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

**Yehuda Dar** YDAR@RICE.EDU
**Richard G. Baraniuk** RICHB@RICE.EDU
*Electrical and Computer Engineering Department, Rice University*

## 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 correlation 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. Specifically, we show that transferring a specific set of parameters that generalizes well on the respective part of the source task can soften the demand on the task correlation level that is required for successful transfer learning. Moreover, 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.

**Keywords:** Overparameterized learning, linear regression, transfer learning, double descent.

## 1. Introduction

Transfer learning (Pan and Yang, 2009) 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., Bengio, 2012; Shin et al., 2016; Long et al., 2017), 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., Rosenstein et al., 2005; Zamir et al., 2018; Kornblith et al., 2019). Moreover, finding a successful transfer learning setting for deep neural networks was shown by Raghu et al. (2019) 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 (e.g., Lampinen and Ganguli, 2019).

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 (Breiman and Freedman, 1983). Recent empirical studies (Spigler et al., 2018; Geiger et al., 2019; Belkin et al., 2019a) 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., by Belkin et al. (2019b); Hastie et al. (2019); Xu and Hsu (2019); Mei and Montanari (2019); Bartlett et al. (2020); Muthukumar et al. (2020). The next stage is to provide corresponding fundamental understanding to learning problems beyond a single fully-supervised regression problem (see, e.g., the study by Dar et al. (2020) 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. 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 correlation 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 that clarifies the fundamental factors affecting the performance of the overall transfer learning approach. 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.

This paper is organized as follows. In Section 2 we define the transfer learning architecture examined in this paper. In Sections 3-4 we present the analytical and empirical results that characterize the generalization errors of the target task, and outline the cases where transfer of parameters is beneficial. Note that Section 3 studies the on-average generalization error in a simplified setting where transferred parameters are chosen uniformly at random, whereas Section 4 examines the generalization error induced by transfer of a single, specific set of parameters. Section 5 concludes the paper. The Appendices 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 $\mathbf{z}\sim\mathcal{N}\left(\mathbf{0},\mathbf{I}_d\right)$ is connected to a response value $v\in\mathbb{R}$ via the noisy linear model

$$v=\mathbf{z}^T\boldsymbol{\theta}+\xi, \tag{1}$$

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 $\mathbf{Z}\triangleq[\mathbf{z}^{(1)},\dots,\mathbf{z}^{(\widetilde{n})}]^T$ and $\mathbf{v}\triangleq[v^{(1)},\dots,v^{(\widetilde{n})}]^T$ that satisfy the relation $\mathbf{v}=\mathbf{Z}\boldsymbol{\theta}+\boldsymbol{\xi}$ where $\boldsymbol{\xi}\triangleq[\xi^{(1)},\dots,\xi^{(\widetilde{n})}]^T$ is an unknown noise vector that its $i^{\text{th}}$ component $\xi^{(i)}$ participates in the relation $v^{(i)}=\mathbf{z}^{(i),T}\boldsymbol{\theta}+\xi^{(i)}$ underlying the $i^{\text{th}}$ 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 (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)

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

where the second equality stems from the data model in (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 $\mathcal{S}\triangleq\{s_1,\dots,s_{\widetilde{p}}\}$ where $1\le s_1<\dots<s_{\widetilde{p}}\le d$ and the complementary set $\mathcal{S}^{\text{c}}\triangleq\{1,\dots,d\}\setminus\mathcal{S}$ 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 ($k=1,\dots,|\mathcal{S}|$) 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., $\mathbf{Q}_{\mathcal{S}^{\text{c}}}$ for $\mathcal{S}^{\text{c}}$) as denoted by the subscript of $\mathbf{Q}$. We now turn to formulate the *source* task using the linear regression form of

$$\widehat{\boldsymbol{\theta}}=\underset{\mathbf{r}\in\mathbb{R}^d}{\arg\min}\left\|\mathbf{v}-\mathbf{Z}\mathbf{r}\right\|_2^2\quad\text{subject to}\quad\mathbf{Q}_{\mathcal{S}^{\text{c}}}\mathbf{r}=\mathbf{0} \tag{3}$$

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

$$\widehat{\boldsymbol{\theta}} = \mathbf{Q}_{\mathcal{S}}^{T} \mathbf{Z}_{\mathcal{S}}^{+} \mathbf{v} \tag{4}$$

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 $i^{\text{th}}$ 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. Moreover, $\widehat{\boldsymbol{\theta}}$ 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 (3) was not explicit in previous studies of non-asymptotic settings (e.g., Breiman and Freedman, 1983; Belkin et al., 2019b), the solution in (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 Breiman and Freedman (1983); Belkin et al. (2019b) 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

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon \tag{5}$$

where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ 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 (1) via

$$\boldsymbol{\theta} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\eta} \tag{6}$$

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 $\boldsymbol{\eta}$, $\mathbf{x}$, $\epsilon$, $\mathbf{z}$ and $\xi$ 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 $i^{\text{th}}$ component $\epsilon^{(i)}$ is involved in the connection $y^{(i)} = \mathbf{x}^{(i),T}\boldsymbol{\beta} + \epsilon^{(i)}$ underlying the $i^{\text{th}}$ 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 (5) 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,

$$\mathcal{E}_{\text{out}} \triangleq \mathbb{E}\left\{\left(\widehat{y} - y^{(\text{test})}\right)^2\right\} = \sigma_{\epsilon}^2 + \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2\right\} \tag{7}$$

is minimized, where $\widehat{y} = \mathbf{x}^{(\text{test}),T}\widehat{\boldsymbol{\beta}}$, and the second equality stems from the data model in (5).

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, $\mathcal{T} \subset \mathcal{S}$ and $t \in \{0, \dots, \widetilde{p}\}$. 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 $k^{\text{th}}$ components of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$, respectively.
- $\ell$ 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

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

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 *precomputed* 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 (3), followed by a single computation of the target task (8) that produces the eventual estimate of interest $\widehat{\boldsymbol{\beta}}$ using the given $\widehat{\boldsymbol{\theta}}$. The min-norm solution of the target task in (8) is (see details in Appendix A.3)

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

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 $\mathbf{X}_{\mathcal{F}} \triangleq \mathbf{X}\mathbf{Q}_{\mathcal{F}}^T$. Note that the desired layout is indeed implemented by the $\widehat{\boldsymbol{\beta}}$ form in (9): 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 (8) 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 (8)–(9).

## 3. Transfer of Random Sets of Parameters

To analytically study the generalization error of the target task we consider, in this section, the *overall layout of coordinate subsets* $\mathcal{L} \triangleq \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ as a random structure that lets us to formulate the expected value (with respect to $\mathcal{L}$) of the generalization error of interest. The simplified settings of this section provide useful insights towards Section 4 where we analyze transfer of a specific set of parameters and formulate the generalization error for a specific layout $\mathcal{L}$ (i.e., there is no expectation over $\mathcal{L}$ in the formulations in Section 4).

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

Recall the relation between the two tasks as provided in (6) and let us denote $\boldsymbol{\beta}^{(\mathbf{H})} \triangleq \mathbf{H}\boldsymbol{\beta}$. Assume that $\boldsymbol{\beta} \neq \mathbf{0}$. The following definitions emphasize crucial aspects in the examined transfer learning framework. The *source task energy* is $\kappa \triangleq \mathbb{E}_{\boldsymbol{\eta}}\left\{ \|\boldsymbol{\theta}\|_2^2 \right\} = \left\| \boldsymbol{\beta}^{(\mathbf{H})} \right\|_2^2 + d\sigma_\eta^2$. Assume that $\boldsymbol{\theta}$ is not deterministically degenerated into a zero vector, hence, $\kappa \neq 0$. The *normalized task correlation* between the two tasks is $\rho \triangleq \frac{\langle \boldsymbol{\beta}^{(\mathbf{H})}, \boldsymbol{\beta} \rangle}{\kappa}$. Let us characterize the expected out-of-sample error of the target task with respect to transfer of uniformly-distributed sets of parameters.

**Theorem 2** *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*

$$\mathbb{E}_{\mathcal{L}}\left\{ \mathcal{E}_{\text{out}} \right\} = \begin{cases} \frac{n-1}{n-p-1}\left( (1-\frac{p}{d}) \|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2 + t \cdot \Delta\mathcal{E}_{\text{transfer}} \right) & \text{for } p \leq n-2, \\ \infty & \text{for } n-1 \leq p \leq n+1, \\ \frac{p-1}{p-n-1}\left( (1-\frac{p}{d}) \|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2 + t \cdot \Delta\mathcal{E}_{\text{transfer}} \right) + \frac{p-n}{d}\|\boldsymbol{\beta}\|_2^2 & \text{for } p \geq n+2, \end{cases} \tag{10}$$

*where*

$$\Delta\mathcal{E}_{\text{transfer}} = \frac{1}{\widetilde{p}} \cdot \left( \mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{ \widetilde{\mathcal{E}}_{\text{out}} \right\} - \sigma_\xi^2 - \kappa \right) - 2\frac{\kappa}{d}(\rho - 1) \times \begin{cases} 1 & \text{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n} \end{cases} \tag{11}$$

*is the expected error difference introduced by each constrained parameter that is transferred from the source task instead of being set to zero. Recall that $\widetilde{\mathcal{E}}_{\text{out}}$ and $\mathcal{E}_{\text{out}}$ are the out-of-sample errors of the source and target tasks, respectively.*

The last theorem is proved using non-asymptotic properties of Wishart matrices (see Appendix B). Negative values of $\Delta\mathcal{E}_{\text{transfer}}$ imply beneficial transfer learning and this occurs, for example, when the *task correlation* $\rho$ is positive and sufficiently large with respect to the *generalization error level in the source task* $\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{ \widetilde{\mathcal{E}}_{\text{out}} \right\}$ that should be sufficiently low (see Corollary 4 below). Note that the out-of-sample error formulation in (10) depends on the parameterization level of the *target* task (i.e., the $p, n$ pair), whereas (11) shows that the error difference $\Delta\mathcal{E}_{\text{transfer}}$ depends on the expected generalization error of the *source* task $\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{ \widetilde{\mathcal{E}}_{\text{out}} \right\}$. Using the explicit expression (see details in Appendix A.2)

$$\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{ \widetilde{\mathcal{E}}_{\text{out}} \right\} = \begin{cases} \frac{\widetilde{n}-1}{\widetilde{n}-\widetilde{p}-1}\left( \left(1-\frac{\widetilde{p}}{d}\right)\kappa + \sigma_\xi^2 \right) & \text{for } \widetilde{p} \leq \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{\widetilde{p}-1}{\widetilde{p}-\widetilde{n}-1}\left( \left(1-\frac{\widetilde{p}}{d}\right)\kappa + \sigma_\xi^2 \right) + \frac{\widetilde{p}-\widetilde{n}}{d}\kappa & \text{for } \widetilde{p} \geq \widetilde{n}+2. \end{cases} \tag{12}$$

we provide the next formula for $\Delta\mathcal{E}_{\text{transfer}}$ that depicts the detailed dependency on the *parameterization level* of the *source* task (i.e., the $\widetilde{p}, \widetilde{n}$ pair). See proof in Appendix B.4.

**Corollary 3** *The expected error difference term $\Delta\mathcal{E}_{\text{transfer}}$ from (11) can be explicitly written as*

$$\Delta\mathcal{E}_{\text{transfer}} = \frac{\kappa}{d} \times \begin{cases} 1 - 2\rho + \frac{d-\widetilde{p}+d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p} \leq \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}}\left( 1 - 2\rho + \frac{d-\widetilde{p}+d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{p}-\widetilde{n}-1} \right) & \text{for } \widetilde{p} \geq \widetilde{n}+2. \end{cases} \tag{13}$$

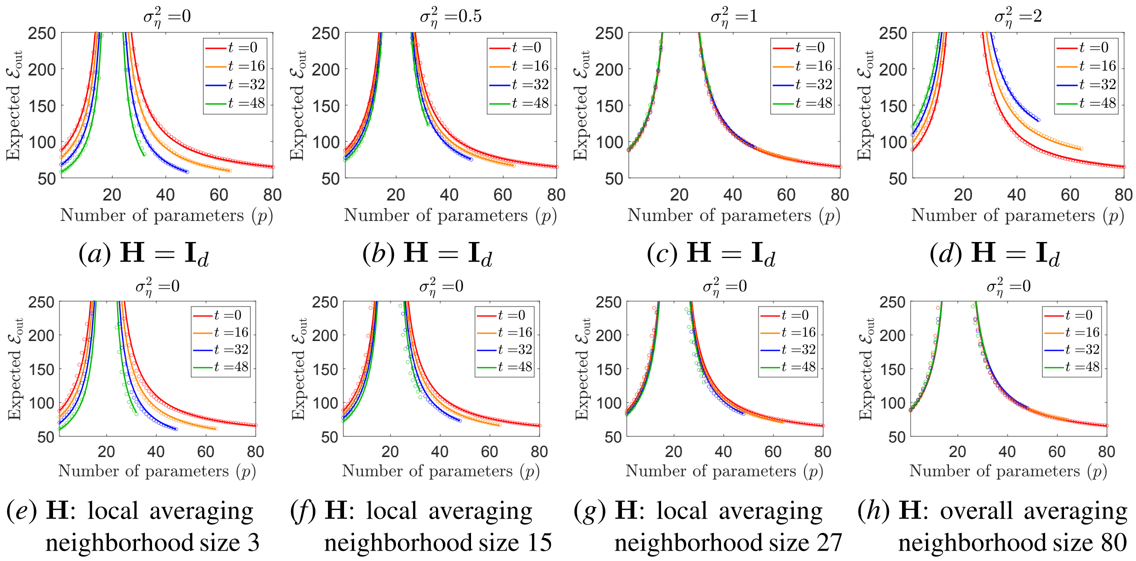
Figure 1 presents the curves of $\mathbb{E}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}\right\}$ with respect to the number of free parameters $p$ in the target task, whereas the source task has $\widetilde{p} = d$ free parameters. In Fig. 1, the solid-line curves correspond to analytical values induced by Theorem 2 and Corollary 3 and the respective empirically computed values are denoted by circles (all the presented results are for $d = 80$, $n = 20$, $\widetilde{n} = 50$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$. See additional details in Appendix D). 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 (this was also the case in settings examined by Belkin et al. (2019b); Dar et al. (2020)).

We can interpret the results in Figure 1 as examples for important cases of transfer learning settings, where each subfigure considers a different task relation with a different pair of noise level $\sigma_\eta^2$ and operator $\mathbf{H}$. Fig. 1(*a*) corresponds to transfer learning between two *identical 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. 1(*b*),1(*e*),1(*f*),1(*g*) correspond to transfer learning between two *related tasks* (although not identical), hence, transfer learning is still *beneficial, but less* than in the former case of identical tasks. Figs. 1(*c*),1(*h*) correspond to transfer learning between two *unrelated tasks* (although not extremely far), hence, transfer learning is *useless*, but not harmful (i.e., for a given $p$, the number of transferred parameters $t$ does not affect the out-of-sample error). Fig. 1(*d*) corresponds 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).

The examined transfer learning approach is simple, which is useful for the purpose of analytical study. Therefore, it is important to compare the examined method with a trivial solution such as the null estimate, i.e., the estimate $\widehat{\boldsymbol{\beta}}$ is all zeros. Note that the out-of-sample error of the null estimate equals to $\|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2$, which equals $d + \sigma_\epsilon^2 = 84$ in all the experiments in this paper (this is due to having $d = 80$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$). Hence, 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 $t > 0$) outperforms the null estimate.

By considering the generalization error formula from Theorem 2 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. 2 and its extended version Fig. 5 where each subfigure is for a different pair of $t$ and $\sigma_\eta^2$. 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 $\widetilde{p} \in \{t, \dots, d\}$ and $p \in \{0, \dots, d - t\}$, 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. 5 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. 2(*d*)).

