create $\widehat{\mathbf{U}}_m$ based on $\widehat{\mathbf{U}}_{m,\mathcal{S}}$ and zeros at rows corresponding to $\mathcal{S}_c$
Output: $\widehat{\mathbf{U}}_m$

In this case, the constraint-projection stage reduces to an operator applied on the singular values of the evolving solution. Specifically, consider a matrix $\mathbf{W}^{(\text{in})} \in \mathbb{R}^{p\times m}$ (where $p \geq m$), with the SVD $\mathbf{W}^{(\text{in})} = \mathbf{\Omega}\mathbf{\Sigma}^{(\text{in})}\mathbf{\Theta}^T$, where $\mathbf{\Omega}$ and $\mathbf{\Theta}$ are $p \times p$ and $m \times m$ real orthonormal matrices, respectively, and $\mathbf{\Sigma}^{(\text{in})}$ is a $p \times m$ real diagonal matrix with m singular values $\{\sigma_i\left(\mathbf{W}^{(\text{in})}\right)\}_{i=1}^{m}$ on its main diagonal. Then, projecting $\mathbf{W}^{(\text{in})}$ onto the hard-orthonormality constraint via

$$\mathbf{W}^{(\text{out})} = \underset{\mathbf{W}\in\mathbb{R}^{p\times m}:\ \mathbf{W}^T\mathbf{W}=\mathbf{I}_m}{\arg\min} \left\|\mathbf{W} - \mathbf{W}^{(\text{in})}\right\|_F^2 \quad (11)$$

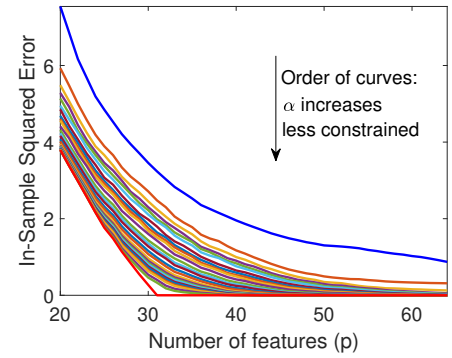
induces the mapping $\mathbf{W}^{(\text{out})} \triangleq T_{\text{hard}}\left(\mathbf{W}^{(\text{in})}\right)$, where $\mathbf{W}^{(\text{out})} = \mathbf{\Omega}\mathbf{\Sigma}^{(\text{out})}\mathbf{\Theta}^T$ and the singular values along the main diagonal of $\mathbf{\Sigma}^{(\text{out})}$ are $\sigma_i\left(\mathbf{W}^{(\text{out})}\right) = 1$ for $i = 1, \ldots, m$. See Appendix B for the proof.

Unlike the unsupervised settings in Section 3, the supervised learning procedures defined here provide estimates $\widehat{\mathbf{U}}_m$ that approximate the mapping from $\mathbf{x} \in \mathbb{R}^d$ to $\mathbf{z} \in \mathbb{R}^m$. This enables us to define the following *supervised* evaluation metrics, considering the in-sample squared error (with respect to the dataset $\mathcal{D}_{\mathcal{S}}^{\text{sup}}$)

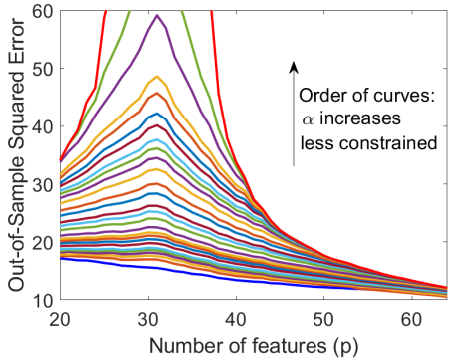
$$\mathcal{E}_{\text{in}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right) \triangleq \frac{1}{n}\sum_{\ell=1}^{n}\left\|\mathbf{z}^{(\ell)} - \widehat{\mathbf{U}}_m^T\mathbf{x}^{(\ell)}\right\|_2^2 \quad (12)$$

and the out-of-sample squared error

$$\mathcal{E}_{\text{out}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right) \triangleq \mathbb{E}\left\|\mathbf{z}_{\text{test}} - \widehat{\mathbf{U}}_m^T\mathbf{x}_{\text{test}}\right\|_2^2 \quad (13)$$



