TRANSFER LEARNING CAN OUTPERFORM THE TRUE PRIOR IN DOUBLE DESCENT REGULARIZATION

YEHUDA DAR∗ AND RICHARD G. BARANIUK†

Abstract. We study a fundamental transfer learning process from source to target linear regression tasks, including overparameterized settings where there are more learned parameters than data samples. The target task learning is addressed by using its training data together with the parameters previously computed for the source task. We define the target task as a linear regression optimization with a regularization on the distance between the to-be-learned target parameters and the already-learned source parameters. This approach can be also interpreted as adjusting the previously learned source parameters for the purpose of the target task, and in the case of sufficiently related tasks this process can be perceived as fine tuning. We analytically characterize the generalization performance of our transfer learning approach and demonstrate its ability to resolve the peak in generalization errors in double descent phenomena of min-norm solutions to ordinary least squares regression. Moreover, we show that for sufficiently related tasks the optimally tuned transfer learning approach can outperform the optimally tuned ridge regression method, even when the true parameter vector conforms with isotropic Gaussian prior distribution. Namely, we demonstrate that transfer learning can beat the minimum mean square error (MMSE) solution of the individual target task.

Key words. Overparameterized machine learning, transfer learning, linear regression, ridge regression.

1. Introduction. Contemporary machine learning models are often overparameterized, meaning that they are more complex (e.g., have more parameters to be learned) than the amount of data available for their training. Deep neural networks are a very successful example for highly overparameterized models that are often trained without explicit regularization. The challenge in such overparameterized learning is to be able to generalize well beyond the given dataset, despite the tendency of overparameterized models to perfectly fit their training data [1].

Recent empirical studies [2, 3, 4] show that generalization errors follow a double descent shape when examined with respect to the complexity of the learned model. In the double descent shape, the generalization error peaks when the learned model becomes sufficiently complex and begins to perfectly fit (i.e., interpolate) the training data. This peak in generalization error reflects poor generalization performance, but when the learned model complexity increases further then the generalization error starts to decrease again and eventually may achieve excellent generalization ability for highly overparameterized models — and this is despite perfectly fitting the training data! The prevalence of double descent phenomena in deep learning motivated a corresponding field of theoretical research where learning of overparameterized models is analytically studied mainly for linear regression problems [5, 6, 7, 8, 9, 10, 11], as well as for other statistical learning problems such as linear subspace learning [12] and linear classification [13]. The existing literature show that double descent phenomena occur in minimum-norm solutions to ordinary least squares regression, i.e., when there is no explicit regularization in the learning process. Moreover, [5, 10] show that ridge regression (i.e., where the learning includes an explicit regularizer) is able to resolve the generalization error peak of the double descent behavior.

Transfer learning [14] is a key approach in the practical training of deep neural networks (DNNs) where learning is conducted not only using a dataset that is
Fig. 1: Qualitative demonstration of the four main cases that determine if the examined transfer learning (TL) generalizes better than ridge regression. The TL and ridge methods are both optimally tuned. The prior distribution of the target task parameters is isotropic Gaussian, therefore the ridge solution corresponds to using the prior of the target task parameters. The four cases are associated with respective portions of generalization error curves in settings where (a) the source tasks is sufficiently related to the target task, and where (b) the source task is not related to the target task. Each of the four cases is presented using a conceptual diagram that compares the uncertainties in the prior distribution of the target task parameters (the green circle area) and the uncertainties in the parameters transferred from the source task (the blue co-centric circle area). The total uncertainties in the transferred parameters are comprised of two parts: inaccuracies in the solution of the source task, and the distance between the source and target tasks.

relatively small compared to the complexity of the DNN, but also using layers of parameters taken from a ready-to-use DNN that was properly trained for a related task [15, 16, 17]. Transfer learning between DNNs can be done by transferring network layers from the source to target models and set them fixed (while other layers are learned), fine tune them (i.e., moderately adjust to the target task data), or use them as initialization for a comprehensive learning process. Clearly, the source task should be sufficiently related to the target task in order to have a useful transfer learning
TRANSFER LEARNING CAN OUTPERFORM THE TRUE PRIOR

Nevertheless, finding successful transfer learning settings is still a fragile task [21] that requires a further understanding — also from theoretical perspectives.

Surprisingly, there are very few analytical theories for transfer learning. Lampinen and Ganguli [22] analyze the optimization dynamics of transfer learning for multi-layer linear networks. In our previous work [23] we study generalization and double descent phenomena in transfer learning between overparameterized linear regression problems, where a subset of parameters is transferred from the source task solution and set fixed in the target task model to be learned. Clearly, there are still more aspects of transfer learning that should be analytically understood, also for linear architectures.

In this paper, we study transfer learning between two linear regression tasks where the transferred parameters from the source task are utilized for the learning of the target task’s parameters. Specifically, we formulate the target task as a linear regression problem that includes regularization on the distance between the transferred source parameters (that were already learned) and the to-be-learned parameters of the target task. One can also perceive the transfer learning approach in this paper as transferring parameters from the source task and adjusting them to the training data of the target task — and for relatively similar source and target tasks the adjustment of parameters is modest and can be interpreted as a fine tuning mechanism. The settings of this paper significantly differ from those in our previous study [23] where only a subset of parameters was transferred and set fixed in the learned model of the target task.

We examine target and source tasks that have a noisy linear relation between their true parameter vectors. Namely, the true parameter vector of the source task is the true parameter vector of the target task that is processed by a linear operator and then immersed in additive white Gaussian noise. Our approach to transfer learning assumes knowledge of the statistical relation between the tasks, i.e., the linear operator and noise level involved in the task relation model are known and utilized in the examined solutions. Our transfer learning approach includes a regularization coefficient that determines the importance of the source parameters in the learning of the target parameters. We characterize the optimally tuned version of our transfer learning approach and study its generalization performance from analytical and empirical perspectives.

We show that our optimally tuned transfer learning can outperform the optimally tuned ridge regression solution of the individual target task. Remarkably, we prove this result also for the case where optimally tuned ridge regression provides the minimum mean square error (MMSE) estimate for the parameters of the individual target task (namely, the case where the true parameters of the target task follow an isotropic Gaussian distribution and the source task solution cannot be utilized). We show that transfer learning outperforms ridge regression if the target and source tasks are sufficiently related and the source task solution generalizes sufficiently well at the source task itself (i.e., the source task solution is sufficiently accurate). Our findings emphasize the following representative cases:

- **Transferred parameters are more informative than the true prior distribution of the target task parameters.** This occurs when the source task is sufficiently related to the target task and the solution of the source task is sufficiently accurate (see Case 1 in Fig. 1).
- **Transferred parameters include more uncertainties than the true prior distribution of the target task parameters, and therefore the true prior distribution is more informative.** This happens when the source task is too far from the target task and/or the solution of the source task is overly inaccurate (see
The main contribution of optimally tuned ridge regression is to resolve the generalization error peak of the ordinary least squares (OLS) solution of the individual target task (see red curves in Fig. 1 where their peaks are located at the point where the target task model shifts from being underparameterized to being overparameterized). Optimally tuned transfer learning from a sufficiently related and accurate source task has the ability to resolve the peak of the OLS solution, but also to achieve significantly lower generalization errors than the ridge solution over a wide range of parameterization levels (see blue curve in Fig. 1a). Importantly, the usefulness of transfer learning based on the OLS solution of the source task has a by-product in the form of another generalization error peak, which is located at the point where the source task model shifts from being underparameterized to being overparameterized (see blue curves in Fig. 1).

This paper is organized as follows. In Section 2 we outline the settings of the source and target tasks, as well as the model for their relation. In Section 3 we present the examined transfer learning approach and study its generalization performance. In Section 4 we compare the examined transfer learning method to ridge regression and characterize the cases where transfer learning generalizes better. Section 5 concludes this paper. The Appendices include all the proofs and additional details.

2. Problem Settings: Two Related Linear Regression Tasks.

2.1. Source Task: Data Model and Solution Form. The source task is a standard linear regression problem with a $d$-dimensional Gaussian input $z \sim \mathcal{N}(0, I_d)$ and a response value $v \in \mathbb{R}$ that is induced by

$$ v = z^T \theta + \xi, $$

where $\xi \sim \mathcal{N}(0, \sigma^2_\xi)$ is a noise variable independent of $z$, and $\theta \in \mathbb{R}^d$ is an unknown parameter vector. Motivation for linear models with random features can be found, e.g., in [5, 24]. While not knowing the true distribution of $(z, v)$, the source learning task is carried out using a dataset $\tilde{D} \triangleq \{(z^{(i)}, v^{(i)})\}_{i=1}^{\tilde{n}}$ that includes $\tilde{n}$ independent and identically distributed (i.i.d.) samples of $(z, v)$. We also denote the $\tilde{n}$ data samples in $\tilde{D}$ using an $\tilde{n} \times d$ input matrix $Z \triangleq [z^{(1)}, \ldots, z^{(\tilde{n})}]^T$ and a $\tilde{n} \times 1$ response vector $v \triangleq [v^{(1)}, \ldots, v^{(\tilde{n})}]^T$. Therefore, $v = Z\theta + \xi$ where $\xi \triangleq [\xi^{(1)}, \ldots, \xi^{(\tilde{n})}]^T$ is an unknown noise vector that its $i^{th}$ component $\xi^{(i)}$ originates in the $i^{th}$ data sample relation $v^{(i)} = z^{(i)^T} \theta + \xi^{(i)}$.