Conceptually, we can observe a *tradeoff between transfer learning and overparamterized learning*: an increased transfer of parameters limits the level of overparameterization applicable in the target task and, in turn, this limits the overall potential gains from the transfer learning. Yet, when



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

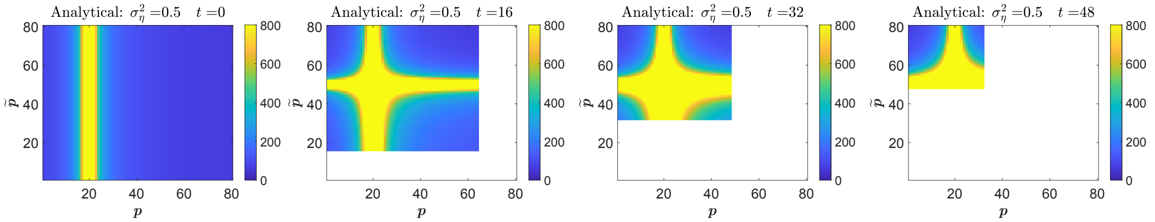
($e$) $\mathbf{H}$: local averaging neighborhood size 3 ($f$) $\mathbf{H}$: local averaging neighborhood size 15 ($g$) $\mathbf{H}$: local averaging neighborhood size 27 ($h$) $\mathbf{H}$: overall averaging neighborhood size 80

Figure 1: The expected generalization error of the target task, $\mathbb{E}_{\mathcal{L}}\left\{\mathcal{E}_{\rm out}\right\}$, with respect to the number of free parameters (in the target task). The analytical values, induced from Theorem 2, are presented using solid-line curves, and the respective empirical results obtained from averaging over 250 experiments are denoted by circle markers. Each subfigure considers a different case of the source-target task relation (6) with a different pair of $\sigma_\eta^2$ and $\mathbf{H}$. The second row of subfigures correspond 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.

the source task is *sufficiently related* to the target task (see, e.g., Figs. 1($a$), 1($b$)), the parameter transfer compensates, at least partially, for an insufficient number of free parameters (in the target task). The last claim is evident in Figures 1($a$), 1($b$) 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. 1($b$) the error achieved by $p = 60$ free parameters and no parameter transfer can be also achieved using $p = 48$ free parameters and $t = 32$ parameters transferred from the source task. In Appendix C we elaborate on two special cases that are induced by setting $t = 0$ or $p = 0$ in the result of Theorem 2.

Let us return to the general case of Theorem 2 and examine the expected generalization error for a given number of free parameters $p$ in the *target* task. We now formulate an analytical condition on the correlation $\rho$ between the two tasks which is required for having a useful parameter transfer.

**Corollary 4** *The term $\Delta\mathcal{E}_{\rm transfer}$, which quantifies the expected error difference due to each parameter being transferred instead of set to zero, satisfies $\Delta\mathcal{E}_{\rm transfer} < 0$ (i.e., **parameter transfer is beneficial** for $p \notin \{n-1, n, n+1\}$) if the correlation between the two tasks is **sufficiently high***



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

Figure 2: Analytical evaluation of 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 $\widetilde{p}$ 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 subfigures are for $\sigma_\eta^2 = 0.5$ and $\mathbf{H} = \mathbf{I}_d$. See Fig. 5 for settings with different values of $\sigma_\eta^2$. See Fig. 6 for the corresponding empirical evaluation.

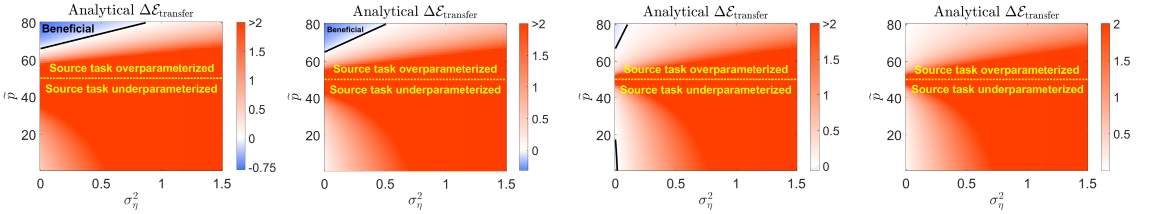
*such that*

$$\rho > 1 + \frac{d}{2\widetilde{p}}\left(\frac{\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\} - \sigma_\xi^2}{\kappa} - 1\right) \times \left(\begin{cases} 1 & \textit{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{p}}{\widetilde{n}} & \textit{for } \widetilde{p} > \widetilde{n}. \end{cases}\right) \tag{14}$$

*Otherwise,* $\Delta\mathcal{E}_{\text{transfer}} \geq 0$ *(i.e., parameter transfer is not beneficial).*

The last corollary is simply proved by using the formula for $\Delta\mathcal{E}_{\text{transfer}}$ from (11) and reorganizing the respective inequality $\Delta\mathcal{E}_{\text{transfer}} < 0$. The last result also emphasizes that the transfer learning performance depends on the interplay between the quality of the solution of the source task (reflected by $\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}$) and the correlation between the two tasks. A source task that generalizes well is important to good transfer learning performance that induces good generalization at the target task.

The number of free parameters $\widetilde{p}$ is a prominent factor that determines the *source* task generalization ability. Accordingly, in Appendix E.1 we use the detailed formulation of $\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}$ to translate the condition (14) to explicitly reflect the interplay between $\widetilde{p}$ and $\rho$. The detailed formulation is provided in Corollary 10 in Appendix E.1) and its main lesson is that *parameter transfer is beneficial* for $p \notin \{n-1, n, n+1\}$ if the *source* task is *sufficiently overparameterized* or sufficiently underparameterized with respect to the correlation between the tasks. This result is in accordance with the shapes of the generalization error curves in our settings that indeed have improved generalization performance at the two extreme cases of under and over parameterization. The results in Fig. 3 show that there are less settings of beneficial transfer learning as the source and target tasks are less related, which is demonstrated in Fig. 3 by higher noise levels $\sigma_\eta^2$ and/or when $\mathbf{H}$ is an averaging operator over a larger neighborhood (e.g., Fig. 3(*c*)), or in the case where $\mathbf{H}$ is a discrete derivative operator (Fig. 3(*d*)). The analytical thresholds for useful transfer learning (Corollaries 4,10) are demonstrated by the black lines in Fig. 3. Our analytical thresholds excellently match the regions where the *empirical* settings yield useful parameter transfer (i.e., where $\Delta\mathcal{E}_{\text{transfer}} < 0$ is



(*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

Figure 3: The analytical values of $\Delta\mathcal{E}_{\text{transfer}}$ defined in Theorem 2 (namely, the expected error difference due to transfer of a parameter from the source to target task) as a function of $\widetilde{p}$ and $\sigma_\eta^2$. 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. The positive values were truncated in the value of 2 for the clarity of visualization. The solid black lines (in all subfigures) denote the analytical thresholds for useful transfer learning as implied by Corollary 10. Each subfigure corresponds to a different task relation model induced by the definitions of **H** as: *(a)-(c)* local averaging operators with different neighborhood sizes, *(d)* discrete derivative. For all the subfigures, $d = 80$, $\widetilde{n} = 50$, $\sigma_\xi^2 = 0.025 \cdot d$, $\|\boldsymbol{\beta}\|_2^2 = d$ where $\boldsymbol{\beta}$ components have a piecewise-constant form (see Fig. 7). See corresponding empirical results in Fig. 8.

empirically satisfied). Additional analytical and empirical results, as well as details on the empirical computation of $\Delta\mathcal{E}_{\text{transfer}}$ are available in Appendix E.2.

## 4. Transfer of a Specific Set of Parameters

Equipped with the fundamental analytical understanding of the key principles formulated in the former section for the case of uniformly-distributed coordinate layout, we now proceed to the analysis of settings that consider a *specific non-random* layout of coordinates $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$. First, we define the next quantities with respect to the *specific* coordinate layout used: The *energy of the transferred parameters* is $\kappa_{\mathcal{T}} \triangleq \left\|\boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})}\right\|_2^2 + t\sigma_\eta^2$. The *normalized task correlation in* $\mathcal{T}$ between the two tasks is defined as $\rho_{\mathcal{T}} \triangleq \frac{\langle \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})}, \boldsymbol{\beta}_{\mathcal{T}} \rangle}{\kappa_{\mathcal{T}}}$ for $t > 0$. The following theorem formulates the generalization error of the target task that is solved using a specific set of transferred parameters indicated by the coordinates in $\mathcal{T}$. See proof in Appendix F.

**Theorem 5** *Let $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ be a **specific, non-random** coordinate subset layout. Then, the out-of-sample error of the target task has the form of*

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

*where*

$$\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = \mathbb{E}\left\{ \left\| \widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}} \right\|_2^2 \right\} - \kappa_{\mathcal{T}} \times \left( 1 + 2\left(\rho_{\mathcal{T}} - 1\right) \times \left( \begin{cases} 1 & \text{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n} \end{cases} \right) \right) \tag{15}$$

*for $t > 0$, and $\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = 0$ for $t = 0$. Here $\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ is the error difference introduced by transferring from the source task the parameters specified in $\mathcal{T}$ instead of setting them to zero.*

The formulation of the error difference term in (15) demonstrates the interplay between the generalization performance in the source task (reflected by $\mathbb{E}\left\{ \left\| \widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}} \right\|_2^2 \right\}$), and the correlation between the source and target tasks (reflected by $\rho_{\mathcal{T}}$) — however, unlike Theorem 2, the current case is affected by source generalization ability and task relation *only in the subset of transferred parameters* $\mathcal{T}$. The following corollary explicitly formulates the error difference term due to transfer learning (see proof in Appendix F.3).

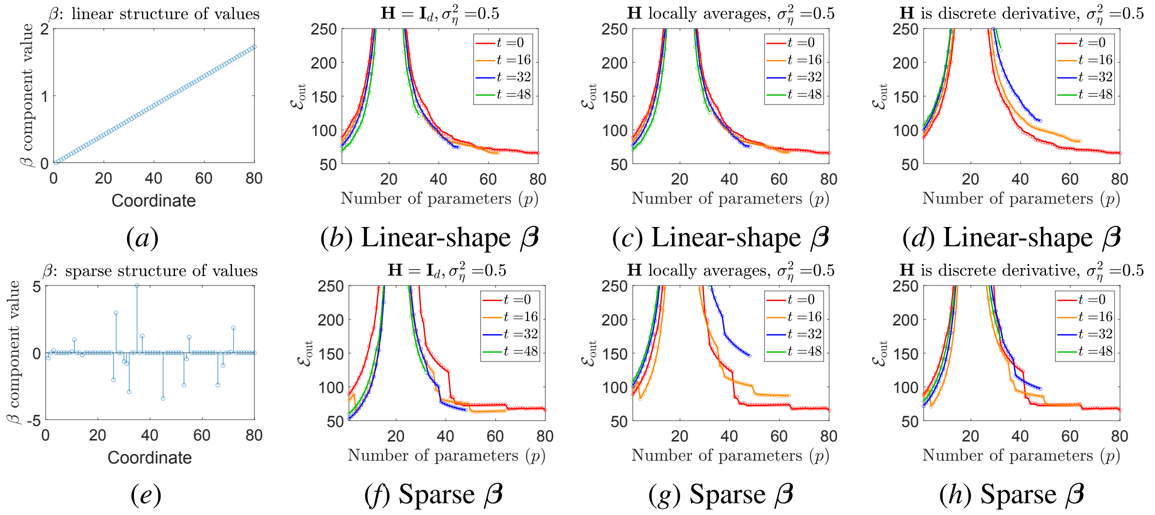
**Corollary 6** *The error difference term $\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ from (15) can be written as*

$$\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = \kappa_{\mathcal{T}} \times \begin{cases} 1 - 2\rho_{\mathcal{T}} + t\frac{\zeta_{\mathcal{S}^c} + \sigma_\xi^2}{(\widetilde{n} - \widetilde{p} - 1)\kappa_{\mathcal{T}}} & \text{for } 1 \leq \widetilde{p} \leq \widetilde{n} - 2, \\ \infty & \text{for } \widetilde{n} - 1 \leq \widetilde{p} \leq \widetilde{n} + 1, \\ \frac{\widetilde{n}}{\widetilde{p}} \left( \frac{(\widetilde{p}^2 - \widetilde{n}\widetilde{p})\psi_{\mathcal{T}} + \widetilde{n}\widetilde{p} - 1}{\widetilde{p}^2 - 1} - 2\rho_{\mathcal{T}} + t\frac{\zeta_{\mathcal{S}^c} + \sigma_\xi^2}{(\widetilde{p} - \widetilde{n} - 1)\kappa_{\mathcal{T}}} \right) & \text{for } \widetilde{p} \geq \widetilde{n} + 2. \end{cases}$$

*Here we used the following definitions. The energy of the zeroed parameters in the source task is $\zeta_{\mathcal{S}^c} \triangleq \left\| \boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} \right\|_2^2 + (d - \widetilde{p})\sigma_\eta^2$. The possibly-transferred to actually-transferred energy ratio of the source task is defined as $\psi_{\mathcal{T}} \triangleq \frac{t}{\widetilde{p}} \cdot \frac{\left\| \boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})} \right\|_2^2 + \widetilde{p}\sigma_\eta^2}{\kappa_{\mathcal{T}}}$ for $t > 0$ and $\widetilde{p} > 0$. We can also interpret $\psi_{\mathcal{T}}^{-1}$ as the utilization of the transfer of $t$ out of the $\widetilde{p}$ parameters of the source task.*

The formulation of error difference due to transfer learning in Corollary 6 shows that the benefits from transfer learning increase for greater positive value of *task correlation in the transferred coordinates* $\rho_{\mathcal{T}}$. Moreover, *higher utilization* $\psi_{\mathcal{T}}^{-1}$ promotes benefits from the transfer learning process.

Figures 4,11,12 show 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. The excellent fit of the analytical results to the empirical values is evident. The effect of the specific coordinate layout utilized is clearly visible by the less-smooth curves (compared to the on-average results for random coordinate layouts in Fig. 1). We examine two different cases for the true $\boldsymbol{\beta}$ (a linearly-increasing (Fig. 4(*a*)) and a sparse (Fig. 4(*e*)) layout of values, with the same norm) and the resulting error curves significantly differ despite the use of the same sequential construction of the coordinate layouts with respect to $p$ (e.g., compare Figs. 4(*f*) and 4(*b*)). The linear 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, also see the complete set of results in Figs. 11-12 in Appendix G). The results clearly show that the interplay among the structures of $\mathbf{H}$, $\boldsymbol{\beta}$, and the coordinate layout significantly affects the generalization performance. Theorem 5 yields the next corollary, which is a direct consequence of (15).



(*a*)

(*b*) Linear-shape $\boldsymbol{\beta}$

(*c*) Linear-shape $\boldsymbol{\beta}$

(*d*) Linear-shape $\boldsymbol{\beta}$

(*e*)

(*f*) Sparse $\boldsymbol{\beta}$

(*g*) Sparse $\boldsymbol{\beta}$

(*h*) Sparse $\boldsymbol{\beta}$

Figure 4: Analytical (solid lines) and empirical (circle markers) values of $\mathcal{E}_{\rm out}^{(\mathcal{L})}$ for specific, non-random coordinate layouts. All subfigures use the same sequential evolution of $\mathcal{L}$ with $p$. Here $\sigma_\eta^2 = 0.5$. See Figures 11-12 for the complete set of results.

**Corollary 7** *Let $\mathcal{S}$ be given. Then, the parameter transfer induced by a specific $\mathcal{T} \subset \mathcal{S}$ is **beneficial** for $p \notin \{n-1, n, n+1\}$ if $\Delta\mathcal{E}_{\rm transfer}^{(\mathcal{T},\mathcal{S})} < 0$, which implies that the correlation in the subset $\mathcal{T}$ between the two tasks should be **sufficiently high** such that*

$$\rho_{\mathcal{T}} > 1 + \frac{1}{2}\left(\frac{1}{\kappa_{\mathcal{T}}}\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} - 1\right) \times \left(\begin{cases} 1 & \textit{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{p}}{\widetilde{n}} & \textit{for } \widetilde{p} > \widetilde{n}. \end{cases}\right) \tag{16}$$

*Otherwise, this specific parameter transfer is not beneficial over zeroing the parameters (i.e., omitting the input features) corresponding to $\mathcal{T}$.*

The last corollary extends Corollary 4 by emphasizing that transferring a specific set of parameters that generalizes well on the respective part of the source task can soften the demand on the task correlation level that is required for successful transfer learning.

