Double Double Descent: On Generalization Errors in Transfer Learning between Linear Regression Tasks

Yehuda Dar* and Richard G. Baraniuk†

Abstract. We study the *transfer learning* process between two linear regression problems. An important and timely special case is when the regressors are *overparameterized* and perfectly *interpolate* their training data. We examine a parameter transfer mechanism whereby a subset of the parameters of the target task solution are constrained to the values learned for a related source task. We analytically characterize the generalization error of the target task in terms of the salient factors in the transfer learning architecture, i.e., the number of examples available, the number of (free) parameters in each of the tasks, the number of parameters transferred from the source to target task, and the relation between the two tasks. Our non-asymptotic analysis shows that the generalization error of the target task follows a *two-dimensional double descent* trend (with respect to the number of free parameters in each of the tasks) that is controlled by the transfer learning factors. Our analysis points to specific cases where the transfer of parameters is beneficial as a substitute for extra overparameterization (i.e., additional free parameters in the target task). Specifically, we show that the usefulness of a transfer learning setting is fragile and depends on a delicate interplay among the set of transferred parameters, the relation between the tasks, and the true solution. We also demonstrate that overparameterized transfer learning is not necessarily more beneficial when the source task is closer or identical to the target task.

Key words. Overparameterized learning, linear regression, transfer learning, double descent.

AMS subject classifications. 62J05, 68Q32

1. Introduction.

Transfer learning [21] is a prominent strategy to address a machine learning task of interest using information and parameters already learned and/or available for a related task. Such designs significantly aid training of overparameterized models like deep neural networks (e.g., [4, 24, 16]), which are inherently challenging due to the vast number of parameters compared to the number of training data examples. There are various ways to integrate the previously-learned information from the source task in the learning process of the target task; often this is done by taking subsets of parameters (e.g., layers in neural networks) learned for the source task and plugging them in the target task model as parameter subsets that can be set fixed, finely tuned, or serve as non-random initialization for a thorough learning process. Obviously, transfer learning is useful only if the source and target tasks are sufficiently related with respect to the transfer mechanism utilized, e.g., [23, 29, 13]. Moreover, finding a successful transfer learning setting for deep neural networks was shown in [22] to be a delicate engineering task. The importance of transfer learning in contemporary practice should motivate fundamental understanding of its main aspects via *analytical* frameworks that may consider linear structures (see, e.g., [14]).

In general, the impressive success of overparameterized architectures for supervised learning have raised fundamental questions on the classical role of the bias-variance tradeoff that guided the traditional designs towards seemingly-optimal underparameterized models [5]. Em-

*Electrical and Computer Engineering Department, Rice University (ydar@rice.edu).
†Electrical and Computer Engineering Department, Rice University (richb@rice.edu).

pirical studies from recent years [25, 9, 2] have demonstrated the phenomenon that overparameterized supervised learning corresponds to a generalization error curve with a *double descent* trend (with respect to the number of parameters in the learned model). This double descent shape means that the generalization error peaks when the learned model starts to interpolate the training data (i.e., to achieve zero training error), but then the error continuously decreases as the overparmeterization increases, often arriving to a global minimum that outperforms the best underparameterized solution. This phenomenon has been studied theoretically from the linear regression perspective in an extensive series of papers, e.g., in [3, 11, 28, 17, 1, 19]. The next stage is to provide corresponding fundamental understanding to learning problems beyond a single fully-supervised regression problem (see, for example, the study in [7] on overparameterized linear subspace fitting in unsupervised and semi-supervised settings).

In this paper we study the fundamentals of the natural meeting point between *overparameterized models* and the *transfer learning* concept. Our analytical framework is based on the least squares solutions to two related linear regression problems: the first is a source task whose solution has been found independently, and the second is a target task that is addressed using the solution already available for the source task. Specifically, the target task is carried out while keeping a subset of its parameters fixed to values transferred from the source task solution. Accordingly, the target task includes three types of parameters: free to-be-optimized parameters, transferred parameters set fixed to values from the source task, and parameters fixed to zeros (which in our case correspond to the elimination of input features). The mixture of the parameter types defines the parameterization level (i.e., the relation between the number of free parameters and the number of examples given) and the transfer-learning level (i.e., the portion of transferred parameters among the solution layout).

We conduct a non-asymptotic statistical analysis of the generalization errors in this transfer learning structure where the minimum ℓ_2-norm solutions are used when the source and/or target tasks are overparameterized. Clearly, since the source task is solved independently, its generalization error follows a regular (one-dimensional) double descent shape with respect to the number of examples and free parameters available in the source task. Hence, our main contribution and interest are in the characterization of the generalization error of the target task that is carried out using the transfer learning approach described above. We show that the generalization error of the target task follows a double descent trend that depends on the double descent shape of the source task and on the transfer learning factors such as the number of parameters transferred and the relation between the source and target tasks. We also examine the generalization error of the target task as a function of two quantities: the number of free parameters in the source task and the number of free parameters in the target task. This interpretation presents the generalization error of the target task as having a *two-dimensional double descent* trend.

We also show how the generalization error of the target task is affected by the *specific set* of transferred parameters and its delicate interplay with the forms of the true solution and the source-target task relation. By that, we provide an analytical theory to the fragile nature of successful transfer learning designs. We demonstrate that the practical approach of arbitrary selection of transferred parameters signifies the fragile nature of successful parameter transfer settings, especially when the true (unknown) parameters have a sparse form in the feature

space.

We characterize the settings where transferring a set of parameters is more beneficial (in the sense of improved generalization in the target task) than defining them as additional free parameters or zeroing them. We also prove that the transfer of parameters from an overparameterized solution of a source task is not necessarily optimal when the source task is closer or identical to the target task.

The majority of this paper is focused on the analytical and empirical study of the minimum ℓ_2-norm solution in our transfer learning setting. Yet, we also empirically examine the utilization of the minimum ℓ_1-norm solution and the ridge regression approach in our transfer learning framework. The minimum ℓ_1-norm (interpolating) solution also induces generalization errors that follow a double descent shape, and can outperform the minimum ℓ_2-norm solution especially if the true parameters are sparse. The ridge regression method includes a regularization term that prevents interpolation and, when properly tuned, eliminates the double descent peak. Ridge regression can outperform the interpolating solutions for a wide range of overparameterization levels. However, at the proximity of maximal overparameterization, the minimum ℓ_2-norm solution performs comparably to the ridge regression. This demonstrates that interpolation at extreme overparameterization levels can substitute properly tuned regularization in our transfer learning setting.

1.1. Related Work. Despite the prevalence of transfer learning in contemporary practice, there are only few analytical theories for this topic. First, Lampinen and Ganguli [14] studied the optimization dynamics in transfer learning of multi-layer linear models, which is a different research objective than our focus on double descent phenomena. Interestingly, when we posted the first version of our work on arXiv in June 2020, there was no literature on double descent phenomena and interpolating solutions in transfer learning. Later on, Dhifallah and Lu [8] studied single-layer nonlinear models that are suitable for both regression and classification problems, where the source task is ridge regularized and therefore prevents interpolation and double descent phenomena. Gerace at al. [10] studied a binary classification problem that is addressed by transfer learning of the first layer in a two-layer model that includes nonlinearities. The learning settings in [10] include regularization on both the source and target task and therefore attenuate some of the double descent behavior. The analysis in [10] requires numerical optimizations and empirical estimation of covariance matrices as inputs for their asymptotic formulations.

Remotely from overparamterized learning but related to transfer learning theory, Obst et al. [20] studied fine-tuning (based on gradient descent) between linear regression problems in underparameterized settings.

To summarize, each of the existing transfer learning theories employs different analytical tools, assumptions, and accordingly presents a unique analytical perspective on the topic. We focus on transfer learning between linear regression problems for Gaussian data and without regularization. This allows us to consider the minimum ℓ_2-norm solution and explicitly characterize the generalization error and its double descent behavior in a detailed, completely closed-form formulation that is unavailable elsewhere. This also contributes to our discussion on beneficial transfer learning and lets us to analytically characterize the optimal source task to transfer from.

1.2. Paper Organization. This paper is organized as follows. In Section 2 we define the transfer learning architecture examined in this paper. In Section 3 we analytically characterize the double descent phenomenon in our transfer learning setting. In Section 4 we analyze the conditions for beneficial transfer of parameters compared to the alternatives of free and zeroed parameters. In Section 5 we characterize the optimal source task to transfer from. Sections 3-5 are focused on the minimum ℓ_2-norm solution in the overparameterized regime of our transfer learning setting; in Section 6 we examine transfer learning in conjunction with the minimum ℓ_1-norm solution and the ridge regression method. Section 7 concludes the paper. The Supplementary Materials include all of the proofs and mathematical developments as well as additional details and results for the empirical part of the paper.

2. Transfer Learning between Linear Regression Tasks: Problem Definition.

2.1. Source Task: Data Model and Solution Form. We start with the *source* data model, where a d-dimensional Gaussian input vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ is connected to a response value $v \in \mathbb{R}$ via the noisy linear model

$$v = \mathbf{z}^T \boldsymbol{\theta} + \xi, \tag{2.1}$$

where $\xi \sim \mathcal{N}\left(0, \sigma_\xi^2\right)$ is a Gaussian noise component independent of $\mathbf{z}$, $\sigma_\xi > 0$, and $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown vector. The data user is unfamiliar with the distribution of $(\mathbf{z}, v)$, however gets a dataset of $\widetilde{n}$ independent and identically distributed (i.i.d.) draws of $(\mathbf{z}, v)$ pairs denoted as $\widetilde{\mathcal{D}} \triangleq \left\{\left(\mathbf{z}^{(i)}, v^{(i)}\right)\right\}_{i=1}^{\widetilde{n}}$. The $\widetilde{n}$ data samples can be rearranged as $\mathbf{Z} \triangleq [\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(\widetilde{n})}]^T$ and $\mathbf{v} \triangleq [v^{(1)}, \ldots, v^{(\widetilde{n})}]^T$ that satisfy the relation $\mathbf{v} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \triangleq [\xi^{(1)}, \ldots, \xi^{(\widetilde{n})}]^T$ is an unknown noise vector that its i^{th} component $\xi^{(i)}$ participates in the relation $v^{(i)} = \mathbf{z}^{(i),T}\boldsymbol{\theta} + \xi^{(i)}$ underlying the i^{th} data sample.

The *source* task is defined for a new (out of sample) data pair $\left(\mathbf{z}^{(\text{test})}, v^{(\text{test})}\right)$ drawn from the distribution induced by (2.1) independently of the $\widetilde{n}$ examples in $\widetilde{\mathcal{D}}$. For a given $\mathbf{z}^{(\text{test})}$, the source task is to estimate the response value $v^{(\text{test})}$ by the value $\widehat{v}$ that minimizes the corresponding out-of-sample squared error (i.e., the generalization error of the *source* task)

$$\widetilde{\mathcal{E}}_{\text{out}} \triangleq \mathbb{E}\left[\left(\widehat{v} - v^{(\text{test})}\right)^2\right] = \sigma_\xi^2 + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2^2\right] \tag{2.2}$$

where the second equality stems from the data model in (2.1) and the corresponding linear form of $\widehat{v} = \mathbf{z}^{(\text{test}),T}\widehat{\boldsymbol{\theta}}$ where $\widehat{\boldsymbol{\theta}}$ estimates $\boldsymbol{\theta}$ based on $\widetilde{\mathcal{D}}$.

To address the source task based on the $\widetilde{n}$ examples, one should choose the number of free parameters in the estimate $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$. Consider a predetermined layout where $\widetilde{p}$ out of the d components of $\widehat{\boldsymbol{\theta}}$ are free to be optimized, whereas the remaining $d - \widetilde{p}$ components are constrained to zero values. The coordinates of the free parameters are specified in the set $\mathcal{S} \triangleq \{s_1, \ldots, s_{\widetilde{p}}\}$ where $1 \le s_1 < \cdots < s_{\widetilde{p}} \le d$ and the complementary set $\mathcal{S}^{\text{c}} \triangleq \{1, \ldots, d\} \setminus \mathcal{S}$ contains the coordinates constrained to be zero valued. We define the $|\mathcal{S}| \times d$ matrix $\mathbf{Q}_{\mathcal{S}}$ as the linear operator that extracts from a d-dimensional vector its $|\mathcal{S}|$-dimensional subvector of components residing at the coordinates specified in $\mathcal{S}$. Specifically, the values of the (k, s_k) components $(k = 1, \ldots, |\mathcal{S}|)$ of $\mathbf{Q}_{\mathcal{S}}$ are ones and the other components of $\mathbf{Q}_{\mathcal{S}}$ are zeros. The

definition given here for $\mathbf{Q}_{\mathcal{S}}$ can be adapted also to other sets of coordinates (e.g., $\mathbf{Q}_{\mathcal{S}^{\rm c}}$ for $\mathcal{S}^{\rm c}$) as denoted by the subscript of $\mathbf{Q}$. We now turn to formulate the *source* task using the linear regression form of

$$\widehat{\boldsymbol{\theta}} = \underset{\mathbf{r}\in\mathbb{R}^d}{\arg\min} \|\mathbf{v}-\mathbf{Z}\mathbf{r}\|_2^2 \quad \text{subject to} \quad \mathbf{Q}_{\mathcal{S}^{\rm c}}\mathbf{r}=\mathbf{0} \tag{2.3}$$

that its minimum ℓ_2-norm solution (see details in Appendix A.1) is

$$\widehat{\boldsymbol{\theta}} = \mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{v} \tag{2.4}$$

where $\mathbf{Z}_{\mathcal{S}}^+$ is the pseudoinverse of $\mathbf{Z}_{\mathcal{S}} \triangleq \mathbf{Z}\mathbf{Q}_{\mathcal{S}}^T$. Note that $\mathbf{Z}_{\mathcal{S}}$ is a $\widetilde{n}\times\widetilde{p}$ matrix that its $i^{\rm th}$ row is formed by the $\widetilde{p}$ components of $\mathbf{z}^{(i)}$ specified by the coordinates in $\mathcal{S}$, namely, only $\widetilde{p}$ out of the d features of the input data vectors are utilized. Hence, in the underparameterized case of $\widetilde{p}\leq\widetilde{n}$ the solution (2.4) almost surely reduces to the unique least squares form of

$$\widehat{\boldsymbol{\theta}} = \mathbf{Q}_{\mathcal{S}}^T\left(\mathbf{Z}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}\right)^{-1}\mathbf{Z}_{\mathcal{S}}^T\mathbf{v}. \tag{2.5}$$

Moreover, in both (2.4)-(2.5), $\widehat{\boldsymbol{\theta}}$ is a d-dimensional vector that may have nonzero values only in the $\widetilde{p}$ coordinates specified in $\mathcal{S}$ (this can be easily observed by noting that for an arbitrary $\mathbf{w}\in\mathbb{R}^{|\mathcal{S}|}$, the vector $\mathbf{u}=\mathbf{Q}_{\mathcal{S}}^T\mathbf{w}$ is a d-dimensional vector that its components satisfy $u_{s_k}=w_k$ for $k=1,...,|\mathcal{S}|$ and $u_j=0$ for $j\notin\mathcal{S}$). While the specific optimization form in (2.3) was not explicit in previous studies of non-asymptotic settings, e.g., [5, 3], the solution in (2.4) coincides with theirs and, thus, the formulation of the generalization error of our *source* task (which is a linear regression problem that, by itself, does not have any transfer learning aspect) is available from [5, 3] and provided in Appendix A.2 in our notations for completeness of presentation.

2.2. Target Task: Data Model and Solution using Transfer Learning. A second data class, which is our main interest, is modeled by $(\mathbf{x},y)\in\mathbb{R}^d\times\mathbb{R}$ that satisfy

$$y = \mathbf{x}^T\boldsymbol{\beta}+\epsilon \tag{2.6}$$

where $\mathbf{x}\sim\mathcal{N}\left(\mathbf{0},\mathbf{I}_d\right)$ is a Gaussian input vector including d features, $\epsilon\sim\mathcal{N}\left(0,\sigma_\epsilon^2\right)$ is a Gaussian noise component independent of $\mathbf{x}$, $\sigma_\epsilon>0$, and $\boldsymbol{\beta}\in\mathbb{R}^d$ is an unknown vector related to the $\boldsymbol{\theta}$ from (2.1) via

$$\boldsymbol{\theta} = \mathbf{H}\boldsymbol{\beta}+\boldsymbol{\eta} \tag{2.7}$$

where $\mathbf{H}\in\mathbb{R}^{d\times d}$ is a deterministic matrix and $\boldsymbol{\eta}\sim\mathcal{N}\left(\mathbf{0},\sigma_\eta^2\mathbf{I}_d\right)$ is a Gaussian noise vector with $\sigma_\eta\geq 0$. Here $\boldsymbol{\eta}$, $\mathbf{x}$, ϵ, $\mathbf{z}$ and ξ are independent. The data user does not know the distribution of $(\mathbf{x},y)$ but receives a small dataset of n i.i.d. draws of $(\mathbf{x},y)$ pairs denoted as $\mathcal{D}\triangleq\left\{\left(\mathbf{x}^{(i)},y^{(i)}\right)\right\}_{i=1}^n$. The n data samples can be organized in a $n\times d$ matrix of input variables $\mathbf{X}\triangleq[\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}]^T$ and a $n\times 1$ vector of responses $\mathbf{y}\triangleq[y^{(1)},\ldots,y^{(n)}]^T$ that together satisfy the relation $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}\triangleq[\epsilon^{(1)},\ldots,\epsilon^{(n)}]^T$ is an unknown noise vector that its $i^{\rm th}$ component $\epsilon^{(i)}$ is involved in the connection $y^{(i)}=\mathbf{x}^{(i),T}\boldsymbol{\beta}+\epsilon^{(i)}$ underlying the $i^{\rm th}$ example pair.

The *target* task considers a new (out of sample) data pair $\left(\mathbf{x}^{(\text{test})}, y^{(\text{test})}\right)$ drawn from the model in (2.6) independently of the training examples in $\mathcal{D}$. Given $\mathbf{x}^{(\text{test})}$, the goal is to establish an estimate $\widehat{y}$ of the response value $y^{(\text{test})}$ such that the out-of-sample squared error, i.e., the generalization error of the *target* task,

$$\mathcal{E}_{\text{out}} \triangleq \mathbb{E}\left[\left(\widehat{y} - y^{(\text{test})}\right)^2\right] = \sigma_\epsilon^2 + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2\right] \tag{2.8}$$

is minimized, where $\widehat{y} = \mathbf{x}^{(\text{test}),T}\widehat{\boldsymbol{\beta}}$, and the second equality stems from the data model in (2.6).

The target task is addressed via linear regression that seeks for an estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$ with a layout including three disjoint sets of coordinates $\mathcal{F}, \mathcal{T}, \mathcal{Z}$ that satisfy $\mathcal{F} \cup \mathcal{T} \cup \mathcal{Z} = \{1, \ldots, d\}$ and correspond to three types of parameters:

- p parameters are *free* to be optimized and their coordinates are specified in $\mathcal{F}$.
- t parameters are *transferred* from the co-located coordinates of the estimate $\widehat{\boldsymbol{\theta}}$ already formed for the source task. Only the free parameters of the *source* task are relevant to be transferred to the target task and, therefore, $\mathcal{T} \subseteq \mathcal{S}$ and $t \in \{0, \ldots, \widetilde{p}\}$. The transferred parameters are taken as is from $\widehat{\boldsymbol{\theta}}$ and set fixed in the corresponding coordinates of $\widehat{\boldsymbol{\beta}}$, i.e., for $k \in \mathcal{T}$, $\widehat{\beta}_k = \widehat{\theta}_k$ where $\widehat{\beta}_k$ and $\widehat{\theta}_k$ are the k^{th} components of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$, respectively.
- ℓ parameters are set to *zeros*. Their coordinates are included in $\mathcal{Z}$ and effectively correspond to ignoring features in the same coordinates of the input vectors.

Clearly, the layout should satisfy $p+t+\ell = d$. Then, the constrained linear regression problem for the target task is formulated as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\arg\min}\, \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 \\ \text{subject to} \quad \mathbf{Q}_{\mathcal{T}}\mathbf{b} = \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} \\ \mathbf{Q}_{\mathcal{Z}}\mathbf{b} = \mathbf{0} \end{aligned} \tag{2.9}$$

where $\mathbf{Q}_{\mathcal{T}}$ and $\mathbf{Q}_{\mathcal{Z}}$ are the linear operators extracting the subvectors corresponding to the coordinates in $\mathcal{T}$ and $\mathcal{Z}$, respectively, from d-dimensional vectors. Here $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^d$ is the *pre-computed* estimate for the source task and considered a constant vector for the purpose of the target task. The examined transfer learning structure includes a single computation of the source task (2.3), followed by a single computation of the target task (2.9) that produces the eventual estimate of interest $\widehat{\boldsymbol{\beta}}$ using the given $\widehat{\boldsymbol{\theta}}$. The minimum ℓ_2-norm solution of the target task in (2.9) is (see details in Appendix A.3)

$$\widehat{\boldsymbol{\beta}} = \mathbf{Q}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}}^+\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) + \mathbf{Q}_{\mathcal{T}}^T\widehat{\boldsymbol{\theta}}_{\mathcal{T}} \tag{2.10}$$

where $\widehat{\boldsymbol{\theta}}_{\mathcal{T}} \triangleq \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}$, $\mathbf{X}_{\mathcal{T}} \triangleq \mathbf{X}\mathbf{Q}_{\mathcal{T}}^T$, and $\mathbf{X}_{\mathcal{F}}^+$ is the pseudoinverse of $\mathbf{X}_{\mathcal{F}} \triangleq \mathbf{X}\mathbf{Q}_{\mathcal{F}}^T$. In the underparameterized case of $p \leq n$ the solution (2.10) almost surely reduces to the unique least squares form of

$$\widehat{\boldsymbol{\beta}} = \mathbf{Q}_{\mathcal{F}}^T\left(\mathbf{X}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}}\right)^{-1}\mathbf{X}_{\mathcal{F}}^T\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) + \mathbf{Q}_{\mathcal{T}}^T\widehat{\boldsymbol{\theta}}_{\mathcal{T}}. \tag{2.11}$$

Note that the desired layout is indeed implemented by the $\widehat{\boldsymbol{\beta}}$ forms in (2.10), (2.11): the components corresponding to $\mathcal{Z}$ are zeros, the components corresponding to $\mathcal{T}$ are taken as is from $\widehat{\boldsymbol{\theta}}$, and only the p coordinates specified in $\mathcal{F}$ are adjusted for the purpose of minimizing the in-sample error in the optimization cost of (2.9) while considering the transferred parameters. In this paper we study the generalization ability of overparameterized solutions (i.e., when $p > n$) to the *target* task formulated in (2.9): In Sections 3-5 we analyze the minimum ℓ_2-norm solution (2.10) and in Section 6 we examine other solutions.

3. The Double Descent Phenomenon in Transfer Learning.

3.1. Analytical Characterization of the Double Descent Phenomenon. Consider an *overall layout of coordinate subsets* $\mathcal{L} \triangleq \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$. The following theorem formulates the generalization error of the target task that is solved using a set of t parameters that are transferred as is from their co-located coordinates (indicated by $\mathcal{T}$) at the source task solution.

Theorem 3.1. *Let $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ be a deterministic (i.e., non-random) coordinate layout. Then, the out-of-sample error of the target task has the form of*

(3.1)
$$\mathcal{E}_{\text{out}}^{(\mathcal{L})} = \begin{cases} \frac{n-1}{n-p-1}\left(\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \sigma_\epsilon^2 + \mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}\right) & \text{for } p \le n-2, \\ \infty & \text{for } n-1 \le p \le n+1, \\ \frac{p-1}{p-n-1}\left(\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \sigma_\epsilon^2 + \mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}\right) + \left(1 - \frac{n}{p}\right)\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2 & \text{for } p \ge n+2, \end{cases}$$

where

(3.2)
$$\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} \triangleq \underbrace{\left\|\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right] - \boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2}_{\textit{transfer bias}} + \underbrace{\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2\right]}_{\textit{transfer variance}}$$

is the error in the transferred coordinates in the target task solution.

The last theorem is proved using non-asymptotic properties of Wishart matrices (see Appendix B).

The error formulation in (3.1) expresses a double descent form in terms of p and n, similarly to the formulation given in [3] for a linear regression task without transfer learning. However, here (3.1) introduces a new term, $\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$, that encapsulates the transfer learning aspect of our setting.

The formulation of the transfer error term in (3.2) demonstrates a bias-variance decomposition where the bias in the transferred parameters is measured with respect to the true parameters of the *target* task, and the variance depends only on the source task. This bias-variance decomposition is affected by the coordinates of the transferred parameters (i.e., $\mathcal{T}$) and also indirectly (through $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$) by the coordinates of the free parameters in the source task (i.e., $\mathcal{S}$), as we will see next.

The following corollary explicitly formulates the transfer bias and variance terms and their detailed dependency on the *parameterization level* of the *source* task, i.e., the $\widetilde{p}, \widetilde{n}$ pair, and the set $\mathcal{S}$ of free parameters in the source task (see proof in Appendix B.3).