Consider an out-of-sample input-response pair $(z^{\text{(test)}}, v^{\text{(test)}})$ that is independently drawn from the $(z, v)$ distribution defined above. Given the input $z^{\text{(test)}}$, the source task goal is to estimate the response value $v^{\text{(test)}}$ by the value $\hat{v} \triangleq z^{\text{(test)}}^T \hat{\theta}$, where $\hat{\theta}$ is learned using $\tilde{D}$. We can assess the generalization performance of the source task using the out-of-sample squared error

$$ E_{\text{src}} \triangleq \mathbb{E}\left\{\left(\hat{v} - v^{\text{(test)}}\right)^2\right\} = \sigma^2_\xi + \mathbb{E}\left\{\|\hat{\theta} - \theta\|_2^2\right\}, $$

where the expectation in the definition of $E_{\text{src}}$ is with respect to the out-of-sample data $(z^{\text{(test)}}, v^{\text{(test)}})$ and the training data $\tilde{D}$. Note that $\hat{\theta}$ is a function of the training data. A lower value of $E_{\text{src}}$ reflects better generalization performance of the source task.
The source task is addressed via ordinary least squares (OLS) regression, namely,

\begin{equation}
\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^d} \| \mathbf{v} - \mathbf{Z}_r \|_2^2 = \mathbf{Z}^+ \mathbf{v}
\end{equation}

where \( \mathbf{Z}^+ \) is the Moore-Penrose pseudoinverse of \( \mathbf{Z} \). The source out-of-sample error of the solution (2.3) can be formulated as

\begin{equation}
\mathcal{E}_{\text{src}} = \begin{cases} 
(1 + \frac{d}{n-\bar{n}-1}) \sigma^2 \xi & \text{for } \bar{n} - 1 \leq d \leq \bar{n} + 1, \\
\infty & \text{for } \bar{n} - 2 \leq d \leq \bar{n} - 1, \\
(1 + \frac{\bar{n}}{n-\bar{n}-1}) \sigma^2 \xi + (1 - \frac{\bar{n}}{n}) \| \theta \|_2^2 & \text{for } d \geq \bar{n} + 2,
\end{cases}
\end{equation}

which is a particular case of the results in [6, 23]. The focus of this paper is on transfer learning and, therefore, we do not further analyze the source out-of-sample error. Yet, it is important to note the peak that occurs in the generalization errors \( \mathcal{E}_{\text{src}} \) of the source task around \( d = \bar{n} \), namely, where the source model starts to be overparameterized.

2.2. Target Task: Data Model and Relation to Source Task. Our interest is in a target task with data \((\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}\) that follow the model

\begin{equation}
y = \mathbf{x}^T \mathbf{\beta} + \epsilon
\end{equation}

where \( \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \) is a \( d \)-dimensional Gaussian input vector, \( \epsilon \sim \mathcal{N}(0, \sigma^2_{\epsilon}) \) is a Gaussian noise independent of \( \mathbf{x} \), and \( \mathbf{\beta} \in \mathbb{R}^d \) is an unknown parameter vector.

The unknown parameter vector of the source task, \( \theta \), is related to the unknown parameter of the target task, \( \mathbf{\beta} \), by the relation

\begin{equation}
\hat{\theta} = \mathbf{H} \mathbf{\beta} + \eta
\end{equation}

where \( \mathbf{H} \in \mathbb{R}^{d \times d} \) is a fixed (non-random) matrix and \( \eta \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\eta} \mathbf{I}_d) \) is a vector of i.i.d. Gaussian noise components with zero mean and variance \( \sigma^2_{\eta} / \sigma^2_{\epsilon} \). The random elements \( \eta, \mathbf{x}, \epsilon, \mathbf{\beta}, \alpha, \xi \) and \( \xi \) are independent. The relation in (2.6) recalls a common data model in inverse problems, which in our case relates to the recovery of the true \( \mathbf{\beta} \) from the true \( \theta \). However, in our settings we do not have the true \( \theta \) but only its estimate \( \hat{\theta} \) that was learned for the source task purposes.

While not knowing the true distribution of \((\mathbf{x}, y)\), the target learning task is performed based on a dataset \( \mathcal{D} \triangleq \{ (\mathbf{x}^{(i)}, y^{(i)}) \}_{i=1}^n \) that contains \( n \) i.i.d. draws of \((\mathbf{x}, y)\) pairs. We denote the \( n \) data samples in \( \mathcal{D} \) using an \( n \times d \) matrix of input variables \( \mathbf{X} \triangleq [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}]^T \) and an \( n \times 1 \) vector of responses \( \mathbf{y} \triangleq [y^{(1)}, \ldots, y^{(n)}]^T \). The training data satisfy \( \mathbf{y} = \mathbf{X} \mathbf{\beta} + \epsilon \) where \( \epsilon \triangleq [\epsilon^{(1)}, \ldots, \epsilon^{(n)}]^T \) is an unknown noise vector that its \( i \)-th component \( \epsilon^{(i)} \) originates in the \( i \)-th data sample relation \( y^{(i)} = \mathbf{x}^{(i)^T} \mathbf{\beta} + \epsilon^{(i)} \).

Consider an out-of-sample input-response pair \((\mathbf{x}^{(\text{test})}, y^{(\text{test})})\) that is independently drawn from the \((\mathbf{x}, y)\) distribution defined above. Given the input \( \mathbf{x}^{(\text{test})} \), the target task aims to estimate the response value \( y^{(\text{test})} \) by the value \( \hat{y} \triangleq \mathbf{x}^{(\text{test})} \mathbf{\hat{\beta}} \), where \( \mathbf{\hat{\beta}} \) is formed using \( \mathcal{D} \) in a transfer learning process that also utilizes the source estimate \( \hat{\theta} \). We evaluate the generalization performance of the target task using the out-of-sample squared error

\begin{equation}
\mathcal{E} \triangleq \mathbb{E} \left\{ \left( \hat{y} - y^{(\text{test})} \right)^2 \right\} = \sigma^2_{\epsilon} + \mathbb{E} \left\{ \| \mathbf{\hat{\beta}} - \mathbf{\beta} \|_2^2 \right\}
\end{equation}
where the expectation in the definition of $E$ is with respect to the out-of-sample data $(\mathbf{x}^{(\text{test})}, y^{(\text{test})})$ of the target task and the training data $\mathcal{D}$, $\mathcal{D}$ of both the target and source tasks. Note that, in a transfer learning process, $\hat{\beta}$ is a function of the training data of both the target and source tasks. A lower value of $E$ reflects better generalization performance of the target task.

In this paper we study the generalization performance of the target task based on $n$ data samples, considering well-specified feature selection, i.e., when all the $d$ features of the data are used in the learning process. Then, we analyze the generalization performance with respect to the parameterization level that is determined by the number of samples $n$ and the data dimension $d$. As explained next, we consider the data dimension $d$ as the resolution at which we examine the transfer learning problem.

### 2.3. The Problem Settings at Different Resolutions.

The target task solution is based on estimating the parameter vector $\beta \in \mathbb{R}^d$. Let us take a Bayesian perspective and assume that $\beta$ is a random vector that follows a Gaussian prior distribution $\mathcal{N}(0, B_d)$ where $B_d$ is the $d \times d$ covariance matrix of $\beta$. For considering the same problem at different resolutions we will assume that $E\left\{\|\beta\|_2^2\right\} = \text{Tr}\{B_d\} = \omega_\beta$ where $\omega_\beta$ is the same positive real constant for all $d, n$. For example, the last assumption is satisfied by $B_d = \frac{1}{d} I_d$ that corresponds to an isotropic Gaussian prior distribution for $\beta$ and a constant $\omega_\beta = 1$.

Next we characterize the way that the parameter vector of the source task and the relation model to the target task behave at the resolution induced by the parameterization level that is determined by the number of samples $n$ and the data dimension $d$. As explained next, we consider the data dimension $d$ as the resolution at which we examine the transfer learning problem.

1. **Transformation to another basis:** $H = \Psi_d^T$ where $\Psi_d$ is a $d \times d$ orthonormal matrix, i.e., $\Psi_d^T \Psi_d = I_d$. Examples for such $\Psi_d$ are the $d \times d$ forms of the identity matrix, discrete cosine transform (DCT) matrix, Hadamard matrix (the case of Hadamard is defined only for $d$ values that satisfy its recursive construction). In this setting, we study the generalization performance versus the dimension $d$ that is coupled with a $d \times d$ orthonormal matrix $\Psi$ of the same type (e.g., DCT).

2. **Circular convolution operation:** In this case, $H$ is a $d \times d$ circulant matrix that can be interpreted as a discrete version of the circular convolution kernel $h_{\text{ker}} : [0, 1] \to \mathbb{R}$, which is defined over the continuous interval $[0, 1]$. The function $h_{\text{ker}}$ is assumed to be smooth and to have finite derivatives. Again, $H$ should be properly scaled to ensure $\frac{1}{d} \|H\|_F^2 = \omega_H$ where $\omega_H$ is a constant independent of $d, n, \tilde{n}$.

Now we can proceed to the definition of a transfer learning procedure and the analysis of its generalization performance.
3. The Examined Transfer Learning Approach.

3.1. Transfer Learning Optimization Form and Its Solution. Let us consider a well-specified model that enables the learning of all the \( d \) parameters of \( \hat{\beta} \in \mathbb{R}^d \). Since the target task is related to the source task by the model (2.6), the optimization of the target task estimate \( \hat{\beta} \) can utilize the source task estimate \( \hat{\theta} \) that was already computed. This means that we consider a learning setting where parameters of the source model are transferred and adjusted for the target model learning, and in many cases this can be conceptually perceived as a fine tuning strategy. Assume that the linear operator \( H \), which connects \( \beta \) to \( \theta \) in (2.6), is known. Then, we suggest to optimize the parameters of \( \hat{\beta} \) via

\[
(3.1) \quad \hat{\beta} = \arg \min_{b \in \mathbb{R}^d} \| y - Xb \|_2^2 + n\alpha_{\text{TL}} \| Hb - \hat{\theta} \|_2^2
\]

where \( \hat{\theta} \) is fixed in this target task optimization. The second term in the last optimization cost evaluates the proximity of the target parameters (after processing by \( H \)) to the source parameters.