Our results also exhibit that a specific 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 (e.g., Fig. 4(*g*) where the orange and red colored curves do not consistently maintain their relative vertical order at the overparameterized range of solutions). Our results exemplify that transfer learning settings are fragile and that finding a successful setting is a delicate engineering task. Hence, our theory qualitatively explains similar practical behavior in deep neural networks (Raghu et al., 2019).

## 5. 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. Our results demonstrate an inherent tradeoff between overparameterized learning and transfer learning, namely, a more extensive transfer of parameters limits the maximal degree of overparameterization in the target task and its potential benefits — nevertheless, in proper settings (e.g., when the source and target tasks are sufficiently related) transfer learning can be a substitute for an increased overparameterization. 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 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 explicit regularization together with transfer learning, and well-specified settings where the task relation model is known 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.

## References

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

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

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

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

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

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.

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

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

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

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

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.

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

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

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

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

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

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.

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

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. *arXiv preprint arXiv:1810.09665*, 2018.

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

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

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

## Appendices

The following appendices support the main paper as follows. Appendix A provides additional details on the mathematical developments leading to the results in Section 2 of the main paper. In Appendix B we present the proofs of Theorem 2 and Corollary 3 that formulate the expected generalization error of the target task in the setting of uniformly-distributed coordinate layouts. Appendix C includes the explicit formulations of two special cases of Theorem 2 that were mentioned in Section 3 of the main paper. Appendix D provides additional empirical results and details for Section 3 of the main paper. In Appendix E we present and prove Corollary 10 as well as related analytical and empirical evaluations and details. In Appendices F-G we refer to the setting from Section 4 of the main text, where a specific set of parameters is transferred. Specifically, in Appendix F we prove Theorem 5 and its Corollary 6, and in Appendix G we provide additional analytical and empirical evaluations.

Note that the indexing of equations and figures in these Appendices continues the corresponding numbering from the main paper.

## Appendix A. Mathematical Developments for Section 2

### A.1. The Estimate $\widehat{\boldsymbol{\theta}}$ in Eq. (4)

Let us solve the optimization problem provided in (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 (3) as

$$\begin{aligned}\widehat{\boldsymbol{\theta}} = &\mathop{\arg\min}_{\mathbf{r}\in\mathbb{R}^d} \|\mathbf{v} - \mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{r} - \mathbf{Z}_{\mathcal{S}^c}\mathbf{Q}_{\mathcal{S}^c}\mathbf{r}\|_2^2 \\ &\text{subject to } \ \mathbf{Q}_{\mathcal{S}^c}\mathbf{r} = \mathbf{0}\end{aligned} \tag{17}$$

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 (17) becomes

$$\begin{aligned}\widehat{\boldsymbol{\theta}} = &\mathop{\arg\min}_{\mathbf{r}\in\mathbb{R}^d} \|\mathbf{v} - \mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{r}\|_2^2 \\ &\text{subject to } \ \mathbf{Q}_{\mathcal{S}^c}\mathbf{r} = \mathbf{0}.\end{aligned} \tag{18}$$

Without the equality constraint, (18) is just an unconstrained least squares problem that its solution is

$$\widehat{\boldsymbol{\theta}} = \mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{v} \tag{19}$$

where $\mathbf{Z}_{\mathcal{S}}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{Z}_{\mathcal{S}}$. Note that $\widehat{\boldsymbol{\theta}}$ in (19) satisfies the equality constraint in (17) and, therefore, (19) is also the solution for the constrained optimization problems in (17), (18), and (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 by Belkin et al. (2019b) for a given coordinate subset (i.e., deterministic $\mathcal{S}$ in our terms). The result of Belkin et al. (2019b) can be written in our notations as

$$\widetilde{\mathcal{E}}_{\text{out}} = \begin{cases} \frac{\widetilde{n}-1}{\widetilde{n}-\widetilde{p}-1}\left(\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2 + \sigma_\xi^2\right) & \text{for } \widetilde{p} \leq \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{\widetilde{p}-1}{\widetilde{p}-\widetilde{n}-1}\left(\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_2^2 + \sigma_\xi^2\right) + \frac{\widetilde{p}-\widetilde{n}}{\widetilde{p}}\|\boldsymbol{\theta}_{\mathcal{S}}\|_2^2 & \text{for } \widetilde{p} \geq \widetilde{n}+2. \end{cases} \tag{20}$$

The analysis in our work considers the coordinate subset $\mathcal{S}$ to be 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

$$\mathbb{E}_{\mathcal{S}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\} = \begin{cases} \frac{\widetilde{n}-1}{\widetilde{n}-\widetilde{p}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\|\boldsymbol{\theta}\|_2^2 + \sigma_\xi^2\right) & \text{for } \widetilde{p} \leq \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{\widetilde{p}-1}{\widetilde{p}-\widetilde{n}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\|\boldsymbol{\theta}\|_2^2 + \sigma_\xi^2\right) + \frac{\widetilde{p}-\widetilde{n}}{d}\|\boldsymbol{\theta}\|_2^2 & \text{for } \widetilde{p} \geq \widetilde{n}+2. \end{cases} \tag{21}$$

The formulation in (21) considers $\boldsymbol{\theta}$ as a deterministic vector. For the analysis of the target task, where the task relation model (6) 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.

$$\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\} = \begin{cases} \frac{\widetilde{n}-1}{\widetilde{n}-\widetilde{p}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\kappa + \sigma_\xi^2\right) & \text{for } \widetilde{p} \leq \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{\widetilde{p}-1}{\widetilde{p}-\widetilde{n}-1}\left(\left(1-\frac{\widetilde{p}}{d}\right)\kappa + \sigma_\xi^2\right) + \frac{\widetilde{p}-\widetilde{n}}{d}\kappa & \text{for } \widetilde{p} \geq \widetilde{n}+2. \end{cases} \tag{22}$$

where $\kappa \triangleq \mathbb{E}_{\boldsymbol{\eta}}\left\{\|\boldsymbol{\theta}\|_2^2\right\} = \left\|\boldsymbol{\beta}^{(\mathbf{H})}\right\|_2^2 + d\sigma_\eta^2$ was defined before Theorem 2 in the main text.

### A.3. The Estimate $\widehat{\boldsymbol{\beta}}$ in Eq. (9)

The optimization problem in (8), 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

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\arg\min}\; & \|\mathbf{y} - \mathbf{X}_{\mathcal{F}}\mathbf{Q}_{\mathcal{F}}\mathbf{b} - \mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{b} - \mathbf{X}_{\mathcal{Z}}\mathbf{Q}_{\mathcal{Z}}\mathbf{b}\|_2^2 \\ \text{subject to}\quad & \mathbf{Q}_{\mathcal{T}}\mathbf{b} = \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} \\ & \mathbf{Q}_{\mathcal{Z}}\mathbf{b} = \mathbf{0} \end{aligned} \tag{23}$$

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 (23) in its optimization cost, the problem (23) can be translated into the form of

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^d}{\arg\min}\; & \left\|\mathbf{y} - \mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} - \mathbf{X}_{\mathcal{F}}\mathbf{Q}_{\mathcal{F}}\mathbf{b}\right\|_2^2 \\ \text{subject to}\quad & \mathbf{Q}_{\mathcal{T}}\mathbf{b} = \mathbf{Q}_{\mathcal{T}}\widehat{\boldsymbol{\theta}} \\ & \mathbf{Q}_{\mathcal{Z}}\mathbf{b} = \mathbf{0}. \end{aligned} \tag{24}$$

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

$$\widehat{\boldsymbol{\beta}} = \mathbf{Q}_{\mathcal{F}}^{T}\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) + \mathbf{Q}_{\mathcal{T}}^{T}\widehat{\boldsymbol{\theta}}_{\mathcal{T}} \tag{25}$$

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. Proof of Theorem 2 and Corollary 3

This section is organized as follows. Section B.1 presents auxiliary results on uniformly-distributed coordinate subsets. Section B.2 provides results on non-asymptotic properties of Gaussian and Wishart matrices. The auxiliary results are utilized in Section B.3 to prove the formulations given in Theorem 2 for the generalization error of the target task in the setting of uniformly-distributed coordinate subset layout.

### B.1. Auxiliary Results on the Uniformly-Distributed Coordinate Subset Layout

Recall Definition 1 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 $\widetilde{p} \in \{1, \dots, d\}$ and $(p, t) \in \{0, \dots, d\} \times \{0, \dots, \widetilde{p}\}$ 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 Theorem 2.

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

$$\mathbb{E}_{\mathcal{L}}\left\{\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Q}_{\mathcal{S}}\right\} = \mathbb{E}_{\mathcal{S}}\left\{\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Q}_{\mathcal{S}}\right\} = \frac{\binom{d-1}{\widetilde{p}-1}}{\binom{d}{\widetilde{p}}}\mathbf{I}_{d} = \frac{\widetilde{p}}{d}\mathbf{I}_{d} \tag{26}$$

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 $j^{\text{th}}$ diagonal component equals 1 if $j \in \mathcal{S}$ and 0 otherwise.

Definition 1 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 $\mathcal{F} \cup \mathcal{T} \cup \mathcal{Z} = \{1, \dots, d\}$ such that $|\mathcal{F}| = p$, $|\mathcal{T}| = t$ and $\mathcal{T} \subset \mathcal{S}$, and $|\mathcal{Z}| = d - p - t$. Accordingly,

$$\begin{aligned}
\mathbb{E}_{\mathcal{L}}\left\{\mathbf{Q}_{\mathcal{T}}^{T}\mathbf{Q}_{\mathcal{T}}\right\} &= \mathbb{E}_{\mathcal{S}}\left\{\mathbb{E}_{\mathcal{L}|\mathcal{S}}\left\{\mathbf{Q}_{\mathcal{T}}^{T}\mathbf{Q}_{\mathcal{T}}\right\}\right\} \\
&= \frac{\binom{\widetilde{p}-1}{t-1}}{\binom{\widetilde{p}}{t}}\mathbb{E}_{\mathcal{S}}\left\{\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Q}_{\mathcal{S}}\right\} \\
&= \frac{t}{\widetilde{p}}\mathbb{E}_{\mathcal{S}}\left\{\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Q}_{\mathcal{S}}\right\} \\
&= \frac{t}{d}\mathbf{I}_{d},
\end{aligned} \tag{27}$$

and similarly

$$\mathbb{E}_{\mathcal{L}}\left\{\mathbf{Q}_{\mathcal{F}}^{T}\mathbf{Q}_{\mathcal{F}}\right\} = \frac{p}{d}\mathbf{I}_{d}, \tag{28}$$

$$\mathbb{E}_{\mathcal{L}}\left\{\mathbf{Q}_{\mathcal{Z}}^{T}\mathbf{Q}_{\mathcal{Z}}\right\} = \frac{d-p-t}{d}\mathbf{I}_{d}. \tag{29}$$

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

$$\begin{aligned}\mathbb{E}_{\mathcal{L}}\Big\{\mathbf{Q}_{\mathcal{S}}\mathbf{Q}_{\mathcal{T}}^T\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\Big\} &= \mathbb{E}_{\mathcal{S}}\Big\{\mathbf{Q}_{\mathcal{S}}\mathbb{E}_{\mathcal{L}|\mathcal{S}}\Big\{\mathbf{Q}_{\mathcal{T}}^T\mathbf{Q}_{\mathcal{T}}\Big\}\mathbf{Q}_{\mathcal{S}}^T\Big\}\\ &= \frac{t}{p}\mathbb{E}_{\mathcal{S}}\Big\{\mathbf{Q}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Q}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}^T\Big\}\\ &= \frac{t}{\widetilde{p}}\mathbf{I}_{\widetilde{p}}.\end{aligned} \tag{30}$$

The results in (27)–(29) imply that

$$\mathbb{E}_{\mathcal{L}}\Big\{\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2\Big\} = \boldsymbol{\beta}^T\mathbb{E}_{\mathcal{L}}\Big\{\mathbf{Q}_{\mathcal{T}}^T\mathbf{Q}_{\mathcal{T}}\Big\}\boldsymbol{\beta} = \frac{t}{d}\,\|\boldsymbol{\beta}\|_2^2\,, \tag{31}$$

$$\mathbb{E}_{\mathcal{L}}\Big\{\|\boldsymbol{\beta}_{\mathcal{F}}\|_2^2\Big\} = \boldsymbol{\beta}^T\mathbb{E}_{\mathcal{L}}\Big\{\mathbf{Q}_{\mathcal{F}}^T\mathbf{Q}_{\mathcal{F}}\Big\}\boldsymbol{\beta} = \frac{p}{d}\,\|\boldsymbol{\beta}\|_2^2\,, \tag{32}$$

$$\mathbb{E}_{\mathcal{L}}\Big\{\|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2\Big\} = \boldsymbol{\beta}^T\mathbb{E}_{\mathcal{L}}\Big\{\mathbf{Q}_{\mathcal{Z}}^T\mathbf{Q}_{\mathcal{Z}}\Big\}\boldsymbol{\beta} = \frac{d-p-t}{d}\,\|\boldsymbol{\beta}\|_2^2\,, \tag{33}$$

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 (31)-(33) hold also for $d$-dimensional deterministic vectors other than $\boldsymbol{\beta}$, e.g., (31)-(33) hold for $\boldsymbol{\beta}^{(\mathbf{H})} \triangleq \mathbf{H}\boldsymbol{\beta}$.

### B.2. 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,

$$\mathbb{E}\Big\{\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}}\Big\} = \mathbf{I}_p \times \begin{cases}1 & \text{for } p\le n,\\ \frac{n}{p} & \text{for } p>n,\end{cases} \tag{34}$$

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,

$$\mathbb{E}\Big\{\big\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}}\mathbf{a}\big\|_2^2\Big\} = \mathbb{E}\Big\{\|\mathbf{a}\|_2^2\Big\} \times \begin{cases}1 & \text{for } p\le n,\\ \frac{n}{p} & \text{for } p>n.\end{cases} \tag{35}$$

Since the components of $\mathbf{X}_{\mathcal{F}}$ are i.i.d. standard Gaussian variables then $\mathbf{X}_{\mathcal{F}}^T\mathbf{X}_{\mathcal{F}} \sim \mathcal{W}_p\left(\mathbf{I}_p, n\right)$ 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

$$\mathbb{E}\Big\{\big(\mathbf{X}_{\mathcal{F}}\mathbf{X}_{\mathcal{F}}^T\big)^{+}\Big\} = \mathbb{E}\Big\{\mathbf{X}_{\mathcal{F}}^{+,T}\mathbf{X}_{\mathcal{F}}^{+}\Big\} = \mathbf{I}_n \times \begin{cases}\frac{1}{n-p-1}\cdot\frac{p}{n} & \text{for } p\le n-2,\\ \infty & \text{for } n-1\le p\le n+1,\\ \frac{1}{p-n-1} & \text{for } p\ge n+2,\end{cases} \tag{36}$$

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 of Breiman and Freedman (1983).

