Phase Transitions in Transfer Learning for High-Dimensional Perceptrons

Oussama Dhifallah and Yue M. Lu

Abstract

Transfer learning seeks to improve the generalization performance of a target task by exploiting the knowledge learned from a related source task. Central questions include deciding what information one should transfer and when transfer can be beneficial. The latter question is related to the so-called negative transfer phenomenon, where the transferred source information actually reduces the generalization performance of the target task. This happens when the two tasks are sufficiently dissimilar. In this paper, we present a theoretical analysis of transfer learning by studying a pair of related perceptron learning tasks. Despite the simplicity of our model, it reproduces several key phenomena observed in practice. Specifically, our asymptotic analysis reveals a phase transition from negative transfer to positive transfer as the similarity of the two tasks moves past a well-defined threshold.

I. Introduction

Transfer learning [1]–[5] is a promising approach to improving the performance of machine learning tasks. It does so by exploiting the knowledge gained from a previously-learned model, referred to as the *source task*, to improve the generalization performance of a related learning problem, referred to as the *target task*. One particular challenge in transfer learning is to avoid the so-called *negative transfer* [6]–[9], where the transferred source information reduces the generalization performance of the target task. Recent literature [6]–[9] shows that negative transfer is closely related to the similarity between the source and target tasks. Transfer learning may hurt the generalization performance if the tasks are sufficiently dissimilar.

In this paper, we present a theoretical analysis of transfer learning by studying a pair of related perceptron learning tasks. Despite the simplicity of our model, it reproduces several key phenomena observed in practice. Specifically, the model reveals a sharp phase transition from negative transfer to positive transfer (i.e. when transfer becomes helpful) as a function of the model similarity.

A. Models and Learning Formulations

We start by describing the models for our theoretical study. We assume that the source task has a collection of training data $\{(\boldsymbol{a}_{s,i}, y_{s,i})\}_{i=1}^{n_s}$, where $\boldsymbol{a}_{s,i} \in \mathbb{R}^p$ is the source feature vector and $y_{s,i} \in \mathbb{R}$ denotes the label corresponding to $\boldsymbol{a}_{s,i}$. Following the standard teacher-student paradigm, we shall assume that the labels $\{y_{s,i}\}_{i=1}^{n_s}$ are generated according to the following model

$$y_{s,i} = \varphi(\boldsymbol{a}_{s,i}^\top \boldsymbol{\xi}_s), \ \forall\ i \in \{1, \dots, n_s\}, \tag{1}$$

where $\varphi(\cdot)$ is a scalar deterministic or probabilistic function and $\boldsymbol{\xi}_s \in \mathbb{R}^p$ is an unknown *source teacher vector*.

Similar to the source task, the target task has access to a different collection of training data $\{(\boldsymbol{a}_{t,i}, y_{t,i})\}_{i=1}^{n_t}$, generated according to

$$y_{t,i} = \varphi(\boldsymbol{a}_{t,i}^\top \boldsymbol{\xi}_t), \ \forall\ i \in \{1, \dots, n_t\}. \tag{2}$$

Here, $\boldsymbol{\xi}_t \in \mathbb{R}^p$ is an unknown *target teacher vector*. We measure the (dis)similarity of the two tasks by

$$\rho \overset{\text{def}}{=} \frac{\boldsymbol{\xi}_t^\top \boldsymbol{\xi}_s}{\|\boldsymbol{\xi}_t\| \|\boldsymbol{\xi}_s\|},$$

O. Dhifallah and Y. M. Lu are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA (e-mails: oussama_dhifallah@g.harvard.edu,yuelu@seas.harvard.edu).

This research was funded by the Harvard FAS Dean's Fund for Promising Scholarship, and by the US National Science Foundations under grants CCF-1718698 and CCF-1910410.

with $\rho = 0$ indicating two uncorrelated tasks, whereas $\rho = 1$ means that the tasks are perfectly aligned.

For the source task, we learn the optimal weight vector $\widehat{\boldsymbol{w}}_s$ by solving a convex optimization problem

$$\widehat{\boldsymbol{w}}_s = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\operatorname{argmin}} \ \frac{1}{p}\sum_{i=1}^{n_s} \ell\left(y_{s,i}; \boldsymbol{a}_{s,i}^\top \boldsymbol{w}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 . \tag{3}$$

Here, $\lambda \geq 0$ is a regularization parameter, and $\ell(.;.)$ denotes some general loss function that can take one of the following two forms

$$\begin{cases} \ell\left(y;x\right) = \widehat{\ell}(y-x), & \text{for regression task} \\ \ell\left(y;x\right) = \widehat{\ell}(yx), & \text{for classification task,} \end{cases} \tag{4}$$

where $\widehat{\ell}(.)$ is a convex function.

In this paper, we consider a common strategy in transfer learning [4] which consists of transferring the optimal source vector, i.e. $\widehat{\boldsymbol{w}}_s$, to the target task. One popular approach is to fix a (random) subset of the target weights to values of the corresponding optimal weights learned during the source training process [10]. In our learning model, this amounts to the following target learning formulation:

$$\widehat{\boldsymbol{w}}_t = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{p}\sum_{i=1}^{n_t} \ell\left(y_{t,i}; \boldsymbol{a}_{t,i}^\top \boldsymbol{w}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \tag{5}$$

$$\text{s.t.} \quad \boldsymbol{Q}\boldsymbol{w} = \boldsymbol{Q}\widehat{\boldsymbol{w}}_s. \tag{6}$$

Here, $\widehat{\boldsymbol{w}}_s$ is the optimal solution of the source learning problem, and $\boldsymbol{Q} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with diagonal entries drawn independently from a Bernoulli distribution with probability $\delta \leq 1$. Thus, on average, we are retaining δp number of entries from the source optimal vector $\widehat{\boldsymbol{w}}_s$. In addition to possible improvement to the generalization performance, this approach can considerably lower the computational complexity of the target learning task by reducing the number of free optimization variables. In what follows, we refer to δ as the *transfer rate* and call (5) the *hard transfer* formulation.