Setting \( \alpha_{\text{TL}} = 0 \) in (3.1) disables the transfer learning aspects and provides an ordinary least squares (OLS) problem that its minimum \( \ell_2 \)-norm solution is \( \hat{\beta}_{\text{OLS}} = X^+ y \) and its corresponding out-of-sample error is

\[
(3.2) \quad \mathcal{E}_{\text{OLS}} = \begin{cases} 
(1 + \frac{d}{n-d-1}) \sigma_i^2 & \text{for } d \leq n - 2, \\
\infty & \text{for } n - 1 \leq d \leq n + 1, \\
(1 + \frac{n}{n-d-1}) \sigma_i^2 + (1 - \frac{n}{2}) \| \beta \|_2^2 & \text{for } d \geq n + 2,
\end{cases}
\]

which is a special case of previous results, e.g., in [6, 23]. Yet, the main focus in this paper is on settings where \( \alpha_{\text{TL}} > 0 \) and the transfer learning aspects in the optimization (3.1) are applied.

Let us consider the next assumption.

**Assumption 1.** \( H \) is a full rank \( d \times d \) matrix.

Then, the closed-form solution of the target task in (3.1) for \( \alpha_{\text{TL}} > 0 \) is

\[
(3.3) \quad \hat{\beta}_{\text{TL}} = \left( X^T X + n\alpha_{\text{TL}} H^T H \right)^{-1} \left( X^T y + n\alpha_{\text{TL}} H^T \hat{\theta} \right).
\]

In this section we study the generalization performance of the target task solution with respect to the parameterization level between \( d \) and \( n \). Accordingly, the learning process is underparameterized when \( d < n \), and overparameterized when \( d > n \).

**Assumption 2** (Isotropic prior distribution). The target parameter vector \( \beta \) is random and has isotropic Gaussian distribution with zero mean and covariance matrix \( B_d = \frac{b}{d} I_d \) for some constant \( b > 0 \). Under Assumption 2, we will evaluate generalization performance using the out-of-sample error with expectation over \( \beta \), i.e., \( \hat{\mathcal{E}}_{\text{TL}} \triangleq \mathbb{E}_\beta \{ \mathcal{E}_{\text{TL}} \} = \sigma_i^2 + \mathbb{E} \{ \| \hat{\beta}_{\text{TL}} - \beta \|_2^2 \} \).

3.2. Analysis for \( H \) of an Orthonormal Matrix Form. We first examine the case of \( H = \Psi^T \) where \( \Psi \) is a \( d \times d \) orthonormal matrix, namely, it is a multi-dimensional rotation operator. This will let us to obtain a complete analytical characterization of the generalization performance. Later on we will proceed to the more intricate case where \( H \) has other forms.
Let us denote \( X_\Psi \equiv X\Psi^T \). Because \( \Psi \) is an orthonormal matrix and the rows of \( X \) are i.i.d. from \( \mathcal{N}(0, I_d) \), \( X_\Psi \) is a \( d \times d \) random matrix with the same distribution as \( X \).

**Lemma 3.1.** Under Assumptions 1-2, and for \( H = \Psi^T \) where \( \Psi \) is a \( d \times d \) orthonormal matrix, the expected out-of-sample error of the solution from (3.3) for \( \alpha_{\text{TL}} > 0 \) can be written as

\[
(3.4) \quad \bar{\mathcal{E}}_{\text{TL}} = \sigma^2_e + \mathbb{E}\left\{ \sum_{k=1}^{d} n^2 \alpha_{\text{TL}}^2 C_{\text{TL}} + \sigma^2_e \cdot \lambda_k \left\{ X_\Psi^T X_\Psi \right\} \right\} / \left\{ \lambda_k \left\{ X_\Psi^T X_\Psi \right\} + n\alpha_{\text{TL}} \right\}^2
\]

where \( \lambda_k \left\{ X_\Psi^T X_\Psi \right\} \) is the \( k \)th eigenvalue of the \( d \times d \) matrix \( X_\Psi^T X_\Psi \), and the transfer learning aspects are included in

\[
(3.5) \quad C_{\text{TL}} \triangleq \begin{cases} \frac{\sigma^2_e}{d} + \frac{\sigma^2_e}{n-d-1} & \text{for } d \leq \bar{n} - 2, \\ \infty & \text{for } \bar{n} - 1 \leq d \leq \bar{n} + 1, \\ \left(1 - \frac{\bar{n}}{d}\right) \frac{\beta}{d} + \frac{\bar{n}}{d} \left( \frac{\sigma^2_e}{d} + \frac{\sigma^2_e}{\sigma^2_e - \bar{n} - 1} \right) & \text{for } d \geq \bar{n} + 2. \end{cases}
\]

The last Lemma is proved in Appendix A.1. Note that the expectation over the sum in (3.4) is only with respect to the eigenvalues of \( X_\Psi^T X_\Psi \). Importantly, note that \( C_{\text{TL}} \) reflects the two aspects that determine the success of the transfer learning process: The first is the distance between the tasks as induced by the noise level \( \sigma^2_e \) in the task relation. The second is the accuracy of the source task solution which is associated with the source data noise level \( \sigma^2_o \) and the source parameterization level corresponding to \( \{\bar{n}, d\} \).

**Theorem 3.2.** Consider Assumptions 1-2 and for \( H = \Psi^T \) where \( \Psi \) is a \( d \times d \) orthonormal matrix. The optimal tuning of the transfer learning solution (i.e., \( \alpha_{\text{TL}} > 0 \)) from (3.3) is achieved for \( d \notin \{\bar{n} - 1, \bar{n}, \bar{n} + 1\} \) by setting \( \alpha_{\text{TL}} \) to

\[
(3.6) \quad \alpha_{\text{TL}}^\text{opt} = -\frac{\sigma_e^2}{nC_{\text{TL}}}
\]

and the corresponding minimal out-of-sample error is

\[
(3.7) \quad \bar{\mathcal{E}}_{\text{TL}}^\text{opt} = \sigma^2_e \left( 1 + \mathbb{E}_\Psi \left\{ \text{Tr} \left\{ \left( X_\Psi^T X_\Psi + n\alpha_{\text{TL}}^\text{opt} I_d \right)^{-1} \right\} \right\} \right)
\]

For \( d \in \{\bar{n} - 1, \bar{n}, \bar{n} + 1\} \), \( \bar{\mathcal{E}}_{\text{TL}} = \infty \) for any \( \alpha_{\text{TL}} > 0 \).

Theorem 3.2 is proved in Appendix A.2.

To develop the optimal out-of-sample error from (3.7) into a more explicit analytical form, we will make use of an asymptotic setting, which is described next.

**Assumption 3 (Asymptotic setting).** The quantities \( d, n, \bar{n} \to \infty \) such that the target task parameterization level \( \frac{d}{n} \to \gamma_{\text{tgt}} \in (0, \infty) \), and the source task parameterization level \( \frac{d}{n} \to \gamma_{\text{src}} \in (0, \infty) \). The task relation model \( \theta = H\beta + \eta \) includes a noise vector \( \eta \sim \mathcal{N}(0, \frac{\sigma^2_e}{\beta} I_d) \) and operator \( H \) that satisfies \( \frac{1}{d} \|H\|^2_F \to \kappa_H \).

Our current case of \( H \) being a transposed orthonormal matrix implies that \( \kappa_H = 1 \).

**Theorem 3.3.** Consider Assumptions 1-3, \( d \notin \{\bar{n} - 1, \bar{n}, \bar{n} + 1\} \), and \( H = \Psi^T \) where \( \Psi \) is a \( d \times d \) orthonormal matrix. Then, the transfer learning form of (3.3)
that its $\alpha_{\text{TL}}>0$ is optimally tuned to minimize the expected out-of-sample error $\bar{\mathcal{E}}_{\text{TL}}$ (i.e., with expectation w.r.t. the isotropic prior on $\beta$) almost surely satisfies

$$\bar{\mathcal{E}}_{\text{TL}}^{\text{opt}} \rightarrow \sigma^2 \left( 1 + \gamma_{\text{tgt}} \cdot m ( -\alpha_{\text{TL},\infty}^{\text{opt}}, \gamma_{\text{tgt}} ) \right)$$

where

$$\alpha_{\text{TL},\infty}^{\text{opt}} = \sigma^2 \gamma_{\text{tgt}} \times \begin{cases} \left( \frac{\sigma^2 + \gamma_{\text{src}} \cdot \sigma^2}{1 - \gamma_{\text{src}}} \right)^{-1} - 1 & \text{for } d \leq \bar{n} - 2, \\ \left( \frac{\gamma_{\text{src}} - 1}{\gamma_{\text{src}}} b + \frac{1}{\gamma_{\text{src}}} \left( \sigma^2 + \gamma_{\text{src}} \cdot \sigma^2 \right) \right)^{-1} - 1 & \text{for } d \geq \bar{n} + 2 \end{cases}$$

is the limiting value of the optimal $\alpha_{\text{TL}}>0$, and

$$m ( -\alpha_{\text{TL},\infty}^{\text{opt}}, \gamma_{\text{tgt}} ) = \frac{1 - \gamma_{\text{tgt}} + \alpha_{\text{TL},\infty}^{\text{opt}}}{2 \gamma_{\text{tgt}} \alpha_{\text{TL},\infty}^{\text{opt}}} \left( 1 - \gamma_{\text{tgt}} + \alpha_{\text{TL},\infty}^{\text{opt}} \right)^2 + 4 \gamma_{\text{tgt}} \alpha_{\text{TL},\infty}^{\text{opt}}$$

is the Stieltjes transform of the Marchenko-Pastur distribution, which is the limiting spectral distribution of the sample covariance associated with $n$ samples that are drawn from a Gaussian distribution $N(0, I_d)$.

The proof outline of the last theorem is provided in Appendix A.3. The blue curves in Fig. 2 show the analytical formula for the out-of-sample error $\bar{\mathcal{E}}_{\text{TL}}^{\text{opt}}$ of the optimally tuned transfer learning under Assumptions 1-3, for instances of the task relation model where $H$ is a transposed orthonormal matrix. Specifically, each subfigure in Fig. 2 considers a different pair of orthonormal $H$ and noise variance $\sigma^2_{\text{eta}}$. The corresponding empirical evaluations of $\bar{\mathcal{E}}_{\text{TL}}^{\text{opt}}$ are denoted in Fig. 2 by circle markers. Evidently, there is a good match between the analytical formula from Theorem 3.3 and the empirical results. Figures 2a-2d consider $H = I_d$, whereas in Figures 2e-2h $H = \Psi^T$ where $\Psi$ is the $d \times d$ discrete cosine transform (DCT) matrix. Comparing the respective results for $H$ being identity and DCT matrices, Fig. 2 demonstrates that the expected generalization performance of the examined transfer learning approach is not affected by high-dimensional rotation operators (i.e., transposed orthonormal $H$) in the task relation model.