Following (36), let $\mathbf{u} \in \mathbb{R}^n$ be a random vector independent of $\mathbf{X}_{\mathcal{F}}$. Then,

$$\mathbb{E}\left\{ \left\| \mathbf{X}_{\mathcal{F}}^{+} \mathbf{u} \right\|_2^2 \right\} = \frac{1}{n} \mathbb{E}\left\{ \left\| \mathbf{u} \right\|_2^2 \right\} \times \begin{cases} \frac{p}{n-p-1} & \text{for } p \leq n-2, \\ \infty & \text{for } n-1 \leq p \leq n+1, \\ \frac{n}{p-n-1} & \text{for } p \geq n+2, \end{cases} \tag{37}$$

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

$$\mathbb{E}\left\{ \left\| \mathbf{X}_{\mathcal{F}}^{+} \mathbf{X}_{\mathcal{F}^c} \boldsymbol{\beta}_{\mathcal{F}^c} \right\|_2^2 \right\} = \mathbb{E}\left\{ \left\| \boldsymbol{\beta}_{\mathcal{F}^c} \right\|_2^2 \right\} \times \begin{cases} \frac{p}{n-p-1} & \text{for } p \leq n-2, \\ \infty & \text{for } n-1 \leq p \leq n+1, \\ \frac{n}{p-n-1} & \text{for } p \geq n+2. \end{cases} \tag{38}$$

The results in (34)-(38) 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 (34)-(38) 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 (34) is

$$\mathbb{E}\left\{ \mathbf{Z}_{\mathcal{S}}^{+} \mathbf{Z}_{\mathcal{S}} \right\} = \mathbf{I}_{\widetilde{p}} \times \begin{cases} 1 & \text{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases} \tag{39}$$

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

### B.3. Proof Outline of Theorem 2

The generalization error $\mathcal{E}_{\text{out}}$ of the target task was formulated in Eq. (7) for a specific coordinate subset layout $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$. Here we prove the formulation given in Theorem 2 for the expecation of $\mathcal{E}_{\text{out}}$ with respect to a $\{\widetilde{p}, p, t\}$-uniformly distributed layout $\mathcal{L}$ (see Definition 1). While in the main text the expectation with respect to $\mathcal{L}$ is explicitly denoted as $\mathbb{E}_{\mathcal{L}}\{\cdot\}$, in the developments below we use the notation of $\mathbb{E}\{\cdot\}$ to refer to the expectation with respect to *all* the random elements (that may also include $\mathcal{L}$) of the expression it is applied on. The developments start with

$$\begin{aligned} \mathbb{E}_{\mathcal{L}}\left\{ \mathcal{E}_{\text{out}} \right\} =& \sigma_\epsilon^2 + \mathbb{E}\left\{ \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|_2^2 \right\} \\ =& \sigma_\epsilon^2 + \mathbb{E}\left\{ \left\| \boldsymbol{\beta}_{\mathcal{Z}} \right\|_2^2 \right\} + \mathbb{E}\left\{ \left\| \mathbf{X}_{\mathcal{F}}^{+} \left( \mathbf{y} - \mathbf{X}_{\mathcal{T}} \widehat{\boldsymbol{\theta}}_{\mathcal{T}} \right) - \boldsymbol{\beta}_{\mathcal{F}} \right\|_2^2 \right\} + \mathbb{E}\left\{ \left\| \widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\beta}_{\mathcal{T}} \right\|_2^2 \right\}, \end{aligned} \tag{40}$$

where the last decomposition shows the expected squared error $\mathbb{E}_{\mathcal{L}}\left\{ \mathcal{E}_{\text{out}} \right\}$ as the sum of the irreducible error $\sigma_\epsilon^2$ and the expected errors corresponding to each of the three subvectors induced by the coordinate subsets $\mathcal{Z}, \mathcal{F}, \mathcal{T}$.

Using the expression for the estimate $\widehat{\boldsymbol{\theta}}$ given in (4) and the relation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the expected error in the subvector induced by $\mathcal{F}$, i.e., the third term in (40), can be developed into

$$\begin{aligned} \mathbb{E}\left\{ \left\| \mathbf{X}_{\mathcal{F}}^{+} \left( \mathbf{y} - \mathbf{X}_{\mathcal{T}} \widehat{\boldsymbol{\theta}}_{\mathcal{T}} \right) - \boldsymbol{\beta}_{\mathcal{F}} \right\|_2^2 \right\} &= \mathbb{E}\left\{ \left\| \mathbf{X}_{\mathcal{F}}^{+} \left( \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}_{\mathcal{T}} \widehat{\boldsymbol{\theta}}_{\mathcal{T}} \right) - \boldsymbol{\beta}_{\mathcal{F}} \right\|_2^2 \right\} \\ = \mathbb{E}\left\{ \left\| \mathbf{X}_{\mathcal{F}}^{+} \left( \mathbf{X}_{\mathcal{F}^c} \boldsymbol{\beta}_{\mathcal{F}^c} + \boldsymbol{\epsilon} \right) \right\|_2^2 \right\} &+ \mathbb{E}\left\{ \left\| \mathbf{X}_{\mathcal{F}}^{+} \mathbf{X}_{\mathcal{T}} \widehat{\boldsymbol{\theta}}_{\mathcal{T}} \right\|_2^2 \right\} \\ + \mathbb{E}\left\{ \left\| \left( \mathbf{I}_p - \mathbf{X}_{\mathcal{F}}^{+} \mathbf{X}_{\mathcal{F}} \right) \boldsymbol{\beta}_{\mathcal{F}} \right\|_2^2 \right\} &- 2\mathbb{E}\left\{ \boldsymbol{\beta}_{\mathcal{T}}^T \left( \mathbf{X}_{\mathcal{F}}^{+} \mathbf{X}_{\mathcal{T}} \right)^T \mathbf{X}_{\mathcal{F}}^{+} \mathbf{X}_{\mathcal{T}} \widehat{\boldsymbol{\theta}}_{\mathcal{T}} \right\}. \end{aligned} \tag{41}$$

The four terms in the decomposition in (41) are further developed as follows. The first term in (41) can be computed using the results in (32), (37) and (38) to receive the form of

$$\mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{F}^c}\boldsymbol{\beta}_{\mathcal{F}^c}+\boldsymbol{\epsilon}\right)\right\|_2^2\right\}=\left(\frac{d-p}{d}\left\|\boldsymbol{\beta}\right\|_2^2+\sigma_\epsilon^2\right)\times\begin{cases}\frac{p}{n-p-1} & \text{for } p\leq n-2,\\ \infty & \text{for } n-1\leq p\leq n+1,\\ \frac{n}{p-n-1} & \text{for } p\geq n+2.\end{cases} \tag{42}$$

The second term in (41) is developed next using (30), (37), 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 $\mathbb{E}\left\{\mathbf{X}_{\mathcal{T}}^T\mathbf{X}_{\mathcal{T}}\right\}=n\mathbf{I}_t$. Then,

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\}\\&=\mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\}\\&=\frac{1}{n}\mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\}\times\begin{cases}\frac{p}{n-p-1} & \text{for } p\leq n-2,\\ \infty & \text{for } n-1\leq p\leq n+1,\\ \frac{n}{p-n-1} & \text{for } p\geq n+2,\end{cases}\\&=\frac{t}{p}\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\}\times\begin{cases}\frac{p}{n-p-1} & \text{for } p\leq n-2,\\ \infty & \text{for } n-1\leq p\leq n+1,\\ \frac{n}{p-n-1} & \text{for } p\geq n+2.\end{cases}\end{aligned} \tag{43}$$

Note that

$$\begin{aligned}&\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\}\\&=\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}-\boldsymbol{\theta}_{\mathcal{S}}\right\|_2^2\right\}+\mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{S}}\right\|_2^2\right\}+2\mathbb{E}\left\{\left(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}-\boldsymbol{\theta}_{\mathcal{S}}\right)^T\boldsymbol{\theta}_{\mathcal{S}}\right\}\\&=\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}-\mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{S}^c}\right\|_2^2\right\}-\sigma_\xi^2+\mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{S}}\right\|_2^2\right\}+2\mathbb{E}\left\{\left(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}-\boldsymbol{\theta}_{\mathcal{S}}\right)^T\boldsymbol{\theta}_{\mathcal{S}}\right\}\\&=\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}-\mathbb{E}\left\{\left\|\boldsymbol{\theta}\right\|_2^2\right\}-\sigma_\xi^2+2\mathbb{E}\left\{\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\boldsymbol{\theta}_{\mathcal{S}}\right\}\\&=\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}-\mathbb{E}\left\{\left\|\boldsymbol{\theta}\right\|_2^2\right\}-\sigma_\xi^2+2\mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{S}}\right\|_2^2\right\}\times\begin{cases}1 & \text{for } \widetilde{p}\leq\widetilde{n},\\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p}>\widetilde{n},\end{cases}\\&=\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}-\kappa-\sigma_\xi^2+2\frac{\widetilde{p}}{d}\kappa\times\begin{cases}1 & \text{for } \widetilde{p}\leq\widetilde{n},\\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p}>\widetilde{n}.\end{cases}\end{aligned} \tag{44}$$

Then, one can combine the formulations in (43) and (44) to receive a formula for the second term in (41).

Next, the third term of (41) can be developed using (32) and (35) into

$$\begin{aligned}\mathbb{E}\left\{\left\|\left(\mathbf{I}_p-\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}}\right)\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right\}&=\mathbb{E}\left\{\left\|\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right\}\times\begin{cases}0 & \text{for } p\leq n,\\ 1-\frac{n}{p} & \text{for } p>n,\end{cases}\\&=\frac{1}{d}\left\|\boldsymbol{\beta}\right\|_2^2\times\begin{cases}0 & \text{for } p\leq n,\\ p-n & \text{for } p>n.\end{cases}\end{aligned} \tag{45}$$

The fourth term of (41) is developed next using the relation $\mathbf{v} = \mathbf{Z}\mathbf{H}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\xi}$ and the independence of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ with the other random variables. Using (36) and (39) one can get

$$\begin{aligned}
&\mathbb{E}\Big\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\left(\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\right)^{T}\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\Big\}\\
&=\mathbb{E}\Big\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\left(\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\right)^{T}\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{v}\Big\}\\
&=\mathbb{E}\Big\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\mathbf{X}_{\mathcal{T}}^{T}\mathbf{X}_{\mathcal{F}}^{+,T}\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})}\Big\}\\
&=\frac{t}{d}\langle\boldsymbol{\beta}^{(\mathbf{H})},\boldsymbol{\beta}\rangle\times\begin{cases}
\frac{p}{n-p-1} & \text{for } p\leq n-2 \text{ and } \widetilde{p}\leq\widetilde{n}-2,\\
\frac{p}{n-p-1}\cdot\frac{\widetilde{n}}{\widetilde{p}} & \text{for } p\leq n-2 \text{ and } \widetilde{p}\geq\widetilde{n}+2,\\
\frac{n}{p-n-1} & \text{for } p\geq n+2 \text{ and } \widetilde{p}\leq\widetilde{n}-2,\\
\frac{n}{p-n-1}\cdot\frac{\widetilde{n}}{\widetilde{p}} & \text{for } p\geq n+2 \text{ and } \widetilde{p}\geq\widetilde{n}+2,\\
\infty & \text{for } n-1\leq p\leq n+1 \text{ or } \widetilde{n}-1\leq\widetilde{p}\leq\widetilde{n}+1.
\end{cases}
\end{aligned}\tag{46}$$

where the last expression can be rewritten using the relation $\langle\boldsymbol{\beta}^{(\mathbf{H})},\boldsymbol{\beta}\rangle = \kappa\cdot\rho$ that stems from the definitions provided before Theorem 2 in the main text for the normalized task correlation and the task energy ratio. At this intermediate stage, one can use the results provided in (41)-(46) to formulate the expected error in the subvector induced by $\mathcal{F}$, i.e., the third term in (40).

We now turn to develop an explicit formula for the expected error in the subvector induced by $\mathcal{T}$, i.e., the fourth term in (40)

$$\begin{aligned}
&\mathbb{E}\Big\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}-\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\Big\}\\
&=\mathbb{E}\Big\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\Big\}+\mathbb{E}\Big\{\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2\Big\}-2\mathbb{E}\Big\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\Big\}\\
&=\mathbb{E}\Big\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\Big\}+\mathbb{E}\Big\{\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2\Big\}-2\mathbb{E}\Big\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}\mathbf{H}\boldsymbol{\beta}\Big\}
\end{aligned}\tag{47}$$

Due to the random coordinate layout, the first term in (47) equals to $\frac{t}{p}\mathbb{E}\Big\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\Big\}$ and then one can use the expression in (44). The second term in (47) is given in (31) as

$$\mathbb{E}\Big\{\|\boldsymbol{\beta}_{\mathcal{T}}\|_2^2\Big\}=\frac{t}{d}\,\|\boldsymbol{\beta}\|_2^2\,.\tag{48}$$

The third term in (47) can be developed using (39) to obtain the expression

$$\begin{aligned}
\mathbb{E}\Big\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}\mathbf{H}\boldsymbol{\beta}\Big\}&=\mathbb{E}\Big\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})}\Big\}\\
&=\frac{t}{d}\langle\boldsymbol{\beta}^{(\mathbf{H})},\boldsymbol{\beta}\rangle\times\begin{cases}
1 & \text{for } \widetilde{p}\leq\widetilde{n},\\
\frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p}>\widetilde{n}.
\end{cases}
\end{aligned}\tag{49}$$

where the last expression can be also written differently using the relation $\langle\boldsymbol{\beta}^{(\mathbf{H})},\boldsymbol{\beta}\rangle = \kappa\cdot\rho$.

Setting the results from (33) and (41)-(49) into (40) leads to the formula given in Theorem 2, namely,

$$\mathbb{E}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}\right\} = \begin{cases} \frac{n-1}{n-p-1}\left(\left(1-\frac{p}{d}\right)\|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2 + t\cdot\Delta\mathcal{E}_{\text{transfer}}\right) & \text{for } p \leq n-2, \\ \infty & \text{for } n-1 \leq p \leq n+1, \\ \frac{p-1}{p-n-1}\left(\left(1-\frac{p}{d}\right)\|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2 + t\cdot\Delta\mathcal{E}_{\text{transfer}}\right) + \frac{p-n}{d}\|\boldsymbol{\beta}\|_2^2 & \text{for } p \geq n+2, \end{cases} \tag{50}$$

where

$$\Delta\mathcal{E}_{\text{transfer}} = \frac{1}{\widetilde{p}}\cdot\left(\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\} - \sigma_\xi^2 - \kappa\right) - 2\frac{\kappa}{d}\left(\rho-1\right)\times\begin{cases} 1 & \text{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}. \end{cases} \tag{51}$$

### B.4. Proof of Corollary 3

One can prove Corollary 3 simply by setting the explicit expression for $\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}$ from (22) in the formula for $\Delta\mathcal{E}_{\text{transfer}}$ from (51) and reorganize the expression to the form presented in Corollary 3.

Let us also outline an alternative proof that does not rely on the formula of $\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}$ from (22). The main aspect in this proof of Corollary 3 is to obtain an explicit formula for $\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\}$ as follows.

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\} = \\
&= \mathbb{E}\left\{\left\|\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{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\left(\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})} + \boldsymbol{\eta}_{\mathcal{S}}\right)\right\|_2^2\right\} + \mathbb{E}\left\{\left\|\mathbf{Z}_{\mathcal{S}}^{+}\left(\mathbf{Z}_{\mathcal{S}^c}\boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} + \mathbf{Z}_{\mathcal{S}^c}\boldsymbol{\eta}_{\mathcal{S}^c} + \boldsymbol{\xi}\right)\right\|_2^2\right\} \\
&= \mathbb{E}\left\{\left\|\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})} + \boldsymbol{\eta}_{\mathcal{S}}\right\|_2^2\right\} \times \left(\begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases}\right) \\
&\quad + \frac{1}{\widetilde{n}}\mathbb{E}\left\{\left\|\mathbf{Z}_{\mathcal{S}^c}\boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} + \mathbf{Z}_{\mathcal{S}^c}\boldsymbol{\eta}_{\mathcal{S}^c} + \boldsymbol{\xi}\right\|_2^2\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) \\
&= \frac{\widetilde{p}}{d} \cdot \left(\left\|\boldsymbol{\beta}^{(\mathbf{H})}\right\|_2^2 + d\sigma_\eta^2\right) \times \left(\begin{cases} 1 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases}\right) \\
&\quad + \left(\frac{d-\widetilde{p}}{d} \cdot \left(\left\|\boldsymbol{\beta}^{(\mathbf{H})}\right\|_2^2 + d\sigma_\eta^2\right) + \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) \\
&= \frac{\kappa}{d} \times \left(\left(\begin{cases} \widetilde{p} & \text{for } \widetilde{p} \le \widetilde{n}, \\ \widetilde{n} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases}\right) + \left(d - \widetilde{p} + d \cdot \kappa^{-1} \cdot \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)\right) \\
&= \frac{\kappa}{d} \times \begin{cases} \widetilde{p}\left(1 + \frac{d-\widetilde{p}+d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{n}-\widetilde{p}-1}\right) & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \widetilde{n}\left(1 + \frac{d-\widetilde{p}+d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{p}-\widetilde{n}-1}\right) & \text{for } \widetilde{p} \ge \widetilde{n}+2. \end{cases} \qquad (52)
\end{aligned}$$

Note that we also assumed that $\left\|\boldsymbol{\beta}^{(\mathbf{H})}\right\|_2^2 + d\sigma_\eta^2 \ne 0$, i.e., that $\kappa \ne 0$.