Another popular approach in transfer learning is to search for target weight vectors in the vicinity of the optimal source weight vector $\widehat{\boldsymbol{w}}_s$. This can be achieved by adding a regularization term to the target formulation [11], [12], which in our model becomes

$$\widehat{\boldsymbol{w}}_t = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{p}\sum_{i=1}^{n_t} \ell\left(y_{t,i}; \boldsymbol{a}_{t,i}^\top \boldsymbol{w}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\left\|\boldsymbol{\Sigma}(\boldsymbol{w} - \widehat{\boldsymbol{w}}_s)\right\|^2 , \tag{7}$$

with $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ denoting some weighting matrix. In what follows, we refer to (7) as the *soft transfer* formulation, since it relaxes the strict equality in (5). In fact, the hard transfer in (5) is just a special case of the soft transfer formulation, if we set $\boldsymbol{\Sigma}$ to be a diagonal matrix whose diagonal entries are either $+\infty$ (with probability δ) or 0 (with probability $1-\delta$).

To measure the performance of the transfer learning methods, we use the generalization error of the target task. Given a new data sample $(\boldsymbol{a}_{t,\text{new}}, y_{t,\text{new}})$ with $y_{t,\text{new}} = \varphi(\boldsymbol{\xi}_t^\top \boldsymbol{a}_{t,\text{new}})$, we assume that the target task predicts the corresponding label as

$$\widehat{y}_{t,\text{new}} = \widehat{\varphi}[\widehat{\boldsymbol{w}}_t^\top \boldsymbol{a}_{t,\text{new}}], \tag{8}$$

where $\widehat{\varphi}(\cdot)$ is a pre-defined scalar function that might be different from $\varphi(\cdot)$. We then calculate the generalization error of the target task as

$$\mathcal{E}_{\text{test}} = \frac{1}{4^\upsilon}\mathbb{E}\left[\left(y_{t,\text{new}} - \widehat{\varphi}(\widehat{\boldsymbol{w}}_t^\top \boldsymbol{a}_{t,\text{new}})\right)^2\right], \tag{9}$$

where the expectation is taken with respect to the new data $(\boldsymbol{a}_{t,\text{new}}, y_{t,\text{new}})$. The variable υ allows us to write a more compact formula: υ is taken to be 0 for a regression problem and $\upsilon = 1$ for a binary classification problem. Finally, we use the training error

$$\mathcal{E}_{\text{train}} = \frac{1}{p}\sum_{i=1}^{n_t} \ell\left(y_{t,i}; \boldsymbol{a}_{t,i}^\top \widehat{\boldsymbol{w}}_t\right) + \frac{1}{2}\left\|\boldsymbol{\Sigma}(\widehat{\boldsymbol{w}}_t - \widehat{\boldsymbol{w}}_s)\right\|^2 ,$$

to quantify the performance of the training process.

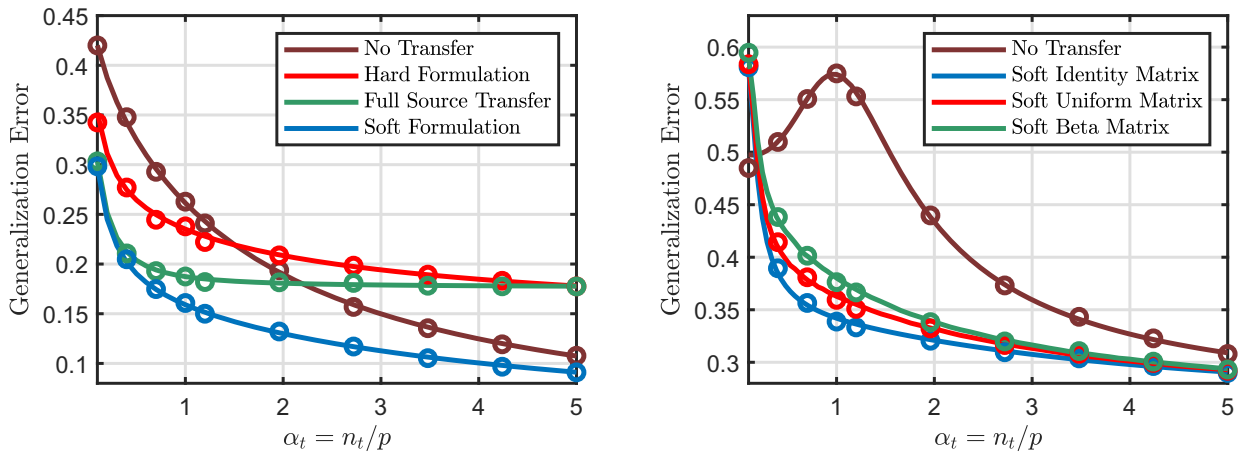

Fig. 1. Theoretical predictions v.s. numerical simulations obtained by averaging over 100 independent Monte Carlo trials with dimension $p = 2500$. **(a)** Binary classification with logistic loss. We take $\alpha_s = 10\alpha_t$, $\lambda = 0.3$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p/\sqrt{5}$ and $\rho = 0.85$, where $\alpha_s = n_s/p$ and $\alpha_t = n_t/p$. The functions $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are both the sign function. For hard transfer, we set the transfer rate to be $\delta = 0.5$. *Full source transfer* corresponds to $\delta = 1.0$, whereas *no transfer* corresponds to $\delta = 0$. **(b)** Nonlinear regression using quadratic loss, where $\varphi(\cdot)$ is the ReLu function and $\widehat{\varphi}(\cdot)$ is the identity function. Soft identity, beta and uniform matrices refer to different choices of the weighting matrix in (7). They correspond to setting $\boldsymbol{\Sigma}$ to be an identity matrix, and a random matrix with diagonal elements drawn from the beta and uniform distributions, respectively. We scale all diagonal elements of $\boldsymbol{\Sigma}$ to have the same mean. We also take $\alpha_s = 10\alpha_t$, $\lambda = 0.1$, and $\rho = 0.8$.

B. Main Contributions

The main contributions of this paper are two-fold, as summarized below:

1) Precise Asymptotic Analysis: We present a precise asymptotic analysis of the transfer learning approaches introduced in (5) and (7) for Gaussian feature vectors. Specifically, we show that, as the dimensions p, n_s, n_t grow to infinity with the ratios $\alpha_s = n_s/p, \alpha_t = n_t/p$ fixed, the generalization errors of the hard and soft formulations can be exactly characterized by the solutions of two low-dimensional *deterministic* optimization problems. (See Theorem 1 and Corollary 1 for details.) Our asymptotic predictions hold for any convex loss functions used in the training process, including the squared loss for regression problems and logistic loss commonly used for binary classification problems.