Let us compare the out-of-sample errors of the examined transfer learning approach with the corresponding errors of OLS regression (i.e., solution without regularization) that appear in red curves in Fig. 2 and correspond to the extension of (3.2) to consider Assumption 2 (see Appendix A.4). One can observe that if the source and target tasks are sufficiently related (for example, see Figs. 2a-2c, 2e-2g, then the transfer learning approach indeed succeeds to resolve the peak that is induced by the OLS solution and also to lower the out-of-sample errors for the majority of parameterization levels. The only exception is that the examined transfer learning solution induces another peak in the generalization errors of the target task, and the location of this peak is determined by the point where the source task shifts from under to over parameterization. This is a side effect of transferring parameters from the source task, which by itself was solved using OLS regression and therefore suffers from a peak of double descent in its own out-of-sample error curves (e.g., recall (2.4)).

3.3. Analysis for $H$ of a General Form. After we provided a detailed analytical characterization for the case where $H$ is a transposed orthonormal matrix, we now proceed to the more intricate case where the matrix $H$ has a general form.
The covariance matrix of $\hat{\theta}$ in (e)-(h) is given by
\begin{equation}
\mathbb{E}\left\{\hat{\theta} - \mathbb{E}\{\hat{\theta} | \beta\}\right\} \mathbb{E}\left\{\hat{\theta} - \mathbb{E}\{\hat{\theta} | \beta\}\right\}^T | \beta\} = \begin{cases} \frac{\sigma_\eta^2}{\overline{n}} & \text{for } d \leq \overline{n}, \\
\frac{\sigma_\eta^2}{\overline{n}} \frac{\sigma_\eta^2}{\overline{n}-d-1} I_d & \text{for } d > \overline{n}. 
\end{cases}
\end{equation}
for \( d \geq \overline{n} + 2 \). For \( d \in \{\overline{n} - 1, \overline{n}, \overline{n} + 1\} \) the covariance matrix is infinite valued.

In (3.13), \( \{H\beta\}_j \) is the \( j \)th component of the vector \( H\beta \). The notation \( \text{diag}(\cdot) \) refers to the \( dx \times d \) diagonal matrix that its main diagonal values are specified as the arguments of \( \text{diag}(\cdot) \). The last proposition is proved in Appendix A.5. Proposition 3.4 demonstrates two very different forms of the covariance of \( \hat{\theta} \) given \( \beta \): For underparameterized source task (i.e., \( d \leq \overline{n} - 2 \)), the covariance (3.12) has a simple, diagonal form that does not reflect \( \beta \). However, for overparameterized source task (i.e., \( d \geq \overline{n} + 2 \)), the covariance (3.13) has a more intricate form that depends on the given \( \beta \). The two covariance matrix forms are consequences of the OLS regression of the source task, which by itself has two different error forms in its under and over parameterized cases.

Our next analytical results for a general \( H \) will mainly focus on the case of underparameterized source task, while the target task can still be overparameterized. The empirical results will consider both under and over parameterized settings of the source task.

Under Assumption 1, let us denote \( XH^{-1} - 1 \). The eigendecomposition

\[
X_{H^{-1}}^T X_{H^{-1}} = \Phi \Lambda_{TL} \Phi^T
\]

where \( \Phi \) is a \( d \times d \) orthonormal matrix with columns being eigenvectors of \( X_{H^{-1}}^T X_{H^{-1}} \) and the corresponding eigenvalues \( \lambda_k \{X_{H^{-1}}^T X_{H^{-1}}\} \), \( k = 1, \ldots, d \), are on the main diagonal of the \( d \times d \) diagonal matrix \( \Lambda_{TL} \).

**Lemma 3.5.** Under Assumptions 1-2, the out-of-sample error of the solution from (3.3) for \( \alpha_{TL} > 0 \) can be written for \( d \leq \overline{n} - 2 \) as

\[
\bar{E}_{TL} = \sigma^2 + \mathbb{E} \left\{ \sum_{k=1}^d g_{kk} \left( n^2 \alpha^2_{TL} C_{TL} + \sigma^2 \lambda_k \{X_{H^{-1}}^T X_{H^{-1}}\} \right) \right\}
\]

where \( g_{kk} \) is the \((k,k)\) component \((k = 1, \ldots, d)\) of the matrix \( G \triangleq (\Phi HH^T \Phi^T)^{-1} \), and

\[
C_{TL} = \frac{\sigma^2}{d} + \frac{\sigma^2}{\overline{n} - d - 1}.
\]

For \( d \in \{\overline{n} - 1, \overline{n}\} \), \( \mathcal{E}_{TL} = \infty \).

The last Lemma is proved in Appendix A.6. Note that the expectation over the sum in (3.15) is only with respect to the eigenvalues \( \lambda_k \{X_{H^{-1}}^T X_{H^{-1}}\} \), \( k = 1, \ldots, d \). Note that for a general \( H \) and underparameterized source task, we get again that \( C_{TL} \) includes the two important aspects for a beneficial transfer learning mechanism: the distance between the tasks as expressed by \( \sigma^2 \), and the accuracy of the source task solution which is related to the interplay among \( \sigma^2, \overline{n}, d \).

**Theorem 3.6.** Under Assumptions 1-2, the optimal tuning of the transfer learning solution (\( \alpha_{TL} > 0 \)) from (3.3) is achieved for \( d \leq \overline{n} - 2 \) by setting \( \alpha_{TL} \) to

\[
\alpha_{TL}^{opt} = \frac{\sigma^2}{nC_{TL}}.
\]
and the corresponding minimal out-of-sample error is

\[ E^\text{opt}_{\text{TL}} = \sigma^2 \left( 1 + \mathbb{E} \left\{ \sum_{k=1}^d \frac{g_{kk}}{\lambda_k \{X^T_{H^{-1}} X_{H^{-1}}\} + n_{\text{TL}}} \right\} \right). \]

For \( d \in \{\tilde{n} - 1, \tilde{n}\} \), \( E_{\text{TL}} = \infty \) for any \( \alpha_{\text{TL}} > 0 \).

Theorem 3.6 is proved in Appendix A.7.

The blue (dashed line) curves in Fig. 3 show the empirical out-of-sample error \( \bar{E}^\text{opt}_{\text{TL}} \) of the optimally tuned transfer learning under Assumptions 1-2 for various instances of the task relation model. Specifically, each subfigure in Fig. 3 considers a different pair of operator \( H \) and noise variance \( \sigma^2 \). In Fig. 3, \( H \) is a circulant matrix that corresponds to circular convolution with the discrete version (\( d \) uniformly spaced samples) of the kernel function \( h_{\text{ker}}(\tau) = e^{-|\tau|/(\pi \sigma^2)} \) defined for \( \tau \in [0,1] \). Note that the discrete convolution kernel is centered to have its peak value at the computed coordinate. Also, the discrete kernel is normalized such that the circulant matrix \( H \) satisfies \( \frac{1}{d} \|H\|^2_F = 1 \) for any \( d \). Each row of subfigures in Fig. 3 corresponds to a different width parameter \( w_{\text{ker}} \). For a larger \( w_{\text{ker}} \) the operator \( H \) averages a larger neighborhood of coordinates and therefore the source task is less related to the target task, accordingly, there are reduced gains from transfer learning (as can be observed in Fig. 3).

4. Transfer Learning versus Ridge Regression. In this section we compare the transfer learning of (3.1) with the standard ridge regression approach, which is independent of the source task and does not involve any transfer learning aspect.

We start in non-asymptotic settings, i.e., without Assumption 3. The standard ridge regression approach can be formulated as

\[
\hat{\beta}_{\text{ridge}} = \arg \min_{b \in \mathbb{R}^d} \|y - Xb\|_2^2 + n\alpha_{\text{ridge}} \|b\|_2^2
\]

\[
= (X^TX + n\alpha_{\text{ridge}} I_d)^{-1} X^Ty.
\]

Here \( \alpha_{\text{ridge}} > 0 \) is a parameter that its optimal value, which minimizes the expected out-of-sample error of the target task, is \( \alpha^\text{opt}_{\text{ridge}} = \frac{d\sigma^2}{n\bar{\epsilon}} \) and the respective out-of-sample error is

\[ E^\text{opt}_{\text{ridge}} = \sigma^2 \left( 1 + \mathbb{E} \left\{ \text{Tr} \left\{ \left( X^TX + n\alpha^\text{opt}_{\text{ridge}} I_d \right)^{-1} \right\} \right\} \right). \]

Related results for optimally tuned ridge regression were provided, e.g., in [10, 25]. See Appendix B.1 for the proof outline in our notations.

Note that the out-of-sample error formulations for the optimally tuned transfer learning for the case of an orthonormal \( H \) in (3.7) and for the optimally tuned ridge regression in (4.3) are the same except for the optimal regularization parameters \( \alpha^\text{opt}_{\text{TL}} \) and \( \alpha^\text{opt}_{\text{ridge}} \), respectively. Accordingly, the cases where transfer learning is better than ridge regression are characterized for non-asymptotic settings as follows (see proof in Appendix B.2).