Corollary 3.2. *The transfer bias term from (3.2) can be written as*

$$\text{Bias}^2_{\mathcal{T}} \triangleq \left\| \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right] - \boldsymbol{\beta}_{\mathcal{T}} \right\|_2^2 = \left\| \mathbf{Q}_{\mathcal{T}} \left(r\mathbf{H} - \mathbf{I}_d \right) \boldsymbol{\beta} \right\|_2^2 \quad \textit{where} \quad r \triangleq \begin{cases} 1 & \textit{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \textit{for } \widetilde{p} > \widetilde{n}. \end{cases} \tag{3.3}$$

The transfer variance term from (3.2) can be formulated as

$$\begin{aligned} &\text{Var}_{\mathcal{T},\mathcal{S}} \triangleq \mathbb{E}\left[\left\| \widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right] \right\|_2^2 \right] \\ &= \begin{cases} t\left(\sigma_\eta^2 + \frac{\zeta_{\mathcal{S}^c} + (d-\widetilde{p})\sigma_\eta^2 + \sigma_\xi^2}{\widetilde{n}-\widetilde{p}-1} \right) & \textit{for } 1 \leq \widetilde{p} \leq \widetilde{n}-2, \\ \infty & \textit{for } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}} \left(\frac{(\widetilde{p}-\widetilde{n})t\zeta_{\mathcal{S}\setminus\mathcal{T}} + \left((\widetilde{p}-\widetilde{n})t - 1 + \frac{\widetilde{n}}{\widetilde{p}}\right)\zeta_{\mathcal{T}}}{\widetilde{p}^2-1} + t\left(\sigma_\eta^2 + \frac{\zeta_{\mathcal{S}^c} + (d-\widetilde{p})\sigma_\eta^2 + \sigma_\xi^2}{\widetilde{p}-\widetilde{n}-1} \right) \right) & \textit{for } \widetilde{p} \geq \widetilde{n}+2, \end{cases} \end{aligned} \tag{3.4}$$

where $\zeta_{\mathcal{T}} \triangleq \|\mathbf{Q}_{\mathcal{T}}\mathbf{H}\boldsymbol{\beta}\|_2^2$, $\zeta_{\mathcal{S}\setminus\mathcal{T}} \triangleq \left\|\mathbf{Q}_{\mathcal{S}\setminus\mathcal{T}}\mathbf{H}\boldsymbol{\beta}\right\|_2^2$, *and* $\zeta_{\mathcal{S}^c} \triangleq \|\mathbf{Q}_{\mathcal{S}^c}\mathbf{H}\boldsymbol{\beta}\|_2^2$.

Note that the out-of-sample error formulation in (3.1) depends on the parameterization level of the *target* task (i.e., the p, n pair) and also on the parameterization level of the *source* task (i.e., the $\widetilde{p}, \widetilde{n}$ pair) via the transfer bias and variance terms that are formulated in (3.3)-(3.4). The formulations in Corollary 3.2 imply that the target error peaks not only around $p = n$, but also around $\widetilde{p} = \widetilde{n}$ when transfer learning is applied ($t > 0$). This induces the *double double descent* behavior that will be demonstrated in the next subsection. Later, in Sections 4-5 we will analyze the formulations in Theorem 3.1 and Corollary 3.2 in more detail and characterize the conditions for beneficial transfer learning.

3.2. On-Average Analysis of Arbitrarily Selected Parameters. In this subsection, we consider the *overall layout of coordinate subsets* $\mathcal{L} \triangleq \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ as a random structure. This will let us to formulate the expected value (with respect to $\mathcal{L}$) of the generalization error of interest. The simplified setting in this subsection provides useful insights towards Section 3.3 where we return to analyze the transfer of a set of parameters which is induced by a single layout $\mathcal{L}$ (i.e., in Section 3.3 we will return to use Theorem 3.1 and Corollary 3.2 where there is no expectation over a random $\mathcal{L}$).

For given d, $\widetilde{p}$, p, and t, we consider a uniform distribution of the coordinate layout $\mathcal{L}$ which is defined as follows.

Definition 3.3. *A coordinate subset layout* $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ *that is* $\{\widetilde{p}, p, t\}$*-uniformly distributed, for* $\widetilde{p} \in \{1, \ldots, d\}$ *and* $(p, t) \in \{0, \ldots, d\} \times \{0, \ldots, \widetilde{p}\}$ *such that* $p + t \leq d$*, satisfies:* $\mathcal{S}$ *is uniformly chosen at random from all the subsets of* $\widetilde{p}$ *unique coordinates of* $\{1, \ldots, d\}$*. Given* $\mathcal{S}$*, the target-task coordinate layout* $\{\mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ *is uniformly chosen at random from all the layouts where* $\mathcal{F}$*,* $\mathcal{T}$*,* $\mathcal{Z}$ *are three disjoint sets of coordinates that satisfy* $\mathcal{F} \cup \mathcal{T} \cup \mathcal{Z} = \{1, \ldots, d\}$ *such that* $|\mathcal{F}| = p$*,* $|\mathcal{T}| = t$ *and* $\mathcal{T} \subseteq \mathcal{S}$*, and* $|\mathcal{Z}| = d - p - t$*.*

Then, the following formulates the out-of-sample error of the target task under expectation with respect to a uniformly distributed $\mathcal{L}$ (more details are provided in Appendix B.4).

Corollary 3.4. *Let $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ be a coordinate subset layout that is $\{\widetilde{p}, p, t\}$-uniformly distributed. Then, the expected out-of-sample error of the target task has the form of*

(3.5)
$$\mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{out}}\right] = \begin{cases} \frac{n-1}{n-p-1}\left(\left(1-\frac{p+t}{d}\right)\|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2 + \mathcal{E}_{\text{transfer}}\right) & \text{for } p \le n-2, \\ \infty & \text{for } n-1 \le p \le n+1, \\ \frac{p-1}{p-n-1}\left(\left(1-\frac{p+t}{d}\right)\|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2 + \mathcal{E}_{\text{transfer}}\right) + \frac{p-n}{d}\|\boldsymbol{\beta}\|_2^2 & \text{for } p \ge n+2, \end{cases}$$

where

(3.6)
$$\mathcal{E}_{\text{transfer}} \triangleq \mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}\right] = \mathbb{E}_{\mathcal{L}}\left[\text{Bias}_{\mathcal{T}}^2\right] + \mathbb{E}_{\mathcal{L}}\left[\text{Var}_{\mathcal{T},\mathcal{S}}\right]$$

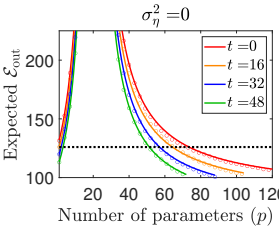
(3.7)
$$\mathbb{E}_{\mathcal{L}}\left[\text{Bias}_{\mathcal{T}}^2\right] = \frac{t}{d}\left\|\left(r\mathbf{H} - \mathbf{I}_d\right)\boldsymbol{\beta}\right\|_2^2 \quad \text{where} \quad r \triangleq \begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}. \end{cases}$$

(3.8)
$$\mathbb{E}_{\mathcal{L}}\left[\text{Var}_{\mathcal{T},\mathcal{S}}\right] = \begin{cases} t\left(\sigma_\eta^2 + \frac{\left(1-\frac{\widetilde{p}}{d}\right)\zeta + (d-\widetilde{p})\sigma_\eta^2 + \sigma_\xi^2}{\widetilde{n}-\widetilde{p}-1}\right) & \text{for } 1 \le \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}}t\left(\frac{(\widetilde{p}-\widetilde{n})\widetilde{p}+\frac{\widetilde{n}}{\widetilde{p}}-1}{d(\widetilde{p}^2-1)}\zeta + \sigma_\eta^2 + \frac{\left(1-\frac{\widetilde{p}}{d}\right)\zeta + (d-\widetilde{p})\sigma_\eta^2 + \sigma_\xi^2}{\widetilde{p}-\widetilde{n}-1}\right) & \text{for } \widetilde{p} \ge \widetilde{n}+2, \end{cases}$$

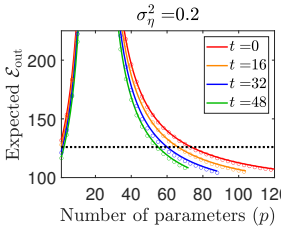
and $\zeta \triangleq \|\mathbf{H}\boldsymbol{\beta}\|_2^2$.

Figure 1 presents the curves of $\mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{out}}\right]$ with respect to the number of free parameters p in the target task, whereas the source task has $\widetilde{p} = d$ free parameters. In Fig. 1, the solid-line curves correspond to analytical values induced by Corollary 3.4, and the respective empirically computed values are denoted by circles (all the presented results are for $d = 120$, $n = 20$, $\widetilde{n} = 50$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$. See additional details in Appendix C. The number of free parameters p is upper bounded by $d - t$ that gets smaller for a larger number of transferred parameters t (see, in Fig. 1, the earlier stopping of the curves when t is larger). Observe that the generalization error peaks at $p = n$ and, then, decreases as p grows in the overparameterized range of $p > n + 1$. We identify this behavior as a *double descent* phenomenon, but without the first descent in the underparameterized range (double descent curves without the first descent are common when the parameters are selected arbitrarily, for example, see the results in [3, 7]).

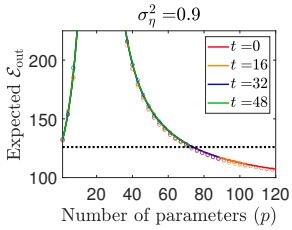
The error of a trivial solution in the form of the null estimate, i.e., the estimate $\widehat{\boldsymbol{\beta}}$ is all zeros, is presented as black dotted horizontal lines in the figures. In various settings where the source and target tasks are sufficiently related and the number of parameters is sufficiently far from the peak of the double descent error curve, the examined transfer learning method (with $t > 0$ arbitrarily selected parameters) outperforms the null estimate.



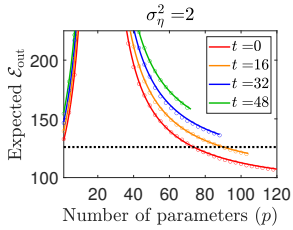
(a) **H**: local averaging neighborhood size 5



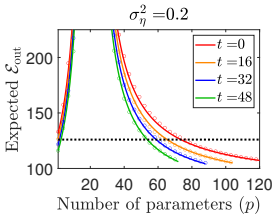
(b) **H**: local averaging neighborhood size 5



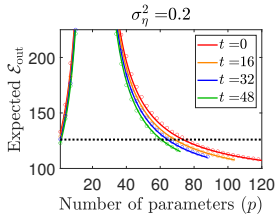
(c) **H**: local averaging neighborhood size 5



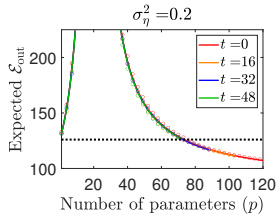
(d) **H**: local averaging neighborhood size 5



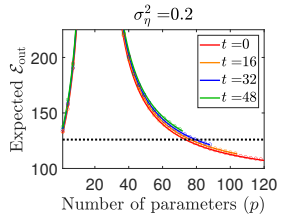
(e) $\mathbf{H} = \mathbf{I}_d$



(f) **H**: local averaging neighborhood size 19



(g) **H**: local averaging neighborhood size 45

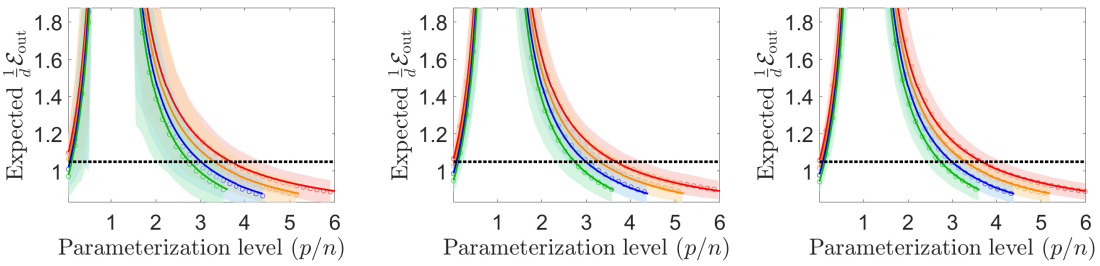


(h) **H**: overall averaging neighborhood size 120

Figure 1: The expected generalization error of the target task, $\mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{out}}\right]$, with respect to the number of free parameters (in the target task). The analytical values, computed using Corollary 3.4, are presented using solid-line curves, and the respective empirical results obtained from averaging over 250 experiments are denoted by circle markers. The horizontal dotted lines denote the error level of the null estimate. Each subfigure considers a different case of the source-target task relation (2.7) with a different pair of σ_η^2 and $\mathbf{H}$. The second row of subfigures corresponds to $\mathbf{H}$ operators that perform local averaging, each subfigure (e)-(h) is w.r.t. a different size of local averaging neighborhood. Each curve color refers to a different number of transferred parameters.

Each subfigure in Fig. 1 considers a different task relation with a different pair of noise level σ_η^2 and operator $\mathbf{H}$. The first row of subfigures in Fig. 1 emphasizes the effect of the noise variance σ_η^2 in the task relation model on the generalization errors in the target task. The second row of subfigures in Fig. 1 emphasizes the effect of the linear operator $\mathbf{H}$ in the task relation model on the generalization errors in the target task.

We can interpret the results in Figure 1 as examples for important cases of transfer learning settings. Figs. 1a,1e correspond to transfer learning between two *highly related tasks*, therefore, transfer learning is *beneficial* in the sense that for a given $p \notin \{n-1, n, n+1\}$ the error decreases as t increases (i.e., as more parameters are transferred instead of being omitted). Figs. 1b,1f correspond to transfer learning between two *moderately related tasks*, hence, transfer learning is still *beneficial, but less* than in the former case of highly related tasks. Figs. 1c,1g correspond to transfer learning between two *unrelated tasks* (although not extremely different), hence, transfer learning is *useless*, but not harmful (i.e., for a given p, the number of transferred



(a) Dimensions as in Fig. 1b (b) Dimensions 3 times larger (c) Dimensions 5 times larger

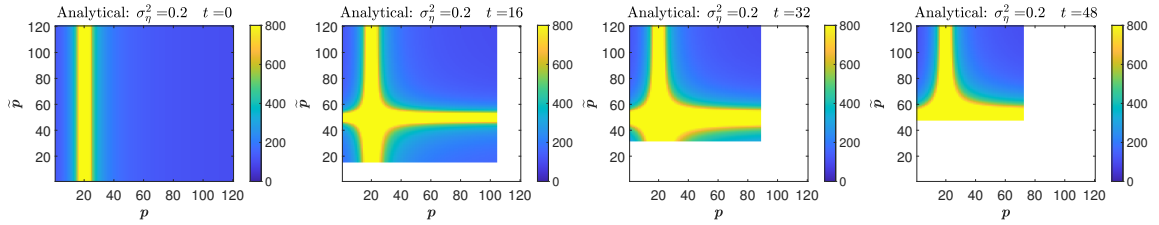
Figure 2: The concentration of the generalization error at three proportional scales of the same problem. The empirical standard deviations are denoted as shaded areas in colors corresponding to the on-average error curves (solid lines and markers denote the analytical and empirical evaluations of the expected error, respectively). Subfigure (a) corresponds to the setting of Fig. 1b. Subfigure (b) corresponds to a setting where all the dimensions and dimension-dependent quantities are 3 times their values in Fig. 1b (see a more detailed explanation in the text). Subfigure (c) corresponds to a setting where all dimensions and dimension-dependent quantities are 5 times their values in Fig. 1b. Lines, markers and areas in red correspond to $t = 0$ (no parameters are transferred); orange corresponds to transferring $t = 16 \times \frac{d}{120}$ parameters; blue corresponds to $t = 32 \times \frac{d}{120}$; green corresponds to $t = 48 \times \frac{d}{120}$. Note that $\frac{d}{120}$ equals to 1, 3, 5, in (a), (b), (c), respectively. The axes in this figure are normalized to be dimension-independent.

parameters t does not affect the out-of-sample error). Figs. 1d,1h correspond to transfer learning between two *very different tasks* and, accordingly, transfer learning *degrades* the generalization performance (namely, for a given p, transferring more parameters increases the out-of-sample error).

Figure 2 demonstrates the empirical standard deviations of the generalization errors (denoted as shaded areas around the curves and markers that denote the average errors). The results show that the empirical errors are more concentrated around their (theoretical and empirical) expectations as the dimensions of the problem (e.g., input dimension, number of data examples and parameters) are proportionally increased. For this example, we consider the setting of Fig. 1b in three proportional scales of the dimensions and dimension-dependent quantities:

- Fig. 2a corresponds to the same dimensions as in Fig. 1b: $d = 120$, $n = 20$, $\widetilde{n} = 50$, $t \in \{0, 16, 32, 48\}$, and local averaging with neighborhood size 5.
- Fig. 2b corresponds to dimensions that are proportionally increased 3 times: $d = 360$, $n = 60$, $\widetilde{n} = 150$, $t \in \{0, 48, 96, 144\}$, and local averaging with neighborhood size 15.
- Fig. 2c corresponds to dimensions that are proportionally increased 5 times: $d = 600$, $n = 100$, $\widetilde{n} = 250$, $t \in \{0, 80, 160, 240\}$, and local averaging with neighborhood size 25.

In all the settings in Figure 2, $\|\boldsymbol{\beta}\|_2^2 = d$ and the noise levels are $\sigma_\epsilon^2 = 0.05 \times d$, $\sigma_\xi^2 = 0.025 \times d$, $\sigma_\eta^2 = 0.2$. More examples are provided in Fig. 12.



(a) (b) (c) (d)

Figure 3: Analytical evaluation of the expected generalization error of the target task, $\mathbb{E}_{\mathcal{L}}[\mathcal{E}_{\text{out}}]$, with respect to the number of free parameters $\widetilde{p}$ and p (in the source and target tasks, respectively). Each subfigure considers a different number of transferred parameters t. The white regions correspond to $(\widetilde{p}, p)$ settings eliminated by the value of t in the specific subfigure. The yellow-colored areas correspond to values greater or equal to 800. All of the subfigures are for $\sigma_\eta^2 = 0.2$, $\mathbf{H}$ a local averaging operator with neighborhood of 5 samples, and $\boldsymbol{\beta}$ that has a piecewise-constant form (Fig. 15). See Fig. 10 for settings with additional values of σ_η^2. See Fig. 11 for the corresponding empirical evaluation.

By considering the generalization error formula from Theorem 3.1 as a function of $\widetilde{p}$ and p (i.e., the number of free parameters in the source and target tasks, respectively) we receive a *two-dimensional double descent* behavior as presented in Fig. 3 and its extended version Fig. 10 in Appendix C.1 that presents results for additional pairs of t and σ_η^2. The results show a double descent trend along the p axis (with a peak at $p = n$) and also, when parameter transfer is applied (i.e., $t > 0$), a double descent trend along the $\widetilde{p}$ axis (with a peak at $\widetilde{p} = \widetilde{n}$). Our solution structure implies that $\widetilde{p} \in \{t, \ldots, d\}$ and $p \in \{0, \ldots, d - t\}$, hence, a larger number of transferred parameters t eliminates a larger portion of the underparameterized range of the source task and also eliminates a larger portion of the overparameterized range of the target task (see in Fig. 3 the white eliminated regions that grow with t). When t is high, the wide elimination of portions from the $(\widetilde{p}, p)$-plane hinders the complete form of the two-dimensional double descent phenomenon (see, e.g., Fig. 3d).

Conceptually, we can observe a *tradeoff between overparamterized learning and transfer learning* where parameters are transferred as is from their co-located coordinates of the source task solution: an increased transfer of parameters limits the level of overparameterization applicable in the target task and, in turn, this may limit the overall potential gains from the transfer learning. Yet, when the source task is *sufficiently related* to the target task (see, e.g., Figs. 1a,1b), the parameter transfer can compensate (sometimes only partially) for an insufficient number of free parameters (in the target task). The last claim is also evident in Figs. 1a,1b,1e,1f where, for $p > n + 1$, there is a range of generalization error values that is achievable by several settings of (p, t) pairs (i.e., specific error levels can be attained by curves of different colors in the same subfigure). E.g., in Fig. 1b the error achieved by $p = 112$ free parameters and no parameter transfer can be also achieved using $p = 70$ free parameters and $t = 48$ parameters transferred from the source task.

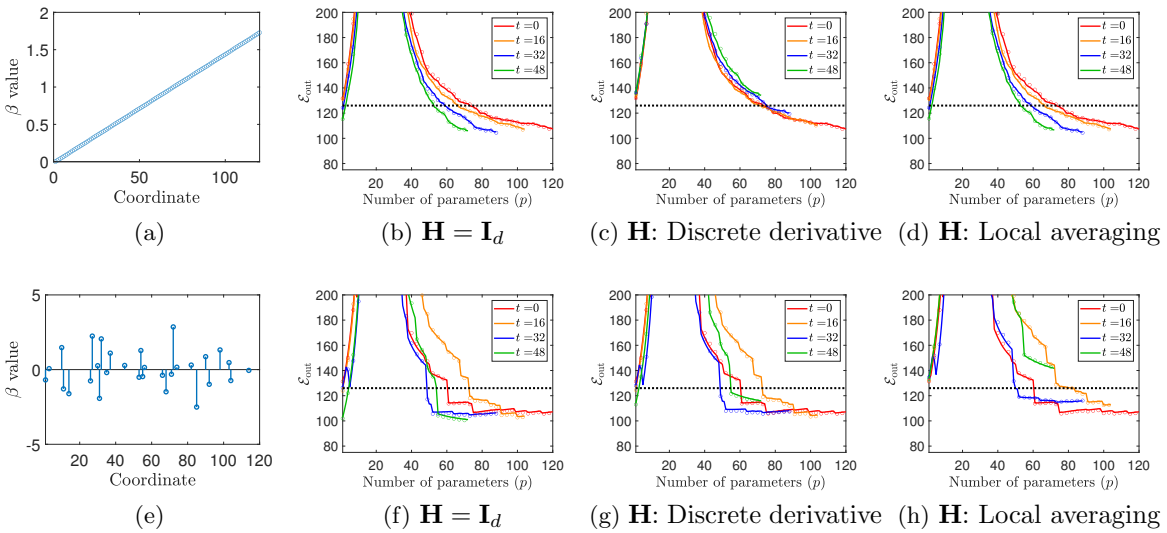
3.3. The Fragile Nature of Transfer Learning: Analysis of a Single Layout of Arbitrarily Selected Parameters. Section 3.2 considers an on-average error for a random coordinate layout. We now turn to discuss the generalization behavior of a *single* coordinate layout that was formed arbitrarily (i.e., without using any knowledge on the problem setting). Hence, we return to consider the generalization error $\mathcal{E}_{\text{out}}^{(\mathcal{L})}$ for a given layout $\mathcal{L}$ using Theorem 3.1 and Corollary 3.2 from Section 3.1.

Figure 4 shows the curves of $\mathcal{E}_{\text{out}}^{(\mathcal{L})}$ for specific coordinate layouts $\mathcal{L}$ that evolve with respect to the number of free parameters p in the target task (for more examples see Figures 13-14 in the Appendices). The excellent fit of the analytical results (that were computed using Theorem 3.1 and Corollary 3.2) to the empirical values (computed by averaging over 250 experiments with the same evolution of $\mathcal{L}$) is evident. The effect of the specific coordinate layout is clearly visible by the less-smooth curves (compared to the on-average results in Fig. 1). We examine two different cases for the true $\boldsymbol{\beta}$: a linearly-increasing (Fig. 4a) and a sparse (Fig. 4e) layout of values, both have the same ℓ_2 norm. The difference in the true $\boldsymbol{\beta}$ forms yields error curves that significantly differ despite the use of the same sequential construction of the coordinate layouts with respect to p (e.g., compare Figs. 4f and 4b). The operator $\mathbf{H}$ in the task relation model greatly affects the generalization error curves as evident from comparing our results for different types of $\mathbf{H}$: an identity, local averaging (with neighborhood size 11), and discrete derivative operators (e.g., compare subfigures within the first row of Fig. 4). The results clearly show that the interplay among the structures of $\mathbf{H}$, $\boldsymbol{\beta}$, and the coordinate layout significantly affects the generalization performance.

Our results also exhibit that an arbitrarily selected set $\mathcal{T}$ of t transferred parameters can be the best setting for a given set $\mathcal{F}$ of p free parameters but not necessarily for an extended set $\mathcal{F}' \supset \mathcal{F}$ of $p' > p$ free parameters. This is especially evident when the true parameters have a sparse form over an unknown support in the feature domain (i.e., true parameters with non-zero values are scarce and their coordinates are unknown). For example, see Fig. 4g where the green colored curve does not consistently maintain its relative vertical order with the red color curve at the overparameterized range of solutions; as a second example see Fig. 4h and observe that the blue and red curves do not maintain their vertical order in the overparameterized range. This exemplifies that, when arbitrary selection of parameters is employed due to unknown task relation, transfer learning settings can be fragile and hence finding a successful setting may require delicate, trial and error engineering. Therefore, our theory qualitatively explains similar practical aspects in deep neural networks (see, for example, [22]).