As illustrated in Figure 1, our theoretical predictions (drawn as solid lines in the figures) reach excellent agreement with the actual performance (shown as circles) of the transfer learning problem. Figure 1(a) considers a binary classification setting with logistic loss, and we plot the generalization errors of different transfer approaches as a function of the target data/dimension ratio $\alpha_t = n_t/p$. We can see that the hard transfer formulation (5) is only useful when α_t is small. In fact, we encounter negative transfer (i.e. hard transfer performing worse than no transfer) when α_t becomes sufficiently large. Moreover, the soft transfer formulation (7) seems to achieve more favorable generalization errors as compared to the hard formulation. In Figure 1(b), we consider a regression setting with a squared loss, and explore the impact of different weighting schemes on the performance of the soft formulation. We can see that the soft formulation indeed considerably improves the generalization performance of the standard learning method (i.e. learning the target task without any knowledge transfer).

2) Phase Transitions: Our asymptotic characterizations reveal a phase transition phenomenon in the hard transfer formulation. Let

$$\delta^\star = \underset{0\le\delta\le 1}{\operatorname{argmin}}\ \mathcal{E}_{\text{test}}(\delta),$$

be the optimal transfer rate that minimizes the generalization error of the target task. Clearly, $\delta^\star = 0$ corresponds to the negative transfer regime, where transferring the knowledge of the source task will actually hurt the performance of the target task. In contract, $\delta^\star > 0$ signifies that we have entered the positive transfer regime, where transfer becomes helpful.

Figure 2(a) illustrates the phase transition from negative to positive transfer regimes in a binary classification setting, as the similarity ρ between the two tasks moves past a critical threshold. Similar phase transition phenomena

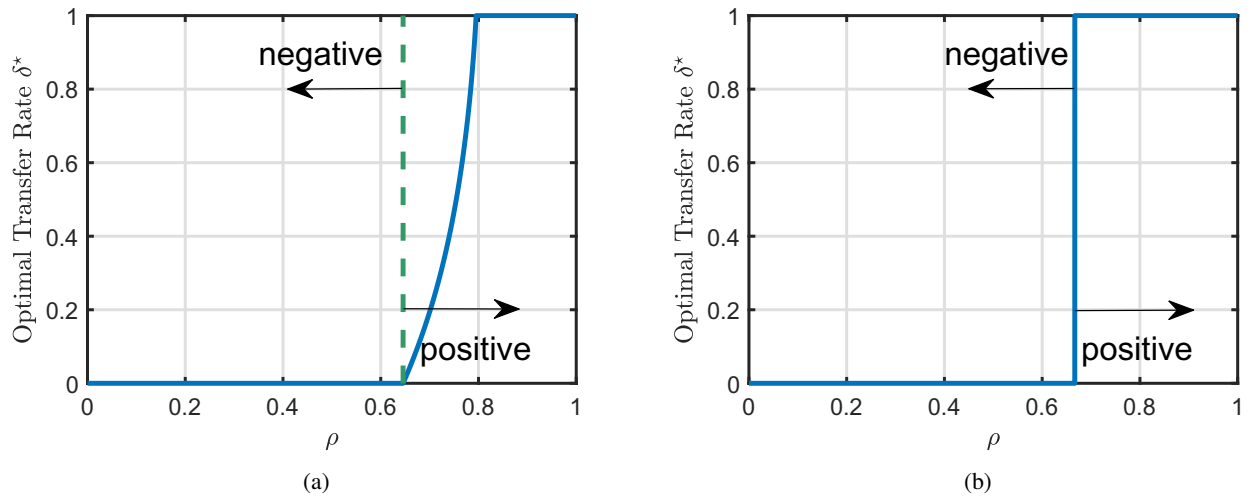

Fig. 2. Phase transitions of the hard transfer formulation. When the similarity ρ between the two tasks is small, we are in the negative transfer regime, where we should not transfer the knowledge from the source task. However, as ρ moves past a critical threshold, we enter the positive transfer regime. **(a)** Binary classification with squared loss, with parameters $\alpha_t = 2$, $\alpha_s = 2\alpha_t$ and $\lambda = 0$. Both $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are the sign function. **(b)** Nonlinear regression with squared loss, with parameters $\alpha_t = 2$, $\alpha_s = 2\alpha_t$, and $\lambda = 0$. $\varphi(.)$ is the ReLu function and $\widehat{\varphi}(.)$ is the identity function.

also appear in nonlinear regression, as shown in Figure 2(b). Interestingly, for this setting the optimal transfer rate *jumps* from $\delta^\star = 0$ to $\delta^\star = 1$ at the transition threshold.

For general loss functions, the exact locations of the phase transitions can only be found numerically by solving the deterministic optimization problems in our asymptotic characterizations. For the special case of squared loss with no regularization, however, we are able to obtain the following simple analytical characterization for the phase transition threshold: We are in the positive transfer regime *if and only if*

$$\rho > \rho_c(\alpha_s, \alpha_t) = 1 - \frac{\mathbb{E}[\varphi^2(z)] - \mathbb{E}^2[z\varphi(z)]}{2\,\mathbb{E}^2[z\varphi(z)]}\Big(\frac{1}{\alpha_t - 1} - \frac{1}{\alpha_s - 1}\Big), \tag{10}$$

where z is a standard Gaussian random variable. This result is shown in Proposition 1.

By the Cauchy-Schwarz inequality, $\mathbb{E}[\varphi^2(z)] \geq \mathbb{E}^2[z\varphi(z)]$. It follows that $\rho_c(\alpha_s, \alpha_t)$ is an increasing function of α_t and a decreasing function of α_s. This property is consistent with our intuition: As we increase α_t, the target task has more training data to work with, and thus we should set a higher bar in terms of when to transfer knowledge; As we increase α_s, the quality of the optimal source vector becomes better, in which case we can start doing the transfer at a lower similarity level. In particular, when $\alpha_t > \alpha_s$, we have $\rho_c(\alpha_s, \alpha_t) > 1$ and thus the inequality in (10) is never satisfied (because $|\rho| \leq 1$ by definition). This indicates that no transfer should be done when the target task has more training data than the source task.