Corollary 4.1. Consider \( H = \Psi^T \) where \( \Psi \) is a \( d \times d \) orthonormal matrix. Then, the out-of-sample error of optimally tuned transfer learning, \( E^\text{TL} \), is lower than the out-of-sample error of optimally tuned standard ridge regression, \( E^\text{ridge} \), if
Fig. 3: The out-of-sample error of the target task under isotropic Gaussian assumption on $\beta$. The matrix $H$ is a $d \times d$ circulant matrix corresponding to the discrete version of the continuous-domain convolution kernel $h_{\text{ker}}(\tau) = e^{-|\tau - 0.5|}$, each row of subfigures is for a different kernel width $w_{\text{ker}}$. Analytical results are presented in solid lines: red curves are induced by OLS solutions, green curves are induced by optimally tuned ridge regression. The corresponding empirical results (errors averaged over 300 experiments) are denoted by markers in the relevant colors. The blue dashed-line curves correspond to empirical results of optimally tuned transfer learning (TL). Number of data samples for the target task is $n = 64$ and for the source task is $\tilde{n} = 128$.

$$\alpha_{\text{TL}}^{\text{opt}} > \alpha_{\text{ridge}}^{\text{opt}}, \text{ which is satisfied if}$$

$$\sigma_{\eta}^2 + \frac{d \cdot \sigma_{\xi}^2}{|d - \tilde{n}| - 1} < b \text{ for } d \notin \{\tilde{n} - 1, \tilde{n}, \tilde{n} + 1\},$$

and never for $d \in \{\tilde{n} - 1, \tilde{n}, \tilde{n} + 1\}$.

The formula for the out-of-sample error of the optimally tuned ridge regression in asymptotic settings, i.e., under Assumptions 2-3 (without our transfer learning and source task aspects) was already provided by [25, 5], for completeness of presentation we provide this formulation in our notations in Appendix B.3. In Figures 2-3 we provide the analytical and empirical evaluations of the out-of-sample error of the optimally tuned ridge regression solution (see green curves and markers). The results in Figures 2-3 exhibit that if the source and target tasks are sufficiently related (e.g., subfigures 2a-2c,2e-2g,3a-3b,3e), then the optimally tuned transfer learning approach outperforms the optimally tuned ridge regression solution for all the parameterization levels besides those in the proximity of the shift of the source task from under to over parameterized domains. Remarkably, whereas ridge regression indeed resolves the peak of the double descent of the target task, the examined transfer learning approach reduces the out-of-sample errors much further and for a wide range of parameterization levels.
Corollary 4.1 characterizes the cases where using a sufficiently related source task is more useful than using the true prior of the desired $\beta$. Moreover, we consider $\beta$ to originate in isotropic Gaussian distribution, hence, the optimally tuned ridge regression solution is the minimum MSE estimate of the target task parameters, i.e., the solution that minimizes the out-of-sample error of the target task when only the sample data $X, y$ are given. This means that we demonstrated a case where even though optimally tuned ridge regression is the best approach for solving the individual target task, it is not necessarily the best approach when there is an option to utilize transferred parameters from a sufficiently-related source task that was already solved in a sufficiently accurate manner (where we associate the solution accuracy of the source task with its generalization performance w.r.t. the source task goal). Our results show that the success of our transfer learning approach depends on four aspects, which were also qualitatively described in Fig. 1:

- The distance between the source and target tasks as expressed by the level of noise $\sigma_r^2$ from the task relation model. In Fig. 1 this distance is part of the source-target task distance (the light blue ring area).

- The accuracy of the source task solution, which is reflected by the parameterization level (i.e., the $\tilde{a}, d$ pair) and the noise level $\sigma_r^2$. In Fig. 1 this aspect is denoted by the blue circle area.

- The uncertainty in the prior distribution of $\beta$: In the case of Corollary 4.1 this is the total variance of the components of $\beta$, namely, $b$. In Fig. 1 this is represented by the radius of the green circle area.

5. Conclusions. We have established a new perspective on transfer learning as a regularizer of overparameterized learning. We defined a transfer learning process between two linear regression tasks such that the target task is optimized with regularization on the distance of its learned parameters from parameters transferred from an already computed source task. We showed that the examined transfer learning method resolves the peak in the generalization errors of the OLS solution to the target task. We demonstrated that if the source task is sufficiently related to the target task and solved in sufficient accuracy, then optimally tuned transfer learning can significantly outperform optimally tuned ridge regression over a wide range of parameterization levels. Future extensions may study other optimization formulations such as hybrid regularizes that merge the ideas of ridge regression and parameter transfer, and additional task relation models and their utilization in the transfer learning process.

Appendix A. Proofs for Section 3.

A.1. Proof of Lemma 3.1. The expected out-of-sample error of the transfer learning solution to the target task is developed as follows.

$$
\bar{E}_{\text{TL}} \triangleq \mathbb{E}_\beta \{ \mathcal{E}_{\text{TL}} \} = \sigma_r^2 + \mathbb{E} \left\{ \left\| \hat{\beta}_{\text{TL}} - \beta \right\|^2 \right\}
$$

$$
= \sigma_r^2 + \mathbb{E} \left\{ \left\| \left( X^T X + n \alpha_{\text{TL}} H^T H \right)^{-1} \left( X^T y + n \alpha_{\text{TL}} H^T \theta \right) - \beta \right\|^2 \right\}
$$

$$
= \sigma_r^2 + \mathbb{E} \left\{ \left\| \left( X^T X + n \alpha_{\text{TL}} H^T H \right)^{-1} \left( X^T \epsilon + n \alpha_{\text{TL}} H^T \left( \eta + \left( \hat{\theta} - \theta \right) \right) \right) \right\|^2 \right\}
$$

(A.1)

where we used the relation $\hat{\theta} = \theta + \left( \hat{\theta} - \theta \right) = H \beta + \eta + \left( \hat{\theta} - \theta \right)$ and that $(X^T X + n \alpha_{\text{TL}} H^T H)$ is always invertible under the full-rank assumption on $H$ (i.e.,
Lemma 3.1 considers the case where $H = \Psi^T$ where $\Psi$ is a $d \times d$ orthonormal matrix. Hence, $\Psi^T \Psi = \Psi \Psi^T = I_d$. We denote $X_{\Psi} \triangleq X \Psi$. Because the rows of $X$ have isotropic Gaussian distributions then their transformations by $\Psi^T$ do not change their distribution. Then, we can develop the error expression for $\tilde{E}_{TL}$ from (A.1) into

$$\tilde{E}_{TL} = \sigma_{\epsilon}^2 + \text{Tr} \left\{ \mathbb{E} \left\{ (X_\Psi^T X_\Psi + n\alpha_{TL} I_d)^{-1} (\sigma_{\epsilon}^2 X_\Psi^T X_\Psi + n^2 \alpha_{TL}^2 \Gamma_{TL}) \right\} \right\}$$

where

$$\Gamma_{TL} \triangleq \mathbb{E} \left\{ \eta \eta^T \right\} + \mathbb{E} \left\{ (\hat{\theta} - \theta) (\hat{\theta} - \theta)^T \right\} + \mathbb{E} \left\{ (\hat{\theta} - \theta) \eta^T \right\} + \mathbb{E} \left\{ \eta (\hat{\theta} - \theta)^T \right\}$$

Now we provide two fundamental results that are useful for the following developments. The first result is about the $n \times d$ matrix $Z$ that has i.i.d. standard Gaussian components, therefore the expectation of the $d \times d$ projection matrix $Z^+ Z$ is formulated (almost surely) as

$$(A.4) \quad \mathbb{E} \left\{ Z^+ Z \right\} = I_d \times \begin{cases} \frac{1}{n} & \text{for } d \leq \tilde{n}, \\ \frac{\sigma^2}{d} & \text{for } d > \tilde{n}. \end{cases}$$

The second fundamental result is on the expectation of the pseudoinverse of the $d \times d$ Wishart matrix $Z^+ Z$ that almost surely satisfies

$$(A.5) \quad \mathbb{E} \left\{ (Z^T Z)^+ \right\} = \mathbb{E} \left\{ Z^+ (Z^+)^T \right\} = I_d \times \begin{cases} \frac{1}{n-d-1} & \text{for } d \leq \tilde{n} - 2, \\ \frac{\sigma^2}{n-d} & \text{for } \tilde{n} - 1 \leq d \leq \tilde{n} + 1, \\ \frac{\tilde{n} \cdot \sigma^2}{d - n - 1} + (1 - \frac{n}{d}) \frac{b + \sigma^2}{d} & \text{for } d \geq \tilde{n} + 2. \end{cases}$$

The last result can be proved using the tools given in Theorem 1.3 of [26].

Using the auxiliary results (A.4)-(A.5) we get that

$$E \left\{ (\hat{\theta} - \theta) (\hat{\theta} - \theta)^T \right\} = E \left\{ (Z^+ Z \theta + Z^+ \xi - \theta) (Z^+ Z \theta + Z^+ \xi - \theta)^T \right\}$$

$$= I_d \times \begin{cases} \frac{\sigma^2}{n-d} & \text{for } d \leq \tilde{n} - 2, \\ \frac{\sigma^2}{d} & \text{for } \tilde{n} - 1 \leq d \leq \tilde{n} + 1, \\ \frac{n \cdot \sigma^2}{d - n - 1} + (1 - \frac{n}{d}) \frac{b + \sigma^2}{d} & \text{for } d \geq \tilde{n} + 2. \end{cases}$$

and

$$E \left\{ (\hat{\theta} - \theta) \eta^T \right\} = E \left\{ \eta (\hat{\theta} - \theta)^T \right\} = \begin{cases} 0 & \text{for } d \leq \tilde{n}, \\ \frac{n}{d} - 1 \frac{\sigma^2}{d} I_d & \text{for } d > \tilde{n}. \end{cases}$$

where we also used the result $E_{\theta, \eta} \left\{ \theta \theta^T \right\} = \frac{b + \sigma^2}{d} I_d$, which is due to the task relation model (2.6), Assumption 2, and because $H = \Psi^T$ where $\Psi$ is a $d \times d$ orthonormal matrix. Hence, the matrix form in (A.6), which is a scaled identity matrix, lets us to express the error formula from (A.2) as

$$\tilde{E}_{TL} = \sigma_{\epsilon}^2 + \sum_{k=1}^d \frac{n^2 \alpha_{TL}^2 C_{TL} + \sigma_{\epsilon}^2 \cdot \lambda_k \left\{ X_{\Psi}^T X_{\Psi} \right\}} \left( \lambda_k \left\{ X_{\Psi}^T X_{\Psi} \right\} + n \alpha_{TL} \right)^2$$
where \( \lambda_k \left\{ X^T \Psi X \right\} \) is the \( k \)th eigenvalue of the \( d \times d \) matrix \( X^T \Psi X \), and

\[
C_{\text{TL}} \triangleq \begin{cases} 
\frac{\sigma^2}{d} + \frac{\sigma^2}{\bar{n} - d + 1} & \text{for } d \leq \bar{n} - 2, \\
\infty & \text{for } \bar{n} - 1 \leq d \leq \bar{n} + 1, \\
(1 - \frac{\bar{n}}{d}) \frac{\sigma^2}{d} + \frac{\bar{n}}{d} \left( \frac{\sigma^2}{\bar{n} - 1} + \frac{\sigma^2}{d - \bar{n} - 1} \right) & \text{for } d \geq \bar{n} + 2.
\end{cases}
\]

This concludes the proof of Lemma 3.1.