4. When is Transfer Learning Beneficial?. The formulation of the generalization error in the target task (Theorem 3.1 and Corollary 3.2) shows that the benefits from parameter transfer depend on various aspects of the learning setting. In this section we characterize the conditions for beneficial transfer from the viewpoint of the following question: Given a setting where $\mathcal{T}$ is the intended set of coordinates for parameter transfer, can avoiding this parameter transfer improve generalization?

4.1. Benefits in Transferred versus Zeroed Parameters. To accurately evaluate the difference in the out-of-sample error of the target task, consider the following definitions. First, recall the coordinate layout $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ where $|\mathcal{S}| = \widetilde{p}$, $|\mathcal{F}| = p$, $|\mathcal{T}| = t$. Second, we define a coordinate layout $\mathcal{L}' = \{\mathcal{S}, \mathcal{F}, \mathcal{T}', \mathcal{Z}'\}$ which is a modified version of $\mathcal{L}$ without trans-



(a) (b) $\mathbf{H} = \mathbf{I}_d$ (c) $\mathbf{H}$: Discrete derivative (d) $\mathbf{H}$: Local averaging

(e) (f) $\mathbf{H} = \mathbf{I}_d$ (g) $\mathbf{H}$: Discrete derivative (h) $\mathbf{H}$: Local averaging

Figure 4: Analytical (solid lines) and empirical (circle markers) values of $\mathcal{E}_{\text{out}}^{(\mathcal{L})}$ for an arbitrary coordinate layout $\mathcal{L}$. In the first row of subfigures the true $\boldsymbol{\beta}$ has feature-domain values that follow a linear form , see (a). In the second row of subfigures the true $\boldsymbol{\beta}$ has a sparse form in the feature domain with non-zero values at 30 coordinates selected randomly out of the $d = 120$, see (e). In each row there are three error plots for different circulant forms of the operator $\mathbf{H}$ (here the local averaging is defined for a neighborhood of 11 samples). Here $\sigma_\eta^2 = 0.2$, $d = 120$, $n = 20$, $\widetilde{n} = 50$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$, and $\widetilde{p} = d$ for all settings.

ferred parameters, specifically, $\mathcal{T}' = \emptyset$ and $\mathcal{Z}' = \mathcal{Z} \cup \mathcal{T}$. Namely, $\mathcal{L}'$ is obtained by zeroing all the parameters that are transferred in $\mathcal{L}$. Then, we define the following error difference due to transferring the parameters in $\mathcal{T}$ *instead of zeroing* them:

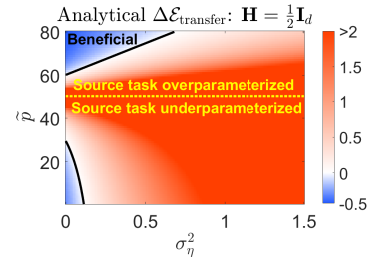
$$\Delta\mathcal{E}_{\text{TvsZ}}^{(\mathcal{L})} \triangleq \mathcal{E}_{\text{out}}^{(\mathcal{L})} - \mathcal{E}_{\text{out}}^{(\mathcal{L}')} \tag{4.1}$$

where $\mathcal{E}_{\text{out}}^{(\mathcal{L})}$ and $\mathcal{E}_{\text{out}}^{(\mathcal{L}')}$ are the out-of-sample errors in the target task for the coordinate layouts $\mathcal{L}$ and $\mathcal{L}'$, respectively. Using Theorem 3.1 we can write (4.1) as

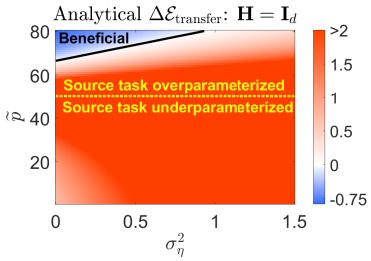
$$\Delta\mathcal{E}_{\text{TvsZ}}^{(\mathcal{L})} = \left(\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} - \|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2\right) \times \begin{cases} \frac{n-1}{n-p-1} & \text{for } p \leq n-2, \\ \frac{p-1}{p-n-1} & \text{for } p \geq n+2, \end{cases} \tag{4.2}$$

where $\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ was defined and formulated in Theorem 3.1 and Corollary 3.2. Note that $\Delta\mathcal{E}_{\text{TvsZ}}^{(\mathcal{L})}$ is undefined for $p \in \{n-1, n, n+1\}$. The definition in (4.1) implies that transferring the parameters in $\mathcal{T}$ is beneficial (over zeroing these coordinates in addition to the coordinates in $\mathcal{Z}$) if $\Delta\mathcal{E}_{\text{TvsZ}}^{(\mathcal{L})} < 0$. We define

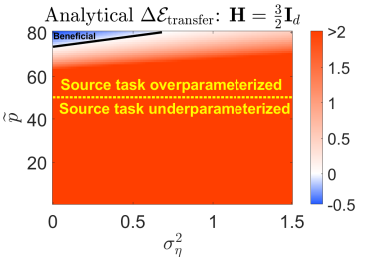
$$\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} \triangleq \mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} - \|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 \,. \tag{4.3}$$



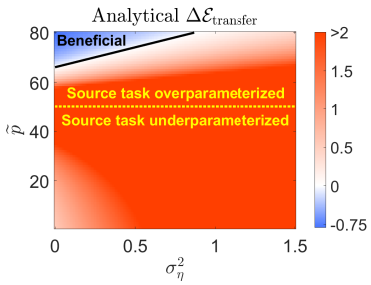
(a) $\mathbf{H} = \frac{1}{2}\mathbf{I}_d$



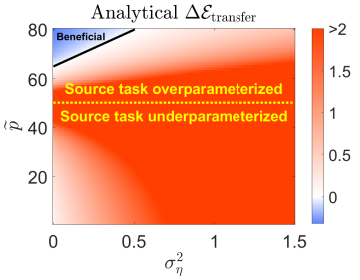
(b) $\mathbf{H} = \mathbf{I}_d$



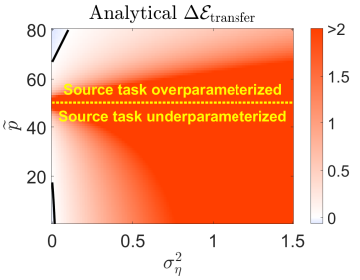
(c) $\mathbf{H} = \frac{3}{2}\mathbf{I}_d$



(d) $\mathbf{H}$: local averaging neighborhood size 3



(e) $\mathbf{H}$: local averaging neighborhood size 15



(f) $\mathbf{H}$: local averaging neighborhood size 59

Figure 5: The analytical values of $\Delta\mathcal{E}_{\text{transfer}}$ (here normalized by t) as a function of $\widetilde{p}$ and σ_η^2. The positive and negative values of $\Delta\mathcal{E}_{\text{transfer}}$ appear in color scales of red and blue, respectively. The regions of negative values (appear in shades of blue) correspond to beneficial transfer of parameters (compared to zeroing them). The positive values were truncated at the value of 2 for the clarity of visualization. Each subfigure corresponds to a different task relation model induced by the definitions of $\mathbf{H}$ as: *(a)-(c)* different scalings of the identity matrix, *(d)-(f)* circulant matrices that correspond to local averaging operators with different neighborhood sizes. For all the subfigures, $d = 80$, $\widetilde{n} = 50$, $\sigma_\xi^2 = 0.025 \cdot d$, $\|\boldsymbol{\beta}\|_2^2 = d$. The components of $\boldsymbol{\beta}$ have a linear form (see Fig. 4a) in *(a)-(c)* and a piecewise-constant form (see Fig. 15) in *(d)-(f)*. See corresponding empirical results in Figs. 16-17.

Then, according to (4.2), $\mathcal{E}_{\text{TvsZ}}^{(\mathcal{L})} < 0$ occurs when $p \notin \{n-1, n, n+1\}$ and $\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} < 0$.

The examples in Fig. 5 show the values of $\Delta\mathcal{E}_{\text{transfer}} \triangleq \mathbb{E}_{\mathcal{L}}\left[\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}\right]$ due to transferring an arbitrarily selected parameter. The error difference is presented with respect to the number $\widetilde{p}$ of free parameters in the source task and the variance σ_η^2 of the noise in the task relation model. Each subfigure in Fig. 5 shows results for a different definition of $\mathbf{H}$. All the subfigures demonstrate that transferring an arbitrarily selected parameter is more beneficial when the solution to the source task is highly overparameterized; indeed, then the generalization error in the source task itself is lower due to its own double descent phenomena (e.g., recall the behavior of the error along the vertical axis in Fig. 3b-3d). As expected, a low level of noise in the task relation is also important for beneficial transfer.

The examples in Fig. 5 also show a somewhat unexpected behavior: parameter transfer

from an *underparameterized* source task can be more beneficial as the tasks are less related, e.g., compare the lower left corners in Figures 5a and 5b. To understand this observation mathematically one may recall the transfer bias and variance formulations in Corollary 3.2 and notice the following effects of $\mathbf{H}$, which in Figs. 5a-5b has the form of $\mathbf{H} = c\mathbf{I}_d$ where $c \in (0, 1)$. First, the transfer bias increases as $c \in (0, 1)$ gets smaller; also note that the transfer bias sums only over the $|\mathcal{T}| = t$ transferred coordinates. Second, the (underparameterized) transfer variance decreases as $c \in (0, 1)$ gets smaller; here note that the transfer variance is affected by $|\mathcal{S}^c| = d - \widetilde{p}$ coordinates of $\mathbf{H} = c\mathbf{I}_d$ in addition to the number t of transferred parameters. In the experiment setting of Fig. 5a, the source task is underparameterized with a sufficiently large $d - \widetilde{p}$ value such that the reduction in the transfer variance dominates the increase in the transfer bias, resulting in beneficial transfer.

More generally, the lesson here is that beneficial transfer learning can be achieved also in non-intuitive settings where the tasks are not necessarily highly related. In Section 5 we will provide additional insights into counter-intuitive beneficial cases.

4.2. Benefits in Transferred versus Free Parameters. Now we turn to discuss the question of when the transfer of the parameters in $\mathcal{T}$ is more beneficial than defining them as free for optimization.

Consider the coordinate layout $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ where, as usual, $|\mathcal{S}| = \widetilde{p}$, $|\mathcal{F}| = p$, $|\mathcal{T}| = t$. Here we also define a second coordinate layout $\mathcal{L}'' = \{\mathcal{S}, \mathcal{F}'', \mathcal{T}'', \mathcal{Z}\}$ which is a modified version of $\mathcal{L}$ without transferred parameters, specifically, $\mathcal{T}'' = \emptyset$ and $\mathcal{F}'' = \mathcal{F} \cup \mathcal{T}$. Hence, the consideration of $\mathcal{L}''$ reflects the case where all the transferred parameters in $\mathcal{L}$ are replaced by free parameters. Then, we define the following error difference term due to transferring the parameters in $\mathcal{T}$ *instead of allocating them as free parameters*:

$$\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}} \triangleq \mathcal{E}^{(\mathcal{L})}_{\text{out}} - \mathcal{E}^{(\mathcal{L}'')}_{\text{out}} \tag{4.4}$$

where $\mathcal{E}^{(\mathcal{L})}_{\text{out}}$ and $\mathcal{E}^{(\mathcal{L}'')}_{\text{out}}$ are the out-of-sample errors in the target task (recall Theorem 3.1) for the coordinate layouts $\mathcal{L}$ and $\mathcal{L}''$, respectively. The definition in (4.4) implies that transferring the parameters in $\mathcal{T}$ is beneficial (over allocating these coordinates as t free parameters in addition to the p free parameters in $\mathcal{F}$) if $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}} < 0$.

We now turn to examine the effects of p and t on beneficial transfer in the sense of $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}} < 0$. For this, note that $\mathcal{L}''$ includes $p'' = p + t$ free parameters and hence is not necessarily in the same parameterization regime as $\mathcal{L}$ that includes p free parameters. The parameterization regimes of $\mathcal{L}$ and $\mathcal{L}''$ affect the behavior of the benefits from parameter transfer, as described by the following corollaries of Theorem 3.1 (these corollaries are simply proved by setting Eq. (3.1) in (4.4) and reorganizing the formulations).

Corollary 4.1. *For $p + t \leq n - 2$, namely, when both $\mathcal{L}$ and $\mathcal{L}''$ correspond to underparameterized settings, the error difference due to transferred versus free parameters is*

$$\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}} = \frac{n-1}{n-p-1}\left(\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}} - \frac{t}{n-p-t-1}\left(\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \sigma_\epsilon^2\right)\right). \tag{4.5}$$

Accordingly, the transfer of the $t > 0$ parameters in $\mathcal{T} \subseteq \mathcal{S}$ is more beneficial than allocating

them as free parameters (i.e., $\Delta\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{TvsF}} < 0$) if

$$\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}} < \frac{t}{n-p-t-1}\left(\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \sigma_\epsilon^2\right). \tag{4.6}$$

Corollary 4.2. *For $p+t \geq n+2$, namely, when $\mathcal{L}''$ corresponds to an overparameterized setting, the error difference due to transferred versus free parameters is*

$$\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}} = \gamma_0\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}} + \gamma_1\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2 + \gamma_2\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 + \gamma_3\left(\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \sigma_\epsilon^2\right) \tag{4.7}$$

where the coefficients γ_0,γ_1,γ_2,γ_3 depend on the parameterization level of $\mathcal{L}$:

- *For an underparameterized $\mathcal{L}$ with $p \leq n-2$:*
  $\gamma_0 = \frac{n-1}{n-p-1}$, $\gamma_1 = -\left(1-\frac{n}{p+t}\right)$, $\gamma_2 = -\left(1-\frac{n}{p+t}\right)$, $\gamma_3 = -\frac{n(n-1)-p(p+t-1)}{(p+t-n-1)(n-p-1)}$.
- *For an overparameterized $\mathcal{L}$ with $p \geq n+2$:*
  $\gamma_0 = \frac{p-1}{p-n-1}$, $\gamma_1 = -\frac{nt}{p(p+t)}$, $\gamma_2 = -\left(1-\frac{n}{p+t}\right)$, $\gamma_3 = \frac{nt}{(p+t-n-1)(p-n-1)}$.

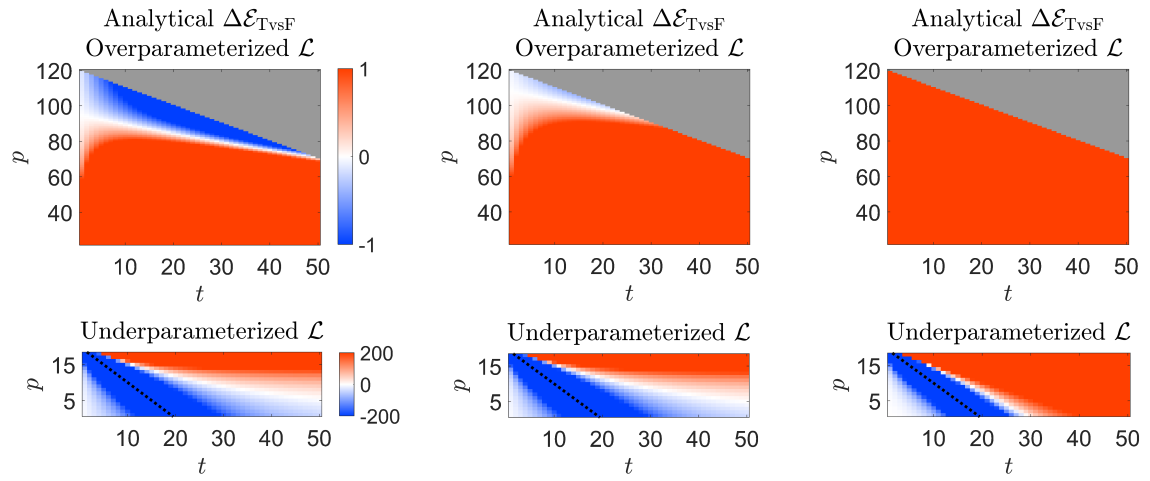
Accordingly, the transfer of the $t>0$ parameters in $\mathcal{T} \subseteq \mathcal{S}$ is more beneficial than allocating them as free parameters (i.e., $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}} < 0$) if

$$\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}} < -\frac{\gamma_1}{\gamma_0}\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2 - \frac{\gamma_2}{\gamma_0}\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2 - \frac{\gamma_3}{\gamma_0}\left(\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \sigma_\epsilon^2\right). \tag{4.8}$$

Clearly, the formulation of $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}}$ is more intricate in the case of overparameterized $\mathcal{L}''$ in Corollary 4.2 than in the case of underparameterized $\mathcal{L}$ and $\mathcal{L}''$ in Corollary 4.1. Specifically, the set of free parameters $\mathcal{F}$ appears in (4.7)-(4.8) in addition to the sets $\mathcal{T}$ and $\mathcal{Z}$.

To intuitively understand the behavior of $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}}$, consider expectation over coordinates that are selected uniformly at random (recall Definition 3.3). Specifically, we define $\Delta\mathcal{E}_{\text{TvsF}} \triangleq \mathbb{E}_{\mathcal{L}}\left[\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}}\right]$ while noting that $\mathcal{L}''$ is deterministically related to $\mathcal{L}$ and hence it is sufficient to consider the expectation over $\mathcal{L}$. Figure 6 shows the value of $\Delta\mathcal{E}_{\text{TvsF}}$ as a function of p and t for several operators $\mathbf{H}$ in the task relation. Figure 6 demonstrates the following typical behaviors:

- For *underparameterized* $\mathcal{L}$ and $\mathcal{L}''$ (i.e., $p+t \leq n-2$, which corresponds to the region to the left of the dotted black lines in the bottom subfigures in Fig. 6): Parameter transfer is more likely to be beneficial, and with larger gains, as $p+t$ increases towards $n-2$. Moreover, when (4.6) is satisfied, it is beneficial to have more transferred than free parameters (i.e., having $t>p$ for a fixed sum of $p+t$). The intuition here is that, when we are constrained to the underparameterized regime in both options (i.e., $\mathcal{L}$ and $\mathcal{L}''$), transferring parameters instead of having more free parameters can keep the error farther away from the peak of the double descent curve, which yields a beneficial transfer. Interestingly, due to the significant peak of the generalization error around the interpolation threshold, this behavior also occurs when the source and target tasks are quite different (see Fig. 6c).
- For *underparameterized* $\mathcal{L}$ and *overparameterized* $\mathcal{L}''$ (i.e., $p \leq n-2$ and $p+t \geq n+2$, which correspond to the region to the right of the dotted black lines in the bottom subfigures in Fig. 6): Parameter transfer is more likely to degrade as p increases



(a) $\mathbf{H}$: local averaging neighborhood size 3

(b) $\mathbf{H}$: local averaging neighborhood size 15

(c) $\mathbf{H} = 5\mathbf{I}_d$

Figure 6: The analytical values of $\Delta\mathcal{E}_{\text{TvsF}}$ (namely, the expected error difference due to transfer of an arbitrarily-selected set of t parameters versus setting them as free parameters) as a function of t and p. The positive and negative values of $\Delta\mathcal{E}_{\text{TvsF}}$ appear in color scales of red and blue, respectively. The regions of negative values (appear in shades of blue) correspond to beneficial transfer of parameters (compared to defining them as free parameters). The gray regions correspond to $p + t > d$ where parameter transfer cannot be performed. For better visual clarity, the underparameterized and overparameterized cases of $\mathcal{L}$ are shown in different subfigures with significantly different range of values. Also, the positive values were truncated at the values of 1 and 200 in the overparameterized and underparameterized cases, respectively. Corresponding truncations are applied on the negative values at the values of -1 and -200. The dotted black line corresponds to $p+t = n$, which is the interpolation threshold of the auxiliary layout $\mathcal{L}''$. Each column of subfigures corresponds to a different task relation model induced by the definitions of $\mathbf{H}$. For all the subfigures, $d = 120$, $\widetilde{p} = d$, $\widetilde{n} = 50$, $\sigma_\xi^2 = 0.025 \cdot d$, $n = 20$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\|\boldsymbol{\beta}\|_2^2 = d$ where $\boldsymbol{\beta}$ components have a piecewise-constant form (see Fig. 15). See corresponding empirical results in Fig. 19 in Appendix D.4.

towards $n - 2$ while $p + t \geq n + 2$. Moreover, increasing the number t of transferred parameters degrades the benefits from transfer learning. The intuition here is that, when the transfer option $\mathcal{L}$ is underparameterized and the no-transfer option $\mathcal{L}''$ is overparameterized, any increase in $p + t$ makes $\mathcal{L}''$ more overparameterized and hence farther away from the double descent peak (hence having an improved generalization ability).

- For *overparameterized* $\mathcal{L}$ and $\mathcal{L}''$ (i.e., $p \geq n + 2$, which corresponds to the upper subfigures in Fig. 6): Here, the important behavior is that, for a given $\mathcal{T}$, parameter transfer improves as the number of free parameters p increases. Of course that this improvement becomes beneficial (at high overparameterization levels) only when the source task is sufficiently related to the target task (see, e.g., Figures 6a and 6b). The

intuition here is that, when both options (i.e., $\mathcal{L}$ and $\mathcal{L}''$) are in the overparameterized regime then the increase in p moves both $\mathcal{L}$ and $\mathcal{L}''$ away from the peak of the double descent. The transfer learning option $\mathcal{L}$ is closer to the double descent peak and hence typically improves more (due to the steeper slope at the corresponding part of the generalization error curve, e.g., see in Fig. 1) than the no-transfer case $\mathcal{L}''$.

5. The Optimal H in a Componentwise Task Relation. Theorem 3.1 shows that the task relation aspect is encapsulated in the term $\mathcal{E}^{(\mathcal{S},\mathcal{T})}_{\text{transfer}}$. Consequently, we demonstrated in Section 4 that $\mathcal{E}^{(\mathcal{S},\mathcal{T})}_{\text{transfer}}$ greatly affects the potential benefits from parameter transfer compared to both zero and free parameters. Specifically, $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsZ}}$ and $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsF}}$ both decrease as $\mathcal{E}^{(\mathcal{S},\mathcal{T})}_{\text{transfer}}$ decreases. This motivates us to characterize the optimal task relation in the sense of minimum $\mathcal{E}^{(\mathcal{S},\mathcal{T})}_{\text{transfer}}$ for a given coordinate layout $\mathcal{L}$. Namely, we characterize the best source task to transfer from when the other aspects of the parameter transfer are fixed.

Consider an operator $\mathbf{H}$ that has the diagonal form of

$$\mathbf{H} = \text{diag}\{\lambda^{(1)}_{\mathbf{H}}, \dots, \lambda^{(d)}_{\mathbf{H}}\} \tag{5.1}$$

where $\{\lambda^{(j)}_{\mathbf{H}}\}_{j=1}^{d} \in \mathbb{R}$. We refer to $\{\lambda^{(j)}_{\mathbf{H}}\}_{j=1}^{d}$ as the eigenvalues of $\mathbf{H}$ due to the fact that applying the same orthonormal rotation on the feature spaces of the source and target tasks can induce a non-diagonal $\mathbf{H}$ that its eigenvalues are the same $\{\lambda^{(j)}_{\mathbf{H}}\}_{j=1}^{d}$ as in (5.1). The diagonal form in (5.1) implies that the task relation in Eq. (2.7) reduces to the componentwise form of

$$\theta^{(j)} = \lambda^{(j)}_{\mathbf{H}}\beta^{(j)} + \eta^{(j)}, \quad j = 1, \dots, d \tag{5.2}$$

where $\theta^{(j)}$ and $\beta^{(j)}$ are the j^{th} components of the true parameter vectors of the source and target tasks, respectively, and $\eta^{(j)}$ is the j^{th} noise component in $\boldsymbol{\eta}$. Then, we can further develop the expressions from Corollary 3.2. First, the transfer bias term from (3.3) can be written as

$$\text{Bias}^2_{\mathcal{T}} = \sum_{j\in\mathcal{T}} \left(r\lambda^{(j)}_{\mathbf{H}} - 1\right)^2 \left(\beta^{(j)}\right)^2 \quad \text{where} \quad r \triangleq \begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}. \end{cases} \tag{5.3}$$

Second, the transfer variance term $\text{Var}_{\mathcal{T},\mathcal{S}}$ from (3.4) can be also rewritten using

$$\zeta_{\mathcal{T}} = \sum_{j\in\mathcal{T}} \left(\lambda^{(j)}_{\mathbf{H}}\beta^{(j)}\right)^2, \quad \zeta_{\mathcal{S}\setminus\mathcal{T}} = \sum_{j\in\mathcal{S}\setminus\mathcal{T}} \left(\lambda^{(j)}_{\mathbf{H}}\beta^{(j)}\right)^2, \quad \zeta_{\mathcal{S}^c} = \sum_{j\in\mathcal{S}^c} \left(\lambda^{(j)}_{\mathbf{H}}\beta^{(j)}\right)^2. \tag{5.4}$$

Consequently, we get the following result (proof is provided in Appendix E.1).