C. Related Work

The idea of transferring informaton between different domains or different tasks was first proposed in [1] and further developed in [2]. It has been attracting significant interest in recent literature [4]–[9], [11], [12]. While most work focuses on the practical aspects of transfer learning, there have been several studies (e.g., [13], [14]) that seek to provide analytical understandings of transfer learning in simplified models. Our work is particularly related to [14], which considers a transfer learning model similar to ours, but for the special case of linear regression. The analysis in this paper is more general as it considers arbitrary convex loss functions. We would also like to mention an interesting recent work that studies a different but related setting referred to as knowledge distillation [15].

In term of technical tools, our asymptotic predictions are derived using the convex Gaussian min-max theorem (CGMT). The CGMT is first introduced in [16] and further developed in [17]. It extends a Gaussian comparison inequality first introduced in [18]. It particularly uses convexity properties to show the equivalence between two Gaussian processes. The CGMT has been successfully used to analyze convex regression formulations [17], [19], [20] and convex classification formulations [21]–[24].

D. Organization

The rest of this paper is organized as follows. Section II states the technical assumptions under which our results are obtained. Section III provides an asymptotic characterization of the soft transfer formulation. The precise analysis of the hard transfer formulation is presented in Section IV. Our theoretical predictions hold for general convex loss functions. We specialize these results to the settings of nonlinear regression and binary classification in Section V, where we also provide additional numerical results to validate our predictions. Section VI provides the detailed proof of the technical statements introduced in Sections III and IV. Section VII concludes the paper. The Appendix provides additional technical details.

II. Technical Assumptions

The theoretical analysis of this paper is carried out under the following assumptions.

Assumption 1 (Gaussian Feature Vectors): The feature vectors $\{\boldsymbol{a}_{s,i}\}_{i=1}^{n_s}$ and $\{\boldsymbol{a}_{t,i}\}_{i=1}^{n_t}$ are drawn independently from a standard Gaussian distribution. The vector $\boldsymbol{\xi}_s \in \mathbb{R}^p$ can be expressed as $\boldsymbol{\xi}_s = \rho\boldsymbol{\xi}_t + \sqrt{1-\rho^2}\boldsymbol{\xi}_r$, where the vectors $\boldsymbol{\xi}_t \in \mathbb{R}^p$ and $\boldsymbol{\xi}_r \in \mathbb{R}^p$ are independent from the feature vectors, and they are generated independently from a uniform distribution on the unit sphere.

Define m as the number of transferred entries in (5). Our results are valid in the high-dimensional asymptotic setting where the dimensions p, n_s, n_t and m grow to infinity at fixed ratios.

Assumption 2 (High-dimensional Asymptotic): The number of samples and the number of transferred components in hard transfer satisfy $n_s = n_s(p)$, $n_t = n_t(p)$ and $m = m(p)$ with $\alpha_{s,p} = n_s(p)/p \to \alpha_s > 0$, $\alpha_{t,p} = n_t(p)/p \to \alpha_t > 0$ and $\delta_p = m(p)/p \to \delta > 0$ as $p \to \infty$.

Assumption 3 (Loss Function): The loss function $\ell(y;.)$ defined in (4) is a proper convex function in $\mathbb{R}$. Moreover, define a random function $\mathcal{L}(\boldsymbol{x}) = \sum_{i=1}^{n_t} \ell(y_i; x_i)$, where $y_i \sim \varphi(z_i)$, with $\{z_i\}$ being a collection of independent standard normal random variables. Denote by $\partial\mathcal{L}$ the sub-differential set of $\mathcal{L}(\boldsymbol{x})$. Then there exists a universal constant $C > 0$ such that

$$\mathbb{P}\Big(\sup_{\|\boldsymbol{v}\| \le C\sqrt{n_t}} \sup_{\boldsymbol{s} \in \partial\mathcal{L}(\boldsymbol{v})} \|\boldsymbol{s}\| \le C\sqrt{n_t}\Big) \xrightarrow{p\to\infty} 1.$$

Furthermore, we consider the following assumption to guarantee that the generalization error defined in (9) concentrates in the large system limit.

Assumption 4 (Regularity Conditions): The data generating function $\varphi(.)$ is independent from the feature vectors. Moreover, the following conditions are satisfied.

- $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are continuous almost everywhere in $\mathbb{R}$. For every $h > 0$ and $z \sim \mathcal{N}(0, h)$, we have $0 < \mathbb{E}[\varphi^2(z)] < +\infty$ and $0 < \mathbb{E}[\widehat{\varphi}^2(z)] < +\infty$.
- For any compact interval $[c, C]$, there exists a function $g(\cdot)$ such that

$$\sup_{h\in[c,C]} \big|\widehat{\varphi}(hx)\big|^2 \le g(x) \quad \text{for all } x \in \mathbb{R}.$$

  Additionally, the function $g(\cdot)$ satisfies $\mathbb{E}[g^2(z)] < +\infty$, where $z \sim \mathcal{N}(0, 1)$.

Finally, we introduce the following assumption to guarantee that the training and generalization errors of the soft formulation can be asymptotically characterized by deterministic optimization problems.

Assumption 5 (Weighting Matrix): Let $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is the weighting matrix in the soft transfer formulation. Let $\sigma_{\max}(\boldsymbol{\Lambda})$ denote its largest eigenvalue, and let $\sigma_{\min,1}(\boldsymbol{\Lambda})$ and $\sigma_{\min,2}(\boldsymbol{\Lambda})$ denote its two smallest eigenvalues. There exist two constants $\mu_{\min} \ge 0$ and $\mu_{\max} \ge 0$ such that

$$\begin{cases} \mathbb{P}(\sigma_{\max}(\boldsymbol{\Lambda}) \le \mu_{\max}) \xrightarrow{n\to\infty} 1 \\ \sigma_{\min,1}(\boldsymbol{\Lambda}) \xrightarrow{p} \mu_{\min} \\ \big|\sigma_{\min,1}(\boldsymbol{\Lambda}) - \sigma_{\min,2}(\boldsymbol{\Lambda})\big| \xrightarrow{p} 0. \end{cases} \tag{11}$$

Moreover, we assume that the empirical distribution of the eigenvalues of the matrix $\boldsymbol{\Lambda}$ converges weakly to a probability distribution $\mathbb{P}_\mu(.)$ supported in $[\mu_{\min}\ \mu_{\max}]$.