**A.2. Proof of Theorem 3.2.** The derivative of the error expression for \( \tilde{\mathcal{E}}_{\text{TL}} \) as given in Lemma 3.1 with respect to \( \alpha_{\text{TL}} \) is

\[
\frac{\partial \tilde{\mathcal{E}}_{\text{TL}}}{\partial \alpha_{\text{TL}}} = 2n \left( \alpha_{\text{TL}} n C_{\text{TL}} - \sigma^2 \right) \cdot \mathbb{E} \left\{ \sum_{k=1}^{d} \frac{\lambda_k \left\{ X^T \Psi X \right\}}{n \alpha_{\text{TL}}} \right\}.
\]

Since we consider \( \alpha_{\text{TL}} > 0 \) then the necessary condition for optimality, \( \frac{\partial \tilde{\mathcal{E}}_{\text{TL}}}{\partial \alpha_{\text{TL}}} = 0 \), yields

\[
\alpha_{\text{TL}}^{\text{opt}} = \frac{\sigma^2}{n C_{\text{TL}}},
\]

which is the optimal value of \( \alpha_{\text{TL}} > 0 \) for our transfer learning process when \( H = \Psi^T \) is a transposed orthonormal matrix and \( d \notin \{ \bar{n} - 1, \bar{n}, \bar{n} + 1 \} \). Next, we set the expression for \( \alpha_{\text{TL}}^{\text{opt}} \) in the error expression for \( \tilde{\mathcal{E}}_{\text{TL}} \) from Lemma 3.1 and using some algebra gives, for \( d \notin \{ \bar{n} - 1, \bar{n}, \bar{n} + 1 \},

\[
\tilde{\mathcal{E}}_{\text{TL}}^{\text{opt}} = \sigma^2 \left( 1 + \mathbb{E} \left\{ \sum_{k=1}^{d} \frac{1}{\lambda_k \left\{ X^T \Psi X \right\} + n \alpha_{\text{TL}}^{\text{opt}}} \right\} \right) = \sigma^2 \left( 1 + \mathbb{E}_{X \Psi} \left\{ \text{Tr} \left\{ \left( X^T \Psi X + n \alpha_{\text{TL}}^{\text{opt}} I_d \right)^{-1} \right\} \right\} \right).
\]

Note that for \( d \in \{ \bar{n} - 1, \bar{n}, \bar{n} + 1 \}, C_{\text{TL}} = \infty \) and based on the error expression in (A.8) we get that \( \tilde{\mathcal{E}}_{\text{TL}} = \infty \) for any \( \alpha_{\text{TL}} > 0 \). This concludes the proof outline for Theorem 3.2.

**A.3. Proof of Theorem 3.3.** In the asymptotic setting (i.e., under Assumption 3), the optimal parameter \( \alpha_{\text{TL}}^{\text{opt}} \) from (A.11) goes to its limiting value

\[
\alpha_{\text{TL}, \infty}^{\text{opt}} = \sigma^2 \gamma_{\text{tgt}} \times \begin{cases} 
(\sigma^2 + \frac{\gamma_{\text{src}} \sigma^2}{1 - \gamma_{\text{src}}})^{-1} & \text{for } d \leq \bar{n} - 2, \\
\left( \frac{\gamma_{\text{src}} - 1}{1 - \gamma_{\text{src}}} b + \frac{1}{\gamma_{\text{src}}} \left( \sigma^2 + \frac{\gamma_{\text{src}} \sigma^2}{1 - \gamma_{\text{src}}} \right) \right)^{-1} & \text{for } d \geq \bar{n} + 2.
\end{cases}
\]

Moreover, note that the error expression of optimally tuned transfer learning in (A.12) includes the form of \( \mathbb{E}_{X \Psi} \left\{ \text{Tr} \left\{ \left( X^T \Psi X + n \alpha_{\text{TL}}^{\text{opt}} I_d \right)^{-1} \right\} \right\} \) where \( X \Psi \) is a \( n \times d \) random matrix of i.i.d. Gaussian variables \( \mathcal{N}(0, 1) \). This form, however with a different parameter than \( \alpha_{\text{TL}}^{\text{opt}} \), appears also in the analysis of optimally tuned ridge regression by Dobriban and Wager [25]. Accordingly, we can readily use the results from [25] in conjunction with the limiting value of our parameter \( \alpha_{\text{TL}, \infty}^{\text{opt}} \) from (A.13) and get that

\[
\tilde{\mathcal{E}}_{\text{TL}}^{\text{opt}} \rightarrow \sigma^2 \left( 1 + \gamma_{\text{tgt}} \cdot m \left( -\alpha_{\text{TL}, \infty}^{\text{opt}}; \gamma_{\text{tgt}} \right) \right)
\]
where
\begin{equation}
(A.15)
m (-\alpha_{\text{TL},\infty}^{\text{opt}}; \gamma_{\text{tgt}}) = \frac{1 - \gamma_{\text{tgt}} + \alpha_{\text{TL},\infty}^{\text{opt}}}{2\gamma_{\text{tgt}}\alpha_{\text{TL},\infty}^{\text{opt}}}
\end{equation}
is the Stieltjes transform of the Marchenko-Pastur distribution, which is the limiting spectral distribution of the sample covariance associated with \(n\) samples that are drawn from a Gaussian distribution \(\mathcal{N}(0, I_d)\). This completes the proof outline for Theorem 3.3.

**A.4. Generalization Error of OLS Regression Under Assumption 2.**

The out-of-sample error of the OLS regression solution of the individual target task was provided in (3.2) for a given parameter vector \(\beta\). Then, the expectation of \(\mathcal{E}_{\text{OLS}}\) from (3.2) with respect to the isotropic Gaussian prior on \(\beta\) (i.e., under Assumption 2) is

\begin{equation}
(A.16) \quad \mathbb{E}_{\beta}\{\mathcal{E}_{\text{OLS}}\} = \begin{cases} 
\left(1 + \frac{d}{n-d-1}\right) \sigma_{\epsilon}^2 & \text{for } d \leq n - 2, \\
\infty & \text{for } n - 1 \leq d \leq n + 1, \\
\left(1 + \frac{n}{d-n-1}\right) \sigma_{\epsilon}^2 + \left(1 - \frac{b}{n}\right) b & \text{for } d \geq n + 2.
\end{cases}
\end{equation}

**A.5. Proof of Proposition 3.4.**

\begin{equation}
(A.17) \quad \mathbb{E}\{\hat{\theta}|\beta\} = \mathbb{E}\{Z^v|\beta\}
= \mathbb{E}\{Z^v (Z\theta + \xi)|\beta\} = \mathbb{E}\{Z^v (H\beta + \eta)|\beta\}
= \mathbb{E}\{Z^v Z\} H\beta
\end{equation}

The last development relies on the fundamental result from (A.4).

Now, we continue to the covariance matrix of \(\hat{\theta}\) given \(\beta\), namely,

\begin{equation}
(A.18) \quad \mathbb{E}\left\{\left(\hat{\theta} - \mathbb{E}\{\hat{\theta}|\beta\}\right) \left(\hat{\theta} - \mathbb{E}\{\hat{\theta}|\beta\}\right)^T\right\} = \mathbb{E}\{\hat{\theta}\hat{\theta}^T|\beta\} - \mathbb{E}\{\hat{\theta}|\beta\} \left(\mathbb{E}\{\hat{\theta}|\beta\}\right)^T.
\end{equation}

Using (A.17) we can easily get that

\begin{equation}
(A.19) \quad \mathbb{E}\{\hat{\theta}|\beta\} \left(\mathbb{E}\{\hat{\theta}|\beta\}\right)^T = H\beta \beta^T H^T \times \begin{cases} 1 & \text{for } d \leq \tilde{n}, \\
\frac{1}{\tilde{n}} & \text{for } d > \tilde{n}.
\end{cases}
\end{equation}

We also need analytical formulation for \(\mathbb{E}\{\hat{\theta}\hat{\theta}^T|\beta\}\), as explained next.