Theorem 5.1. *Consider a task relation operator of the form* $\mathbf{H} = \text{diag}\{\lambda^{(1)}_{\mathbf{H}}, \dots, \lambda^{(d)}_{\mathbf{H}}\}$. *Also,* $\widetilde{p} \notin \{\widetilde{n}-1, \widetilde{n}, \widetilde{n}+1\}$. *Then, for given coordinate sets* $\mathcal{S}$ *(of size* $\widetilde{p}$*) and* $\mathcal{T} \subseteq \mathcal{S}$ *(of size* $t \le \widetilde{p}$*), the transfer error term* $\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$ *attains its minimal value with respect to the eigenvalues of* $\mathbf{H}$

at

$$\text{(5.5)} \qquad \text{for } j \in \mathcal{T},\ \beta^{(j)} \neq 0: \qquad \lambda_{\mathbf{H}}^{(j)} = \begin{cases} 1 & \text{for } \widetilde{p} \leq \widetilde{n} - 2, \\ \frac{\widetilde{p}^2 - 1}{\widetilde{n}\widetilde{p} - 1 + t(\widetilde{p} - \widetilde{n})} & \text{for } \widetilde{p} \geq \widetilde{n} + 2, \end{cases}$$

$$\text{(5.6)} \qquad \text{for } j \in \mathcal{S} \setminus \mathcal{T},\ \beta^{(j)} \neq 0: \qquad \lambda_{\mathbf{H}}^{(j)} = \begin{cases} \text{any value} & \text{for } \widetilde{p} \leq \widetilde{n} - 2, \\ 0 & \text{for } \widetilde{p} \geq \widetilde{n} + 2, \end{cases}$$

$$\text{(5.7)} \qquad \text{for } j \in \mathcal{S}^c,\ \beta^{(j)} \neq 0: \qquad \lambda_{\mathbf{H}}^{(j)} = 0.$$

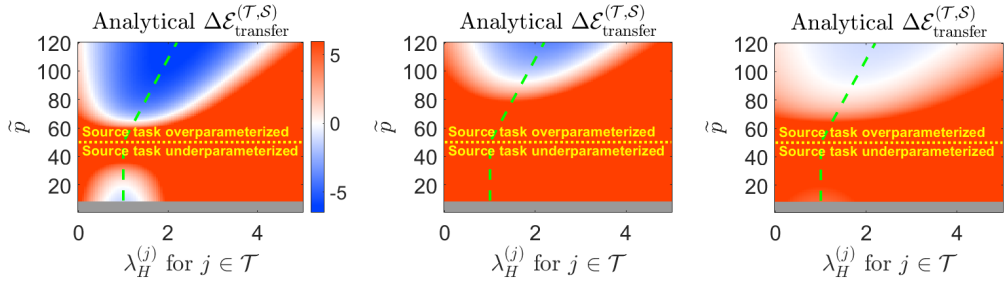
For $j \in \{1, \ldots, d\}$ where $\beta^{(j)} = 0$, $\lambda_{\mathbf{H}}^{(j)}$ can have any value.

Let us interpret the meaning of Theorem 5.1 for the case of $\beta^{(j)} \neq 0$ for $j \in \{1, \ldots, d\}$. The theorem shows that the linear operator in the optimal task relation has two different characterizations depending on whether the *source* task is under or over parameterized:

- For an *underparameterized source* task it is best that the true parameters in the transferred coordinates of the source task are the *same* (up to the additive noise terms from $\boldsymbol{\eta}$) as the corresponding true parameters of the target task (see (5.5) and recall (5.2)).
- For an *overparameterized source* task it is best that the true parameters in the transferred coordinates of the source task are *amplified* versions by a factor of $\frac{\widetilde{p}^2 - 1}{\widetilde{n}\widetilde{p} - 1 + t(\widetilde{p} - \widetilde{n})}$ (see (5.5)) of the corresponding true parameters of the target task. This amplification intends to partially compensate for the increased transfer bias in the case of overparameterized source task (see the $\widetilde{p} > \widetilde{n}$ case in (5.3)). Indeed, the optimal amplification increases as the source task is more overparameterized (i.e., as p increases towards d). While this amplification reduces the transfer bias, it increases the transfer variance (see (3.4) and (5.4)) and hence the amplification is somewhat restrained. Also, the optimal amplification $\frac{\widetilde{p}^2 - 1}{\widetilde{n}\widetilde{p} - 1 + t(\widetilde{p} - \widetilde{n})}$ decreases as the number t of transferred parameters increases. Specifically, when $t = \widetilde{p}$ (i.e., all the free parameters of the source task are transferred to the target task) the optimal $\lambda_{\mathbf{H}}^{(j)}$ values for $j \in \mathcal{T}$ become 1 and there is no amplification.

Eq. (5.6) considers the coordinates of the free parameters of the source task that are *not* transferred (i.e., $\mathcal{S} \setminus \mathcal{T}$) and shows that the optimal true parameters of the source task in these coordinates are zeros in the overparameterized regime due to the dependency of the transfer variance on the true parameters of the source task at $\mathcal{S} \setminus \mathcal{T}$ (see (3.4) and (5.4)). In the underparameterized regime of the source task, there is no such dependency on $\boldsymbol{\theta}_{\mathcal{S} \setminus \mathcal{T}}$ and hence these true parameters may have any value without affecting the minimization of $\mathcal{E}_{\text{transfer}}^{(\mathcal{T}, \mathcal{S})}$.

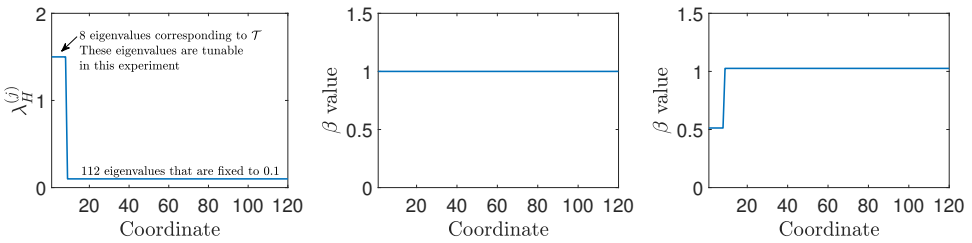
According to Eq. (5.7), in both the under and over parameterized cases, it is best that the true parameters of the source task are all zeros in the coordinates that do not participate in the source task solution (i.e., $\mathcal{S}^c$). This optimal form reduces the misspecification in the solution of the source task to originate only in the task relation noise $\boldsymbol{\eta}_{\mathcal{S}^c}$. Avoiding these misspecifications makes the estimate of each free source parameter closer to its true value (i.e., instead of also trying to compensate for omitted parameters that are informative) and, hence, this improves the relevance of the parameters $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ that are transferred "as is" to the



(a) $\sigma_\eta^2 = 0.2$; $\boldsymbol{\beta}$ has the form in Fig. 7e

(b) $\sigma_\eta^2 = 1$; $\boldsymbol{\beta}$ has the form in Fig. 7e

(c) $\sigma_\eta^2 = 0.2$; $\boldsymbol{\beta}$ has the form in Fig. 7f



(d) $\mathbf{H}$ eigenvalues

(e) $\boldsymbol{\beta}$ with all components ones

(f) $\boldsymbol{\beta}$ with components in $\mathcal{T}^c$ are x2 larger than in $\mathcal{T}$

Figure 7: The analytical values of $\Delta\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$ as a function of $\widetilde{p}$ and a value that determines the eigenvalues of $\mathbf{H}$ in $\mathcal{T}$. Here $\mathcal{T} = \{1, \ldots, 8\}$. The operator $\mathbf{H}$ is diagonal with a main diagonal in the form described in subfigure *(d)*. In subfigures *(a)-(c)*, the positive and negative values of $\Delta\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$ appear in color scales of red and blue, respectively. The color scales are the same for all the subfigures in the first row. The regions of negative values (appear in shades of blue) correspond to beneficial transfer of parameters. The positive values were truncated at the value of 6 for the clarity of visualization. The gray regions correspond to $\widetilde{p} < t$ where parameter transfer cannot be performed. The dashed green lines (in all subfigures) denote the optimal eigenvalues (corresponding to $\mathcal{T}$) as formulated in Theorem 5.1. Here, $d = 120$, $\widetilde{n} = 50$, $\sigma_\xi^2 = 0.025 \cdot d$. Subfigure *(e)* shows the form of $\boldsymbol{\beta}$ in the experiments of subfigures *(a),(b)*. Subfigure *(f)* shows the form of $\boldsymbol{\beta}$ in the experiments of subfigure *(c)*. The corresponding empirical evaluations are provided in Fig. 20 in Appendix E.2.

target task from the co-located coordinates in the source task.

Theorem 5.1 describes the eigenvalues of $\mathbf{H}$ that minimize $\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$. However, $\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$ also depends on additional factors such as the noise level in the task relation and the true parameters $\boldsymbol{\beta}$ of the target task, which determine whether transferring the parameters in $\mathcal{T}$ is preferred, e.g., over zeroing them. Thus, evaluation of the sign of $\Delta\mathcal{E}^{(\mathcal{L})}_{\text{TvsZ}}$ is required for understanding whether the optimal eigenvalues of $\mathbf{H}$ induce beneficial transfer and, if so, how much can the eigenvalues deviate from their optimal values while still having a beneficial transfer.

In Fig. 7 we show the analytical values of $\Delta\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$ (recall the definition in (4.3)) as a function of $\widetilde{p}$ and the eigenvalues of $\mathbf{H}$ that correspond to the transferred parameters (see Fig. 20 in Appendix E.2 for the corresponding empirical evaluations). Specifically, we consider a family of $\mathbf{H}$ operators that are diagonal in the feature domain and have the following step-shaped structure for their eigenvalues (see Fig. 7d): the first 8 eigenvalues are all equal to a value that varies among the $\mathbf{H}$ operators in this family; the next 112 eigenvalues are 0.1 for any $\mathbf{H}$ operator in this family. Here we consider transfer of the first eight parameters, i.e., $\mathcal{T} = \{1,\ldots,8\}$ that correspond to the eigenvalues with a tunable value in our evaluations (this is reflected by the horizontal axes in Figs. 7a-7c). The results in Fig. 7 demonstrate that the range of eigenvalues that induce beneficial transfer increases together with overparameterization (see the wider blue regions around the green dashed line that denotes the optimal eigenvalue from Theorem 5.1). Fig. 7a shows that benefits can be also obtained in underparameterized settings where the eigenvalues (in $\mathcal{T}$) are around 1, however, these eigenvalue regions are smaller and produce lower benefits than their overparameterized alternatives. The effect of the noise in the task model is also apparent by comparing Fig. 7a to Fig. 7b that shows results for the same settings but with a higher σ^2_η. Specifically, the higher noise level in the task relation reduces the size of the beneficial regions and their gains in the overparameterized range, and leaves the underparameterized range without any beneficial regions. Next, compare Fig. 7a to Fig. 7c that shows results for the same settings but with $\boldsymbol{\beta}$ of the less favorable structure in Fig. 7f instead of the structure in Fig. 7e. This shows how the structure of the true parameters in $\boldsymbol{\beta}$ can also affect the parameter transfer performance.

6. Additional Linear Regression Methods. The previous sections presented analytical and empirical results on transfer learning between two least squares solutions that take their minimum ℓ_2-norm forms (2.4), (2.10) in the overparameterized regime. In this section we examine the same transfer learning mechanism (i.e., transferring co-located parameters from the already-learned source model to the to-be-learned target model) for two more kinds of solutions to linear regression: least squares with the minimum ℓ_1-norm form in the overparameterized regime, and ridge regression.

6.1. The Minimum ℓ_1-Norm Interpolator. This setting also emerges from the optimization problems formulated in (2.3) and (2.9) for the source and target tasks, respectively. In the underparameterized regime, the source and target tasks almost surely have the unique least squares solutions as provided in (2.5) and (2.11).

For an overparameterized source task, i.e., $\widetilde{p} > \widetilde{n}$, the problem (2.3) has infinite interpolating solutions among which we choose here the minimum ℓ_1-norm solution:

$$\begin{aligned} \widehat{\boldsymbol{\theta}} = \underset{\mathbf{r}\in\mathbb{R}^d}{\arg\min} \|r\|_1 \\ \text{subject to } \mathbf{Q}_{\mathcal{S}^c}\mathbf{r} = \mathbf{0} \\ \mathbf{Z}\mathbf{r} = \mathbf{v}. \end{aligned} \tag{6.1}$$

This solution equals to $\widehat{\boldsymbol{\theta}}$ where $\widehat{\boldsymbol{\theta}}_{\mathcal{S}^c} = \mathbf{0}$ and $\widehat{\boldsymbol{\theta}}_{\mathcal{S}} = \arg\min_{\mathbf{k}\in\mathbb{R}^{\widetilde{p}}} \|\mathbf{k}\|_1$ subject to $\mathbf{Z}_{\mathcal{S}}\mathbf{k} = \mathbf{v}$, which is a basis pursuit problem [6] that can be addressed via linear programming methods.

For an overparameterized target task, i.e., $p > n$, the problem (2.9) has infinite interpolating solutions, from which we choose here the minimum ℓ_1-norm solution

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\arg\min} \|\mathbf{b}\|_1 \\ \text{subject to} \quad \mathbf{Q}_{\mathcal{T}}\mathbf{b} = \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} \\ \mathbf{Q}_{\mathcal{Z}}\mathbf{b} = \mathbf{0} \\ \mathbf{X}\mathbf{b} = \mathbf{y} \end{aligned} \tag{6.2}$$

that can be formulated also as $\widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}}_{\mathcal{Z}} = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}_{\mathcal{T}} = \widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ and

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\mathcal{F}} = \underset{\mathbf{f}\in\mathbb{R}^p}{\arg\min} \|\mathbf{f}\|_1 \\ \text{subject to} \quad \mathbf{X}_{\mathcal{F}}\mathbf{f} = \mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}, \end{aligned} \tag{6.3}$$

which is a basis pursuit problem that can be solved via linear programming.

The empirical out-of-sample errors of the minimum ℓ_1-norm transfer learning are shown in Figure 8 as blue curves where the shade of blue denotes the number of transferred parameters. The results demonstrate that the double descent behavior occurs also for the minimum ℓ_1-norm solution. Double descent phenomena for minimum ℓ_1-norm solutions to linear regression (without transfer learning) were studied in [19, 18]. The asymptotic analyses in [15, 27] show that a triple descent can be observed if the true parameters are sufficiently sparse. Here we do not clearly observe an additional error peak in the overparameterized regime, which could possibly be due to the misspecification strategy (i.e., zeroing parameters in predetermined coordinates) that we use for controlling the parameterization levels in our non-asymptotic setting. For $\boldsymbol{\beta}$ of a linear shape the minimum ℓ_2-norm solution is better than the minimum ℓ_1-norm solution in the entire overparameterized range (see Figs. 8a-8c). Nevertheless, for a sparse $\boldsymbol{\beta}$ the minimum ℓ_1-norm solution can provide the best performance in the high overparameterization levels due to its sparsity-promoting nature (see Figs. 8d, 8g-8i).

6.2. Ridge Regression. Now we turn to formulate our transfer learning approach for ridge regression. So far we considered the optimization problems (2.3) and (2.9) that do not include explicit regularization terms in their cost functions. Hence, the ridge regression extension of (2.3) is

$$\widehat{\boldsymbol{\theta}} = \underset{\mathbf{r}\in\mathbb{R}^d}{\arg\min} \|\mathbf{v} - \mathbf{Z}\mathbf{r}\|_2^2 + \widetilde{\alpha}\,\|\mathbf{r}\|_2^2 \quad \text{subject to} \quad \mathbf{Q}_{\mathcal{S}^c}\mathbf{r} = \mathbf{0}, \tag{6.4}$$

where $\widetilde{\alpha} > 0$ determines the level of ridge regularization. The formulation in (6.4) is equivalent to $\widehat{\boldsymbol{\theta}}$ where $\widehat{\boldsymbol{\theta}}_{\mathcal{S}^c} = 0$ and

$$\widehat{\boldsymbol{\theta}}_{\mathcal{S}} = \underset{\mathbf{k}\in\mathbb{R}^{\widetilde{p}}}{\arg\min} \|\mathbf{v} - \mathbf{Z}_{\mathcal{S}}\mathbf{k}\|_2^2 + \widetilde{\alpha}\,\|\mathbf{k}\|_2^2\,. \tag{6.5}$$

The optimization in (6.5) has a standard ridge regression form and, thus, its closed-form solution is

$$\widehat{\boldsymbol{\theta}}_{\mathcal{S}} = \left(\mathbf{Z}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}} + \widetilde{\alpha}\mathbf{I}_{\widetilde{p}}\right)^{-1}\mathbf{Z}_{\mathcal{S}}^T\mathbf{v}. \tag{6.6}$$

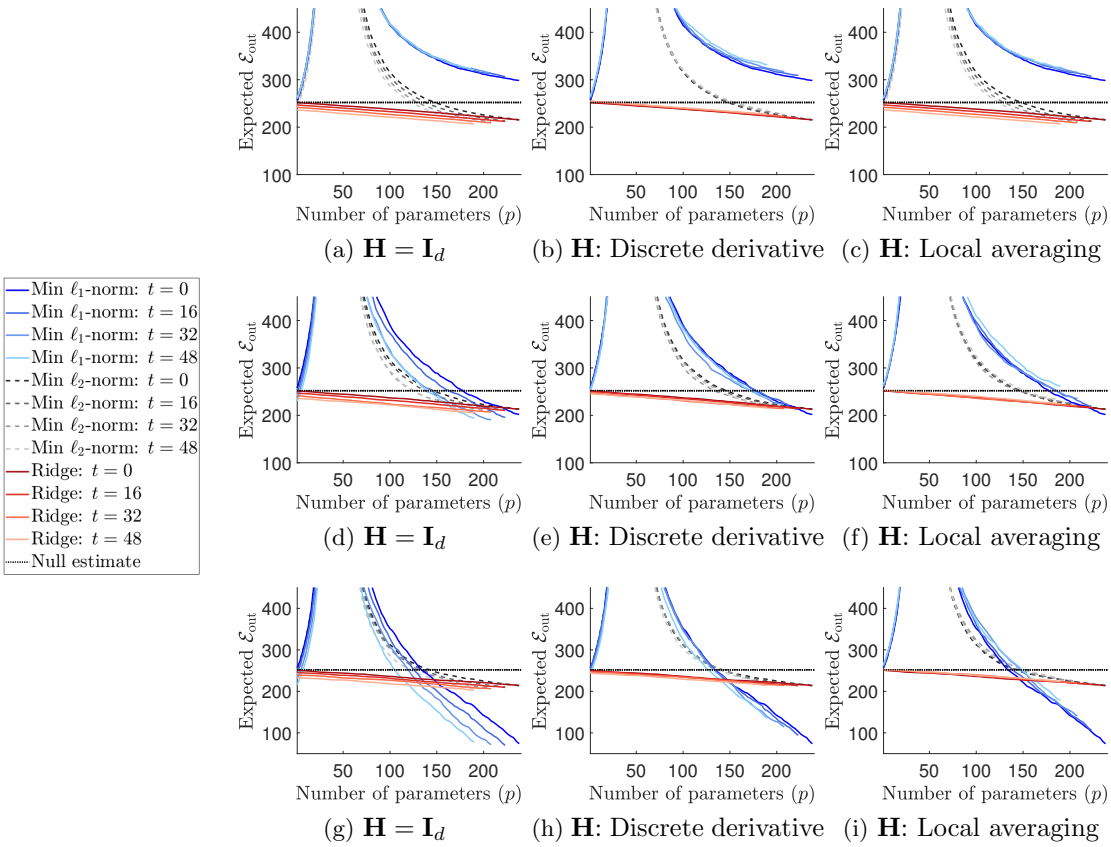


Figure 8: The expected generalization error of the target task, $\mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{out}}\right]$, compared for minimum ℓ_2-norm, minimum ℓ_1-norm, and ridge solutions. In this figure all the errors are empirically evaluated. The errors of the minimum ℓ_1-norm solution are shown in curves of blue shades. The errors of the ridge regression are shown in curves of red shades. The errors of the minimum ℓ_2-norm solution are shown in dashed curves of gray shades. For each of the solution types the darkest shade corresponds to no transfer ($t = 0$), a lighter shade denotes more transferred parameters (larger t), and the lightest shade corresponds to $t = 48$. In the first row of subfigures the true $\boldsymbol{\beta}$ has values that follow a linear form. In the second row of subfigures the true $\boldsymbol{\beta}$ has a sparse form with non-zero values at 25% of the coordinates (selected randomly out of the $d = 240$). In the third row of subfigures the true $\boldsymbol{\beta}$ is sparse with only 5% non-zero values. Examples for linear and sparse $\boldsymbol{\beta}$ of a smaller dimension are provided in Figs. 4a, 4e. In each row there are three error plots for different circulant forms of the operator $\mathbf{H}$ (here the local averaging is defined for a neighborhood of 11 samples). Here $\sigma_\eta^2 = 0.2$, $d = 240$, $n = 40$, $\widetilde{n} = 100$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$, and $\widetilde{p} = d$ for all settings.

In appendix F.1 we formulate the out-of-sample error of the ridge regression solution to the source task, and show that optimal tuning is provided by $\widetilde{\alpha} = \frac{\widetilde{p}\left(\sigma_\xi^2 + \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2\right)}{\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2}$. In our experiments

we assume that $\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2, \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2$ are unknown and only $\|\boldsymbol{\theta}\|_2^2$, σ_ξ^2 are known about the source task. Thus, we assume $\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2 \approx \frac{\widetilde{p}}{d}\|\boldsymbol{\theta}\|_2^2, \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2 \approx \left(1-\frac{\widetilde{p}}{d}\right)\|\boldsymbol{\theta}\|_2^2$ that let us to approximate the optimal tuning using $\widetilde{\alpha} = \frac{d\sigma_\xi^2}{\|\boldsymbol{\theta}\|_2^2} + d - \widetilde{p}$.

Proceeding to the target task, the ridge regression extension of (2.9) is

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\arg\min} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \alpha\|b\|_2^2 \\ \text{subject to } \mathbf{Q}_{\mathcal{T}}\mathbf{b} = \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} \\ \mathbf{Q}_{\mathcal{Z}}\mathbf{b} = \mathbf{0}, \end{aligned} \tag{6.7}$$

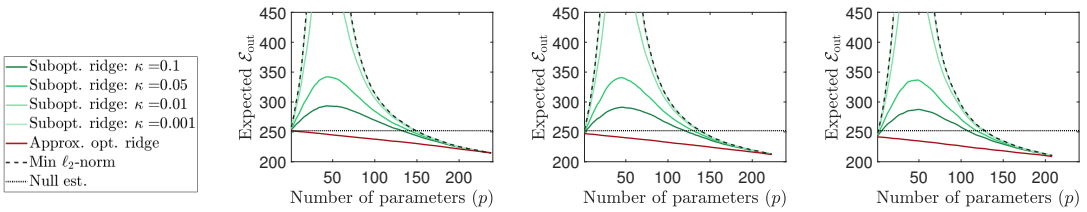
for $\alpha > 0$. The solution of (6.7) has the closed form of $\widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}}_{\mathcal{Z}} = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}_{\mathcal{T}} = \widehat{\boldsymbol{\theta}}_{\mathcal{T}}$, and

$$\widehat{\boldsymbol{\beta}}_{\mathcal{F}} = \left(\mathbf{X}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}} + \alpha\mathbf{I}_p\right)^{-1}\mathbf{X}_{\mathcal{F}}^T\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right). \tag{6.8}$$

In appendix F.2 we provide further details on (6.8), the corresponding out-of-sample error, and show that optimal tuning is given by $\alpha = \frac{p\left(\sigma_\epsilon^2 + \|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \mathbb{E}\left[\left\|\boldsymbol{\beta}_{\mathcal{T}} - \widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right]\right)}{\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2}$. In our experiments we approximate the optimal tuning by setting $\alpha = \frac{d\sigma_\epsilon^2}{\|\boldsymbol{\beta}\|_2^2} + d - p - t + t\frac{\|(\mathbf{I}_d - \mathbf{H})\boldsymbol{\beta}\|_2^2 + d\sigma_\eta^2}{\|\boldsymbol{\beta}\|_2^2}$ that assumes the knowledge of $\|\boldsymbol{\beta}\|_2^2$, $\|(\mathbf{I}_d - \mathbf{H})\boldsymbol{\beta}\|_2^2$, σ_η, σ_ϵ. See Appendix F.2 for more details.

The empirical out-of-sample errors for the ridge regression transfer learning appear in Figure 8 as red curves where the shade of red denotes the number of transferred parameters. As expected, the ridge regularization (which is approximately optimally tuned) resolves the error peak and eliminates the double descent shape of the minimum norm interpolating solutions. In the case of suboptimally tuned ridge regularization the double descent peak may not be fully resolved (see Fig. 9). For $\boldsymbol{\beta}$ with a linear shape, the ridge regularization suits best among the examined methods at any parameterization level, except for the maximal overparameterization levels where the minimum ℓ_2-norm solution performs comparably (see Figs. 8a-8c). However, for a sparse $\boldsymbol{\beta}$ the minimum ℓ_1-norm outperforms the ridge solution for high overparameterization levels (see Figs. 8d-8f).