III. Sharp Asymptotic Analysis of the Soft Transfer Formulation

In this section, we study the asymptotic properties of the soft transfer formulation. Specifically, we provide a precise characterization of the training and generalization errors corresponding to (7).

The asymptotic performance of the source formulation defined in (3) has been studied in the literature [24]. In particular, it has been shown that the asymptotic limit of the source formulation in (3) can be quantified by the following deterministic optimization problem:

$$\min_{q_s,r_s\geq 0}\sup_{\sigma>0}\ \alpha_s\mathbb{E}\Big[\mathcal{M}_{\ell(Y_s,.)}\Big(r_sH_s+q_sS_s;\frac{r_s}{\sigma}\Big)\Big]-\frac{r_s\sigma}{2}+\frac{\lambda}{2}(q_s^2+r_s^2). \tag{12}$$

Here, $Y_s=\varphi(S_s)$, and H_s and S_s are two independent standard Gaussian random variables. Furthermore, the function $\mathcal{M}_{\ell(Y_s,.)}$ introduced in the scalar optimization problem (12) is the Moreau envelope function defined as

$$\mathcal{M}_{\ell(y,.)}(a;b)=\min_{c\in\mathbb{R}}\ell(y;c)+\frac{1}{2b}(c-a)^2. \tag{13}$$

The expectation in (12) is taken over the random variables H_s, S_s.

In our work, we focus on the target problem with soft transfer, as formulated in (7). It turns out that the asymptotic performance of the target problem can also be characterized by a deterministic optimization problem:

$$\begin{aligned}\min_{q_t,r_t\geq 0}\sup_{\sigma>-\mu_{\min}}&-\frac{\sigma r_t^2}{2}+\frac{1}{2}\Big((1-\rho^2)(q_s^\star)^2+(r_s^\star)^2\Big)T_2(\sigma)\\&+\alpha_t\mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_tH_t+q_tS_t;T_1(\sigma)\Big)\Big]+\frac{\lambda}{2}(q_t^2+r_t^2)\\&-\frac{1}{2}\left(q_t-\rho q_s^\star\right)^2\left(\sigma-1/T_1(\sigma)\right).\end{aligned} \tag{14}$$

Here, $Y_t=\varphi(S_t)$, and H_t and S_t are independent standard Gaussian random variables. Additionally, $\mu_{\min}$ represents the minimum value of the random variable with distribution $\mathbb{P}_\mu(.)$ as defined in Assumption 5. In the formulation (14), the constants $q_s^\star$ and $r_s^\star$ are the optimal solutions of the asymptotic formulation given in (12). Moreover, the functions $T_1(.)$ and $T_2(.)$ are defined as follows:

$$T_1(\sigma)=\mathbb{E}_\mu[1/(\mu+\sigma)],\ T_2(\sigma)=\mathbb{E}_\mu\left[\mu\sigma/(\mu+\sigma)\right],$$

where the expectations are taken over the probability distribution $\mathbb{P}_\mu(.)$ defined in Assumption 5.

Theorem 1 (Precise Analysis of the Soft Transfer): Suppose that the Assumptions 1, 2, 3, 4, and 5 are satisfied. Then, the training error corresponding to the soft transfer formulation in (7) converges in probability as follows

$$\mathcal{E}_{\text{train}}\xrightarrow{p\to\infty}C_t^\star-\frac{\lambda}{2}\big((q_t^\star)^2+(r_t^\star)^2\big), \tag{15}$$

where $C_t^\star$, $q_t^\star$ and $r_t^\star$ are the optimal objective value and the optimal solution of the scalar formulation in (14), respectively. Moreover, the generalization error introduced in (9) corresponding to the soft transfer formulation converges in probability as follows

$$\mathcal{E}_{\text{test}}\xrightarrow{p\to\infty}\frac{1}{4^v}\mathbb{E}\left[\left(\varphi(\nu_1)-\widehat{\varphi}(\nu_2)\right)^2\right], \tag{16}$$

where ν_1 and ν_2 are two jointly Gaussian random variables with zero mean and a covariance matrix given by

$$\begin{bmatrix}1 & q_t^\star\\ q_t^\star & (q_t^\star)^2+(r_t^\star)^2\end{bmatrix}.$$

The proof of Theorem 1 is based on the CGMT framework [17, Theorem 6.1]. The detailed proof is provided in Section VI-C. The statements in Theorem 1 are valid for a general convex loss function and general learning models that can be expressed as in (1) and (2). The analysis in Section VI-C shows that the deterministic problems in (12) and (14) are the asymptotic limits of the source and target formulations given in (3) and (7), respectively. Moreover, it shows that the deterministic problems (12) and (14) are strictly convex in the minimization variables and concave in the maximization variables. This implies the uniqueness of the optimal solutions of the minimization problems.

Remark 1: The results of the theorem show that the training and generalization errors corresponding to the soft transfer formulation can be fully characterized using the optimal solutions of the scalar formulation in (14). Moreover, from its definition, (14) depends on the optimal solutions of the scalar formulation in (12) of the source task. This shows that the precise asymptotic performance of the soft transfer formulation can be characterized after solving two scalar deterministic problems.

IV. Sharp Asymptotic Analysis of Hard Transfer Formulation

In this section, we study the asymptotic properties of the hard transfer formulation. We then use these predictions to rigorously prove the existence of phase transitions from negative to positive transfer.

A. Asymptotic Predictions

As mentioned earlier, the hard transfer formulation can be recovered from (7) as a special case where the eigenvalues of the matrix $\mathbf{\Lambda}$ are $+\infty$ with probability δ and 0 otherwise. Thus, we obtain the following result as a simple consequence of Theorem 1.

Corollary 1: Suppose that the Assumptions 1, 2, 3 and 4 are satisfied. Then, the asymptotic limit of the hard formulation defined in (5) is given by the following deterministic formulation

$$\begin{aligned}\min_{q_t, r_t \geq 0} \sup_{\sigma > 0} \; & \frac{\lambda}{2}(q_t^2 + r_t^2) + \frac{\sigma\delta}{2}\left[(1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\right] \\ & + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_t H_t + q_t S_t; \frac{1-\delta}{\sigma}\Big)\Big] - \frac{\sigma r_t^2}{2} \\ & + \frac{\sigma\delta}{2(1-\delta)}\left(q_t - \rho q_s^\star\right)^2. \end{aligned} \tag{17}$$

Additionally, the training and generalization errors associated with the hard formulation converge in probability to the limits given in (15) and (16), respectively.