(a)



(b)

Figure 3. The **(a) in-sample** errors $\mathcal{E}_{\text{in}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ and **(b) out-of-sample** errors $\mathcal{E}_{\text{out}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ of fully-supervised learning versus the number of parameters p. The errors are **averaged** over 10 experiments with different sequential orders of adding coordinates to $\mathcal{S}$. Here $d = 64$, $m = 20$ and $n = 32$. Each curve presents the results for a different level α of orthonormality constraints. The results here correspond to problems located along the yellow-colored border line in Fig. 1. The colors of the curves in this figure are arbitrary and not related to the colors in Fig. 1.

where the expectation is over $(\mathbf{x}_{\text{test}}, \mathbf{z}_{\text{test}}) \sim P_{\mathbf{x},\mathbf{z}}$ as induced by (1).

Our results (see the bottom blue-colored curve of out-of-sample errors in Fig. 3b and Appendix B for more details) show that there is no double-descent behavior in this setting, despite the fact the learning is fully supervised. Moreover, the corresponding in-sample error curve (see the upper blue-colored curve in Fig. 3a) shows that, under strict orthonormality constraints, interpolation is not achieved, even not by solutions corresponding to $p > n$.

4.2. The Regression Approach: A Supervised, Unconstrained Setting

The problem defined in (10) recalls the usual regression form, except for the constraint on the matrix estimate. This motivates us to extend the range of problems we consider to include a standard regression problem for the purpose of estimating $\mathbf{U}_m$ without constraining its structure. This problem is located at the green coordinate in the corner of the supervision-orthonormality plane in Fig. 1. This setting is simply obtained by removing the constraint from (10), namely,

$$\widehat{\mathbf{U}}_{m,\mathcal{S}} = \underset{\mathbf{W}\in\mathbb{R}^{p\times m}}{\arg\min} \frac{1}{n}\left\|\mathbf{Z} - \mathbf{W}^T\mathbf{X}_{\mathcal{S}}\right\|_F^2 \tag{14}$$

which has a closed-form solution $\widehat{\mathbf{U}}_{m,\mathcal{S}} = \left(\mathbf{Z}\mathbf{X}_{\mathcal{S}}^{+}\right)^T$, where $\mathbf{X}_{\mathcal{S}}^{+}$ is the pseudoinverse of $\mathbf{X}_{\mathcal{S}}$. Similar to the previous settings, the matrix $\widehat{\mathbf{U}}_m$ is formed based on $\widehat{\mathbf{U}}_{m,\mathcal{S}}$ in addition to zeros at the rows corresponding to indices in $\mathcal{S}_c$. Again, the relevant evaluation metrics are $\mathcal{E}_{\text{in}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ and $\mathcal{E}_{\text{out}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ as defined in (12) and (13), respectively.

Note that in this setting, which does not include strict orthonormality constraints on the columns of $\widehat{\mathbf{U}}_m$, one can construct estimates also for $p < m$. However, since our scope includes also problems with strict or soft orthonormality constraints, all the results in this paper are presented only for $p \geq m$.

Our results (see the upper red-colored curve in Fig. 3b and Appendix B for more details) demonstrate that the generalization error follows a double-descent behavior. Note that the "first descent" in the underparameterized range is missing due to the constructions from Section 2.2 (this is also the case in (Belkin et al., 2019b)). The corresponding in-sample error curve (see the bottom red-colored curve in Fig. 3a) shows that all the *unconstrained* overparameterized solutions interpolate, i.e., zero in-sample error is achieved for $p \geq n - 1$ (this range is defined by $n - 1$ and not n due to data centering). This specific result is a consequence of the pure regression setting we examine in this subsection. In our next steps below we explore settings that are not standard regression problems and, for them, studying the existence of double descent phenomena is of interest.