7. Conclusions. In this work we have established an analytical framework for the fundamental study of transfer learning in conjunction with overparameterized models. We used least squares solutions to linear regression problems for shedding clarifying light on the generalization performance induced for a target task addressed using parameters transferred from an already completed source task. We formulated the generalization error of the target task and presented its two-dimensional double descent shape as a function of the number of free parameters individually available in the source and target tasks. We characterized the conditions for a beneficial transfer of parameters and demonstrated its high sensitivity to the delicate interaction among crucial aspects such as the source-target task relation, the specific choice of transferred parameters, and the form of the true solution. We importantly showed that overparameterized transfer learning is not necessarily improved by using a source task which is closer or identical to the target task. Our focus was mainly on the analytical and empirical study of the minimum ℓ_2-norm solution to overparameterized transfer learning. We



(a) $t = 0$ (b) $t = 16$ (c) $t = 32$

Figure 9: The expected generalization error of the target task, $\mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{out}}\right]$, for transfer learning with suboptimally tuned ridge regularization. Each of the subfigures corresponds to a different number of transferred parameters t, and shows the error curves for four suboptimal ridge tunings where the parameters α and $\widetilde{\alpha}$ are κ times their (approximately) optimal values. Here $\boldsymbol{\beta}$ has linear shape, $\mathbf{H}$ is a local averaging operator (over a neighborhood size 11), $\sigma_\eta^2 = 0.2$, $d = 240$, $n = 40$, $\widetilde{n} = 100$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$, and $\widetilde{p} = d$.

also empirically examined the performance of the minimum ℓ_1-norm solution and ridge regression in our transfer learning framework. We believe that our work opens a new research direction for the fundamental understanding of the generalization ability of transfer learning designs. Future work may study the theory and practice of additional transfer learning layouts such as fine tuning of the transferred parameters, inclusion of various regularization methods, well-specified and other settings where the task relation model is known (to some extent) and utilized in the actual learning process.

Acknowledgments. This work was supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014-18-12571, N00014-20-1-2534, and MURI N00014-20-1-2787; AFOSR grant FA9550-18-1-0478; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047.

Appendices. The following appendices support the main paper as follows. Appendix A provides additional details on the mathematical developments leading to the formulations in Section 2 of the main paper. In Appendix B we present the proofs of Theorem 3.1 and Corollaries 3.2, 3.4 from Section 3, which formulate the generalization error of the target task. Appendix C provides additional empirical results and details for Section 3 of the main paper. In Appendices D, E we provide analytical proofs and empirical results for Sections 4, 5 of the main paper. In Appendix F we provide the mathematical developments for the ridge regression setting of our transfer learning problem from Section 6.2 of the main paper.

Appendix A. Mathematical Developments for Section 2.

A.1. The Estimate $\widehat{\theta}$ in Eq. (2.4). Let us solve the optimization problem provided in (2.3). Using the relation $\mathbf{Q}_{\mathcal{S}}^T\mathbf{Q}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}^c}^T\mathbf{Q}_{\mathcal{S}^c} = \mathbf{I}_d$ we can rewrite (2.3) as

$$\begin{aligned} \widehat{\boldsymbol{\theta}} = \underset{\mathbf{r}\in\mathbb{R}^d}{\arg\min} \, & \|\mathbf{v} - \mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{r} - \mathbf{Z}_{\mathcal{S}^c}\mathbf{Q}_{\mathcal{S}^c}\mathbf{r}\|_2^2 \\ \text{subject to} \quad & \mathbf{Q}_{\mathcal{S}^c}\mathbf{r} = \mathbf{0} \end{aligned} \tag{A.1}$$

where $\mathbf{Z}_{\mathcal{S}} \triangleq \mathbf{Z}\mathbf{Q}_{\mathcal{S}}^T$ and $\mathbf{Z}_{\mathcal{S}^c} \triangleq \mathbf{Z}\mathbf{Q}_{\mathcal{S}^c}^T$. By setting the equality constraint in the optimization cost, the problem in (A.1) becomes

$$\widehat{\boldsymbol{\theta}} = \underset{\mathbf{r}\in\mathbb{R}^d}{\arg\min} \|\mathbf{v} - \mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{r}\|_2^2 \quad \text{subject to} \quad \mathbf{Q}_{\mathcal{S}^c}\mathbf{r} = \mathbf{0}. \tag{A.2}$$

Without the equality constraint, (A.2) is just an unconstrained least squares problem that its minimum ℓ_2-norm solution is

$$\widehat{\boldsymbol{\theta}} = \mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{v} \tag{A.3}$$

where $\mathbf{Z}_{\mathcal{S}}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{Z}_{\mathcal{S}}$. Note that $\widehat{\boldsymbol{\theta}}$ in (A.3) satisfies the equality constraint in (A.1) and, therefore, (A.3) is also the solution for the constrained optimization problems in (A.1), (A.2), and (2.3).

A.2. The Double Descent Formulation for the Generalization Error of the Source Task. The generalization error of a single linear regression problem (that includes noise) in non-asymptotic settings is provided in [3] for a given coordinate subset (i.e., deterministic $\mathcal{S}$ in our terms). The result from [3] can be written in our notations as

$$\widetilde{\mathcal{E}}_{\text{out}} = \begin{cases} \frac{\widetilde{n}-1}{\widetilde{n}-\widetilde{p}-1}\left(\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2 + \sigma_\xi^2\right) & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{p}-1}{\widetilde{p}-\widetilde{n}-1}\left(\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2 + \sigma_\xi^2\right) + \frac{\widetilde{p}-\widetilde{n}}{\widetilde{p}}\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2 & \text{for } \widetilde{p} \ge \widetilde{n}+2. \end{cases} \tag{A.4}$$

In case that the coordinate subset $\mathcal{S}$ is uniformly chosen at random from all the subsets of $\widetilde{p} \in \{1,\dots,d\}$ unique coordinates of $\{1,\dots,d\}$, then we get that $\mathbb{E}_{\mathcal{S}}\left[\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2\right] = \frac{\widetilde{p}}{d}\|\boldsymbol{\theta}\|_2^2$ and $\mathbb{E}_{\mathcal{S}}\left[\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2\right] = \frac{d-\widetilde{p}}{d}\|\boldsymbol{\theta}\|_2^2$. Accordingly, the expectation over $\mathcal{S}$ of the generalization error of the source task leads to the following result

$$\mathbb{E}_{\mathcal{S}}\left[\widetilde{\mathcal{E}}_{\text{out}}\right] = \begin{cases} \frac{\widetilde{n}-1}{\widetilde{n}-\widetilde{p}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\|\boldsymbol{\theta}\|_2^2 + \sigma_\xi^2\right) & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{p}-1}{\widetilde{p}-\widetilde{n}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\|\boldsymbol{\theta}\|_2^2 + \sigma_\xi^2\right) + \frac{\widetilde{p}-\widetilde{n}}{d}\|\boldsymbol{\theta}\|_2^2 & \text{for } \widetilde{p} \ge \widetilde{n}+2. \end{cases} \tag{A.5}$$

The formulation in (A.5) considers $\boldsymbol{\theta}$ as a deterministic vector. For the analysis of the target task, where the task relation model (2.7) is assumed to hold, it is also useful to formulate the expectation of the out-of-sample error of the source task with respect to both $\mathcal{S}$ and the noise vector $\boldsymbol{\eta}$ from the task relation model. This leads us to to consider $\boldsymbol{\theta}$ as a random vector and to formulate the following expectation.

$$\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left[\widetilde{\mathcal{E}}_{\text{out}}\right] = \begin{cases} \frac{\widetilde{n}-1}{\widetilde{n}-\widetilde{p}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\kappa + \sigma_\xi^2\right) & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{p}-1}{\widetilde{p}-\widetilde{n}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\kappa + \sigma_\xi^2\right) + \frac{\widetilde{p}-\widetilde{n}}{d}\kappa & \text{for } \widetilde{p} \ge \widetilde{n}+2. \end{cases} \tag{A.6}$$

where $\kappa \triangleq \mathbb{E}_{\boldsymbol{\eta}}\left[\|\boldsymbol{\theta}\|_2^2\right] = \|\mathbf{H}\boldsymbol{\beta}\|_2^2 + d\sigma_\eta^2$.

A.3. The Estimate $\widehat{\beta}$ in Eq. (2.10). The optimization problem in (2.9), given for the target task, can be addressed using the relation $\mathbf{Q}_{\mathcal{F}}^T\mathbf{Q}_{\mathcal{F}} + \mathbf{Q}_{\mathcal{T}}^T\mathbf{Q}_{\mathcal{T}} + \mathbf{Q}_{\mathcal{Z}}^T\mathbf{Q}_{\mathcal{Z}} = \mathbf{I}_d$ and rewritten as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\arg\min} \, & \|\mathbf{y} - \mathbf{X}_{\mathcal{F}}\mathbf{Q}_{\mathcal{F}}\mathbf{b} - \mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{b} - \mathbf{X}_{\mathcal{Z}}\mathbf{Q}_{\mathcal{Z}}\mathbf{b}\|_2^2 \\ \text{subject to} \quad & \mathbf{Q}_{\mathcal{T}}\mathbf{b} = \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} \\ & \mathbf{Q}_{\mathcal{Z}}\mathbf{b} = \mathbf{0} \end{aligned} \tag{A.7}$$

where $\mathbf{X}_{\mathcal{F}} \triangleq \mathbf{X}\mathbf{Q}_{\mathcal{F}}^T$, $\mathbf{X}_{\mathcal{T}} \triangleq \mathbf{X}\mathbf{Q}_{\mathcal{T}}^T$, and $\mathbf{X}_{\mathcal{Z}} \triangleq \mathbf{X}\mathbf{Q}_{\mathcal{Z}}^T$. By setting the equality constraints of (A.7) in its optimization cost, the problem (A.7) can be translated into the form of

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\arg\min} \, & \left\|\mathbf{y} - \mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{X}_{\mathcal{F}}\mathbf{Q}_{\mathcal{F}}\mathbf{b}\right\|_2^2 \\ \text{subject to} \quad & \mathbf{Q}_{\mathcal{T}}\mathbf{b} = \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} \\ & \mathbf{Q}_{\mathcal{Z}}\mathbf{b} = \mathbf{0}. \end{aligned} \tag{A.8}$$

The last optimization is a restricted least squares problem that can be solved using the method of Lagrange multipliers to show that

$$\widehat{\boldsymbol{\beta}} = \mathbf{Q}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}}^+\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) + \mathbf{Q}_{\mathcal{T}}^T\widehat{\boldsymbol{\theta}}_{\mathcal{T}} \tag{A.9}$$

where $\widehat{\boldsymbol{\theta}}_{\mathcal{T}} \triangleq \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}$ and $\mathbf{X}_{\mathcal{F}}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{X}_{\mathcal{F}}$.

Appendix B. Proofs for Section 3.

In this section we outline the proof of Theorem 3.1 for the generalization error of the target task in the setting where a specific coordinate subset layout $\mathcal{L}$ determines the transferred set of parameters. We start in Section B.1 by providing auxiliary results that use non-asymptotic properties of Gaussian and Wishart matrices. Then, in Section B.2 we prove Theorem 3.1, in Section B.3 we prove Corollary 3.2, and in Section B.4 we prove Corollary 3.4.

B.1. Auxiliary Results using Non-Asymptotic Properties of Gaussian and Wishart Matrices. The random matrix $\mathbf{X}_{\mathcal{F}} \triangleq \mathbf{X}\mathbf{Q}_{\mathcal{F}}^T$ is of size $n\times p$ and all its components are i.i.d. standard Gaussian variables. Then, almost surely,

$$\mathbb{E}\left[\mathbf{X}_{\mathcal{F}}^+\mathbf{X}_{\mathcal{F}}\right] = \mathbf{I}_p \times \begin{cases} 1 & \text{for } p \le n, \\ \frac{n}{p} & \text{for } p > n, \end{cases} \tag{B.1}$$

where $\mathbf{X}_{\mathcal{F}}^+\mathbf{X}_{\mathcal{F}}$ is the $p \times p$ projection operator onto the range of $\mathbf{X}_{\mathcal{F}}$. Accordingly, let $\mathbf{a} \in \mathbb{R}^p$ be a random vector independent of the matrix $\mathbf{X}_{\mathcal{F}}$ and, then,

$$\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^+\mathbf{X}_{\mathcal{F}}\mathbf{a}\right\|_2^2\right] = \mathbb{E}\left[\|\mathbf{a}\|_2^2\right] \times \begin{cases} 1 & \text{for } p \le n, \\ \frac{n}{p} & \text{for } p > n. \end{cases} \tag{B.2}$$

The components of $\mathbf{X}_{\mathcal{F}}$ are i.i.d. standard Gaussian variables, hence $\mathbf{X}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}} \sim \mathcal{W}_p\left(\mathbf{I}_p, n\right)$ is a $p \times p$ Wishart matrix with n degrees of freedom, and $\mathbf{X}_{\mathcal{F}}\mathbf{X}_{\mathcal{F}}^T \sim \mathcal{W}_n\left(\mathbf{I}_n, p\right)$ is a $n \times n$

Wishart matrix with p degrees of freedom. The pseudoinverse of the $n \times n$ Wishart matrix (almost surely) satisfies

$$\mathbb{E}\left[\left(\mathbf{X}_{\mathcal{F}}\mathbf{X}_{\mathcal{F}}^{T}\right)^{+}\right] = \mathbb{E}\left[\mathbf{X}_{\mathcal{F}}^{+,T}\mathbf{X}_{\mathcal{F}}^{+}\right] = \mathbf{I}_n \times \begin{cases} \frac{1}{n-p-1} \cdot \frac{p}{n} & \text{for } p \le n-2, \\ \infty & \text{for } n-1 \le p \le n+1, \\ \frac{1}{p-n-1} & \text{for } p \ge n+2, \end{cases} \tag{B.3}$$

where the result for $p \ge n+2$ corresponds to the common case of inverse Wishart matrix with more degrees of freedom than its dimension, and the result for $p \le n-2$ is based on constructions provided in the proof of Theorem 1.3 in [5].

Following (B.3), let $\mathbf{u} \in \mathbb{R}^n$ be a random vector independent of $\mathbf{X}_{\mathcal{F}}$. Then,

$$\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{u}\right\|_2^2\right] = \frac{1}{n}\mathbb{E}\left[\|\mathbf{u}\|_2^2\right] \times \begin{cases} \frac{p}{n-p-1} & \text{for } p \le n-2, \\ \infty & \text{for } n-1 \le p \le n+1, \\ \frac{n}{p-n-1} & \text{for } p \ge n+2, \end{cases} \tag{B.4}$$

that specifically for $\mathbf{u} = \mathbf{X}_{\mathcal{F}^c}\boldsymbol{\beta}_{\mathcal{F}^c}$ becomes

$$\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}^c}\boldsymbol{\beta}_{\mathcal{F}^c}\right\|_2^2\right] = \mathbb{E}\left[\|\boldsymbol{\beta}_{\mathcal{F}^c}\|_2^2\right] \times \begin{cases} \frac{p}{n-p-1} & \text{for } p \le n-2, \\ \infty & \text{for } n-1 \le p \le n+1, \\ \frac{n}{p-n-1} & \text{for } p \ge n+2. \end{cases} \tag{B.5}$$

The results in (B.1)-(B.5) are presented using notions of the *target* task, specifically, using the data matrix $\mathbf{X}$ and the coordinate subset $\mathcal{T}$. One can obtain the corresponding results for the *source* task by updating (B.1)-(B.5) by replacing $\mathbf{X}$, $\mathcal{T}$, n and p with $\mathbf{Z}$, $\mathcal{S}$, $\widetilde{n}$ and $\widetilde{p}$, respectively. For example, the result corresponding to (B.1) is

$$\mathbb{E}\left[\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\right] = \mathbf{I}_{\widetilde{p}} \times \begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases} \tag{B.6}$$

where $\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}$ is the $\widetilde{p} \times \widetilde{p}$ projection operator onto the range of $\mathbf{Z}_{\mathcal{S}}$.

The next auxiliary results consider a coordinate subset layout $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ which is specific, i.e., non random, and therefore the induced operators such as $\mathbf{Q}_{\mathcal{S}}$, $\mathbf{Q}_{\mathcal{F}}$, $\mathbf{Q}_{\mathcal{T}}$, $\mathbf{Q}_{\mathcal{Z}}$ are also fixed and do not have any random aspect. Recall that $\mathbf{Q}_{\mathcal{S}}^T\mathbf{Q}_{\mathcal{S}}$ is a $d \times d$ diagonal matrix with its j^{th} diagonal component equals 1 if $j \in \mathcal{S}$ and 0 otherwise. Similarly holds for the other coordinate subsets. Accordingly, here the norms of vector forms such as $\boldsymbol{\beta}_{\mathcal{T}} \triangleq \mathbf{Q}_{\mathcal{T}}\boldsymbol{\beta}$, $\boldsymbol{\beta}_{\mathcal{F}} \triangleq \mathbf{Q}_{\mathcal{F}}\boldsymbol{\beta}$, and $\boldsymbol{\beta}_{\mathcal{Z}} \triangleq \mathbf{Q}_{\mathcal{Z}}\boldsymbol{\beta}$, are directly referred to as $\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2$, $\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2$, $\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2$, respectively.

Recall that $\mathcal{T} \subseteq \mathcal{S}$. Then, for a deterministic vector $\mathbf{w} \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{w}\right\|_2^2\right] = \begin{cases} \|\mathbf{w}_{\mathcal{T}}\|_2^2 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}(\widetilde{p}+1)}\left(\left(\widetilde{n} + \frac{\widetilde{n}-1}{\widetilde{p}-1}\right)\|\mathbf{w}_{\mathcal{T}}\|_2^2 + \left(1 - \frac{\widetilde{n}-1}{\widetilde{p}-1}\right)t\|\mathbf{w}_{\mathcal{S}}\|_2^2\right) & \text{for } \widetilde{p} > \widetilde{n}. \end{cases} \tag{B.7}$$

$$\text{(B.8)}\qquad \mathbb{E}\left[\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}^c}\mathbf{Q}_{\mathcal{S}^c}\mathbf{w}\right\|_2^2\right]=\frac{t}{\widetilde{p}}\left\|\mathbf{w}_{\mathcal{S}^c}\right\|_2^2\times\begin{cases}\frac{\widetilde{p}}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p}\le\widetilde{n}-2,\\ \infty & \text{for } \widetilde{n}-1\le\widetilde{p}\le\widetilde{n}+1,\\ \frac{\widetilde{n}}{\widetilde{p}-\widetilde{n}-1} & \text{for } \widetilde{p}\ge\widetilde{n}+2.\end{cases}$$

For two deterministic vectors $\mathbf{w},\mathbf{a}\in\mathbb{R}^d$,

$$\text{(B.9)}\qquad \mathbb{E}\left[\left\langle\mathbf{Q}_{\mathcal{T}}\mathbf{a},\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{w}\right\rangle\right]=\left\langle\mathbf{a}_{\mathcal{T}},\mathbf{w}_{\mathcal{T}}\right\rangle\times\begin{cases}1 & \text{for } \widetilde{p}\le\widetilde{n},\\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p}>\widetilde{n}.\end{cases}$$

For a deterministic vector $\mathbf{r}\in\mathbb{R}^{\widetilde{n}}$,

$$\text{(B.10)}\qquad \mathbb{E}\left[\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{r}\right\|_2^2\right]=\frac{t}{\widetilde{n}\widetilde{p}}\left\|\mathbf{r}\right\|_2^2\times\begin{cases}\frac{\widetilde{p}}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p}\le\widetilde{n}-2,\\ \infty & \text{for } \widetilde{n}-1\le\widetilde{p}\le\widetilde{n}+1,\\ \frac{\widetilde{n}}{\widetilde{p}-\widetilde{n}-1} & \text{for } \widetilde{p}\ge\widetilde{n}+2.\end{cases}$$

In our case we have the $\widetilde{n}\times\widetilde{p}$ matrix $\mathbf{Z}_{\mathcal{S}}$ that its components are i.i.d. standard Gaussian variables, thus, $\mathbf{Z}_{\mathcal{S}}$ can be decomposed into a form that involves an independent Haar-distributed matrix, i.e., a random orthonormal matrix that is uniformly distributed over the set of orthonormal matrices of the relevant size. This lets us to prove the results in (B.7)-(B.10) using some algebra and the non-asymptotic properties of random Haar-distributed matrices, see examples for such properties in Lemma 2.5 in [26] and also in Proposition 1.2 in [12].

B.2. Proof Outline of Theorem 3.1. The generalization error $\mathcal{E}_{\text{out}}$ of the target task was expressed in its basic form in Eq. (2.8) for a specific coordinate subset layout $\mathcal{L}=\{\mathcal{S},\mathcal{F},\mathcal{T},\mathcal{Z}\}$. Please note that the expectations below do not include any expectation with respect to $\mathcal{L}$ or its components, which are non-random here.

We start with the relevant decomposition of the error expression, namely,

$$\begin{aligned}\mathcal{E}_{\text{out}}&=\sigma_\epsilon^2+\mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|_2^2\right]\\ &=\sigma_\epsilon^2+\left\|\boldsymbol{\beta}_{\mathcal{Z}}\right\|_2^2+\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{y}-\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right)-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right]+\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\right].\end{aligned}\qquad\text{(B.11)}$$

Then, we use the expression for the estimate $\widehat{\boldsymbol{\theta}}$ given in (2.4) and the relation $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, to

decompose the third term in (B.11) as follows

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{y}-\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right)-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] \\
&=\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{y}-\mathbf{X}_{\mathcal{T}}\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right)-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right]+\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right)\right\|_2^2\right] \\
&=\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}}+\mathbf{X}_{\mathcal{T}^{\mathrm{c}}}\boldsymbol{\beta}_{\mathcal{T}^{\mathrm{c}}}+\boldsymbol{\epsilon}-\mathbf{X}_{\mathcal{T}}\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right)-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right]+\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right)\right\|_2^2\right] \\
&=\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{T}^{\mathrm{c}}}\boldsymbol{\beta}_{\mathcal{T}^{\mathrm{c}}}+\boldsymbol{\epsilon}\right)-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right]+\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\left(\boldsymbol{\beta}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right)\right\|_2^2\right] \\
&\quad+\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right)\right\|_2^2\right]
\end{aligned} \tag{B.12}$$

We further develop the last expression using the result in (B.4) and that $\mathbf{X}_{\mathcal{T}}^{T}\mathbf{X}_{\mathcal{T}}\sim\mathcal{W}_t\left(\mathbf{I}_t,n\right)$ is a Wishart matrix with mean $\mathbb{E}\left[\mathbf{X}_{\mathcal{T}}^{T}\mathbf{X}_{\mathcal{T}}\right]=n\mathbf{I}_t$, and get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{y}-\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right)-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] \\
&=\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{T}^{\mathrm{c}}}\boldsymbol{\beta}_{\mathcal{T}^{\mathrm{c}}}+\boldsymbol{\epsilon}\right)-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] \\
&+\left(\left\|\boldsymbol{\beta}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2+\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2\right]\right)\times\begin{cases}\frac{p}{n-p-1} & \text{for } p\leq n-2,\\ \infty & \text{for } n-1\leq p\leq n+1,\\ \frac{n}{p-n-1} & \text{for } p\geq n+2.\end{cases}
\end{aligned} \tag{B.13}$$

We proceed to the fourth error term in (B.11), i.e., the error in the subvector induced by the specific $\mathcal{T}$ of interest, and develop its formulation as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\right]= \\
&=\left\|\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]-\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2+\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2\right]+2\mathbb{E}\left[\left(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right)^{T}\left(\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]-\boldsymbol{\beta}_{\mathcal{T}}\right)\right] \\
&=\left\|\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]-\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2+\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2\right]
\end{aligned} \tag{B.14}$$

Then, setting (B.13) and (B.14) in (B.11) gives

$$\mathcal{E}_{\text{out}} = \sigma_\epsilon^2 + \|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{T}^c}\boldsymbol{\beta}_{\mathcal{T}^c} + \boldsymbol{\epsilon}\right) - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] + \left(\left\|\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right] - \boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2 + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2\right]\right)\left(1 + \begin{cases} \frac{p}{n-p-1} & \text{for } p \le n-2, \\ \infty & \text{for } n-1 \le p \le n+1, \\ \frac{n}{p-n-1} & \text{for } p \ge n+2. \end{cases}\right). \tag{B.15}$$

Also,

$$\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{T}^c}\boldsymbol{\beta}_{\mathcal{T}^c} + \boldsymbol{\epsilon}\right) - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] = \mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}}\boldsymbol{\beta}_{\mathcal{F}} - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] + \mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{Z}}\boldsymbol{\beta}_{\mathcal{Z}} + \boldsymbol{\epsilon}\right)\right\|_2^2\right] \tag{B.16}$$

where the first term can be developed using (B.2) into

$$\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}}\boldsymbol{\beta}_{\mathcal{F}} - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] = \|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2 \times \begin{cases} 0 & \text{for } p \le n, \\ 1 - \frac{n}{p} & \text{for } p > n, \end{cases} \tag{B.17}$$

and the second term in (B.16) can be rewritten using the result in (B.4) and that $\mathbf{X}_{\mathcal{Z}}^T\mathbf{X}_{\mathcal{Z}} \sim \mathcal{W}_{d-p-t}\left(\mathbf{I}_{d-p-t}, n\right)$ is a Wishart matrix with mean $\mathbb{E}\left[\mathbf{X}_{\mathcal{Z}}^T\mathbf{X}_{\mathcal{Z}}\right] = n\mathbf{I}_{d-p-t}$:

$$\mathbb{E}\left[\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{Z}}\boldsymbol{\beta}_{\mathcal{Z}} + \boldsymbol{\epsilon}\right)\right\|_2^2\right] = \left(\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \sigma_\epsilon^2\right) \times \begin{cases} \frac{p}{n-p-1} & \text{for } p \le n-2, \\ \infty & \text{for } n-1 \le p \le n+1, \\ \frac{n}{p-n-1} & \text{for } p \ge n+2. \end{cases} \tag{B.18}$$

Hence, (B.16)-(B.18) let us write (B.15) in the form which is provided in Theorem 3.1.