We can use the explicit expression for $\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\}$ from (52) together with the result from (44) that connects $\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{S}}\right\|_2^2\right\}$ to $\mathbb{E}_{\mathcal{S},\boldsymbol{\eta}}\left\{\widetilde{\mathcal{E}}_{\text{out}}\right\}$, and this lets us to develop the formula for $\Delta\mathcal{E}_{\text{transfer}}$ from Theorem 2 (and also (51)) into

$$\Delta\mathcal{E}_{\text{transfer}} = \frac{\kappa}{d} \times \begin{cases} 1 - 2\rho + \frac{d-\widetilde{p}+d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}}\left(1 - 2\rho + \frac{d-\widetilde{p}+d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{p}-\widetilde{n}-1}\right) & \text{for } \widetilde{p} \ge \widetilde{n}+2 \end{cases} \qquad (53)$$

that appears in Corollary 3.

## Appendix C. Special Cases of Theorem 2

Let us emphasize two special cases of the general result in Theorem 2. The first case considers the solution of the target task *without transfer learning*, i.e., $t = 0$. The corresponding generalization error is formulated in the following corollary that shows that our general formula from Theorem 2 reduces for $t = 0$ into the expectation (over $\mathcal{L}$) of the standard double descent formula (in non-asymptotic settings such as by Belkin et al. (2019b)). The red-colored curves in the subfigures of Fig. 1 are identical and correspond to the formula of this case.

**Corollary 8 (No transfer learning)** *For $t = 0$, i.e., no parameters are transferred from the source task to the target task. Then, for $p \in \{0, \dots, d\}$,*

$$\mathbb{E}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}^{(t=0)}\right\} = \begin{cases} \left(1 + \frac{p}{n-p-1}\right)\left(\left(1 - \frac{p}{d}\right)\|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2\right) & \textit{for } p \leq n-2, \\ \infty & \textit{for } n-1 \leq p \leq n+1, \\ \left(1 + \frac{n}{p-n-1}\right)\left(\left(1 - \frac{p}{d}\right)\|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2\right) + \frac{p-n}{d}\|\boldsymbol{\beta}\|_2^2 & \textit{for } p \geq n+2. \end{cases} \tag{54}$$

The second special case corresponds to transferring parameters from the source task to the target task *without applying any additional learning*. This case is induced by setting $p = 0$ in the formula of Theorem 2 (also using Corollary 3), as explicitly formulated in the following corollary. The corresponding generalization errors are demonstrated in the left-most vertical slices of each of the subfigures in Fig. 5 that are clearly affected by the relation between the source and target tasks and the number of transferred parameters.

**Corollary 9 (Parameter transfer without additional learning in target problem)** *For $p = 0$, i.e., no parameters are learned in the target problem using $\mathcal{D}$. The target estimate $\widehat{\boldsymbol{\beta}}$ includes only $t \in \{0, \dots, \widetilde{p}\}$ parameters transferred from the source task, and the remaining components are set to zeros.*

$$\mathbb{E}_{\mathcal{L}}\left\{\mathcal{E}_{\text{out}}^{(p=0)}\right\} = \|\boldsymbol{\beta}\|_2^2 + \sigma_\epsilon^2 + t\frac{\kappa}{d} \times \begin{cases} 1 - 2\rho + \frac{d - \widetilde{p} + d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{n} - \widetilde{p} - 1} & \textit{for } \widetilde{p} \leq \widetilde{n} - 2, \\ \infty & \textit{for } \widetilde{n} - 1 \leq \widetilde{p} \leq \widetilde{n} + 1, \\ \frac{\widetilde{n}}{\widetilde{p}}\left(1 - 2\rho + \frac{d - \widetilde{p} + d\cdot\kappa^{-1}\cdot\sigma_\xi^2}{\widetilde{p} - \widetilde{n} - 1}\right) & \textit{for } \widetilde{p} \geq \widetilde{n} + 2. \end{cases} \tag{55}$$

## Appendix D. Empirical Results for Section 3: Additional Details and Demonstrations

In Fig. 6 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. 6 (and also the values denoted by circle markers in Fig. 1) 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. 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 $d = 80$, $\widetilde{n} = 50$, $n = 20$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$. The deterministic $\boldsymbol{\beta} \in \mathbb{R}^d$ used in the experiments has an increasing linear form (see Fig. 4(*a*)) that starts at zero and satisfies $\|\boldsymbol{\beta}\|_2^2 = d$. Since we consider averaging over uniformly distributed layout of coordinate subsets, the results depend only on $\|\boldsymbol{\beta}\|_2^2$ and not on the shape of the sequence of values in $\boldsymbol{\beta}$.

One can observe the excellent match between the empirical results in Fig. 6 and the analytical results provided in Fig. 5. This establishes further the formulations given in Theorem 2 and Corollary 3.

## Appendix E. Additional Details on Parameter Transfer Usefulness in the Setting of Uniformly-Distributed Coordinate Layouts

Here we form analytical conditions on the required number of free parameters $\widetilde{p}$ in the *source* task to get a useful parameter transfer.

In this section we define the signal-to-noise ratio of the source task as $\Gamma_{\text{src}} \triangleq \frac{\kappa}{\sigma_\xi^2}$.

**Corollary 10** *Consider $\widetilde{p} \in \{1, \dots, d\}$. Then, the term $\Delta\mathcal{E}_{\text{transfer}}$, which quantifies the expected error difference due to each parameter being transferred instead of set to zero, satisfies $\Delta\mathcal{E}_{\text{transfer}} < 0$ (i.e.,* ***parameter transfer is beneficial*** *for $p \notin \{n-1, n, n+1\}$) if the source task is* ***sufficiently overparameterized*** *such that*

$$\widetilde{p} > \widetilde{n} + 1 + \left(d - \widetilde{n} + d \cdot \Gamma_{\text{src}}^{-1} - 1\right)/2\rho \quad \textit{for} \;\; \rho > 0,\; \widetilde{n} < d - 1, \tag{56}$$

*or* ***sufficiently underparameterized*** *such that*

$$\widetilde{p} < \widetilde{n} - 1 - \frac{d - \widetilde{n} + d \cdot \Gamma_{\text{src}}^{-1} + 1}{2(\rho - 1)} \quad \textit{for} \;\; \rho > 1,\; \widetilde{n} \le d\left(1 + \Gamma_{\text{src}}^{-1}\right) + 1, \tag{57}$$

$$\textit{or} \quad \widetilde{n} - 1 - \frac{d - \widetilde{n} + d \cdot \Gamma_{\text{src}}^{-1} + 1}{2(\rho - 1)} < \widetilde{p} \le d \quad \textit{for} \;\; 0 \le \rho < 1,\; \widetilde{n} > d\left(1 + \Gamma_{\text{src}}^{-1}\right) + 1, \tag{58}$$

$$\textit{or} \quad 1 \le \widetilde{p} \le d \quad \textit{for} \;\; \rho \ge 1,\; \widetilde{n} > d\left(1 + \Gamma_{\text{src}}^{-1}\right) + 1. \tag{59}$$

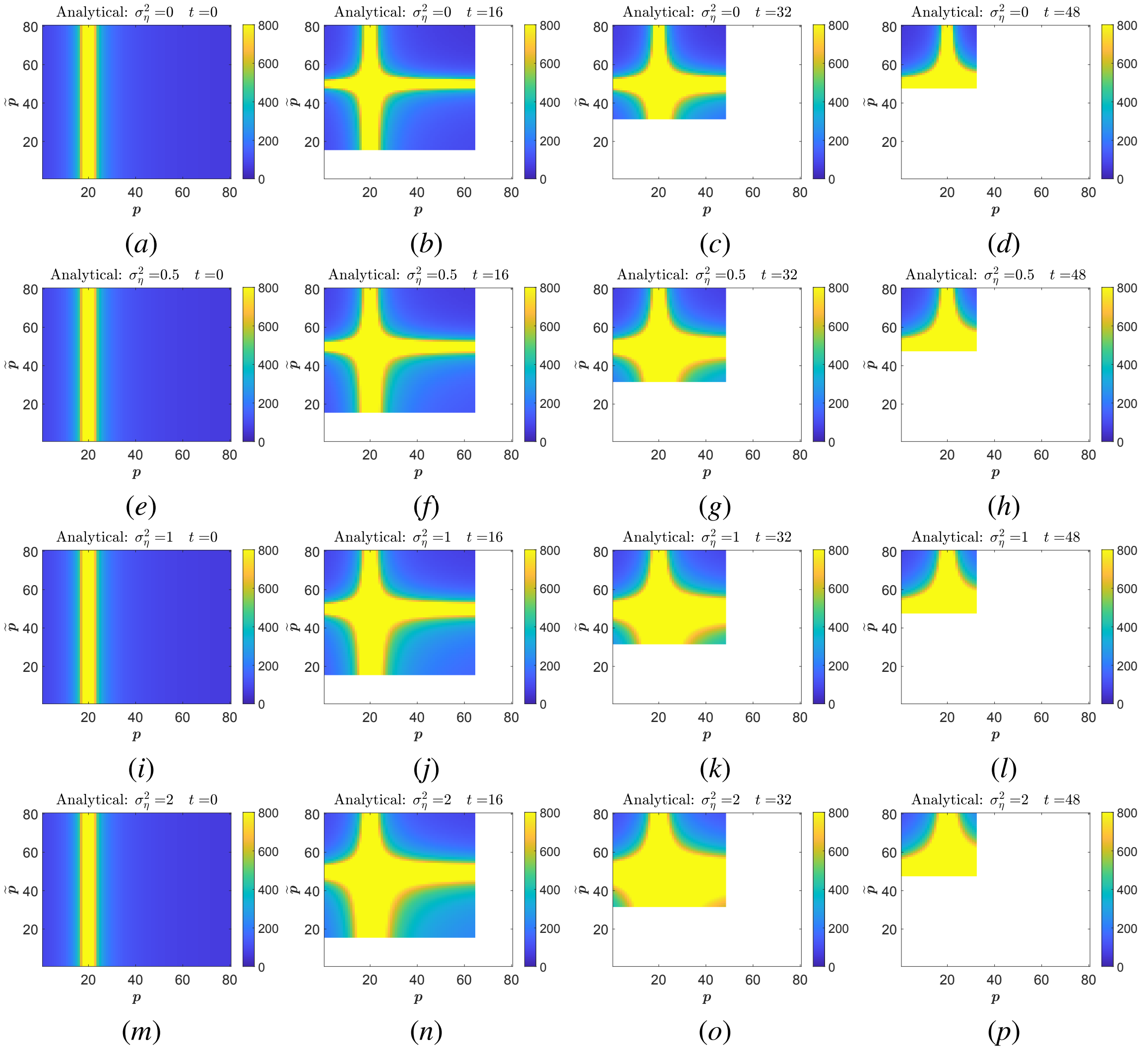
*Otherwise, $\Delta\mathcal{E}_{\text{transfer}} \ge 0$ (i.e., parameter transfer is not beneficial).*

The proof of the last corollary is provided next.

### E.1. Proof of Corollary 10

Recall that Eq. (13) in Theorem 2 formulates $\Delta\mathcal{E}_{\text{transfer}}$ and defines it as the expected error difference introduced by each constrained parameter that is transferred from the source task instead of being set to zero. Accordingly, Corollary 10 presents the conditions on the number of free parameters $\widetilde{p}$ in the source task, such that the parameter transfer is useful, namely, $\Delta\mathcal{E}_{\text{transfer}} < 0$ that is relevant when $p \notin \{n-1, n, n+1\}$ (recall that for $p \in \{n-1, n, n+1\}$ the target task generalization error becomes infinite). The conditions provided in Corollary 10 separately refer to the case where the source task is overparameterized (i.e., $\widetilde{p} > \widetilde{n}$) and the case where the source task is underparameterized (i.e., $\widetilde{p} < \widetilde{n}$).

Recall that in this section we define the signal-to-noise ratio of the source task as $\Gamma_{\text{src}} \triangleq \frac{\kappa}{\sigma_\xi^2}$.



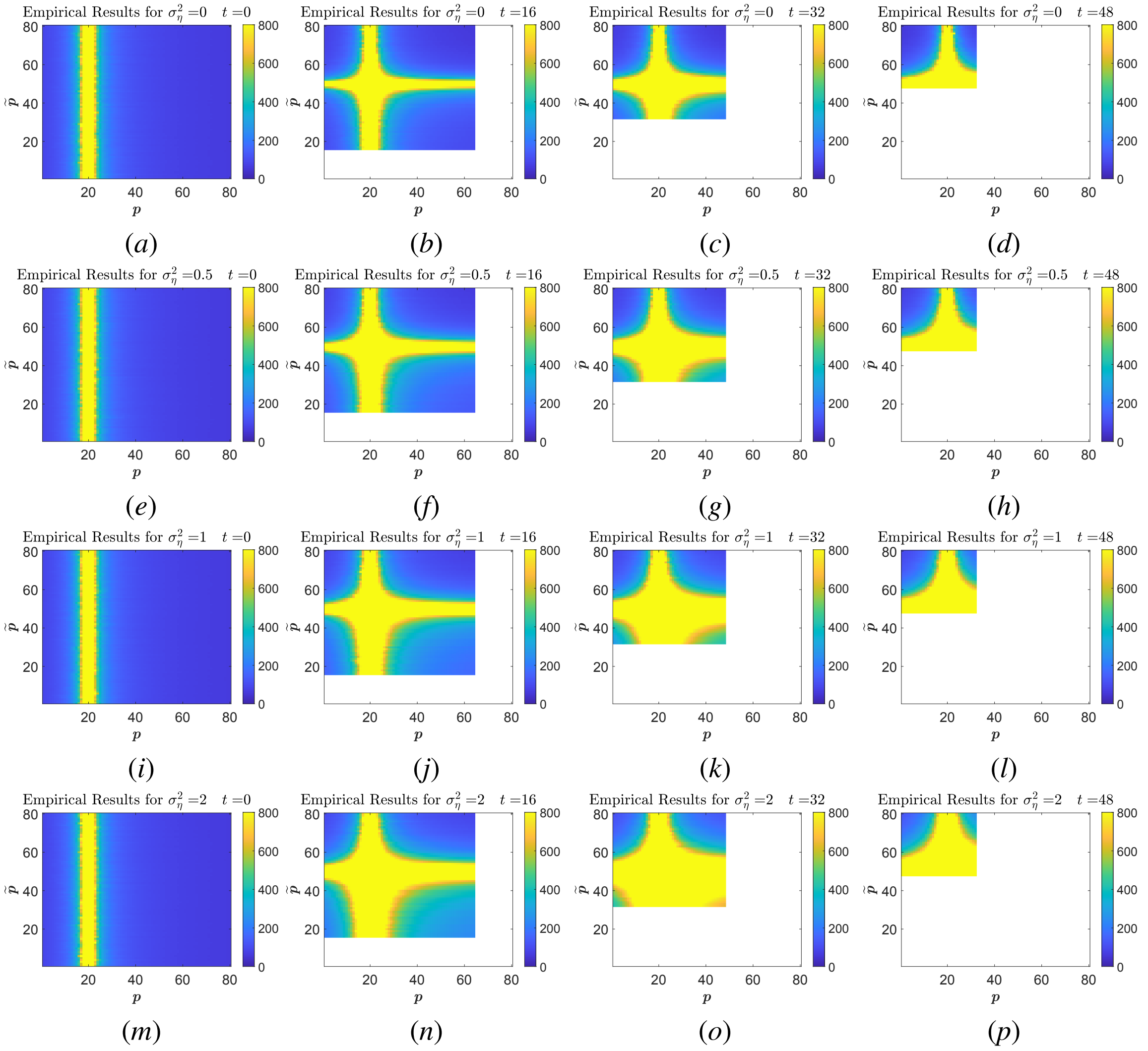
(*a*) (*b*) (*c*) (*d*)

(*e*) (*f*) (*g*) (*h*)

(*i*) (*j*) (*k*) (*l*)

(*m*) (*n*) (*o*) (*p*)

Figure 5: **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 (6) between the source and target tasks in the form of a different noise variance $\sigma_\eta^2$ whereas $\mathbf{H} = \mathbf{I}_d$ for all. Each column of subfigures considers a different number of transferred parameters $t$. Here $d = 80$, $\widetilde{n} = 50$, $n = 20$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$. 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. 6 for the corresponding empirical evaluation.