4.3. Supervised Learning with Soft Orthonormality Constraints

The two supervised problems defined in (10) and (14) correspond to the extreme cases of strict orthonormality constraints and no constraints at all, respectively. We observed that, while the unconstrained problem yields generalization errors following the double-descent behavior, the strictly constrained problem does not (despite the fact it is also fully supervised). This motivates us to explore the entire range of supervised problems connecting (10) and (14) via orthonormality constraints that can be progressively softened. This range of problems is denoted by the yellow line in Fig. 1.

The following constructions rely on the fact that a tall (rectangular) matrix has orthonormal columns if and only if all of its singular values equal 1. This statement is proved in Appendix B. Accordingly, we formulate the soft-constraint problem (for $p \geq m$) as

$$\widehat{\mathbf{U}}_{m,\mathcal{S}} = \underset{\mathbf{W}\in\mathbb{R}^{p\times m}}{\arg\min} \frac{1}{n} \left\|\mathbf{Z} - \mathbf{W}^T\mathbf{X}_{\mathcal{S}}\right\|_F^2 \quad (15)$$
$$\text{subject to } |\sigma_i^2(\mathbf{W}) - 1| \leq \alpha \text{ for } i = 1, ..., m$$

where $\sigma_i(\mathbf{W})$ is the i^{th} singular value of $\mathbf{W}$, and the constant $\alpha \geq 0$ defines the softness of the constraints. Note that for $\alpha = 0$ the demand becomes a hard constraint of orthonormality and, then, (15) reduces to (10). When $\alpha \to \infty$ the problem converges to the unconstrained regression form of (14).

Due to the constraints, the problem (15) does not have a closed-form solution. Hence, we propose again a procedure based on the projected gradient descent technique. Nicely, the constraint-projection step takes the form of a thresholding operation applied on the singular values of the evolving solution, as explained next (see details in Appendix B). Consider a matrix $\mathbf{W}^{(\text{in})} \in \mathbb{R}^{p\times m}$ (where $p \geq m$), with the SVD $\mathbf{W}^{(\text{in})} = \mathbf{\Omega}\mathbf{\Sigma}^{(\text{in})}\mathbf{\Theta}^T$, where $\mathbf{\Omega}$ and $\mathbf{\Theta}$ are $p \times p$ and $m \times m$ real orthonormal matrices, respectively, and $\mathbf{\Sigma}^{(\text{in})}$ is a $p \times m$ real diagonal matrix with m singular values $\{\sigma_i(\mathbf{W}^{(\text{in})})\}_{i=1}^m$ on its main diagonal (recall that, by definition, singular values are non-negative). Projecting $\mathbf{W}^{(\text{in})}$ on the soft-orthonormality constraints via

$$\mathbf{W}^{(\text{out})} = \underset{\mathbf{W}\in\mathbb{R}^{p\times m}}{\arg\min} \left\|\mathbf{W} - \mathbf{W}^{(\text{in})}\right\|_F^2 \quad (16)$$
$$\text{subject to } |\sigma_i^2(\mathbf{W}) - 1| \leq \alpha \text{ for } i = 1, ..., m$$

is equivalent to the thresholding mapping $\mathbf{W}^{(\text{out})} \triangleq T_\alpha(\mathbf{W}^{(\text{in})})$ where $\mathbf{W}^{(\text{out})} = \mathbf{\Omega}\mathbf{\Sigma}^{(\text{out})}\mathbf{\Theta}^T$ and the singular values along the main diagonal of $\mathbf{\Sigma}^{(\text{out})}$ are