B.3. Proof of Corollary 3.2. To prove Corollary 3.2 we first formulate the expectation of the transferred parameters:

$$\begin{aligned} \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right] = \mathbb{E}\left[\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}\mathbf{H}\boldsymbol{\beta}\right] &= \mathbb{E}\left[\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{H}\boldsymbol{\beta}\right] \\ &= \mathbf{Q}_{\mathcal{T}}\mathbf{H}\boldsymbol{\beta} \times \begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases} \end{aligned} \tag{B.19}$$

where the last equality stems from (B.6). Consequently, we use (B.19) to formulate the transfer bias term from Theorem 3.1 as

$$\text{Bias}_{\mathcal{T}}^2 = \left\|\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right] - \boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2 = \left\|\mathbf{Q}_{\mathcal{T}}\left(r\mathbf{H} - \mathbf{I}_d\right)\boldsymbol{\beta}\right\|_2^2 \quad \text{where} \quad r \triangleq \begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases} \tag{B.20}$$

which corresponds to the formulation in Corollary 3.2.

Now we turn to prove the transfer variance formulation from Corollary 3.2. For a start, note that

$$\begin{aligned}\mathrm{Var}_{\mathcal{T},\mathcal{S}} &= \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2\right] = \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right] - \left\|\mathbb{E}\left[\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right]\right\|_2^2 \\ &= \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right] - \|\mathbf{Q}_{\mathcal{T}}\mathbf{H}\boldsymbol{\beta}\|_2^2 \times \begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \left(\frac{\widetilde{n}}{\widetilde{p}}\right)^2 & \text{for } \widetilde{p} > \widetilde{n}. \end{cases}\end{aligned} \tag{B.21}$$

To further develop the last expression, we will now explicitly formulate $\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right]$. Using the auxiliary notation $\boldsymbol{\beta}^{(\mathbf{H})} \triangleq \mathbf{H}\boldsymbol{\beta}$ we get

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right] &= \mathbb{E}\left[\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\left(\mathbf{Z}\mathbf{H}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\xi}\right)\right\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{Z}_{\mathcal{S}}\left(\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})} + \boldsymbol{\eta}_{\mathcal{S}}\right)\right\|_2^2\right] \\ &\quad + \mathbb{E}\left[\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{Z}_{\mathcal{S}^c}\left(\boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} + \boldsymbol{\eta}_{\mathcal{S}^c}\right)\right\|_2^2\right] + \mathbb{E}\left[\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\boldsymbol{\xi}\right\|_2^2\right]\end{aligned} \tag{B.22}$$

that using (B.7)-(B.8), (B.10) leads to

$$\begin{aligned}&\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right] = \\ &= \left(\begin{cases} \left\|\boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})}\right\|_2^2 + t\sigma_\eta^2 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}(\widetilde{p}+1)}\left(\left(\widetilde{n} + \frac{\widetilde{n}-1}{\widetilde{p}-1}\right)\left(\left\|\boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})}\right\|_2^2 + t\sigma_\eta^2\right) + \left(1 - \frac{\widetilde{n}-1}{\widetilde{p}-1}\right)t\left(\left\|\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})}\right\|_2^2 + \widetilde{p}\sigma_\eta^2\right)\right) & \text{for } \widetilde{p} > \widetilde{n}, \end{cases}\right) \\ &\quad + \frac{t}{\widetilde{p}}\left(\left\|\boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})}\right\|_2^2 + (d-\widetilde{p})\sigma_\eta^2 + \sigma_\xi^2\right) \times \left(\begin{cases} \frac{\widetilde{p}}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}-\widetilde{n}-1} & \text{for } \widetilde{p} \ge \widetilde{n}+2, \end{cases}\right) \\ &= \begin{cases} \zeta_{\mathcal{T}} + t\sigma_\eta^2 + t\cdot\frac{\zeta_{\mathcal{S}^c} + (d-\widetilde{p})\sigma_\eta^2 + \sigma_\xi^2}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}}\left(\frac{(\widetilde{p}-\widetilde{n})t\left(\zeta_{\mathcal{S}} + \widetilde{p}\sigma_\eta^2\right) + (\widetilde{n}\widetilde{p}-1)\left(\zeta_{\mathcal{T}} + t\sigma_\eta^2\right)}{\widetilde{p}^2-1} + t\cdot\frac{\zeta_{\mathcal{S}^c} + (d-\widetilde{p})\sigma_\eta^2 + \sigma_\xi^2}{\widetilde{p}-\widetilde{n}-1}\right) & \text{for } \widetilde{p} \ge \widetilde{n}+2. \end{cases}\end{aligned} \tag{B.23}$$

where $\zeta_{\mathcal{T}} \triangleq \|\mathbf{Q}_{\mathcal{T}}\mathbf{H}\boldsymbol{\beta}\|_2^2$, $\zeta_{\mathcal{S}^c} \triangleq \|\mathbf{Q}_{\mathcal{S}^c}\mathbf{H}\boldsymbol{\beta}\|_2^2$ and $\zeta_{\mathcal{S}} \triangleq \|\mathbf{Q}_{\mathcal{S}}\mathbf{H}\boldsymbol{\beta}\|_2^2$. Then, we set (B.23) in (B.21) and using some algebra obtain the transfer variance formulation in Corollary 3.2.

B.4. Proof of Corollary 3.4. Corollary 3.4 formulates the generalization error of the target task under the expectation over a coordinate layout $\mathcal{L}$ which is chosen uniformly at

random. Recall Definition 3.3 in the main text that characterizes a coordinate subset layout $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ that is $\{\widetilde{p}, p, t\}$-uniformly distributed, for $\widetilde{p} \in \{1, \ldots, d\}$ and $(p, t) \in \{0, \ldots, d\} \times \{0, \ldots, \widetilde{p}\}$ such that $p + t \leq d$. Here we provide several auxiliary results that are induced by this random structure and utilized in the proof of Corollary 3.4.

For $\mathcal{S}$ that is uniformly chosen at random from all the subsets of $\widetilde{p}$ unique coordinates of $\{1, \ldots, d\}$, we get that the mean of the projection operator $\mathbf{Q}_{\mathcal{S}}^T \mathbf{Q}_{\mathcal{S}}$ is

$$\mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{S}}^T \mathbf{Q}_{\mathcal{S}}\right] = \mathbb{E}_{\mathcal{S}}\left[\mathbf{Q}_{\mathcal{S}}^T \mathbf{Q}_{\mathcal{S}}\right] = \frac{\binom{d-1}{\widetilde{p}-1}}{\binom{d}{\widetilde{p}}} \mathbf{I}_d = \frac{\widetilde{p}}{d} \mathbf{I}_d \tag{B.24}$$

where we used the structure of $\mathbf{Q}_{\mathcal{S}}^T \mathbf{Q}_{\mathcal{S}}$ that is a $d \times d$ diagonal matrix with its j^{th} diagonal component equals 1 if $j \in \mathcal{S}$ and 0 otherwise.

Definition 3.3 also specifies that, given $\mathcal{S}$, the target-task coordinate layout $\{\mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ is uniformly chosen at random from all the layouts where $\mathcal{F}$, $\mathcal{T}$, and $\mathcal{Z}$ are three disjoint sets of coordinates that satisfy $\mathcal{F} \cup \mathcal{T} \cup \mathcal{Z} = \{1, \ldots, d\}$ such that $|\mathcal{F}| = p$, $|\mathcal{T}| = t$ and $\mathcal{T} \subseteq \mathcal{S}$, and $|\mathcal{Z}| = d - p - t$. Accordingly,

$$\begin{aligned}
\mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{T}}^T \mathbf{Q}_{\mathcal{T}}\right] &= \mathbb{E}_{S}\left[\mathbb{E}_{\mathcal{L}|\mathcal{S}}\left[\mathbf{Q}_{\mathcal{T}}^T \mathbf{Q}_{\mathcal{T}}\right]\right] \\
&= \frac{\binom{\widetilde{p}-1}{t-1}}{\binom{\widetilde{p}}{t}} \mathbb{E}_{S}\left[\mathbf{Q}_{\mathcal{S}}^T \mathbf{Q}_{\mathcal{S}}\right] \\
&= \frac{t}{\widetilde{p}} \mathbb{E}_{S}\left[\mathbf{Q}_{\mathcal{S}}^T \mathbf{Q}_{\mathcal{S}}\right] \\
&= \frac{t}{d} \mathbf{I}_d, \qquad \text{(B.25)}
\end{aligned}$$

and similarly

$$\mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{F}}^T \mathbf{Q}_{\mathcal{F}}\right] = \frac{p}{d} \mathbf{I}_d, \tag{B.26}$$

$$\mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{Z}}^T \mathbf{Q}_{\mathcal{Z}}\right] = \frac{d - p - t}{d} \mathbf{I}_d. \tag{B.27}$$

Another useful auxiliary result, based on the relation $\mathbf{Q}_{\mathcal{S}} \mathbf{Q}_{\mathcal{S}}^T = \mathbf{I}_{\widetilde{p}}$ (carefully note the transpose appearance), is provided by

$$\begin{aligned}
\mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{S}} \mathbf{Q}_{\mathcal{T}}^T \mathbf{Q}_{\mathcal{T}} \mathbf{Q}_{\mathcal{S}}^T\right] &= \mathbb{E}_{S}\left[\mathbf{Q}_{\mathcal{S}} \mathbb{E}_{\mathcal{L}|\mathcal{S}}\left[\mathbf{Q}_{\mathcal{T}}^T \mathbf{Q}_{\mathcal{T}}\right] \mathbf{Q}_{\mathcal{S}}^T\right] \\
&= \frac{t}{\widetilde{p}} \mathbb{E}_{S}\left[\mathbf{Q}_{\mathcal{S}} \mathbf{Q}_{\mathcal{S}}^T \mathbf{Q}_{\mathcal{S}} \mathbf{Q}_{\mathcal{S}}^T\right] \\
&= \frac{t}{\widetilde{p}} \mathbf{I}_{\widetilde{p}}. \qquad \text{(B.28)}
\end{aligned}$$

The results in (B.25)–(B.27) imply that

$$\mathbb{E}_{\mathcal{L}}\left[\left\|\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\right] = \boldsymbol{\beta}^T \mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{T}}^T \mathbf{Q}_{\mathcal{T}}\right] \boldsymbol{\beta} = \frac{t}{d} \left\|\boldsymbol{\beta}\right\|_2^2, \tag{B.29}$$

$$\mathbb{E}_{\mathcal{L}}\left[\left\|\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right] = \boldsymbol{\beta}^T \mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{F}}^T \mathbf{Q}_{\mathcal{F}}\right] \boldsymbol{\beta} = \frac{p}{d} \left\|\boldsymbol{\beta}\right\|_2^2, \tag{B.30}$$

$$\mathbb{E}_{\mathcal{L}}\left[\left\|\boldsymbol{\beta}_{\mathcal{Z}}\right\|_2^2\right] = \boldsymbol{\beta}^T \mathbb{E}_{\mathcal{L}}\left[\mathbf{Q}_{\mathcal{Z}}^T \mathbf{Q}_{\mathcal{Z}}\right] \boldsymbol{\beta} = \frac{d - p - t}{d} \left\|\boldsymbol{\beta}\right\|_2^2, \tag{B.31}$$

where $\boldsymbol{\beta}_{\mathcal{T}} \triangleq \mathbf{Q}_{\mathcal{T}}\boldsymbol{\beta}$, $\boldsymbol{\beta}_{\mathcal{F}} \triangleq \mathbf{Q}_{\mathcal{F}}\boldsymbol{\beta}$, and $\boldsymbol{\beta}_{\mathcal{Z}} \triangleq \mathbf{Q}_{\mathcal{Z}}\boldsymbol{\beta}$. Note that the expressions in (B.29)-(B.31) hold also for d-dimensional deterministic vectors other than $\boldsymbol{\beta}$, e.g., (B.29)-(B.31) hold for $\boldsymbol{\beta}^{(\mathbf{H})} \triangleq \mathbf{H}\boldsymbol{\beta}$.

Then, the auxiliary results in (B.24)-(B.31) can be utilized to formulate the expectation over $\mathcal{L}$ of the analytical results in Theorem 3.1 and, by that, proving Corollary 3.4.

Appendix C. Additional Results for Section 3.

C.1. Additional Results for the On-Average Analysis in Section 3.2. In Fig. 11 we present the empirically computed values of the out-of-sample squared error of the target task, $\mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{out}}\right]$, with respect to the number of free parameters $\widetilde{p}$ and p (in the source and target tasks, respectively). The empirical values in Fig. 11 (and also the values denoted by circle markers in Fig. 1 in Section 3.2) were obtained by averaging over 250 experiments where each experiment was carried out based on new realizations of the data matrices, noise components, and the sequential order of adding coordinates to subsets (such as $\mathcal{S}$) for the gradual increase of $\widetilde{p}$ and p within each experiment. Note that the results in Fig. 4 do not include averaging over the sequential order of adding coordinates to subsets. Each single evaluation of the expectation of the squared error for an out-of-sample data pair $\left(\mathbf{x}^{(\text{test})}, y^{(\text{test})}\right)$ was empirically carried out by averaging over a set of 1000 out-of-sample realizations of data pairs. Here $d = 120$, $\widetilde{n} = 50$, $n = 20$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$. The deterministic $\boldsymbol{\beta} \in \mathbb{R}^d$ used in the experiments satisfies $\|\boldsymbol{\beta}\|_2^2 = d$.

One can observe the excellent match between the empirical results in Fig. 11 and the analytical results provided in Fig. 10. This further establishes the formulations given in Corollary 3.4.

C.2. Additional Results for the Single-Layout Analysis in Section 3.3. The following results are for two different forms of the true solution $\boldsymbol{\beta}$: the first is a form with linearly increasing values (Fig. 4a), the second is a form with sparse values where only 25% of coordinates have non-zero value (Fig. 4e). Note that both forms satisfy $\|\boldsymbol{\beta}\|_2^2 = d$.

The three types of linear operator $\mathbf{H}$ in the evaluations in Section 3.3 are as follows. First, $\mathbf{H} = \mathbf{I}_d$ that is the identity operator. Second, is the circulant matrix $\mathbf{H}$ that corresponds to a shift-invariant local averaging operator that uniformly considers 11-coordinates neighborhood around the computed coordinate (note that in other parts of this paper we consider also averaging operators with neighborhood sizes other than 11). Third, is the circulant matrix $\mathbf{H}$ that corresponds to discrete derivative operator based on the convolution kernel $[-0.5, 0.5]$.

Figures 13-14 present the analytical and empirical values of the generalization error of the target task with respect to specific coordinate layouts $\mathcal{L}$ that evolve with respect to the value of p (this evolution of $\mathcal{L}$ is the same in each of the subfigures and it is not particularly designed to any of the combinations of the true $\boldsymbol{\beta}$, $\mathbf{H}$, and σ_η^2). It is clear from Figures 13-14 that the increase in σ_η^2, which by its definition corresponds to less related source and target tasks, reduces the benefits or even increases the harm due to transfer of parameters (one can observe that in Figs. 13-14 by comparing the error curves among subfigures in the same row).

The effect of $\mathbf{H}$ with respect to the true $\boldsymbol{\beta}$ is also evident. First, the identity operator $\mathbf{H} = \mathbf{I}_d$ does not reduce the relation between the source and target tasks and therefore does not degrade the parameter transfer performance by itself (i.e., for $\mathbf{H} = \mathbf{I}_d$, only the

additive noise level σ_η^2 can reduce the relation between the tasks). Second, when $\mathbf{H}$ is a local averaging operator it does not reduce the benefits from transfer learning (e.g., compare second to first row of subfigures in Figs. 13-14) in the case of linearly-increasing $\boldsymbol{\beta}$ shape (because local averaging does not affect a linear function, except to the few first and last coordinates where the periodic averaging is applied), in contrast, the local averaging operator significantly degrades the parameter transfer performance in the case of the sparse $\boldsymbol{\beta}$ form. Lastly, when $\mathbf{H}$ is a discrete derivative operator it renders transfer learning harmful in the case of linearly-increasing $\boldsymbol{\beta}$ shape (e.g., compare third to first row of subfigures in Fig. 13). In the case of the sparse $\boldsymbol{\beta}$ form the discrete derivative reduces the potential benefits of the parameter transfer but does not eliminate them completely in case these benefits exist for $\mathbf{H} = \mathbf{I}_d$ (e.g., compare third to first row of subfigures in Fig. 14).

Appendix D. Additional Empirical Results on Parameter Transfer Usefulness.

D.1. Details on the Empirical Evaluation of $\Delta\mathcal{E}_{\text{transfer}}$. The analytical formula for $\Delta\mathcal{E}_{\text{transfer}} \triangleq \mathbb{E}_{\mathcal{L}}\left[\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}\right]$, which is based on (4.3) and Corollary 3.4, measures the (normalized) expected difference in the generalization error (of the target task) due to transferring parameters instead of setting them to zero. Accordingly, the empirical evaluation of $\Delta\mathcal{E}_{\text{transfer}}$ for a given $\widetilde{p}$ can be computed by

$$\widehat{\Delta}\mathcal{E}_{\text{transfer}} = \frac{1}{d-3} \sum_{p=1,\ldots,n-2,n+2,\ldots,d} \frac{\widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}^{(\widetilde{p},p,t=m)}\right\} - \widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}^{(\widetilde{p},p,t=0)}\right\}}{m \cdot \alpha(p)} \tag{D.1}$$

where

$$\alpha(p) \triangleq \begin{cases} 1 + \frac{p}{n-p-1} & \text{for } p \leq n-2, \\ 1 + \frac{n}{p-n-1} & \text{for } p \geq n+2 \end{cases} \tag{D.2}$$

is a normalization factor required for independence from p. The value measured in (D.1) is also normalized by the number of transferred parameters. Here $\widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}^{(\widetilde{p},p,t=m)}\right\}$ is the out-of-sample error of the target task that is *empirically* computed for m transferred parameters, p free parameters in the target task, and $\widetilde{p}$ free parameters in the source task. Correspondingly, $\widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}^{(\widetilde{p},p,t=0)}\right\}$ is the empirically computed error induced by avoiding parameter transfer. Therefore, the formula in (D.1) empirically measures the average error difference for a single transferred parameter by averaging over the various settings induced by different values of p while $\widetilde{p}$ is kept fixed. To obtain a good numerical accuracy with averaging over a moderate number of experiments we use the value $m = 5$.

Each empirical evaluation of $\widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}^{(\widetilde{p},p,t)}\right\}$, for a specific set of values $\widetilde{p}, p, t$ corresponds to averaging over 500 experiments where each experiment was conducted for new realizations of the data matrices, noise components, and the sequential order of adding coordinates to subsets. Each single evaluation of the expectation of the squared error for an out-of-sample data pair $\left(\mathbf{x}^{(\text{test})}, y^{(\text{test})}\right)$ was empirically computed by averaging over 1000 out-of-sample realizations of data pairs.

D.2. Results for $\widetilde{n} < d$. In Figures 16-17 we present analytical and empirical values of $\frac{1}{t}\Delta\mathcal{E}_{\text{transfer}}$ induced by settings where $\widetilde{n} < d$ (specifically, $\widetilde{n} = 50$ and $d = 80$), which naturally enable the corresponding overparameterized (i.e., $\widetilde{n} < \widetilde{p} < d$) and underparameterized (i.e., $\widetilde{p} < \widetilde{n} < d$) settings of the source task. In Fig. 16 we provide the analytical and empirical results for cases where $\mathbf{H}$ is local averaging and discrete derivative operators. In the main text only the analytical results were provided and here we show them again near their empirical counterparts that excellently match them (up to the resolution of the empirical settings).

In Figure 17 we provide additional results for cases where the operator $\mathbf{H}$ is a scaled identity matrix.

D.3. Results for $\widetilde{n} > d$. Here we provide in Fig. 18 the analytical and empirical evaluations of $\Delta\mathcal{E}_{\text{transfer}}$ that correspond to settings where $\widetilde{n} > d$, we specifically consider $\widetilde{n} = 150$ and $d = 80$. Note that $\widetilde{n} > d$ implies that, by the definition of $\widetilde{p}$, the corresponding settings (of the source task) are underparameterized with $\widetilde{p} \leq d < \widetilde{n}$. Like in Fig. 17, the results in Fig. 18 show the excellent match between the analytical and empirical results.

D.4. Empirical Results for Benefits in Transferred versus Free Parameters. We provide in Figure 19 the empirical evaluation of $\Delta\mathcal{E}_{\text{TvsF}}$ that corresponds to the analytical evaluation in Figure 6 in the main text. The empirical evaluation was conducted by averaging over 750 experiments where, in each of them, $\Delta\mathcal{E}_{\text{TvsF}}$ was evaluated based on its definition in (4.4) and using a different realization of $\mathcal{L}$ (from the uniform distribution that we use) and its corresponding $\mathcal{L}''$.

Appendix E. The Optimal Componentwise Task Relation: Additional Analytical and Empirical Details.

E.1. Proof of Theorem 5.1. The proof outline for Theorem 5.1, which characterizes the optimal $\mathbf{H}$ in a componentwise task relation, is as follows. Recall that in this theorem we consider $\mathbf{H} = \text{diag}\{\lambda_{\mathbf{H}}^{(1)}, \ldots, \lambda_{\mathbf{H}}^{(d)}\}$. The proof starts by setting the expressions from (5.3)-(5.4) in the formulations of the transfer bias and variance from Corollary 3.2, then we can also reformulate the expression of $\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ from Theorem 3.1.

In the case of an underparameterized source task where $1 \leq \widetilde{p} \leq \widetilde{n} - 2$, we get

(E.1)
$$\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = t\sigma_\eta^2 + \sum_{j\in\mathcal{T}} \left(\lambda_{\mathbf{H}}^{(j)} - 1\right)^2 \left(\beta^{(j)}\right)^2 + \frac{t}{\widetilde{n} - \widetilde{p} - 1} \left((d - \widetilde{p})\,\sigma_\eta^2 + \sum_{j\in\mathcal{S}^c} \left(\lambda_{\mathbf{H}}^{(j)} \beta^{(j)}\right)^2 + \sigma_\xi^2 \right)$$

Recall that $\mathcal{T} \subseteq \mathcal{S}$ and hence $\mathcal{T} \cap \mathcal{S}^c = \emptyset$. Accordingly, the two sums in (E.1) include distinct sets of coordinates, which simplify the derivative of $\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ with respect to $\lambda_{\mathbf{H}}^{(k)}$ for a particular k. Hence, in this case of an underparameterized source task and $\beta^{(k)} \neq 0$, the condition $\frac{\partial \mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}}{\partial \lambda_{\mathbf{H}}^{(k)}} = 0$ is satisfied by $\lambda_{\mathbf{H}}^{(k)} = 1$ for $k \in \mathcal{T}$ and by $\lambda_{\mathbf{H}}^{(k)} = 0$ for $k \in \mathcal{S}^c$, which minimize $\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ due to convexity. Note that the eigenvalues $\{\lambda_{\mathbf{H}}^{(k)}\}_{k\in\mathcal{S}\setminus\mathcal{T}}$ do not appear in (E.1) and hence they can have any value without affecting the minimization of $\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ for $\widetilde{p} \leq \widetilde{n} - 2$.

In the case of an overparameterized source task where $\widetilde{p} \geq \widetilde{n} + 2$, we get

$$\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}} = \frac{\widetilde{n}}{\widetilde{p}} \left(t\sigma_\eta^2 + \frac{\widetilde{p}-\widetilde{n}}{\widetilde{p}^2-1} t \sum_{j \in \mathcal{S} \setminus \mathcal{T}} \left(\lambda^{(j)}_{\mathbf{H}} \beta^{(j)} \right)^2 + \sum_{j \in \mathcal{T}} \left(\frac{(\widetilde{p}-\widetilde{n})\, t + \widetilde{n}\widetilde{p} - 1}{\widetilde{p}^2 - 1} \lambda^{(j)}_{\mathbf{H}} - 2 \right) \lambda^{(j)}_{\mathbf{H}} \left(\beta^{(j)} \right)^2 + \frac{t}{\widetilde{p}-\widetilde{n}-1} \left((d-\widetilde{p})\, \sigma_\eta^2 + \sum_{j \in \mathcal{S}^c} \left(\lambda^{(j)}_{\mathbf{H}} \beta^{(j)} \right)^2 + \sigma_\xi^2 \right) \right) \tag{E.2}$$

where the three sums refer to disjoint sets of coordinates. Accordingly, in this case of an overparameterized source task and $\beta^{(k)} \neq 0$, the condition $\frac{\partial \mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}}{\partial \lambda^{(k)}_{\mathbf{H}}} = 0$ is satisfied by $\lambda^{(k)}_{\mathbf{H}} = \frac{\widetilde{p}^2-1}{\widetilde{n}\widetilde{p}-1+t(\widetilde{p}-\widetilde{n})}$ for $k \in \mathcal{T}$ and by $\lambda^{(k)}_{\mathbf{H}} = 0$ for $k \in \mathcal{T}^c$, which minimize $\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$ due to convexity.