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

(*e*) (*f*) (*g*) (*h*)

(*i*) (*j*) (*k*) (*l*)

(*m*) (*n*) (*o*) (*p*)

Figure 6: **Empirical** 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). The presented values obtained by averaging over 250 experiments. Each row of subfigures considers a different case of the relation (6) between the source and target tasks in the form of a different noise variance $\sigma_\eta^2$ whereas $\mathbf{H} = \mathbf{I}_d$ for all. Each column of subfigures considers a different number of transferred parameters $t$. Here $d = 80$, $\widetilde{n} = 50$, $n = 20$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\epsilon^2 = 0.05 \cdot d$, $\sigma_\xi^2 = 0.025 \cdot d$. 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.

### E.1.1. PROOF FOR THE OVERPARAMETERIZED CASE

Let us start by proving the condition in Eq. (56) that refers to an overparameterized source task, i.e., $\widetilde{p} > \widetilde{n}$. Then, according to (13), $\Delta\mathcal{E}_{\text{transfer}} < 0$ is possible only when $\widetilde{p} \geq \widetilde{n} + 2$ and

$$1 - 2\rho + \frac{d - \widetilde{p} + d \cdot \Gamma_{\text{src}}^{-1}}{\widetilde{p} - \widetilde{n} - 1} < 0. \tag{60}$$

For $\rho > 0$, the last inequality can be rewritten as

$$\widetilde{p} > \widetilde{n} + 1 + \frac{d - \widetilde{n} + d \cdot \Gamma_{\rm src}^{-1} - 1}{2\rho}. \tag{61}$$

The overparameterization condition of $\widetilde{p} \geq \widetilde{n} + 2$ implies $d \geq \widetilde{p} \geq \widetilde{n} + 2$ and, thus, $d - \widetilde{n} - 1 \geq 1$. Then, together with $\rho > 0$ (and because $\Gamma_{\rm src}^{-1}$ is always non-negative) we get that the term $\frac{d-\widetilde{n}+d\cdot\Gamma_{\rm src}^{-1}-1}{2\rho}$ in (61) is positive valued and, hence, the lower bound of $\widetilde{p}$ in (61) is at least $n+1$ and when the SNR of the source task or the task correlation are lower then a greater overparameterization is required in the source task (i.e., larger $\widetilde{p}$) for having a useful transfer of parameters. At this intermediate stage we finished to prove the condition given in Eq. (56) of Corollary 10 and now we turn to prove that this is the only *overparameterized* case that enables $\Delta\mathcal{E}_{\rm transfer} < 0$.

For $\rho = 0$, the condition in (60) induces the requirement of $\widetilde{n} > d + d \cdot \Gamma_{\rm src}^{-1} - 1$ and, because $\Gamma_{\rm src}^{-1}$ is non-negative by its definition, this requirement also implies that $\widetilde{n} > d - 1$ that contradicts the basic overparameterization relations of $d \geq \widetilde{p} \geq \widetilde{n} + 2$. Hence, one cannot obtain $\Delta\mathcal{E}_{\rm transfer} < 0$ when $\widetilde{p} > \widetilde{n}$ and $\rho = 0$.

For $\rho < 0$, the condition in (60) means that

$$\widetilde{p} < \widetilde{n} + 1 + \frac{d - \widetilde{n} + d \cdot \Gamma_{\rm src}^{-1} - 1}{2\rho} \tag{62}$$

where $d - \widetilde{n} + d \cdot \Gamma_{\rm src}^{-1} - 1 > 0$ again due to $\widetilde{p} \geq \widetilde{n} + 2$. However, here $\rho < 0$ makes (62) to imply that $\widetilde{p} < \widetilde{n} + 1$ that clearly contradicts the overparameterized case of $\widetilde{p} \geq \widetilde{n} + 2$. Accordingly, $\Delta\mathcal{E}_{\rm transfer} < 0$ is impossible when $\widetilde{p} > \widetilde{n}$ and $\rho < 0$. This completes the proof for the overparameterized part of Corollary 10.

### E.1.2. PROOF FOR THE UNDERPARAMETERIZED CASE

We now turn to prove the conditions in Eq. (57)-(59) that refer to underparameterized settings of the source task, i.e., $\widetilde{p} < \widetilde{n}$. Then, by (13), $\Delta\mathcal{E}_{\rm transfer} < 0$ is possible only when $\widetilde{p} \leq \widetilde{n} - 2$ and

$$1 - 2\rho + \frac{d - \widetilde{p} + d \cdot \Gamma_{\rm src}^{-1}}{\widetilde{n} - \widetilde{p} - 1} < 0. \tag{63}$$

For $\rho > 1$, the last inequality can be also formulated as

$$\widetilde{p} < \widetilde{n} - 1 - \frac{d - \widetilde{n} + d \cdot \Gamma_{\rm src}^{-1} + 1}{2\left(\rho - 1\right)} \tag{64}$$

which, due to the required intersection with $\widetilde{p} \leq \widetilde{n} - 2$, remains in its form of (64) only for $\widetilde{n} \leq d + d \cdot \Gamma_{\rm src}^{-1} + 1$. For $\widetilde{n} > d + d \cdot \Gamma_{\rm src}^{-1} + 1$, the condition in (64) becomes $\widetilde{p} \leq d$ (because then the right hand side of (64) is at least $\widetilde{n} - 1$ whereas $\widetilde{p} \leq d < \widetilde{n} - 1$).

For $\rho = 1$, the condition in (63) can be developed into $\widetilde{n} > d + d \cdot \Gamma_{\rm src}^{-1} + 1$ that naturally conforms with the underparameterization condition $\widetilde{p} \leq \widetilde{n} - 2$ (this is because always $\widetilde{p} \leq d$ and $\Gamma_{\rm src}^{-1} \geq 0$ by their definitions). Then, for $\rho = 1$ and $\widetilde{n} > d + d \cdot \Gamma_{\rm src}^{-1} + 1$ we get $1 \leq \widetilde{p} \leq d$.

For $0 \leq \rho < 1$, the condition in (63) can be rewritten as

$$\widetilde{p} > \widetilde{n} - 1 - \frac{d - \widetilde{n} + d \cdot \Gamma_{\rm src}^{-1} + 1}{2\left(\rho - 1\right)}. \tag{65}$$

If $\widetilde{n} \leq d + d \cdot \Gamma_{\rm src}^{-1} + 1$ then (65) implies $\widetilde{p} > \widetilde{n} - 1$, which contradicts the underparameterized case of $\widetilde{p} \leq \widetilde{n} - 2$. Hence, for $0 \leq \rho < 1$ and $\widetilde{n} \leq d + d \cdot \Gamma_{\rm src}^{-1} + 1$ it is impossible to get $\Delta\mathcal{E}_{\rm transfer} < 0$. If $\widetilde{n} > d + d \cdot \Gamma_{\rm src}^{-1} + 1$ then the right hand side of (65) is lower than $\widetilde{n} - 1$ and, thus, the condition

$$\widetilde{n} - 1 - \frac{d - \widetilde{n} + d \cdot \Gamma_{\rm src}^{-1} + 1}{2\left(\rho - 1\right)} < \widetilde{p} \leq d \tag{66}$$

can be feasible (for $\widetilde{n} > d + d \cdot \Gamma_{\rm src}^{-1} + 1$) in underparameterized settings.

For $\rho < 0$, the condition in (63) is equivalent to

$$1 + \frac{d - \widetilde{p} + d \cdot \Gamma_{\rm src}^{-1}}{\widetilde{n} - \widetilde{p} - 1} < 0 \tag{67}$$

which implies

$$\widetilde{p} > \frac{1}{2}\left(\widetilde{n} + d + d \cdot \Gamma_{\rm src}^{-1} - 1\right). \tag{68}$$

Then, for $\widetilde{n} > d + d \cdot \Gamma_{\rm src}^{-1} + 1$, the inequality in (68) leads to

$$\widetilde{p} > d + d \cdot \Gamma_{\rm src}^{-1} > d \tag{69}$$

that clearly contradicts the basic demand $\widetilde{p} \leq d$ in our general settings. If $\widetilde{n} \leq d + d \cdot \Gamma_{\rm src}^{-1} + 1$ then the form of (65) is also relevant for $\rho < 0$ and yields that $\widetilde{p} > \widetilde{n} - 1$, which contradicts the underparameterized case of $\widetilde{p} \leq \widetilde{n} - 2$. Hence, we showed that for $\rho < 0$ and any $\widetilde{n}$ it is impossible to get $\Delta\mathcal{E}_{\rm transfer} < 0$. This completes the proof of all the conditions in Corollary 10.

### E.2. Details on the Empirical Evaluation of $\Delta\mathcal{E}_{\rm transfer}$

The analytical formula for $\Delta\mathcal{E}_{\rm transfer}$, given in Theorem 2, measures the expected difference in the generalization error (of the target task) due to each parameter that is transferred instead of being set to zero. Accordingly, the empirical evaluation of $\Delta\mathcal{E}_{\rm transfer}$ for a given $\widetilde{p}$ can be computed by

$$\widehat{\Delta\mathcal{E}}_{\rm transfer} = \frac{1}{d-3} \sum_{p=1,\dots,n-2,n+2,\dots,d} \frac{\widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\rm out}^{(\widetilde{p},p,t=m)}\right\} - \widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\rm out}^{(\widetilde{p},p,t=0)}\right\}}{m \cdot \alpha(p)} \tag{70}$$

where

$$\alpha(p) \triangleq \frac{1}{d}\left\|\boldsymbol{\beta}\right\|_2^2 \times \begin{cases} 1 + \frac{p}{n-p-1} & \text{for } p \leq n-2, \\ 1 + \frac{n}{p-n-1} & \text{for } p \geq n+2 \end{cases} \tag{71}$$

is a normalization factor required for the accurate correspondence to the analytical definition of $\Delta\mathcal{E}_{\rm transfer}$ provided in Theorem 2 in a form independent of $p$. Here $\widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\rm out}^{(\widetilde{p},p,t=m)}\right\}$ 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, $\widehat{\mathbb{E}}_{\mathcal{L}}\left\{\mathcal{E}_{\rm out}^{(\widetilde{p},p,t=0)}\right\}$ is the empirically computed error induced by avoiding parameter transfer. Therefore, the formula in (70) 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 $m = 5$.

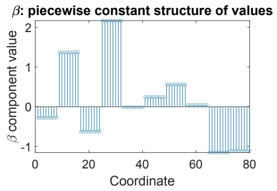


Figure 7: The piecewise-constant structure of $\boldsymbol{\beta}$ that was used in part of the experiments.

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.

## E.3. Empirical Results for Parameter Transfer Usefulness in the Setting of Uniformly-Distributed Coordinate Layouts

### E.3.1. SETTINGS WHERE $\widetilde{n} < d$

In Figures 8-9 we present analytical and empirical values of $\Delta\mathcal{E}_{\text{transfer}}$ 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. 8 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).

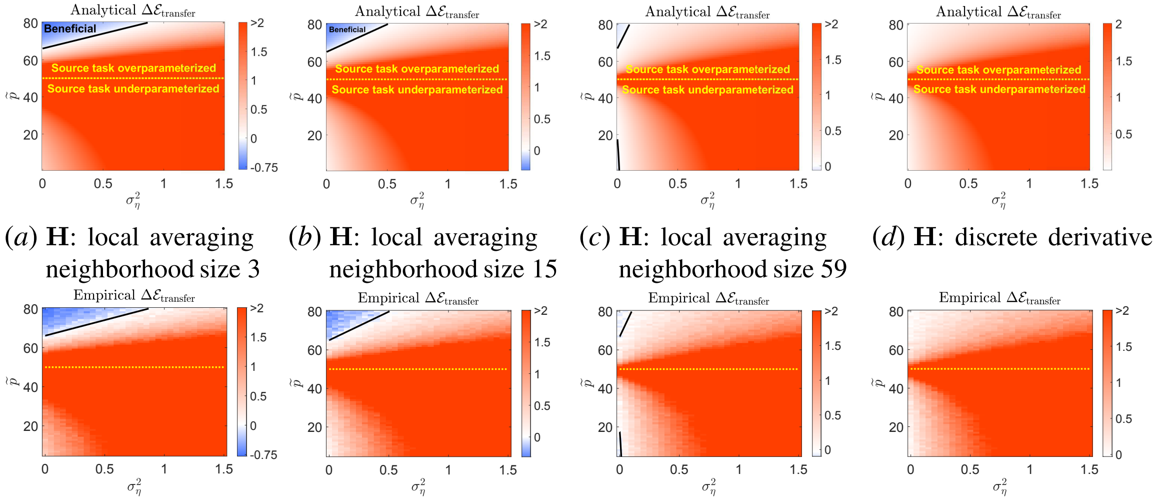
In Figure 9 we provide additional results for cases where the operator $\mathbf{H}$ is a scaled identity matrix.

### E.3.2. SETTINGS WHERE $\widetilde{n} > d$

Here we provide in Fig. 10 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 $\widetilde{p} \le d < \widetilde{n}$. Like in Fig. 9, the results in Fig. 10 show the excellent match between the analytical and empirical results and, specifically, the accuracy of the analytical thresholds (from Corollary 10) in determining the empirical settings where parameter transfer is beneficial.