$$\sigma_i\left(\mathbf{W}^{(\text{out})}\right) = \quad (17)$$
$$\begin{cases} \sigma_i(\mathbf{W}^{(\text{in})}), & \text{if } \sigma_i(\mathbf{W}^{(\text{in})}) \in [\tau_\alpha^{\text{low}}, \tau_\alpha^{\text{high}}] \\ \tau_\alpha^{\text{low}}, & \text{if } \sigma_i(\mathbf{W}^{(\text{in})}) < \tau_\alpha^{\text{low}} \\ \tau_\alpha^{\text{high}}, & \text{if } \sigma_i(\mathbf{W}^{(\text{in})}) > \tau_\alpha^{\text{high}} \end{cases}$$

for $i = 1, ..., m$, where the threshold levels are defined by $\tau_\alpha^{\text{low}} \triangleq \sqrt{\max\{0, 1-\alpha\}}$ and $\tau_\alpha^{\text{high}} \triangleq \sqrt{1+\alpha}$. The entire optimization process is like in Algorithm 1, except that the projections onto the constraint are done using the soft thresholding T_α defined using (17) (instead of the hard thresholding T_{hard}). See Appendix B for details.

The empirical demonstration in Fig. 3b shows the generalization errors (as function of p) corresponding to a range of problem settings where α gradually increases from 0 (i.e., strictly constrained setting) to ∞ (i.e., practically unconstrained, standard regression problem). This demonstrates that the double-descent trend emerges in the fully supervised setting as the orthonormality constraints are relaxed (and eventually removed). The evolution of the corresponding in-sample error curves in Fig. 3a shows that the range of interpolating solutions gradually increases as the orthonormality constraints are relaxed. Specifically, for a given α, the interpolation occurs for $p \geq p_\alpha$ where $p_\alpha \geq n-1$ is a threshold that monotonically decreases together with the increase in the constraint level α. Eventually, when the orthonormality constraint is completely removed (i.e., $\alpha \to \infty$), the range of interpolating solutions becomes the full range of overparameterized solutions (i.e., $p \geq n-1$). Interestingly, the peaks of the double descent trends of the out-of-sample error curves are still obtained at $p = n-1$ even if $p_\alpha > n-1$. In the few supervised settings where the orthonormality is nearly or exactly strictly constrained, the curves do not arrive to accurate interpolation ability and accordingly the double descent shape is not apparent (or apparent in very weak forms) in the matching out-of-sample error curves.

Our findings for fully supervised settings with varying orthonormality constraints can be also examined in the future for other formulations of the optimization cost and constraints, and different optimization techniques.

5. Semi-Supervised Subspace Fitting

The fully supervised problem (15), enabling flexible orthonormality constraint levels, demonstrated the important dependency of the double-descent behavior on the constraints. Now we turn to explore the supervision level as the additional crucial factor for the existence of double descent in subspace estimation tasks. Here, we essentially establish the ability to explore estimation problems induced *anywhere* on the supervision-orthonormality plane (Fig. 1).

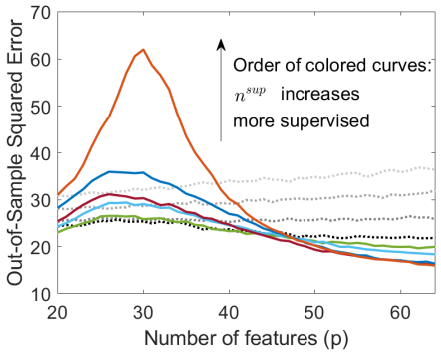
We define a learning problem with an arbitrary level of supervision, implemented as described next. The data model is again as specified in Section 2.1. However, now, the provided dataset of n examples is $\mathcal{D}^{\text{semisup}} \triangleq \widetilde{\mathcal{D}}^{\text{sup}} \cup \widetilde{\mathcal{D}}^{\text{unsup}}$, where $\widetilde{\mathcal{D}}^{\text{sup}} \triangleq \left\{(\mathbf{x}^{(\ell)}, \mathbf{z}^{(\ell)})\right\}_{\ell=1}^{n^{\text{sup}}} \in \mathbb{R}^d \times \mathbb{R}^m$ is a set of $n^{\text{sup}} \in \{0, \ldots, n\}$ i.i.d. samples of $(\mathbf{x}, \mathbf{z})$ pairs satisfying (1), and $\widetilde{\mathcal{D}}^{\text{unsup}} \triangleq \left\{\mathbf{x}^{(\ell)}\right\}_{\ell=n^{\text{sup}}+1}^{n} \in \mathbb{R}^d$ contains additional $n^{\text{unsup}} \triangleq n - n^{\text{sup}}$ i.i.d. samples of $\mathbf{x}$. Again, the learning goal is to estimate a linear operator $\widehat{\mathbf{U}}_m$, where only the p features (specified in $\mathcal{S}$) of $\mathbf{x}$ are used in the actual learning. Note the extreme cases of $n^{\text{sup}} = 0$ and $n^{\text{sup}} = n$ where the setting reduces to unsupervised and fully-supervised forms, respectively. For any $n^{\text{sup}} \in \{1, ..., n-1\}$, the problem is semi-supervised at a level that grows with n^{sup}.