\begin{equation}
(A.20) \quad \mathbb{E}\{\hat{\theta}\hat{\theta}^T|\beta\} = \mathbb{E}\{Z^v (Z (H\beta + \eta) + \xi) (Z (H\beta + \eta) + \xi)^T (Z^v)^T\|\beta\}
= \mathbb{E}\{Z^v Z H\beta \beta^T Z^v (Z^v)^T\|\beta\} + \mathbb{E}\{Z^v Z \eta \eta^v (Z^v)^T\} + \mathbb{E}\{Z^v \xi \xi^v (Z^v)^T\}
\end{equation}
where the second term in the last expression can be explicitly formulated using (A.4). The third term in (A.20) requires the fundamental result from (A.5). The first term in (A.20) is an instance of the more general form \( \mathbb{E}\{Z^+Zaa^T(Z^+Z)^T\} \), where \( a \in \mathbb{R}^d \) is a non-random vector. For \( d \leq \bar{n} \), we almost surely have that \( Z^+Z = I_d \) and therefore
\[
\mathbb{E}\{Z^+Zaa^T(Z^+Z)^T\} = aa^T.
\]
For \( d > \bar{n} \), consider the decomposition \( Z^+Z = RR^T \) where \( R \) is a \( d \times \bar{n} \) matrix with \( \bar{n} \) orthonormal columns that are taken from a random orthonormal matrix that is uniformly distributed over the set of \( d \times d \) orthonormal matrices (i.e., the Haar distribution of matrices). Then, using the non-asymptotic properties of Haar-distributed matrices (see, e.g., Lemma 2.5 in [27] and Proposition 1.2 in [28]) and some algebra, one can prove that, for \( d > \bar{n} \),
\[
\begin{align*}
\mathbb{E}\{Z^+Zaa^T(Z^+Z)^T\} &= \frac{\bar{n}}{d} \left( \frac{\bar{n}^2 + 1}{d + 1} aa^T + \frac{d - \bar{n}}{d(d - 1)} \text{diag} \left( \{||a||_2^2 - (a_j)^2\}_{j=1,\ldots,d} \right) \right) \\
\end{align*}
\]
where \( a_j \) is the \( j \)-th component of the vector \( a \). Based on the described proof outline, one can use (A.20) to develop (A.18) into the form
\[
\begin{align*}
\mathbb{E}\left( \hat{\theta} - \mathbb{E}\{\hat{\theta}|\beta\}\right) \left( \hat{\theta} - \mathbb{E}\{\hat{\theta}|\beta\}\right)^T | \beta \rangle = \left( \frac{\sigma_n^2}{d} + \frac{\sigma_x^2}{\bar{n} - d - 1} \right) I_d
\end{align*}
\]
for \( d \leq \bar{n} - 2 \), and
\[
\begin{align*}
\mathbb{E}\left( \hat{\theta} - \mathbb{E}\{\hat{\theta}|\beta\}\right) \left( \hat{\theta} - \mathbb{E}\{\hat{\theta}|\beta\}\right)^T | \beta \rangle = \\
\bar{n} \left( \frac{d - \bar{n}}{d(d + 1)} H\beta \beta^T H^T + \frac{d - \bar{n} + 1}{d(d - 1)} \text{diag} \left( \{||H\beta||_2^2 - (H\beta)_j^2\}_{j=1,\ldots,d} \right) + \left( \frac{\sigma_n^2}{d} + \frac{\sigma_x^2}{\bar{n} - d - 1} \right) I_d \right)
\end{align*}
\]
for \( d \geq \bar{n} + 2 \). In (A.23), \( (H\beta)_j \) is the \( j \)-th component of the vector \( H\beta \). For \( d \in \{\bar{n} - 1, \bar{n}, \bar{n} + 1\} \) the covariance matrix is infinite valued as a result of the infinite valued \( \mathbb{E}\{\{Z^T Z\}^+\} \), see (A.5).

\textbf{A.6. Proof of Lemma 3.5.} Similarly to (A.1)-(A.2) that were given above for the case of orthonormal \( H \), one can express the expected error for the case of a general form of \( H \) as
\[
\tilde{\xi}_{TL} = \sigma^2 + \mathbb{E}\left( \parallel X^T X + n\alpha_{TL} H^T H^{-1} \left( X^T \epsilon + n\alpha_{TL} H^T \left( \eta + (\hat{\theta} - \theta) \right) \right) \parallel^2 \right)
\]
Due to Assumption 1, \( H \) is always invertible and this lets us to define \( X_{H^{-1}} \equiv XH^{-1} \). Moreover, the covariance matrix of \( \hat{\theta} \) given \( \beta \) is a scaled identity matrix (see Proposition 3.4, thus, for \( d \leq \bar{n} - 2 \) we can express the error as
\[
\tilde{\xi}_{TL} = \sigma^2 + \mathbb{E}\{\{H(H^T)^{-1}(X_{H^{-1}},X_{H^{-1}}+n\alpha_{TL}L)^{-1}\}{(\sigma^2\epsilon_{X_{H^{-1}},X_{H^{-1}}+n\alpha_{TL}L}^2 + \sigma_x^2)^\frac{1}{2}}L\} \}
\]
Consider the eigendecomposition \( X_{H^{-1}}^T X_{H^{-1}} = \Phi A_{TL} \Phi^T \) where \( \Phi \) is an orthonormal matrix with the eigenvectors of \( X_{H^{-1}}^T X_{H^{-1}} \) as its columns and \( A_{TL} \) is a diagonal matrix with the respective eigenvalues \( \lambda_k \) \{\( X_{H^{-1}}^T X_{H^{-1}} \}, k = 1, \ldots, d \), on its main diagonal. Then, we can rewrite (A.25) as
\[
\tilde{\xi}_{TL} = \sigma^2 + \mathbb{E}\left( \text{tr} \left( \Phi (HH^{-1})^{-1} \Phi (A_{TL} + n\alpha_{TL}L)^{-1} \left( \sigma^2 A_{TL} + n\alpha_{TL}^2 \left( \frac{\sigma^2}{d} + \frac{\sigma_x^2}{\bar{n} - d - 1} \right) I_d \right) (A_{TL} + n\alpha_{TL}L)^{-1} \right) \right)
\]

\[Y.\) DAR AND R. G. BARANIUK\]
that, by denoting $G \triangleq \Phi^T (HH^T)^{-1} \Phi$, can be also written as
\[ (A.27) \]
\[ \tilde{E}_{TL} = \sigma^2 + \mathbb{E}\left\{ \text{Tr}\left\{ G (A_{TL} + n\alpha_{TL}I_d)^{-1} \left( \frac{\sigma^2}{\tau} + \frac{\sigma^2}{n-d-1} \right) I_d \right\} (A_{TL} + n\alpha_{TL}I_d)^{-1} \right\}. \]

The trace operator in the last expression is applied on the product of the matrix $G$ and a diagonal matrix form, hence, \((A.27)\) can be written as
\[ (A.28) \]
\[ \tilde{E}_{TL} = \sigma^2 + \mathbb{E}\left\{ \sum_{k=1}^{d} \frac{g_{kk}}{\lambda_k} \left( \frac{X_{H-1}^T X_{H-1}}{\lambda_k} + n\alpha_{TL}I_d \right)^2 \right\}. \]

where $g_{kk}$ is the $(k, k)$ component $(k = 1, \ldots, d)$ of the matrix $G$ and

\[ (A.29) \]
\[ C_{TL} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n-d-1}. \]

Recall that the results in \((A.28)-(A.29)\) are for $d \leq \tilde{n} - 2$.

For $d \in \{ \tilde{n} - 1, \tilde{n} \}$, we got in Proposition 3.4 that $\mathbb{E}\left\{ (\hat{\theta} - \theta) (\hat{\theta} - \theta)^T \right\}$ is infinite, therefore $\tilde{E}_{TL} = \infty$.

**A.7. Proof of Theorem 3.6.** In the case where $H$ has a general form, the derivative of the error expression for $\tilde{E}_{TL}$ as given in Lemma 3.5 with respect to $\alpha_{TL}$ is
\[ (A.30) \]
\[ \frac{\partial \tilde{E}_{TL}}{\partial \alpha_{TL}} = 2n (\alpha_{TL} n C_{TL} - \sigma^2) \cdot \mathbb{E}\left\{ \sum_{k=1}^{d} g_{kk} \frac{X_{H-1}^T X_{H-1}}{\lambda_k} \left( \frac{X_{H-1}^T X_{H-1}}{\lambda_k} + n\alpha_{TL}I_d \right)^2 \right\}. \]

We consider $\alpha_{TL} > 0$ and therefore the necessary condition for optimality, $\frac{\partial \tilde{E}_{TL}}{\partial \alpha_{TL}} = 0$, yields
\[ (A.31) \]
\[ \alpha_{opt}^{TL} = \frac{\sigma^2}{n C_{TL}}, \]

which is the optimal value of $\alpha_{TL} > 0$ for the examined transfer learning mechanism when $H$ has a general form and $d \leq \tilde{n} - 2$. Setting the expression for $\alpha_{opt}^{TL}$ in the error expression for $\tilde{E}_{TL}$ from Lemma 3.5 shows that, for $d \leq \tilde{n} - 2$,
\[ (A.32) \]
\[ \tilde{E}_{TL}^{opt} = \sigma^2 \left( 1 + \mathbb{E}\left\{ \sum_{k=1}^{d} \frac{g_{kk}}{\lambda_k} \left( \frac{X_{H-1}^T X_{H-1}}{\lambda_k} + n\alpha_{opt}^{TL}I_d \right)^2 \right\} \right). \]

**Appendix B. Proofs and Details for Section 4.**

**B.1. Generalization Error of Ridge Regression in Non-Asymptotic Settings.** The expected out-of-sample error of the ridge regression solution of the (indi-
individual target task can be developed as outlined next.

\[
\bar{E}_{\text{ridge}} \triangleq \mathbb{E}_{\beta} \left\{ \mathbb{E}_{\text{ridge}} \right\} = \sigma_{\epsilon}^2 + \mathbb{E} \left\{ \left\| \beta_{\text{ridge}} - \beta \right\|_2^2 \right\}
\]

\[
= \sigma_{\epsilon}^2 + \mathbb{E} \left\{ \left\| \left( X^T + n\alpha_{\text{ridge}} I_d \right)^{-1} X^T \mathbf{y} - \beta \right\|_2^2 \right\}
\]

\[
= \sigma_{\epsilon}^2 + \mathbb{E} \left\{ \left\| \left( X^T + n\alpha_{\text{ridge}} I_d \right)^{-1} X^T \mathbf{e} \right\|_2^2 \right\}
\]

\[
+ \mathbb{E} \left\{ \left\| \left( X^T + n\alpha_{\text{ridge}} I_d \right)^{-1} X^T \mathbf{X} - \mathbf{I}_d \right) \beta \right\|_2^2 \right\}
\]

(B.1)

where we used the fact that \( \mathbf{e} \) is independent of \( X \). Consider the eigendecomposition