# Appendix F. Transfer of Specific Sets of Parameters: Proof of Theorem 5 and Corollary 6

In this section we outline the proof of Theorem 5 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. The proof of Theorem 5 resembles the one of Theorem 2 given in Appendix B with the important difference that now we cannot use the simplified constructions that were provided in Section B.1 for uniformlly-distributed coordinate subsets. Hence, we start in Section F.1 by providing auxiliary



(*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

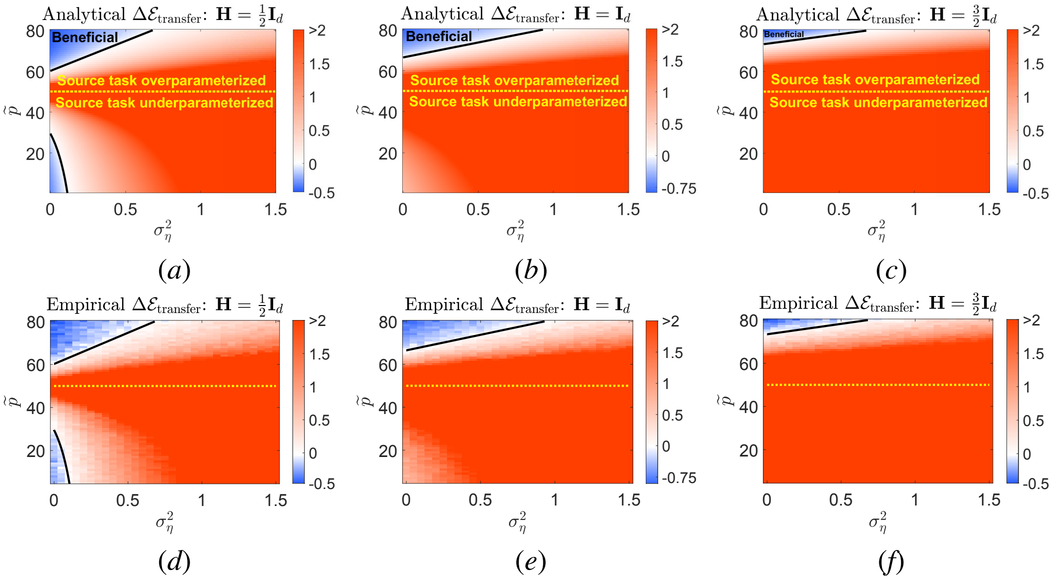
(*e*) **H**: local averaging neighborhood size 3 (*f*) **H**: local averaging neighborhood size 15 (*g*) **H**: local averaging neighborhood size 59 (*h*) **H**: discrete derivative

Figure 8: The analytical and empirical values of $\Delta\mathcal{E}_{\text{transfer}}$ defined in Theorem 2 (namely, the expected error difference due to transfer of a parameter from the source to target task) as a function of $\widetilde{p}$ and $\sigma_\eta^2$. 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. The positive values were truncated in the value of 2 for the clarity of visualization. The solid black lines (in all subfigures) denote the analytical thresholds for useful transfer learning as implied by Corollary 10. 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, $d = 80$, $\widetilde{n} = 50$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\xi^2 = 0.025 \cdot d$.

results that use non-asymptotic properties of random orthonormal matrices (that in our case originate in decompositions of the random matrices $\mathbf{Z}$ and $\mathbf{X}$ that have i.i.d. Gaussian components) in order to formulate important quantities for the setting of specific coordinate subset layouts. We also use our results that were provided in Section B.2 based on non-asymptotic properties of Gaussian and Wishart matrices and without any necessary aspect of random coordinate subset layouts. Then, in Section F.2 we prove Theorem 5 and in Section F.3 we prove Corollary 6. This proof has a similar general structure as the proof given for 2, but with several important modifications. Therefore, it is recommended to read first the proof of Theorem 2 in Appendix B before getting into the details of the following proof for the case of a specific coordinate layout.

### F.1. Auxiliary Results for the Specific Coordinate Subset Layout Setting

Here the coordinate subset layout $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$ 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 $j^{\text{th}}$ diagonal component equals 1 if $j \in \mathcal{S}$ and 0 otherwise. Similarly holds for the other coordinate subsets. Accordingly, here the norms of



(*a*) (*b*) (*c*)

(*d*) (*e*) (*f*)

Figure 9: The analytical and empirical values of $\Delta\mathcal{E}_{\rm transfer}$ defined in Theorem 2 (namely, the expected error difference due to transfer of a parameter from the source to target task) as a function of $\widetilde{p}$ and $\sigma_\eta^2$. The positive and negative values of $\Delta\mathcal{E}_{\rm transfer}$ 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. The solid black lines (in all subfigures) denote the analytical thresholds for useful transfer learning as implied by Corollary 10. 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, $d=80$, $\widetilde{n}=50$, $\|\boldsymbol{\beta}\|_2^2=d$, $\sigma_\xi^2=0.025\cdot d$.

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.

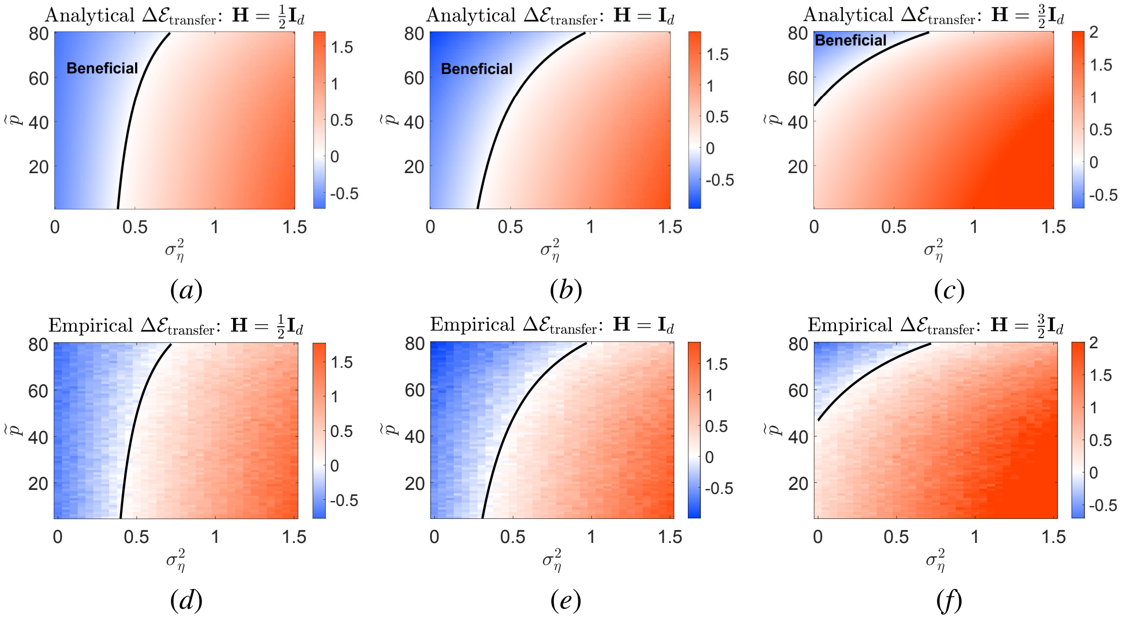
Recall that $\mathcal{T}\subset\mathcal{S}$. Then, for a deterministic vector $\mathbf{w}\in\mathbb{R}^d$,

$$\mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{w}\right\|_2^2\right\}=\begin{cases}\|\mathbf{w}_{\mathcal{T}}\|_2^2 & \text{for } \widetilde{p}\le\widetilde{n},\\ \frac{\widetilde{n}}{\widetilde{p}(\widetilde{p}+1)}\left(\left(\widetilde{n}+\frac{\widetilde{n}-1}{\widetilde{p}-1}\right)\|\mathbf{w}_{\mathcal{T}}\|_2^2+\left(1-\frac{\widetilde{n}-1}{\widetilde{p}-1}\right)t\|\mathbf{w}_{\mathcal{S}}\|_2^2\right) & \text{for } \widetilde{p}>\widetilde{n}.\end{cases} \tag{72}$$

$$\mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{Z}_{\mathcal{S}^c}\mathbf{Q}_{\mathcal{S}^c}\mathbf{w}\right\|_2^2\right\}=\frac{t}{\widetilde{p}}\|\mathbf{w}_{\mathcal{S}^c}\|_2^2\times\begin{cases}\frac{\widetilde{p}}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p}\le\widetilde{n}-2,\\ \infty & \text{for } \widetilde{n}-1\le\widetilde{p}\le\widetilde{n}+1,\\ \frac{\widetilde{n}}{\widetilde{p}-\widetilde{n}-1} & \text{for } \widetilde{p}\ge\widetilde{n}+2.\end{cases} \tag{73}$$

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

$$\mathbb{E}\left\{\left\langle\mathbf{Q}_{\mathcal{T}}\mathbf{a},\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^+\mathbf{Z}_{\mathcal{S}}\mathbf{Q}_{\mathcal{S}}\mathbf{w}\right\rangle\right\}=\langle\mathbf{a}_{\mathcal{T}},\mathbf{w}_{\mathcal{T}}\rangle\times\begin{cases}1 & \text{for } \widetilde{p}\le\widetilde{n},\\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p}>\widetilde{n}.\end{cases} \tag{74}$$



(*a*) (*b*) (*c*)

(*d*) (*e*) (*f*)

Figure 10: The analytical (top row of subfigures) and empirical (bottom row of subfigures) values of $\Delta\mathcal{E}_{\text{transfer}}$ defined in Theorem 2 (namely, the expected error difference due to transfer of a parameter from the source to target task) as a function of $\widetilde{p}$ and $\sigma_\eta^2$. 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. The positive values were truncated in the value of 2 for the clarity of visualization. The solid black lines (in all subfigures) denote the analytical thresholds for useful transfer learning as implied by Corollary 10. 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, $d = 80$, $\widetilde{n} = 150$, $\|\boldsymbol{\beta}\|_2^2 = d$, $\sigma_\xi^2 = 0.025 \cdot d$. Note that the results in this figure are for $\widetilde{n} > d$.

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

$$\mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{r}\right\|_2^2\right\} = \frac{t}{\widetilde{n}\widetilde{p}}\|\mathbf{r}\|_2^2 \times \begin{cases} \frac{\widetilde{p}}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p} \leq \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}-\widetilde{n}-1} & \text{for } \widetilde{p} \geq \widetilde{n}+2. \end{cases} \tag{75}$$

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 (72)-(75) using some algebra and the non-asymptotic properties of random Haar-distributed matrices, see examples for such properties in Lemma 2.5 by Tulino and Verdú (2004) and also in Proposition 1.2 by Hiai and Petz (2000).

### F.2. Proof Outline of Theorem 5 and Corollary 6

The generalization error $\mathcal{E}_{\text{out}}$ of the target task was expressed in its basic form in Eq. (7) for a specific coordinate subset layout $\mathcal{L} = \{\mathcal{S}, \mathcal{F}, \mathcal{T}, \mathcal{Z}\}$. Please note that, unlike in the proof of Theorem 2 in Section B, the expectations below do not include the expectation with respect to $\mathcal{L}$, which is non-random here.

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

$$\begin{aligned}\mathcal{E}_{\text{out}} &= \sigma_\epsilon^2 + \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2\right\}\\ &= \sigma_\epsilon^2 + \|\boldsymbol{\beta}_{\mathcal{Z}}\|_2^2 + \mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right\} + \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\right\}.\end{aligned} \tag{76}$$

Then, we use the expression for the estimate $\widehat{\boldsymbol{\theta}}$ given in (4) and the relation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, to decompose the third term in (76) as follows

$$\begin{aligned}&\mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{y} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right\}\\ &= \mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right) - \boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right\}\\ &= \mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{F}^c}\boldsymbol{\beta}_{\mathcal{F}^c} + \boldsymbol{\epsilon}\right)\right\|_2^2\right\} + \mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\}\\ &\quad + \mathbb{E}\left\{\left\|\left(\mathbf{I}_p - \mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}}\right)\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right\} - 2\mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^T\left(\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\right)^T\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\}.\end{aligned} \tag{77}$$

The four terms in (77) are further developed as follows. The first term in (77) can be computed using the results in (37) and (38) to receive the form of

$$\mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\left(\mathbf{X}_{\mathcal{F}^c}\boldsymbol{\beta}_{\mathcal{F}^c} + \boldsymbol{\epsilon}\right)\right\|_2^2\right\} = \left(\|\boldsymbol{\beta}_{\mathcal{F}^c}\|_2^2 + \sigma_\epsilon^2\right) \times \begin{cases}\frac{p}{n-p-1} & \text{for } p \leq n-2,\\ \infty & \text{for } n-1 \leq p \leq n+1,\\ \frac{n}{p-n-1} & \text{for } p \geq n+2.\end{cases} \tag{78}$$

The second term in (77) is developed next using the result in (37) 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 $\mathbb{E}\left\{\mathbf{X}_{\mathcal{T}}^T\mathbf{X}_{\mathcal{T}}\right\} = n\mathbf{I}_t$. Then,

$$\mathbb{E}\left\{\left\|\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\} = \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\} \times \begin{cases}\frac{p}{n-p-1} & \text{for } p \leq n-2,\\ \infty & \text{for } n-1 \leq p \leq n+1,\\ \frac{n}{p-n-1} & \text{for } p \geq n+2.\end{cases} \tag{79}$$

Then, using (74) and the definition of $\kappa_{\mathcal{T}}$,

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\} \\
&= \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} + \mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} + 2\mathbb{E}\left\{\left(\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}}\right)^T \boldsymbol{\theta}_{\mathcal{T}}\right\} \\
&= \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} - \mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} + 2\mathbb{E}\left\{\widehat{\boldsymbol{\theta}}_{\mathcal{T}}^T \boldsymbol{\theta}_{\mathcal{T}}\right\} \\
&= \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} - \mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} + 2\mathbb{E}\left\{\left\|\boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} \times \begin{cases} 1 & \text{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}, \end{cases} \\
&= \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} - \kappa_{\mathcal{T}} \left(1 - 2 \times \begin{cases} 1 & \text{for } \widetilde{p} \leq \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p} > \widetilde{n}. \end{cases}\right)
\end{aligned}
\tag{80}
$$

One can join the formulations in (79) and (80) to obtain an expression for the second term in (77).

Now, the third term of (77) can be developed using (35) into

$$
\mathbb{E}\left\{\left\|\left(\mathbf{I}_p - \mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{F}}\right)\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2\right\} = \left\|\boldsymbol{\beta}_{\mathcal{F}}\right\|_2^2 \times \begin{cases} 0 & \text{for } p \leq n, \\ 1 - \frac{n}{p} & \text{for } p > n, \end{cases}
\tag{81}
$$

The fourth term of (77) is developed next using the relation $\mathbf{v} = \mathbf{Z}\mathbf{H}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\xi}$ and the independence of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ with the other random variables. Using (36) and (74) we get

$$
\begin{aligned}
&\mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^T \left(\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\right)^T \mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\} \\
&= \mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^T \left(\mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\right)^T \mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{v}\right\} \\
&= \mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^T \mathbf{X}_{\mathcal{T}}^T \mathbf{X}_{\mathcal{F}}^{+,T} \mathbf{X}_{\mathcal{F}}^{+}\mathbf{X}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^T\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})}\right\} \\
&= \langle \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})}, \boldsymbol{\beta}_{\mathcal{T}} \rangle \times \begin{cases} \infty & \text{for } n-1 \leq p \leq n+1 \\ & \text{ or } \widetilde{n}-1 \leq \widetilde{p} \leq \widetilde{n}+1, \\ \frac{p}{n-p-1} & \text{for } p \leq n-2 \text{ and } \widetilde{p} \leq \widetilde{n}-2, \\ \frac{p}{n-p-1} \cdot \frac{\widetilde{n}}{\widetilde{p}} & \text{for } p \leq n-2 \text{ and } \widetilde{p} \geq \widetilde{n}+2, \\ \frac{n}{p-n-1} & \text{for } p \geq n+2 \text{ and } \widetilde{p} \leq \widetilde{n}-2, \\ \frac{n}{p-n-1} \cdot \frac{\widetilde{n}}{\widetilde{p}} & \text{for } p \geq n+2 \text{ and } \widetilde{p} \geq \widetilde{n}+2. \end{cases}
\end{aligned}
\tag{82}
$$

At this intermediate stage, one can use the results provided in (77)-(82) to formulate the error in the subvector induced by the specific $\mathcal{F}$ of interest, namely, this part of the error is represented by the third term in (76).