We define the learning task by extending (15) into

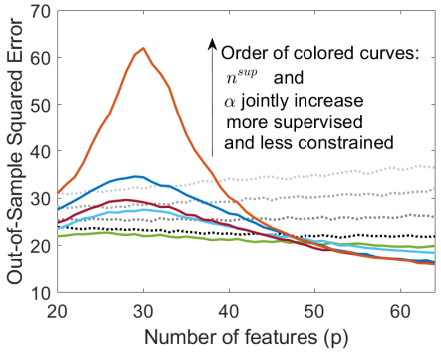
$$\widehat{\mathbf{U}}_{m,\mathcal{S}} = \underset{\mathbf{W}\in\mathbb{R}^{p\times m}}{\arg\min} \left\{ \left\| \mathbf{Z}^{\text{sup}} - \mathbf{W}^T \mathbf{X}_{\mathcal{S}}^{\text{sup}} \right\|_F^2 + \left\| \left(\mathbf{I}_p - \mathbf{W}\mathbf{W}^T \right) \mathbf{X}_{\mathcal{S}}^{\text{unsup}} \right\|_F^2 \right\}$$
$$\text{subject to } |\sigma_i^2(\mathbf{W}) - 1| \leq \alpha \text{ for } i = 1, ..., m \qquad (18)$$

where $\mathbf{X}_{\mathcal{S}}^{\text{sup}} \triangleq \left[\mathbf{x}_{\mathcal{S}}^{(1)}, \ldots, \mathbf{x}_{\mathcal{S}}^{(n^{\text{sup}})}\right]$, $\mathbf{Z}^{\text{sup}} \triangleq \left[\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n^{\text{sup}})}\right]$, $\mathbf{X}_{\mathcal{S}}^{\text{unsup}} \triangleq \left[\mathbf{x}_{\mathcal{S}}^{(n^{\text{sup}}+1)}, \ldots, \mathbf{x}_{\mathcal{S}}^{n}\right]$, and α determines the orthonormality constraint level. The optimization cost in (18) naturally blends the supervised and unsupervised metrics in proportions induced by the n^{sup} to n^{unsup} ratio. We address (18) using a projected gradient descent approach. Since (18) extends (15) only with respect to the optimization cost, the current optimization procedure extends Algorithm 1 by using the soft-threshold projection T_α from (17), replacing the gradient descent stage with the one suitable to the new cost function in (18), and initializing the optimization process with a random matrix (of i.i.d. Gaussian components with zero mean and variance $1/p$) that is projected onto the relevant orthonormality constraint. See Appendix C for details.