In both cases (E.1) and (E.2), if $k \in \{1, \ldots, d\}$ corresponds to $\beta^{(k)} = 0$ then $\lambda^{(k)}_{\mathbf{H}}$ may have any value without affecting the minimization of $\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$. By this we complete the proof outline for Theorem 5.1.

E.2. Empirical Results for The Optimal $\mathbf{H}$ in a Componentwise Task Relation. In Figure 20 we provide the empirical evaluations that correspond to the analytical results in Figure 7. The empirical values were computed by averaging over 500 experiments.

Appendix F. Additional Details and Results for Section 6.

F.1. Ridge Regression: Error Expression and Optimal Tuning for the Source Task. The out-of-sample error of the ridge solution (6.5)-(6.6) to the source task is developed as follows. First, due to the layout of free parameters in $\widehat{\boldsymbol{\theta}}$ we get that

$$\begin{aligned} \widetilde{\mathcal{E}}_{\text{out}} &= \sigma_\xi^2 + \mathbb{E}\left[\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|_2^2 \right] \\ &= \sigma_\xi^2 + \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2 + \mathbb{E}\left[\left\| \widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S}} \right\|_2^2 \right]. \end{aligned} \tag{F.1}$$

Now we set the closed-form ridge solution from (6.6) to develop the third term in (F.1):

$$\begin{aligned} &\mathbb{E}\left[\left\| \widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S}} \right\|_2^2 \right] = \mathbb{E}\left[\left\| \left(\mathbf{Z}_{\mathcal{S}}^T \mathbf{Z}_{\mathcal{S}} + \widetilde{\alpha} \mathbf{I}_{\widetilde{p}} \right)^{-1} \mathbf{Z}_{\mathcal{S}}^T \mathbf{v} - \boldsymbol{\theta}_{\mathcal{S}} \right\|_2^2 \right] \\ &= \mathbb{E}\left[\left\| \left(\mathbf{Z}_{\mathcal{S}}^T \mathbf{Z}_{\mathcal{S}} + \widetilde{\alpha} \mathbf{I}_{\widetilde{p}} \right)^{-1} \mathbf{Z}_{\mathcal{S}}^T \left(\mathbf{Z}_{\mathcal{S}^c} \boldsymbol{\theta}_{\mathcal{S}^c} + \boldsymbol{\xi} \right) \right\|_2^2 \right] + \mathbb{E}\left[\left\| \left(\left(\mathbf{Z}_{\mathcal{S}}^T \mathbf{Z}_{\mathcal{S}} + \widetilde{\alpha} \mathbf{I}_{\widetilde{p}} \right)^{-1} \mathbf{Z}_{\mathcal{S}}^T \mathbf{Z}_{\mathcal{S}} - \mathbf{I}_{\widetilde{p}} \right) \boldsymbol{\theta}_{\mathcal{S}} \right\|_2^2 \right]. \end{aligned} \tag{F.2}$$

Next, note that

$$\mathbb{E}\left[\left\| \left(\mathbf{Z}_{\mathcal{S}}^T \mathbf{Z}_{\mathcal{S}} + \widetilde{\alpha} \mathbf{I}_{\widetilde{p}} \right)^{-1} \mathbf{Z}_{\mathcal{S}}^T \left(\mathbf{Z}_{\mathcal{S}^c} \boldsymbol{\theta}_{\mathcal{S}^c} + \boldsymbol{\xi} \right) \right\|_2^2 \right] = \left(\sigma_\xi^2 + \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2 \right) \mathbb{E}\left[\operatorname{Tr}\left\{ \left(\mathbf{Z}_{\mathcal{S}}^T \mathbf{Z}_{\mathcal{S}} + \widetilde{\alpha} \mathbf{I}_{\widetilde{p}} \right)^{-2} \mathbf{Z}_{\mathcal{S}}^T \mathbf{Z}_{\mathcal{S}} \right\} \right] \tag{F.3}$$

Also, we use the eigendecomposition $\mathbf{Z}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}} = \mathbf{\Phi}_{\mathbf{Z}_{\mathcal{S}}}\mathbf{\Lambda}_{\mathbf{Z}_{\mathcal{S}}}\mathbf{\Phi}_{\mathbf{Z}_{\mathcal{S}}}^T$ where $\mathbf{\Phi}_{\mathbf{Z}_{\mathcal{S}}}$ is a $\widetilde{p}\times\widetilde{p}$ orthonormal matrix and $\mathbf{\Lambda}_{\mathbf{Z}_{\mathcal{S}}} \triangleq \operatorname{diag}\{\lambda_{\mathbf{Z}_{\mathcal{S}},1},\dots,\lambda_{\mathbf{Z}_{\mathcal{S}},\widetilde{p}}\}$ is the $\widetilde{p}\times\widetilde{p}$ diagonal matrix formed by the eigenvalues of $\mathbf{Z}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}$. By using this eigendecomposition and (F.3), we develop (F.2) into the form of

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_{\mathcal{S}}\right\|_2^2\right] = \\
&= \left(\sigma_\xi^2 + \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2\right)\mathbb{E}\left[\operatorname{Tr}\left\{\left(\mathbf{\Lambda}_{\mathbf{Z}_{\mathcal{S}}} + \widetilde{\alpha}\mathbf{I}_{\widetilde{p}}\right)^{-2}\mathbf{\Lambda}_{\mathbf{Z}_{\mathcal{S}}}\right\}\right] + \frac{\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2}{\widetilde{p}}\widetilde{\alpha}^2\mathbb{E}\left[\operatorname{Tr}\left\{\left(\mathbf{\Lambda}_{\mathbf{Z}_{\mathcal{S}}} + \widetilde{\alpha}\mathbf{I}_{\widetilde{p}}\right)^{-2}\right\}\right]. \\
&= \mathbb{E}\left[\sum_{k=1}^{\widetilde{p}} \frac{\left(\sigma_\xi^2 + \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2\right)\lambda_{\mathbf{Z}_{\mathcal{S}},k} + \frac{\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2}{\widetilde{p}}\widetilde{\alpha}^2}{\left(\lambda_{\mathbf{Z}_{\mathcal{S}},k} + \widetilde{\alpha}\right)^2}\right].
\end{aligned} \tag{F.4}$$

The optimal tuning is achieved by the $\widetilde{\alpha}$ value that satisfies $\frac{\partial\widetilde{\mathcal{E}}_{\text{out}}}{\partial\widetilde{\alpha}} = 0$, which by using (F.1) and (F.4) is $\widetilde{\alpha} = \frac{\widetilde{p}\left(\sigma_\xi^2 + \|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2\right)}{\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2}$. In the main text we explain how we approximate the optimal $\widetilde{\alpha}$ in our experiments.

F.2. Ridge Regression: Error Expression and Optimal Tuning for the Target Task.

Let us start by explaining in more detail the ridge solution in (6.8). For this, note that the optimization constraints in (6.7) imply $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 = \left\|\mathbf{y} - \mathbf{X}_{\mathcal{F}}\mathbf{b}_{\mathcal{F}} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2$. Hence, the solution of (6.7) is equivalent to $\widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}}_{\mathcal{Z}} = 0$, $\widehat{\boldsymbol{\beta}}_{\mathcal{T}} = \widehat{\boldsymbol{\theta}}_{\mathcal{T}}$, and

$$\widehat{\boldsymbol{\beta}}_{\mathcal{F}} = \underset{\mathbf{f}\in\mathbb{R}^p}{\arg\min} \left\|\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) - \mathbf{X}_{\mathcal{F}}\mathbf{f}\right\|_2^2 + \alpha\|\mathbf{f}\|_2^2. \tag{F.5}$$

The optimization in (F.5) has a standard ridge regression form and, accordingly, its closed-form solution is provided in (6.8).

Let us develop the out-of-sample error expression of the target task. Based on the layout of free, transferred, and zeroed parameters in $\widehat{\boldsymbol{\beta}}$ we have

$$\begin{aligned}
\mathcal{E}_{\text{out}} &= \sigma_\epsilon^2 + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2\right] \\
&= \sigma_\epsilon^2 + \|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\right] + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{F}} - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right].
\end{aligned} \tag{F.6}$$

Then, similar to the proof given for the source task in Section F.1, we get that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{F}}-\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right]= \\
&=\left(\sigma_\epsilon^2+\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2+\mathbb{E}\left[\left\|\boldsymbol{\beta}_{\mathcal{T}}-\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right]\right)\mathbb{E}\left[\mathrm{Tr}\left\{\left(\boldsymbol{\Lambda}_{\mathbf{X}_{\mathcal{F}}}+\alpha\mathbf{I}_p\right)^{-2}\boldsymbol{\Lambda}_{\mathbf{X}_{\mathcal{F}}}\right\}\right] \\
&\quad+\frac{\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2}{p}\alpha^2\mathbb{E}\left[\mathrm{Tr}\left\{\left(\boldsymbol{\Lambda}_{\mathbf{X}_{\mathcal{F}}}+\alpha\mathbf{I}_p\right)^{-2}\right\}\right]. \\
&=\mathbb{E}\left[\sum_{k=1}^{p}\frac{\left(\sigma_\epsilon^2+\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2+\mathbb{E}\left[\left\|\boldsymbol{\beta}_{\mathcal{T}}-\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right]\right)\lambda_{\mathbf{X}_{\mathcal{F}},k}+\frac{\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2}{p}\alpha^2}{\left(\lambda_{\mathbf{X}_{\mathcal{F}},k}+\alpha\right)^2}\right],
\end{aligned} \tag{F.7}$$

where we use the eigendecomposition $\mathbf{X}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}}=\boldsymbol{\Phi}_{\mathbf{X}_{\mathcal{F}}}\boldsymbol{\Lambda}_{\mathbf{X}_{\mathcal{F}}}\boldsymbol{\Phi}_{\mathbf{X}_{\mathcal{F}}}^T$ where $\boldsymbol{\Phi}_{\mathbf{X}_{\mathcal{F}}}$ is a $p\times p$ orthonormal matrix and $\boldsymbol{\Lambda}_{\mathbf{X}_{\mathcal{F}}}\triangleq\mathrm{diag}\{\lambda_{\mathbf{X}_{\mathcal{F}},1},\ldots,\lambda_{\mathbf{X}_{\mathcal{F}},p}\}$ is the $p\times p$ diagonal matrix of the eigenvalues of $\mathbf{X}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}}$.

Here, the optimal tuning is obtained by the α value that provides $\frac{\partial\mathcal{E}_{\text{out}}}{\partial\alpha}=0$. Then, (F.6) and (F.7) imply that the optimal tuning is given by

$$\alpha=\frac{p\left(\sigma_\epsilon^2+\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2+\mathbb{E}\left[\left\|\boldsymbol{\beta}_{\mathcal{T}}-\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right]\right)}{\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2}. \tag{F.8}$$

In our experiments we assume that only $\|\boldsymbol{\beta}\|_2^2$, $\|\boldsymbol{\beta}-\boldsymbol{\theta}\|_2^2$, σ_ϵ, p, t, and d are known; thus, we use the approximations $\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2\approx\frac{p}{d}\|\boldsymbol{\beta}\|_2^2$, $\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2\approx\left(1-\frac{p+t}{d}\right)\|\boldsymbol{\beta}\|_2^2$, and

$$\mathbb{E}\left[\left\|\boldsymbol{\beta}_{\mathcal{T}}-\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right]\approx\frac{t}{d}\mathbb{E}\left[\|\boldsymbol{\beta}-\boldsymbol{\theta}\|_2^2\right]=\frac{t}{d}\mathbb{E}\left[\|\boldsymbol{\beta}-(\mathbf{H}\boldsymbol{\beta}+\boldsymbol{\eta})\|_2^2\right]=\frac{t}{d}\left\|\left(\mathbf{I}_d-\mathbf{H}\right)\boldsymbol{\beta}\right\|_2^2+t\sigma_\eta^2,$$

to approximate (F.8) as $\alpha=\frac{d\sigma_\epsilon^2}{\|\boldsymbol{\beta}\|_2^2}+d-p-t+t\frac{\|(\mathbf{I}_d-\mathbf{H})\boldsymbol{\beta}\|_2^2+d\sigma_\eta^2}{\|\boldsymbol{\beta}\|_2^2}$ in our experiments.

REFERENCES

[1] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, *Benign overfitting in linear regression*, Proc. Natl. Acad. Sci. USA, 117 (2020), pp. 30063–30070.

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

[3] M. Belkin, D. Hsu, and J. Xu, *Two models of double descent for weak features*, SIAM J. Math. Data Sci., 2 (2020), pp. 1167–1180.

[4] Y. Bengio, *Deep learning of representations for unsupervised and transfer learning*, in ICML Workshop on Unsupervised and Transfer Learning, 2012, pp. 17–36.

[5] L. Breiman and D. Freedman, *How many variables should be entered in a regression equation?*, J. Amer. Statist. Assoc., 78 (1983), pp. 131–136.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Rev., 43 (2001), pp. 129–159.

[7] 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), 2020.
[8] O. Dhifallah and Y. M. Lu, *Phase transitions in transfer learning for high-dimensional perceptrons*, Entropy, 23 (2021).
[9] 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*, J. Stat. Mech., 2 (2020), p. 023401.
[10] F. Gerace, L. Saglietti, S. S. Mannelli, A. Saxe, and L. Zdeborová, *Probing transfer learning with a model of synthetic correlated datasets*, Mach. Learn.: Sci. Technol., 3 (2022), p. 015030.
[11] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, Ann. Statist., 50 (2022), pp. 949 – 986.
[12] F. Hiai and D. Petz, *Asymptotic freeness almost everywhere for random matrices*, Acta Sci. Math. Szeged, 66 (2000), pp. 801–826.
[13] S. Kornblith, J. Shlens, and Q. V. Le, *Do better imagenet models transfer better?*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2661–2671.
[14] 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.
[15] Y. Li and Y. Wei, *Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent*, preprint, arXiv:2110.09502, (2021).
[16] M. Long, H. Zhu, J. Wang, and M. I. Jordan, *Deep transfer learning with joint adaptation networks*, in International Conference on Machine Learning (ICML), 2017, pp. 2208–2217.
[17] S. Mei and A. Montanari, *The generalization error of random features regression: Precise asymptotics and the double descent curve*, Comm. Pure Appl. Math., 75 (2022), pp. 667–766.
[18] P. P. Mitra, *Understanding overfitting peaks in generalization error: Analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation*, preprint, arXiv:1906.03667, (2019).
[19] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, *Harmless interpolation of noisy data in regression*, IEEE J. Sel. Areas Inf. Theory, (2020).
[20] D. Obst, B. Ghattas, J. Cugliari, G. Oppenheim, S. Claudel, and Y. Goude, *Transfer learning for linear regression: a statistical test of gain*, preprint, arXiv:2102.09504, (2021).
[21] S. J. Pan and Q. Yang, *A survey on transfer learning*, IEEE Trans. Knowl. Data Eng., 22 (2009), pp. 1345–1359.
[22] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, *Transfusion: Understanding transfer learning for medical imaging*, in Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 3347–3357.
[23] 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.
[24] 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 Trans. Med. Imag., 35 (2016), pp. 1285–1298.
[25] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, *A jamming transition from under-to over-parametrization affects loss landscape and generalization*, J. Phys. A, 52 (2019), p. 474001.
[26] A. M. Tulino and S. Verdú, *Random matrix theory and wireless communications*, Foundations and Trends in Communications and Information Theory, (2004), pp. 1–182.
[27] G. Wang, K. Donhauser, and F. Yang, *Tight bounds for minimum ℓ_1-norm interpolation of noisy data*, in International Conference on Artificial Intelligence and Statistics (AISTATS), 2022, pp. 10572–10602.
[28] 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), 2019, pp. 5095–5104.
[29] 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), 2018, pp. 3712–3722.

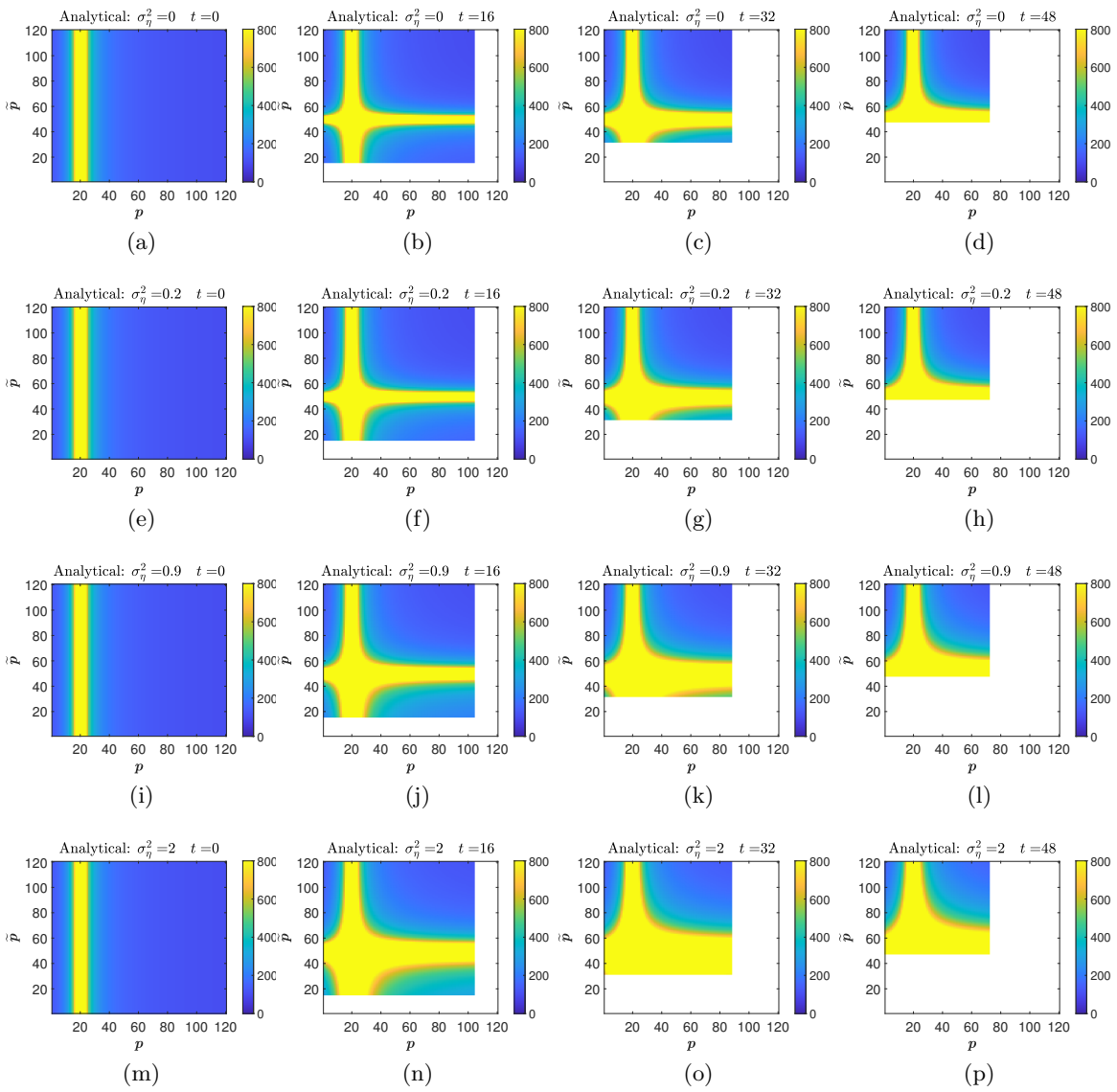


Figure 10: **Analytical** evaluation of the expected out-of-sample squared error of the target task, $\mathbb{E}_{\mathcal{L}}\left[\mathcal{E}_{\text{out}}\right]$, with respect to the number of free parameters $\widetilde{p}$ and p (in the source and target tasks, respectively). Each row of subfigures considers a different case of the relation (2.7) between the source and target tasks in the form of a different noise variance σ_η^2 whereas $\mathbf{H}$ is a local averaging operator with neighborhood size 5 for all. Each column of subfigures considers a different number of transferred parameters t. Here $d = 120$, $\widetilde{n} = 50$, $n = 20$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$. The white regions correspond to $(\widetilde{p}, p)$ settings eliminated by the value of t in the specific subfigure. The yellow-colored areas correspond to values greater or equal to 800. See Fig. 11 for the corresponding empirical evaluation.

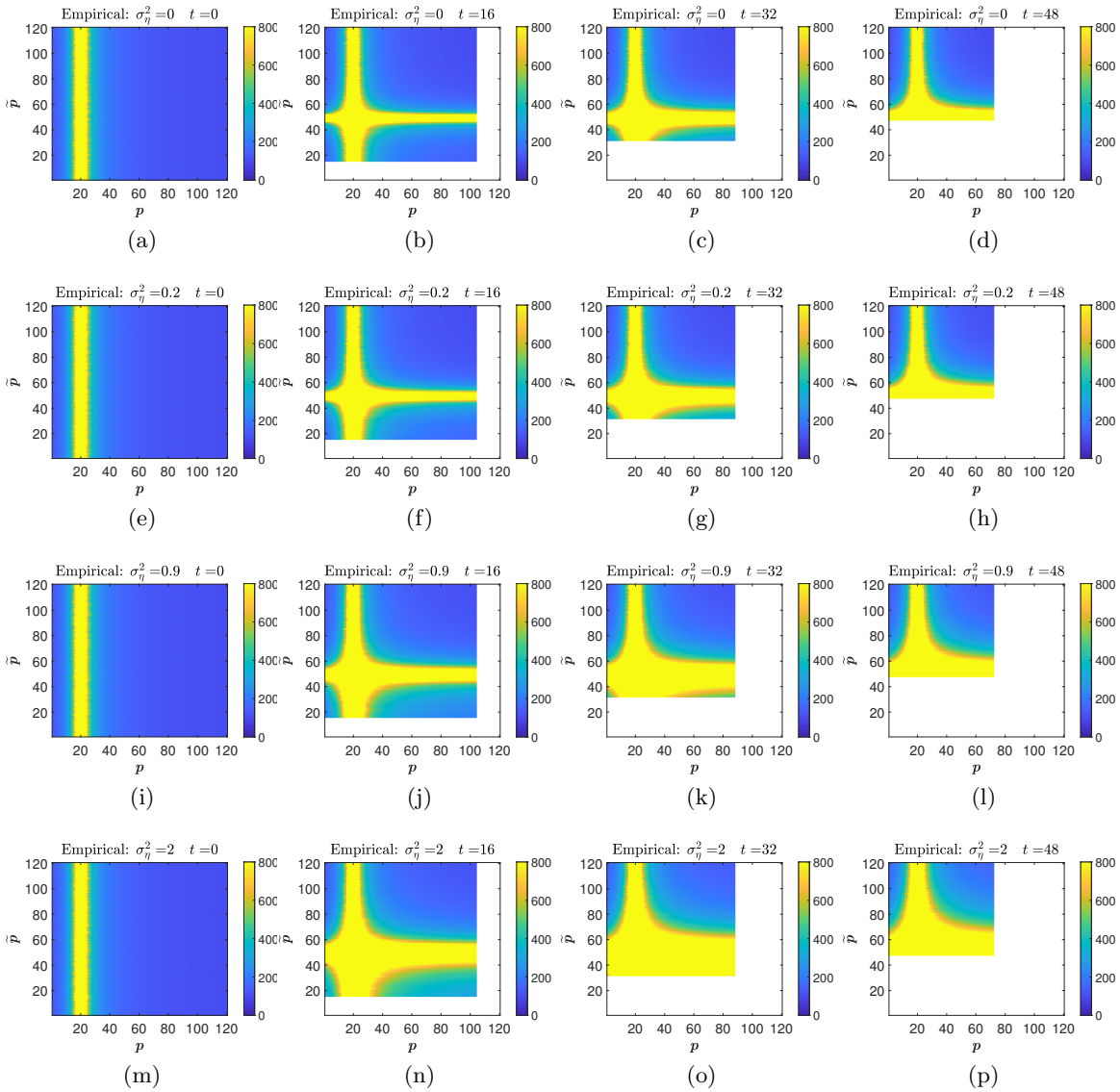
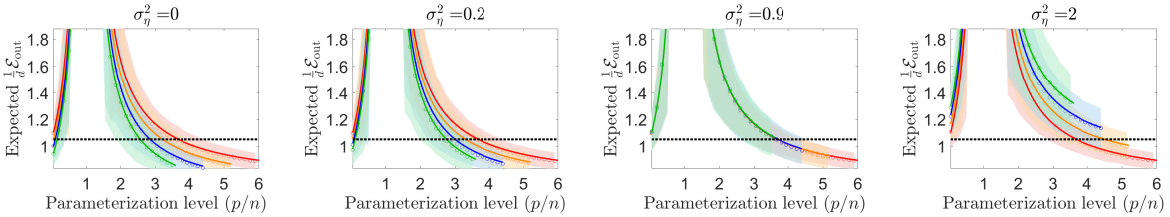
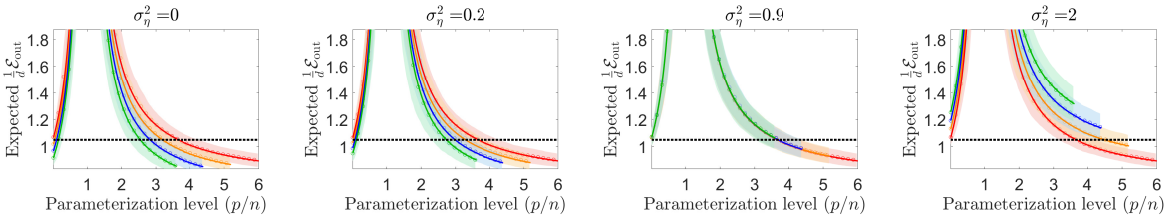


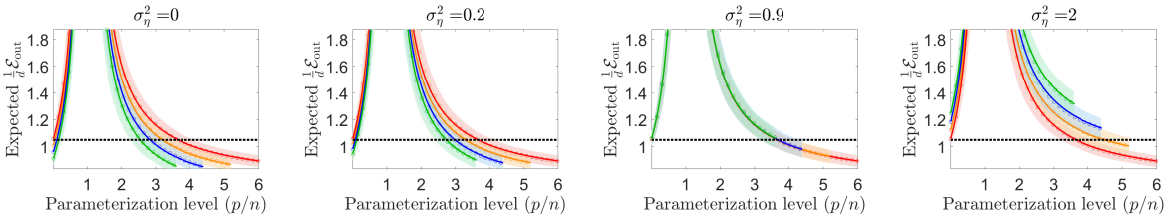
Figure 11: **Empirical** evaluation of the expected out-of-sample squared error of the target task, $\mathbb{E}_{\mathcal{L}}[\mathcal{E}_{\text{out}}]$, with respect to the number of free parameters $\widetilde{p}$ and p (in the source and target tasks, respectively). The presented values obtained by averaging over 250 experiments. The figures here have settings as in the corresponding figures in the analytical evaluation in Fig. 10.