B. Phase Transitions

As illustrated in Figure 2, there is a phase transition phenomenon in the hard transfer formulation, where the problem moves from negative transfer to positive transfer as the similarity of the source and target tasks increases. For general loss functions, the exact location of the phase transition boundary can only be determined by numerically solving the scalar optimization problem in (17).

For the special case of squared loss, however, we are able to obtain analytical expressions. For the rest of this section, we restrict our discussions to the following special settings:

(a) The loss function $\ell(\cdot,\cdot)$ in (3) and (5) is the squared loss, i.e. $\ell(y,x) = \frac{1}{2}(y-x)^2$.
(b) The regularization strength $\lambda = 0$ in the source and target formulations (3) and (5).
(c) The data/dimension ratios α_s and α_t satisfy $\alpha_s > 1$ and $\alpha_t > 1$.

We first consider a nonlinear regression task, where the function $\varphi(\cdot)$ in the generative models (1) and (2) can be arbitrary, and the function $\widehat{\varphi}(\cdot)$ in (8) is the identity function.

Proposition 1 (Regression Phase Transition): In addition to the conditions (a)–(c) introduced above, assume that the pre-defined function $\widehat{\varphi}(\cdot)$ in (8) is the identity function. Let $\delta^\star$ be the optimal transfer rate that leads to the lowest generalization error in the hard formulation (5). Then,

$$\delta^\star = \begin{cases} 0 & \text{if } \rho < \rho_c(\alpha_s, \alpha_t) \\ 1 & \text{if } \rho > \rho_c(\alpha_s, \alpha_t), \end{cases} \tag{18}$$

where $\rho_c(\alpha_s, \alpha_t)$ is defined in (10).

The result of Proposition 1, whose proof can be found in Section VI-D1, shows that $\rho_c(\alpha_s, \alpha_t)$ is the phase transition boundary separating the negative transfer regime from the positive transfer regime. When the similarity metric $\rho < \rho_c(\alpha_s, \alpha_t)$, the optimal transfer ratio $\delta^\star = 0$, indicating that we should not transfer any source knowledge. Transfer becomes helpful only when ρ moves past the threshold. Note that for this particular model, there is also

an interesting feature that the optimal $\delta^\star$ jumps to 1 in the positive transfer phase, meaning that we should fully copy the source weight vector.

Next, we consider a binary classification task, where the nonlinear functions $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are both the sign function. We first define a function

$$g(\alpha_t, \alpha_s) = 1 - \frac{(1-\frac{2}{\pi})\alpha_t(\alpha_s - \alpha_t)}{(\alpha_s - 1)\left[\frac{4}{\pi}(\alpha_t - 1)\alpha_t + 2(1-\frac{2}{\pi})(\alpha_t - 1)\right]}. \tag{19}$$

Proposition 2 (Classification): Assume that the conditions (a)-(c) introduced above hold, and both $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are the sign function. Then

$$\delta^\star > 0 \quad \text{if} \quad \rho > g(\alpha_t, \alpha_s). \tag{20}$$

We prove this result at the end of Section VI. Unlike (18), the result in (20) only provides a *sufficient* condition for when the hard transfer is beneficial. Nevertheless, our numerical simulations show that the sufficient condition in (20) is actually the correct phase transition boundary for the majority of parameter settings for α_t, α_s.

V. Additional Simulation Results

In this section, we provide additional simulation examples to confirm our asymptotic analysis and illustrate the phase transition phenomenon. In our experiments, we focus on the regression and classification models.

A. Model Assumptions

For the regression model, we assume that the source, target and test data are generated according to

$$y_i = \max(\boldsymbol{a}_i^\top \boldsymbol{\xi}, 0), \ \forall i \in \{1, \dots, n\}. \tag{21}$$

The data $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^n$ can be the training data of the source or target tasks. In this regression model, we assume that the function $\widehat{\varphi}(.)$ is the identity function, i.e. $\widehat{\varphi}(x) = x$. Then, the generalization error corresponding to the soft formulation converges in probability as follows

$$\mathcal{E}_{\text{test}} \xrightarrow{p\to\infty} v - 2cq_t^\star + ((q_t^\star)^2 + (r_t^\star)^2),$$

where c and v are defined as follows

$$c = \mathbb{E}[z \max(z, 0)], \ \ v = \mathbb{E}[\max(z, 0)^2].$$

Here, z is a standard Gaussian random variable and $q_t^\star$ and $r_t^\star$ are defined in Theorem 1. Additionally, the asymptotic limit of the generalization error corresponding to the hard formulation can be expressed in a similar fashion.

For the binary classification model, we assume that the source, target and test data labels are binary and generated as follows:

$$y_i = \text{sign}(\boldsymbol{a}_i^\top \boldsymbol{\xi}), \ \forall i \in \{1, \dots, n\}. \tag{22}$$

Here, the data $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^n$ can be the training data of the source and target tasks. In this classification model, the objective is to predict the correct sign of any unseen sample y_{new}. Then, we fix the function $\widehat{\varphi}(.)$ to be the sign function. Following Theorem 1, it can be easily shown that the generalization error corresponding to the soft formulation given in (7) converges in probability as follows

$$\mathcal{E}_{\text{test}} \xrightarrow{p\to\infty} \frac{1}{\pi} \cos^{-1}\left(\frac{q_t^\star}{\sqrt{(q_t^\star)^2 + (r_t^\star)^2}}\right).$$

Here, $q_t^\star$ and $r_t^\star$ are the optimal solutions of the target scalar formulation given in (14). The generalization error corresponding to the hard formulation given in (5) can be expressed in a similar fashion.