Equipped with the problem defined in (18), we are able to generate a subspace estimation problem at any point of the supervision-orthonormality plane (recall Fig. 1) and empirically evaluate the corresponding generalization errors as function of the number of features p used in the actual learning. We start by evaluating the range of problems that are unconstrained (i.e., $\alpha \to \infty$) and their supervision level gradually varies from unsupervised ($n^{\text{sup}} = 0$) to fully supervised ($n^{\text{sup}} = n$). This set of problems is located along the right, blue-colored border line of the supervision-orthonormality plane in Fig. 1. Figure 4a clearly demonstrates the emergence of the double descent trend together with the increase in supervision level. This shows that double descent can occur in problems that are semi-supervised and deviate from the ordinary regression form. Our concluding demonstration evaluates the range of problems on the diagonal trajectory (on the supervision-orthonormality plane) connecting the standard subspace fitting and the pure regression settings (see the purple-colored trajectory in Fig. 1). Here we jointly increase α (from 0 to ∞) and n^{sup} (from 0 to n). The observed generalization errors (Fig. 4b) clearly exhibit the rise of the double descent phenomena with the joint increase in supervision level and decrease in orthonormality level.



(a)



(b)

Figure 4. The out-of-sample errors, $\mathcal{E}_{\text{out}}^{\text{sup}}\left(\widehat{\mathbf{U}}_m\right)$ versus the number of parameters p. The errors are averaged over 25 experiments with different sequential orders of adding coordinates to $\mathcal{S}$. Here $d = 64$, $m = 20$, $n = 32$. **(a) Unconstrained** settings ($\alpha \to \infty$): Each curve is for a different supervision level, $n^{\text{sup}} \in \{0, 4, 8, 12, 16, 20, 24, 28, n = 32\}$. **(b)** Problems residing at the supervision-orthonormality plane along the **diagonal trajectory** connecting the standard subspace fitting and the pure regression. Each curve is for a different pair of supervision and orthonormality constraint levels that jointly increase. The gray dotted curves correspond to $n^{\text{sup}} \in \{0, 4, 8, 12\}$.

6. Conclusions

In this work we have opened up a new avenue of research on linear subspace estimation problems. We defined a family of linear subspace estimation problems that reside over a supervision-orthonormality plane (where each coordinate induces a unique problem setting). This class of problems connects the standard subspace fitting and the pure regression problems. We proposed an optimization procedure, based on the projected gradient descent technique, to evaluate any problem instance on the supervision-orthonormality plane. Then, we explored problems defined along various trajectories of the supervision-orthonormality plane, and showed that the double-descent phenomena is more evident as the problems are more supervised and less orthonormally constrained. We believe that our findings open a new direction of theoretical and practical research of the generalization ability of overparameterized models learned in diverse supervision levels (i.e., including semi-supervised settings) and various optimization constraints.

Acknowledgments

This work was supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014-18-12571 and N00014-17-1-2551; AFOSR grant FA9550-18-1-0478; DARPA grant G001534-7500; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019b.

Breiman, L. and Freedman, D. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983.

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

Gower, J. C., Dijksterhuis, G. B., et al. *Procrustes Problems*, volume 30. Oxford University Press on Demand, 2004.

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

Johnstone, I. M. and Lu, A. Y. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

Jolliffe, I. T. Discarding variables in a principal component analysis. i: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2):160–173, 1972.

Jolliffe, I. T. Discarding variables in a principal component analysis. ii: Real data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(1):21–31, 1973.

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

Nie, F., Xu, D., Tsang, I. W.-H., and Zhang, C. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing*, 19(7):1921–1932, 2010.

Paul, D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pp. 1617–1642, 2007.

Shen, D., Shen, H., and Marron, J. A general framework for consistency of principal component analysis. *The Journal of Machine Learning Research*, 17(1):5218–5251, 2016.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018.

Sugiyama, M. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 905–912, 2006.

Ulfarsson, M. O. and Solo, V. Vector l_0 sparse variable PCA. *IEEE Transactions on Signal Processing*, 59(5): 1949–1958, 2011.

Xu, J. and Hsu, D. J. On the number of variables to use in principal component regression. In *Advances in Neural Information Processing Systems*, pp. 5095–5104, 2019.

Yang, X., Fu, H., Zha, H., and Barlow, J. Semi-supervised nonlinear dimensionality reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 1065–1072, 2006.

Zhang, D., Zhou, Z.-H., and Chen, S. Semi-supervised dimensionality reduction. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 629–634. SIAM, 2007.