(a) Dimensions as in Fig. 1a (b) Dimensions as in Fig. 1b (c) Dimensions as in Fig. 1c (d) Dimensions as in Fig. 1d



(e) Dimensions 3 times those in Fig. 1a (f) Dimensions 3 times those in Fig. 1b (g) Dimensions 3 times those in Fig. 1c (h) Dimensions 3 times those in Fig. 1d



(i) Dimensions 5 times those in Fig. 1a (j) Dimensions 5 times those in Fig. 1b (k) Dimensions 5 times those in Fig. 1c (l) Dimensions 5 times those in Fig. 1d

Figure 12: Demonstration of error concentration. Each row of subfigures corresponds to the settings of Figures 1a-1d but with a different proportional increase of dimensions and dimension-dependent quantities (see explanation in Section 3.2 in the main text). The empirical standard deviations are denoted as shaded areas in colors corresponding to the on-average error curves (solid lines and markers denote the analytical and empirical evaluations of the expected error, respectively). Lines, markers and areas in red correspond to $t = 0$ (no parameters are transferred); orange corresponds to transferring $t = 16 \times \frac{d}{120}$ parameters; blue corresponds to $t = 32 \times \frac{d}{120}$; green corresponds to $t = 48 \times \frac{d}{120}$. Note that $\frac{d}{120}$ equals to 1, 3, 5, in the first, second, third rows of subfigures, respectively. The axes in this figure are normalized to be dimension-independent.

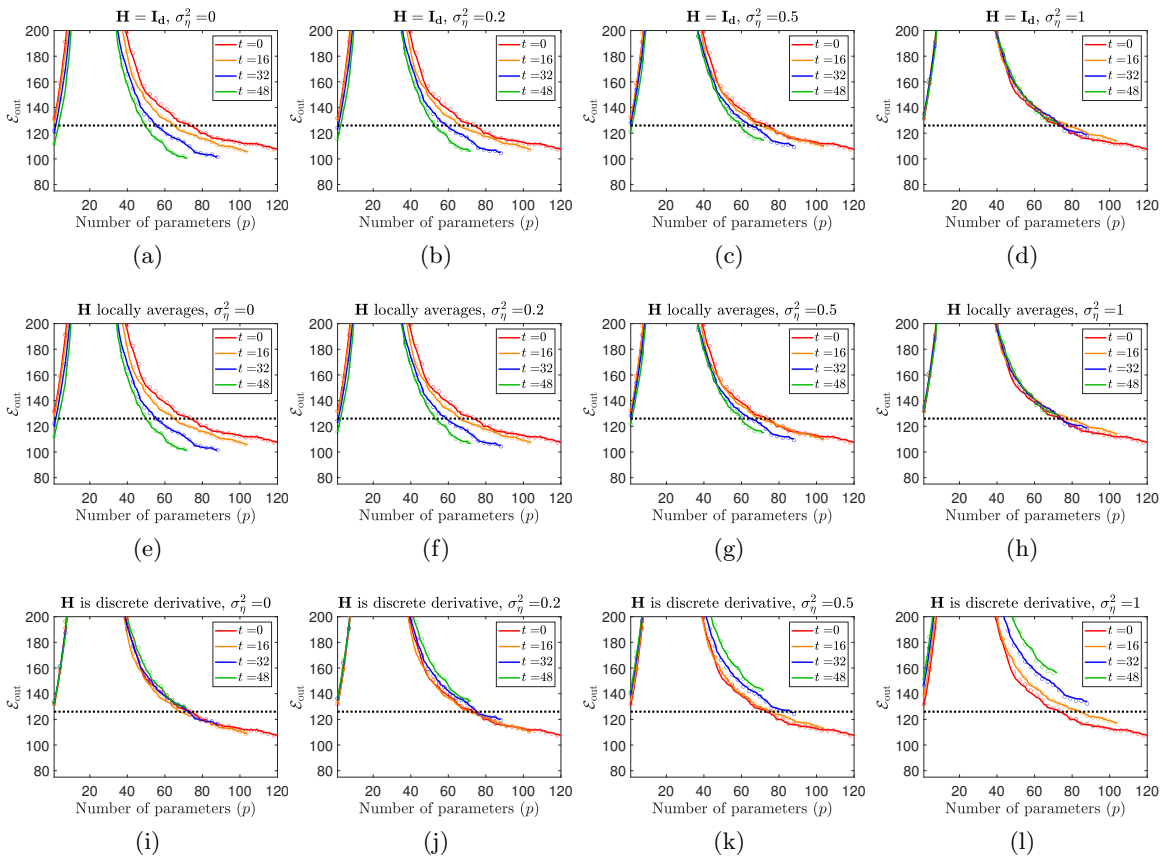


Figure 13: Analytical (solid lines) and empirical (circle markers) values of $\mathcal{E}_{\text{out}}^{(\mathcal{L})}$ for specific, non-random coordinate layouts. **The true solution $\boldsymbol{\beta}$ has linearly-increasing values.** All subfigures use the same sequential evolution of $\mathcal{L}$ with p. Each subfigure considers a different case of the relation (2.7) between the source and target tasks: each column of subfigures has a different σ_η^2 value, and each row of subfigures corresponds to a different linear operator $\mathbf{H}$. The analytical values, computed using Theorem 3.1, are presented using solid-line curves, and the respective empirical results obtained from averaging over 250 experiments are denoted by circle markers. Each curve color refers to a different number of transferred parameters.

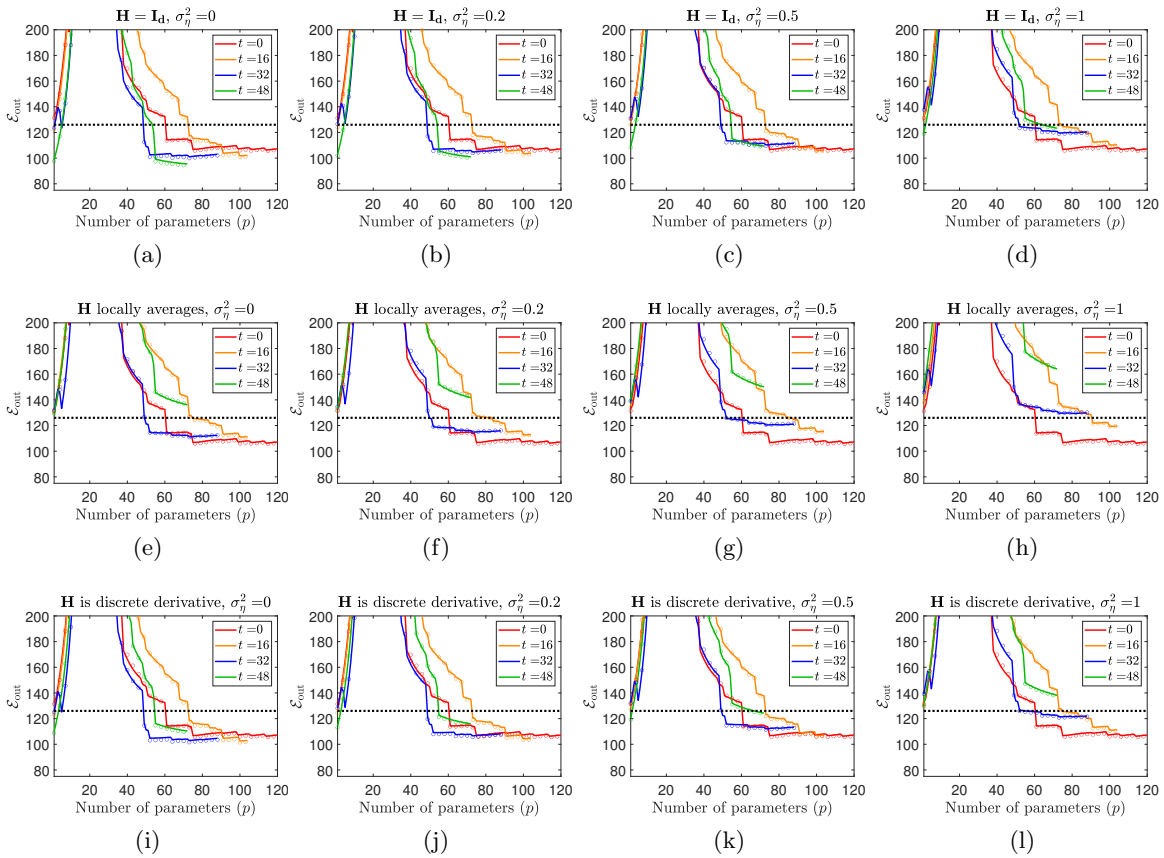


Figure 14: Analytical (solid lines) and empirical (circle markers) values of $\mathcal{E}_{\text{out}}^{(\mathcal{L})}$ for specific, non-random coordinate layouts. **The true solution $\boldsymbol{\beta}$ has a sparse form of values.** All subfigures use the same sequential evolution of $\mathcal{L}$ with p. Each subfigure considers a different case of the relation (2.7) between the source and target tasks: each column of subfigures has a different σ_η^2 value, and each row of subfigures corresponds to a different linear operator $\mathbf{H}$. The analytical values, computed using Theorem 3.1, are presented using solid-line curves, and the respective empirical results obtained from averaging over 250 experiments are denoted by circle markers. Each curve color refers to a different number of transferred parameters.

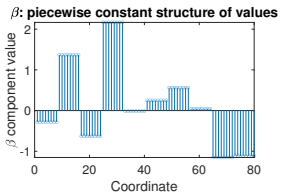
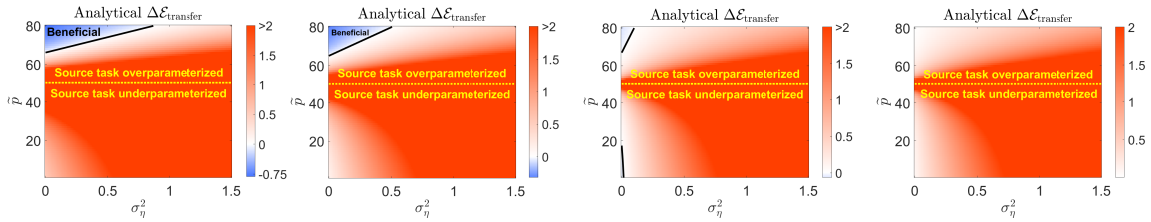
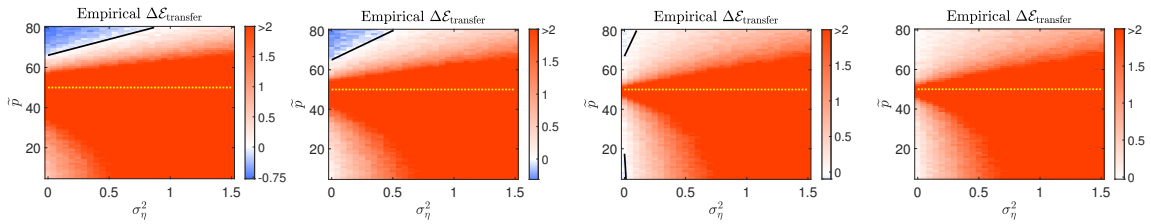


Figure 15: The piecewise-constant structure of $\boldsymbol{\beta}$ that was used in part of the experiments.



(a) **H**: local averaging neighborhood size 3 (b) **H**: local averaging neighborhood size 15 (c) **H**: local averaging neighborhood size 59 (d) **H**: discrete derivative



(e) **H**: local averaging neighborhood size 3 (f) **H**: local averaging neighborhood size 15 (g) **H**: local averaging neighborhood size 59 (h) **H**: discrete derivative

Figure 16: The analytical and empirical values of $\Delta\mathcal{E}_{\text{transfer}}$ defined in Corollary 3.4 (here, normalized by t, namely, the expected error difference due to transfer of a parameter from the source to target task) as a function of $\widetilde{p}$ and σ_η^2. The positive and negative values of $\frac{1}{t}\Delta\mathcal{E}_{\text{transfer}}$ appear in color scales of red and blue, respectively. The regions of negative values (appear in shades of blue) correspond to beneficial transfer of parameters. The positive values were truncated in the value of 2 for the clarity of visualization. Each subfigure corresponds to a different task relation model induced by the definitions of $\mathbf{H}$ as: *(a)-(c),(e)-(g)* local averaging operators with different neighborhood sizes, *(d),(h)* discrete derivative. For all the subfigures, $d = 80$, $\widetilde{n} = 50$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\xi^2 = 0.025 \cdot d$. Here, all the subfigures correspond to a $\boldsymbol{\beta}$ vector with a piecewise-constant form (see Fig. 15).

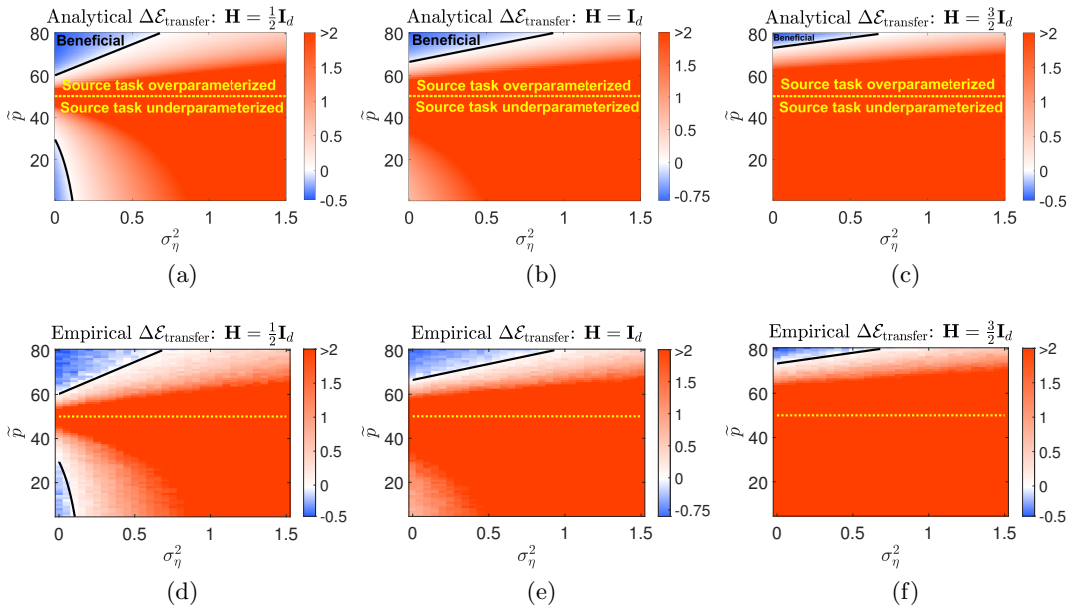


Figure 17: The analytical and empirical values of $\Delta\mathcal{E}_{\text{transfer}}$ defined in Corollary 3.4 (here, normalized by t, namely, the expected error difference due to transfer of a parameter from the source to target task) as a function of $\widetilde{p}$ and σ_η^2. The positive and negative values of $\frac{1}{t}\Delta\mathcal{E}_{\text{transfer}}$ appear in color scales of red and blue, respectively. The regions of negative values (appear in shades of blue) correspond to beneficial transfer of parameters. The positive values were truncated in the value of 2 for the clarity of visualization. Each subfigure corresponds to a different task relation model induced by the definitions of $\mathbf{H}$ as $\mathbf{H} = \frac{1}{2}\mathbf{I}_d$, $\mathbf{H} = \mathbf{I}_d$, and $\mathbf{H} = \frac{3}{2}\mathbf{I}_d$. For all the subfigures, $d = 80$, $\widetilde{n} = 50$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\xi^2 = 0.025 \cdot d$. Here, all the subfigures correspond to a $\boldsymbol{\beta}$ vector with a linear form (see Fig. 4a).

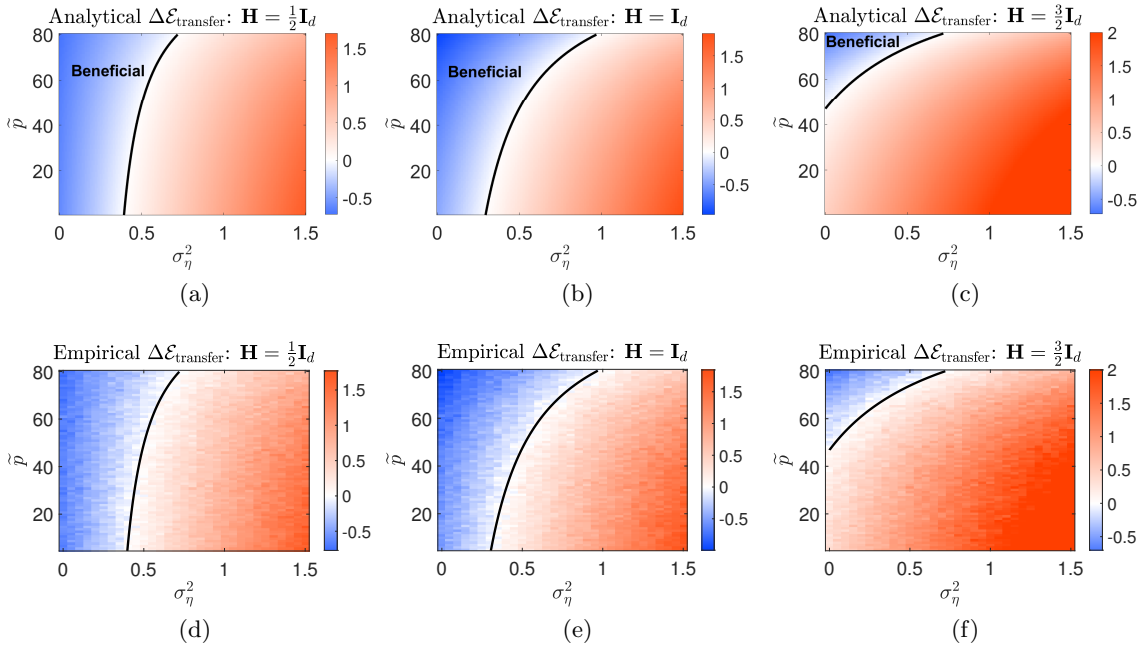
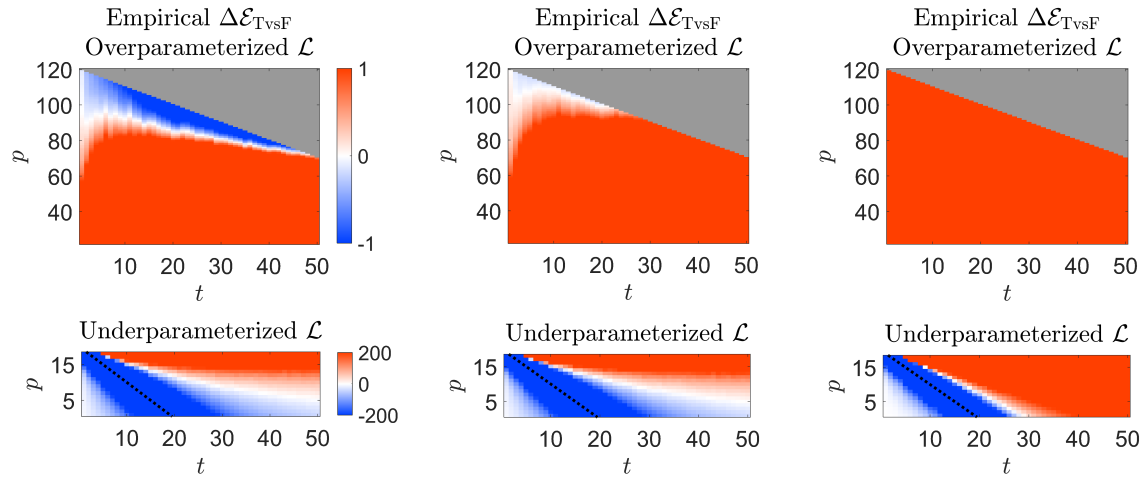


Figure 18: The analytical (top row of subfigures) and empirical (bottom row of subfigures) values of $\Delta\mathcal{E}_{\text{transfer}}$ defined in Corollary 3.4 (here, normalized by t, namely, the expected error difference due to transfer of a parameter from the source to target task) as a function of $\widetilde{p}$ and σ_η^2. The positive and negative values of $\frac{1}{t}\Delta\mathcal{E}_{\text{transfer}}$ appear in color scales of red and blue, respectively. The regions of negative values (appear in shades of blue) correspond to beneficial transfer of parameters. The positive values were truncated in the value of 2 for the clarity of visualization. Each column of subfigures correspond to a different task relation model induced by the definitions of $\mathbf{H}$ as $\mathbf{H} = \frac{1}{2}\mathbf{I}_d$, $\mathbf{H} = \mathbf{I}_d$, and $\mathbf{H} = \frac{3}{2}\mathbf{I}_d$. For all the subfigures, $d = 80$, $\widetilde{n} = 150$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\xi^2 = 0.025 \cdot d$. Here, all the subfigures correspond to a $\boldsymbol{\beta}$ vector with a linear form (see Fig. 4a). Note that the results in this figure are for $\widetilde{n} > d$.

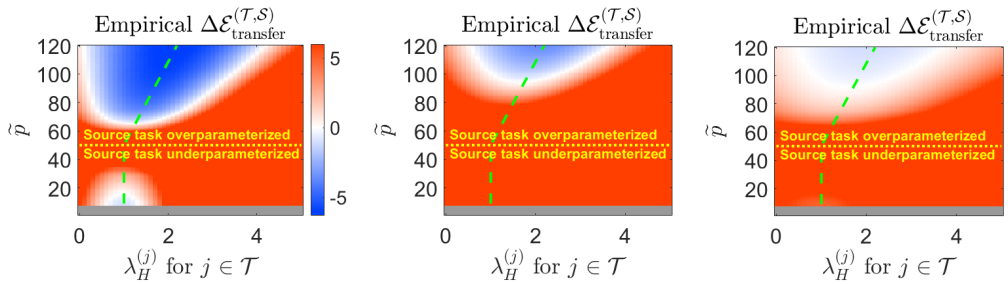


(a) $\mathbf{H}$: local averaging neighborhood size 3

(b) $\mathbf{H}$: local averaging neighborhood size 15

(c) $\mathbf{H} = 5\mathbf{I}_d$

Figure 19: The empirical values of $\Delta\mathcal{E}_{\text{TvsF}}$ (namely, the average error difference due to transfer of an arbitrarily-selected set of t parameters versus setting them as free parameters) as a function of t and p. The settings and visualization of results are as in the corresponding analytical results in Fig. 6.



(a) $\sigma_\eta^2 = 0.2$; $\boldsymbol{\beta}$ has the form in Fig. 7e

(b) $\sigma_\eta^2 = 1$; $\boldsymbol{\beta}$ has the form in Fig. 7e

(c) $\sigma_\eta^2 = 0.2$; $\boldsymbol{\beta}$ has the form in Fig. 7f

Figure 20: The empirical evaluation of $\Delta\mathcal{E}^{(\mathcal{T},\mathcal{S})}_{\text{transfer}}$ as a function of $\widetilde{p}$ and a value that determines the eigenvalues of $\mathbf{H}$ in $\mathcal{T}$. The settings and visualization of results are as in the corresponding analytical results in Fig. 7.