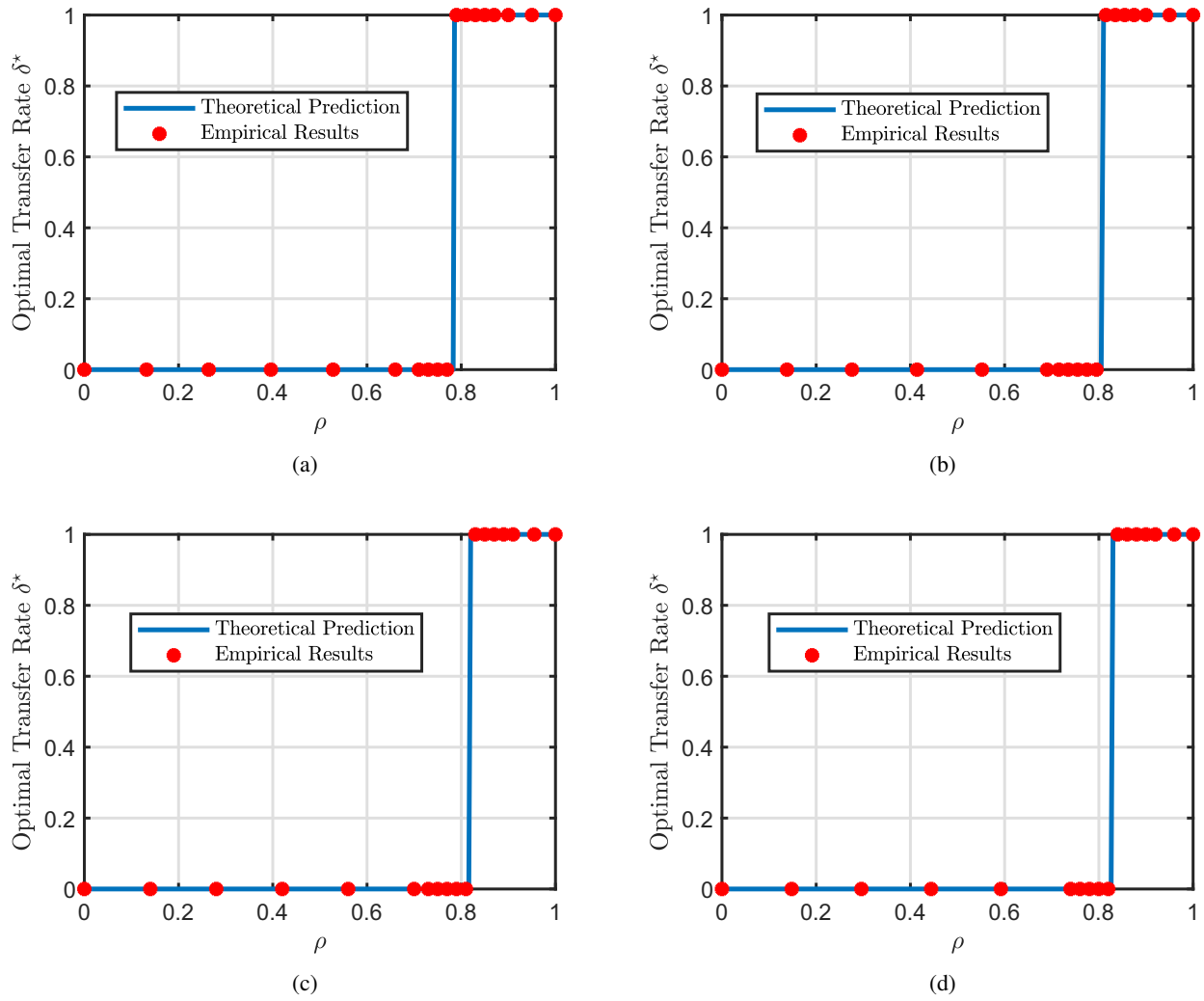

Fig. 3. Additional illustrations of the phase transition phenomenon. **(a)** Regression (squared loss, $\alpha_t = 0.5$, and $\alpha_s = 3\alpha_t$) **(b)** Regression (squared loss, $\alpha_t = 2$, and $\alpha_s = 2\alpha_t$) **(c)** Binary classification (squared loss, $\alpha_t = 1.5$, and $\alpha_s = 3\alpha_t$) **(d)** Binary classification (hinge loss, $\alpha_t = 1.5$, and $\alpha_s = 3\alpha_t$). In all the experiments, we set the regularization strength to be $\lambda = 0.1$. The blue line represents our theoretical predictions of the optimal transfer rate obtained by solving our asymptotic results in Section IV for multiple values of δ. The empirical results are averaged over 100 independent Monte Carlo trials with $p = 2500$.

B. Phase Transitions in the Hard Formulation

In Section IV, we have presented analytical formulas for the phase transition phenomenon, but only for the special case of squared loss with no regularization. The main purpose of this experiment, shown in Figure 3, is to demonstrate that the phase transition phenomenon still takes place in more general settings with different loss functions and regularization strengths.

In all the cases shown in Figure 3, the transition from negative to positive transfer is a discontinuous jump from standard learning (i.e. no transfer) to full source transfer. Additionally, Figures 3(c) and 3(d) show that the loss function has a small effect on the phase transition boundary.

C. Soft Transfer: Impact of the Weighting Matrix and Regularization Strength

In this experiment, we empirically explore the impact of the weighting matrix $\boldsymbol{\Sigma}$ on the generalization error corresponding to the soft formulation. We focus on the binary classification problem with logistic loss. The weighting matrix in (7) takes the following form