(B.2)

\[
X^T X = \Phi_X \Lambda_X \Phi_X^T
\]

where \( \Phi_X \) is a \( d \times d \) orthonormal matrix with columns being eigenvectors of \( X^T X \) and the corresponding eigenvalues \( \lambda_k \{ X^T X \}, k = 1, \ldots, d \), are on the main diagonal of the \( d \times d \) diagonal matrix \( \Lambda_X \). Then, we can continue to develop (B.1) as follows.

\[
\bar{E}_{\text{ridge}} = \sigma_{\epsilon}^2 + \sigma_{\epsilon}^2 \mathbb{E} \left\{ \text{Tr} \left\{ \left( \Lambda_X + n\alpha_{\text{ridge}} I_d \right)^{-2} \Lambda_X \right\} \right\}
\]

\[
+ \frac{b}{d} \mathbb{E} \left\{ \text{Tr} \left\{ \left( \left( \Lambda_X + n\alpha_{\text{ridge}} I_d \right)^{-1} \Lambda_X - I_d \right)^2 \right\} \right\}
\]

(B.3)

By equating the derivative (w.r.t. \( \alpha_{\text{ridge}} \)) of the expression in (B.3) to zero, one can obtain the \( \alpha_{\text{ridge}} > 0 \) that minimizes the error \( \bar{E}_{\text{ridge}} \). The suggested calculations show that \( \alpha_{\text{ridge}}^{\text{opt}} = \frac{d\sigma_{\epsilon}^2}{nb} \). By setting \( \alpha_{\text{ridge}}^{\text{opt}} \) back in (B.3) one can show that the minimal expected out-of-sample error for ridge regression is

(B.4)

\[
\bar{E}_{\text{opt}}^{\text{ridge}} = \sigma_{\epsilon}^2 \left( 1 + \mathbb{E}_X \left\{ \text{Tr} \left\{ \left( X^T X + n\alpha_{\text{ridge}}^{\text{opt}} I_d \right)^{-1} \right\} \right\} \right).
\]

**B.2. Proof Outline for Corollary 4.1.** Consider \( H = \Psi^T \) and \( \Psi \) is an orthonormal matrix. The main case to be proved is for \( d \notin \{ \tilde{n} - 1, \tilde{n}, \tilde{n} + 1 \} \). Then, according to Theorem 3.2 and (A.12), the out-of-sample error of optimally tuned transfer learning can be written as

(B.5)

\[
\bar{E}_{\text{TL}}^{\text{opt}} = \sigma_{\epsilon}^2 \left( 1 + \sum_{k=1}^{d} \mathbb{E}_{\lambda_k} \left\{ \frac{1}{\lambda_k \{ X^T \Psi X\} + n\alpha_{\text{TL}}^{\text{opt}}} \right\} \right)
\]

where \( X_{\Psi} \) is a \( n \times d \) matrix of i.i.d. standard Gaussian variables. Note that the eigenvalues \( \lambda_k \{ X^T \Psi X\} \) are i.i.d. random variables.

The optimally tuned ridge regression solution has the out-of-sample error (B.4) that can be also expressed as

(B.6)

\[
\bar{E}_{\text{ridge}}^{\text{opt}} = \sigma_{\epsilon}^2 \left( 1 + \sum_{k=1}^{d} \mathbb{E}_{\lambda_k} \left\{ \frac{1}{\lambda_k \{ X^T X\} + n\alpha_{\text{ridge}}^{\text{opt}}} \right\} \right)
\]
where $X$ is a $n \times d$ matrix of i.i.d. standard Gaussian variables. Note that the eigenvalues $\lambda_k \{X^T X\}$ are i.i.d. random variables.

Because $X$ and $X\Psi$ have the same distribution, then their eigenvalues $\lambda_k \{X^T X\}$ and $\lambda_k \{X^T \Psi X\}$ are also identically distributed. Therefore, the only difference between (B.5) and (B.6) is the respective values of $\alpha_{\text{TL}}^{\text{opt}}$ and $\alpha_{\text{ridge}}^{\text{opt}}$. Then, according to the forms in (B.5)-(B.6), $\bar{E}_{\text{TL}}^{\text{opt}} < \bar{E}_{\text{ridge}}^{\text{opt}}$ when $\alpha_{\text{TL}}^{\text{opt}} > \alpha_{\text{ridge}}^{\text{opt}}$. According to Theorem 3.2, the condition $\alpha_{\text{TL}}^{\text{opt}} > \alpha_{\text{ridge}}^{\text{opt}}$ is satisfied when $\frac{\sigma^2}{nC_{\text{TL}}^2} > \frac{d\sigma^2}{nb}$ where $C_{\text{TL}}$ is defined in Lemma 3.1 for the case of $H = \Psi^T$. This leads to the condition

$$\frac{\sigma^2}{n} + \frac{d \cdot \sigma^2}{|d - \tilde{n}| - 1} < b$$

for $d \notin \{\tilde{n} - 1, \tilde{n}, \tilde{n} + 1\}$.

Theorem 3.2 states that the transfer learning error is infinite for $d \in \{\tilde{n} - 1, \tilde{n}, \tilde{n} + 1\}$. Hence, $\bar{E}_{\text{TL}}^{\text{opt}} < \bar{E}_{\text{ridge}}^{\text{opt}}$ is never satisfied for $d \in \{\tilde{n} - 1, \tilde{n}, \tilde{n} + 1\}$.

**B.3. Generalization Error of Ridge Regression in Asymptotic Settings.**

Previous studies [25, 5] already provided the analytical formula for the expected out-of-sample error of ridge regression when the true parameter vector ($\beta$ in our case) originates at isotropic Gaussian distribution and the sample data matrix ($X$ in our case) has i.i.d. Gaussian $\mathcal{N}(0, 1)$ components. Then, translating the results from [25, 5] to our notations shows that

$$\bar{E}_{\text{ridge}}^{\text{opt}} \rightarrow \sigma^2 \left( 1 + \gamma_{\text{tgt}} \cdot m \left( -\alpha_{\text{ridge}, \infty}^{\text{opt}}; \gamma_{\text{tgt}} \right) \right)$$

where $\alpha_{\text{ridge}, \infty}^{\text{opt}} = \frac{\gamma_{\text{tgt}}^{\infty} \sigma^2}{b}$ is the limiting value of $\alpha_{\text{ridge}}^{\text{opt}}$, and

$$m \left( -\alpha_{\text{ridge}, \infty}^{\text{opt}}; \gamma_{\text{tgt}} \right) = -\frac{(1 - \gamma_{\text{tgt}} + \alpha_{\text{ridge}, \infty}^{\text{opt}})}{2\gamma_{\text{tgt}} \alpha_{\text{ridge}, \infty}^{\text{opt}}} - \sqrt{\frac{(1 - \gamma_{\text{tgt}} + \alpha_{\text{ridge}, \infty}^{\text{opt}})^2 + 4\gamma_{\text{tgt}} \alpha_{\text{ridge}, \infty}^{\text{opt}}}{2\gamma_{\text{tgt}} \alpha_{\text{ridge}, \infty}^{\text{opt}}}}$$

is the Stieltjes transform of the Marchenko-Pastur distribution. For more details see [25, 5].

**REFERENCES**

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In ICLR, 2017.

[2] S. Spigler, M. Geiger, S. d’Ascoli, L. Sagun, G. Biroli, and M. Wyart. A jamming transition from under-to over-parametization affects loss landscape and generalization. arXiv preprint arXiv:1810.09665, 2018.

[3] 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. arXiv preprint arXiv:1901.01608, 2019.

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

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

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

[7] J. Xu and D. J. Hsu. On the number of variables to use in principal component regression. In Advances in Neural Information Processing Systems (NeurIPS), pages 5095–5104, 2019.
[8] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. arXiv preprint arXiv:1908.05355, 2019.

[9] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. IEEE Journal on Selected Areas in Information Theory, 2020.

[10] P. Nakkiran, P. Venkat, S. Kakade, and T. Ma. Optimal regularization can mitigate double descent. arXiv preprint arXiv:2003.01897, 2020.

[11] S. D’Ascoli, M. Refinetti, G. Biroli, and F. Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 2280–2290. PMLR, 13–18 Jul 2020.

[12] Y. Dar, P. Mayer, L. Luzi, and R. G. Baraniuk. Subspace fitting meets regression: The effects of supervision and orthonormality constraints on double descent of generalization errors. In International Conference on Machine Learning (ICML), pages 2366–2375, 2020.

[13] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? arXiv preprint arXiv:2005.08054, 2020.

[14] S. J. Pan and Q. Yang. A survey on transfer learning. IEEE transactions on knowledge and data engineering, 22(10):1345–1359, 2009.

[15] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In ICML workshop on unsupervised and transfer learning, pages 17–36, 2012.

[16] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5):1285–1298, 2016.

[17] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In International Conference on Machine Learning (ICML), pages 2208–2217, 2017.

[18] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich. To transfer or not to transfer. In NIPS workshop on transfer learning, 2005.

[19] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In IEEE conference on computer vision and pattern recognition (CVPR), pages 3712–3722, 2018.

[20] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In IEEE conference on computer vision and pattern recognition (CVPR), pages 2661–2671, 2019.

[21] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In Advances in neural information processing systems, pages 3347–3357, 2019.

[22] A. K. Lampinen and S. Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In International Conference on Learning Representations (ICLR), 2019.

[23] Y. Dar and R. G. Baraniuk. Double double descent: On generalization errors in transfer learning between linear regression tasks. arXiv preprint arXiv:2006.07002, 2020.

[24] L. Chizat and F. Bach. A note on lazy training in supervised differentiable programming. arXiv preprint arXiv:1812.07956, 1, 2018.

[25] E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. The Annals of Statistics, 46(1):247–279, 2018.

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

[27] A. M. Tulino and S. Verdú. Random matrix theory and wireless communications. Now Publishers Inc, 2004.

[28] F. Hiai and D. Petz. Asymptotic freeness almost everywhere for random matrices. Acta Sci. Math. Szeged, 66:801–826, 2000.