We proceed to the formulation of the error in the subvector induced by the specific $\mathcal{T}$ of interest, i.e., the fourth error term in (76)

$$
\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\right\} = \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\} + \left\|\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2 - 2\mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^T\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\}
\tag{83}
$$

Note that the first term in (83) was already developed in (80) into a more explicit form. The second term in (83) remains as it is because $\mathcal{T}$ is fixed in the current setting. The third term in (83) can be developed using (74) into

$$\mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\} = \mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}\mathbf{H}\boldsymbol{\beta}\right\} = \mathbb{E}\left\{\boldsymbol{\beta}_{\mathcal{T}}^{T}\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})}\right\}$$
$$= \langle\boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})},\boldsymbol{\beta}_{\mathcal{T}}\rangle \times \begin{cases}1 & \text{for } \widetilde{p}\le\widetilde{n},\\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p}>\widetilde{n}.\end{cases} \tag{84}$$

Also recall the definition of the *normalized task correlation in* $\mathcal{T}$ between the two tasks that was defined in the main text before Theorem 5 as $\rho_{\mathcal{T}} \triangleq \frac{\langle\boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})},\boldsymbol{\beta}_{\mathcal{T}}\rangle}{\kappa_{\mathcal{T}}}$ for $t>0$ and that $\kappa_{\mathcal{T}} = \left\|\boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})}\right\|_2^2 + t\sigma_{\eta}^2$.

Setting the results from (77)-(84) into (76) provides the formulations in Theorem 5, namely,

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

where

$$\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = \mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}}\right\|_2^2\right\} - \kappa_{\mathcal{T}} \times \left(1 + 2\left(\rho_{\mathcal{T}}-1\right)\times\left(\begin{cases}1 & \text{for } \widetilde{p}\le\widetilde{n},\\ \frac{\widetilde{n}}{\widetilde{p}} & \text{for } \widetilde{p}>\widetilde{n}\end{cases}\right)\right) \tag{85}$$

for $t>0$, and $\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = 0$ for $t=0$.

### F.3. Proof of Corollary 6

The main part in the proof of Corollary 6 is to develop a more explicit form for $\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\}$, as presented next.

$$\mathbb{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\right\|_2^2\right\} = \mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\left(\mathbf{Z}\mathbf{H}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\xi}\right)\right\|_2^2\right\} \tag{86}$$
$$= \mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}}\left(\boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})} + \boldsymbol{\eta}_{\mathcal{S}}\right)\right\|_2^2\right\}$$
$$+ \mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\mathbf{Z}_{\mathcal{S}^c}\left(\boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} + \boldsymbol{\eta}_{\mathcal{S}^c}\right)\right\|_2^2\right\} + \mathbb{E}\left\{\left\|\mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\boldsymbol{\xi}\right\|_2^2\right\}$$

that using (72)-(73), (75) leads to

$$\mathbb{E}\left\{ \left\Vert \mathbf{Q}_{\mathcal{T}}\mathbf{Q}_{\mathcal{S}}^{T}\mathbf{Z}_{\mathcal{S}}^{+}\left(\mathbf{Z}\mathbf{H}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\eta}+\boldsymbol{\xi}\right)\right\Vert_2^2 \right\} = \tag{87}$$

$$= \left( \begin{cases} \left\Vert \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})} \right\Vert_2^2 + t\sigma_{\eta}^2 & \text{for } \widetilde{p} \le \widetilde{n}, \\ \frac{\widetilde{n}}{\widetilde{p}(\widetilde{p}+1)} \left( \left( \widetilde{n} + \frac{\widetilde{n}-1}{\widetilde{p}-1} \right) \left( \left\Vert \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})} \right\Vert_2^2 + t\sigma_{\eta}^2 \right) + \left( 1 - \frac{\widetilde{n}-1}{\widetilde{p}-1} \right) t \left( \left\Vert \boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})} \right\Vert_2^2 + \widetilde{p}\sigma_{\eta}^2 \right) \right) & \text{for } \widetilde{p} > \widetilde{n}, \end{cases} \right)$$

$$+ \frac{t}{\widetilde{p}} \left( \left\Vert \boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} \right\Vert_2^2 + (d-\widetilde{p})\sigma_{\eta}^2 + \sigma_{\xi}^2 \right) \times \left( \begin{cases} \frac{\widetilde{p}}{\widetilde{n}-\widetilde{p}-1} & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}-\widetilde{n}-1} & \text{for } \widetilde{p} \ge \widetilde{n}+2, \end{cases} \right)$$

$$= \left( \left\Vert \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})} \right\Vert_2^2 + t\sigma_{\eta}^2 \right) \times \begin{cases} 1 + t \cdot \frac{\left\Vert \boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} \right\Vert_2^2 + (d-\widetilde{p})\sigma_{\eta}^2 + \sigma_{\xi}^2}{(\widetilde{n}-\widetilde{p}-1)\left( \left\Vert \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})} \right\Vert_2^2 + t\sigma_{\eta}^2 \right)} & \text{for } \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}} \left( \frac{\left(\widetilde{p}^2 - \widetilde{n}\widetilde{p}\right)\psi_{\mathcal{T}} + \widetilde{n}\widetilde{p} - 1}{\widetilde{p}^2-1} + t \cdot \frac{\left\Vert \boldsymbol{\beta}_{\mathcal{S}^c}^{(\mathbf{H})} \right\Vert_2^2 + (d-\widetilde{p})\sigma_{\eta}^2 + \sigma_{\xi}^2}{(\widetilde{p}-\widetilde{n}-1)\left( \left\Vert \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})} \right\Vert_2^2 + t\sigma_{\eta}^2 \right)} \right) & \text{for } \widetilde{p} \ge \widetilde{n}+2. \end{cases} \tag{88}$$

where $\psi_{\mathcal{T}} \triangleq \frac{t}{\widetilde{p}} \frac{\left\Vert \boldsymbol{\beta}_{\mathcal{S}}^{(\mathbf{H})} \right\Vert_2^2 + \widetilde{p}\sigma_{\eta}^2}{\left\Vert \boldsymbol{\beta}_{\mathcal{T}}^{(\mathbf{H})} \right\Vert_2^2 + t\sigma_{\eta}^2}$ was defined for $t > 0$ and $\widetilde{p} > 0$ before Theorem 5 in the main text.

Then, we use the last result together with the relation between $\mathbb{E}\left\{ \left\Vert \widehat{\boldsymbol{\theta}}_{\mathcal{T}} \right\Vert_2^2 \right\}$ and $\mathbb{E}\left\{ \left\Vert \widehat{\boldsymbol{\theta}}_{\mathcal{T}} - \boldsymbol{\theta}_{\mathcal{T}} \right\Vert_2^2 \right\}$ from (80) and the formulation of $\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})}$ from (85) to get that

$$\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = \kappa_{\mathcal{T}} \times \begin{cases} 1 - 2\rho_{\mathcal{T}} + t\frac{\zeta_{\mathcal{S}^c} + \sigma_{\xi}^2}{(\widetilde{n}-\widetilde{p}-1)\kappa_{\mathcal{T}}} & \text{for } 1 \le \widetilde{p} \le \widetilde{n}-2, \\ \infty & \text{for } \widetilde{n}-1 \le \widetilde{p} \le \widetilde{n}+1, \\ \frac{\widetilde{n}}{\widetilde{p}} \left( \frac{\left(\widetilde{p}^2 - \widetilde{n}\widetilde{p}\right)\psi_{\mathcal{T}} + \widetilde{n}\widetilde{p} - 1}{\widetilde{p}^2-1} - 2\rho_{\mathcal{T}} + t\frac{\zeta_{\mathcal{S}^c} + \sigma_{\xi}^2}{(\widetilde{p}-\widetilde{n}-1)\kappa_{\mathcal{T}}} \right) & \text{for } \widetilde{p} \ge \widetilde{n}+2 \end{cases}$$

for $t > 0$, and $\Delta\mathcal{E}_{\text{transfer}}^{(\mathcal{T},\mathcal{S})} = 0$ for $t = 0$. The last expression utilizes the definitions provided in the main text before Theorem 5.

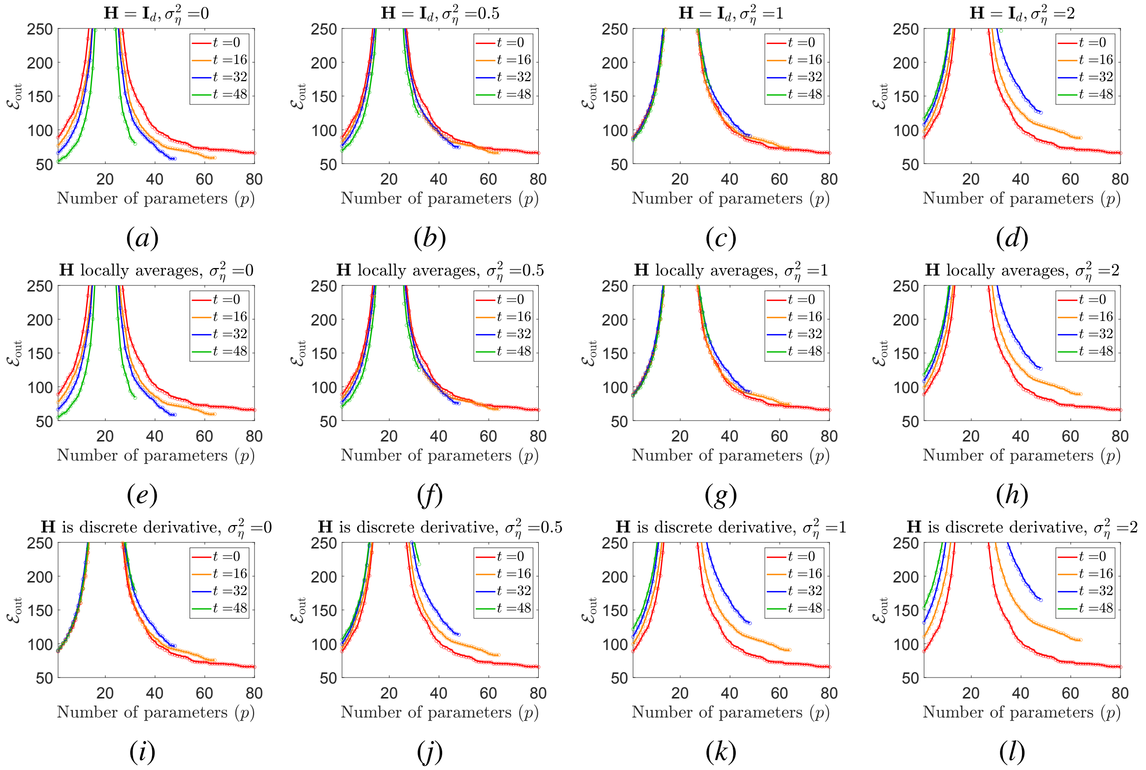
## Appendix G. Transfer of Specific Sets of Parameters: Additional Analytical and Empirical Evaluations of Generalization Error Curves

The following results are for two different forms of the true solution $\boldsymbol{\beta}$: the first is a form with linearly increasing values (Fig. 4(*a*)), the second is a form with sparse values where only 25% of coordinates have non-zero value (Fig. 4(*e*)). Note that both forms satisfy $\Vert\boldsymbol{\beta}\Vert_2^2 = d$.

The three types of linear operator $\mathbf{H}$ in the evaluations of Section 4 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 11-12 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 $\sigma_\eta^2$). It is clear from Figures 11-12 that the increase in $\sigma_\eta^2$, 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. 11-12 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 $\sigma_\eta^2$ 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. 11-12) 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. 11). 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 the case these benefits exist for $\mathbf{H} = \mathbf{I}_d$ (e.g., compare third to first row of subfigures in Fig. 12).


$\mathbf{H} = \mathbf{I}_d, \sigma_\eta^2 = 0$ $\mathbf{H} = \mathbf{I}_d, \sigma_\eta^2 = 0.5$ $\mathbf{H} = \mathbf{I}_d, \sigma_\eta^2 = 1$ $\mathbf{H} = \mathbf{I}_d, \sigma_\eta^2 = 2$

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

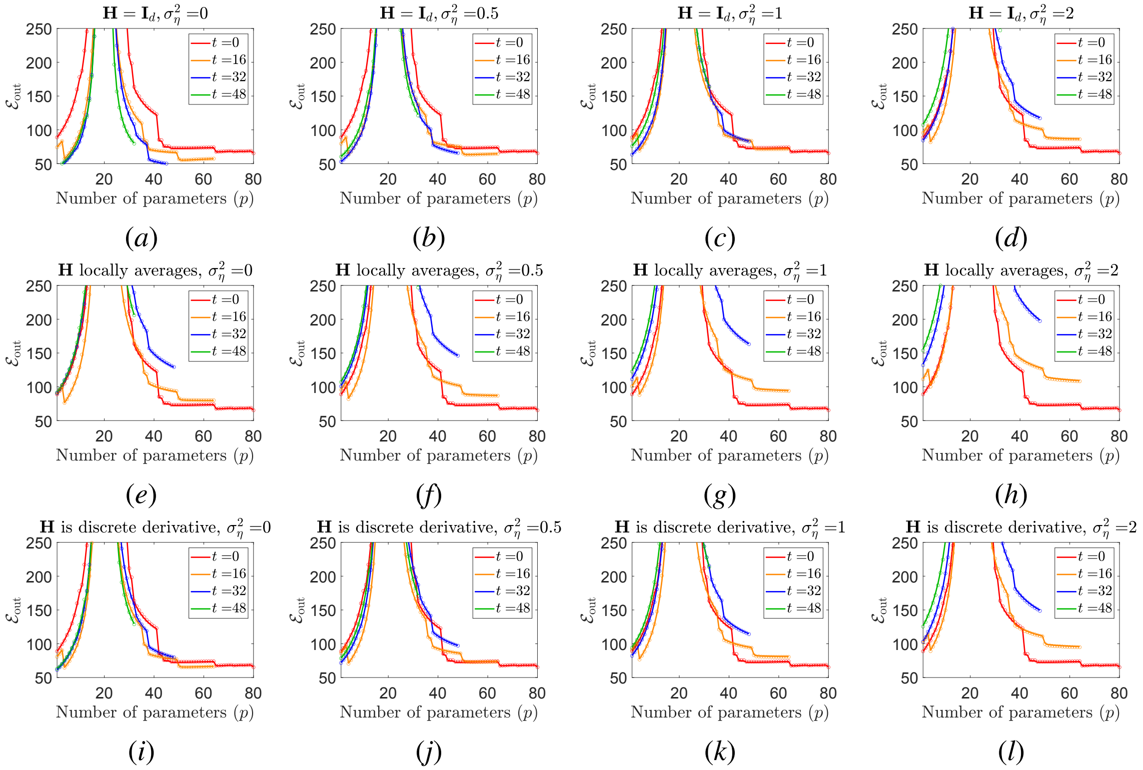
$\mathbf{H}$ locally averages, $\sigma_\eta^2 = 0$ $\mathbf{H}$ locally averages, $\sigma_\eta^2 = 0.5$ $\mathbf{H}$ locally averages, $\sigma_\eta^2 = 1$ $\mathbf{H}$ locally averages, $\sigma_\eta^2 = 2$

(*e*) (*f*) (*g*) (*h*)

$\mathbf{H}$ is discrete derivative, $\sigma_\eta^2 = 0$ $\mathbf{H}$ is discrete derivative, $\sigma_\eta^2 = 0.5$ $\mathbf{H}$ is discrete derivative, $\sigma_\eta^2 = 1$ $\mathbf{H}$ is discrete derivative, $\sigma_\eta^2 = 2$

(*i*) (*j*) (*k*) (*l*)

Figure 11: 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 (6) between the source and target tasks: each column of subfigures has a different $\sigma_\eta^2$ value, and each row of subfigures corresponds to a different linear operator $\mathbf{H}$. The analytical values, induced from Theorem 5, are presented using solid-line curves, and the respective empirical results obtained from averaging over 750 experiments are denoted by circle markers. Each curve color refers to a different number of transferred parameters.



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

(*e*) (*f*) (*g*) (*h*)

(*i*) (*j*) (*k*) (*l*)

Figure 12: 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 (6) between the source and target tasks: each column of subfigures has a different $\sigma_\eta^2$ value, and each row of subfigures corresponds to a different linear operator $\mathbf{H}$. The analytical values, induced from Theorem 5, are presented using solid-line curves, and the respective empirical results obtained from averaging over 750 experiments are denoted by circle markers. Each curve color refers to a different number of transferred parameters.