$$\boldsymbol{\Sigma} = \sqrt{\beta_t}\boldsymbol{V}, \tag{23}$$

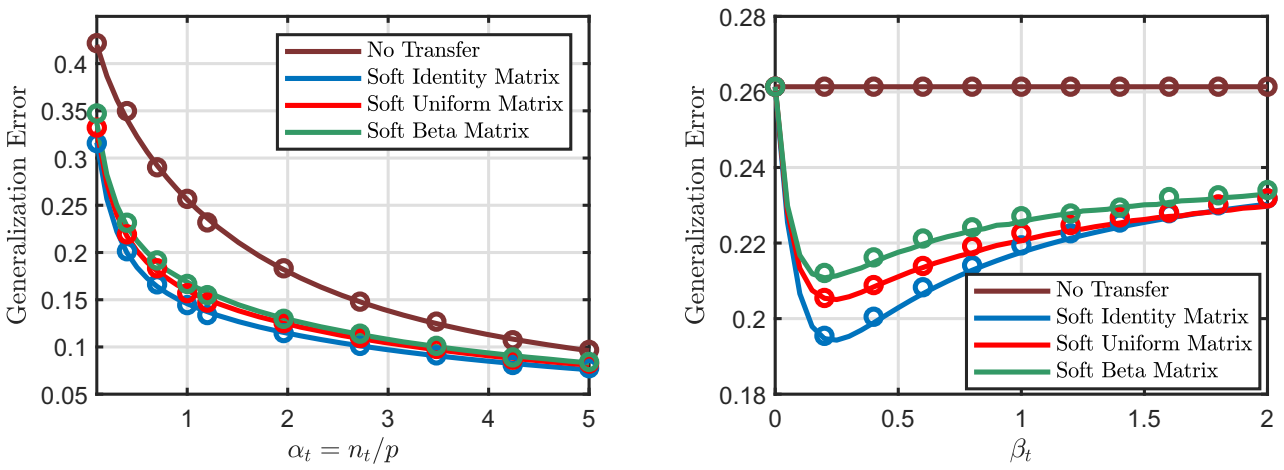

Fig. 4. Continuous line: Theoretical predictions. Circles: numerical simulations. **(a)** $\alpha_s = 6\alpha_t$, $\lambda = 0.1$, $\beta_t = 1/10$ and $\rho = 0.9$. **(b)** $\alpha_t = 1$, $\alpha_s = 5\alpha_t$, $\lambda = 0.3$ and $\rho = 0.75$. In all the experiments, we consider the binary classification problem with the logistic loss function. The empirical results are averaged over 50 independent Monte Carlo trials and we set $p = 1000$.

where $\boldsymbol{V}$ is a diagonal matrix generated in three different ways. (1) *Soft Identity*: $\boldsymbol{V}$ is an identity matrix; (2) *Soft Uniform*: the diagonal entries of $\boldsymbol{V}$ are drawn independently from the uniform distribution and then scaled to have their mean equal to 1; (3): *Soft Beta*: similar to (2), but with the diagonal entries drawn from the beta distribution, followed by rescaling to unit mean.

Figure 4(a) shows that the considered weighting matrix choices have similar generalization performance, with the identity matrix being slightly better than the other alternatives. Moreover, Figure 4(b) illustrates the effects of the parameter β_t in (23) on the generalization performance. It points to the interesting possibility of "designing" the optimal weight matrix to minimize the generalization error.

D. Soft and Hard Transfer Comparison

In the last simulation example, we consider the regression model and compare the performance of the hard and soft transfer formulations as a function of α_t and ρ.

Figure 5(a) shows that the soft formulation provides the best generalization performance for all values of α_t. Moreover, we can see that the hard transfer formulation is only useful for small values α_t. Figure 5(b) shows that the performance of the soft and hard transfer formulations depend on the similarity between the source and target tasks. Specifically, the generalization performances of different transfer approaches all improve as we increase the similarity measure ρ. We can also see that the full source transfer approach provides the lowest generalization error when the similarity measure is close to 1, while the soft transfer method leads to the best generalization performance at moderate values of the similarity measure. At very small values of ρ, which means that the two tasks share little resemblance, the standard learning method (i.e. no transfer) is the best scheme one should use.

VI. Technical Details

In this section, we provide a detailed proof of Theorem 1, Corollary 1 and Proportions 1 and 2. Specifically, we focus on analyzing the generalized formulation in (7) using the CGMT framework introduced in the following part.

A. Technical Tool: Convex Gaussian Min-Max Theorem

The CGMT provides an asymptotic equivalent formulation of primary optimization (PO) problems of the following form

$$\Phi_p(\boldsymbol{G}) = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \boldsymbol{u}^\top \boldsymbol{G}\boldsymbol{w} + \psi(\boldsymbol{w},\boldsymbol{u}). \tag{24}$$

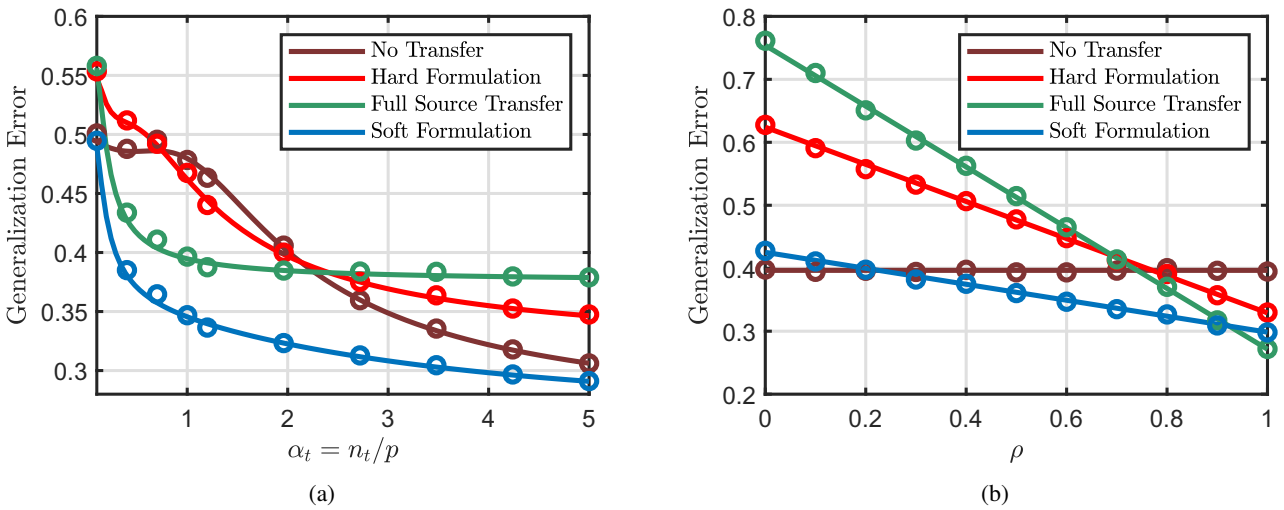

Fig. 5. Continuous line: Theoretical predictions. Circles: numerical simulations. **(a)** $\alpha_s = 12\alpha_t$, $\lambda = 0.2$ and $\rho = 0.75$. **(b)** $\alpha_t = 1.5$, $\alpha_s = 8\alpha_t$ and $\lambda = 0.4$. In all the experiments, we consider the regression setting with a squared loss. The hard transfer formulation uses $\delta = 0.5$, and the soft transfer formulation uses an identity weighting matrix. The empirical results are averaged over 50 independent Monte Carlo trials and we set $p = 1000$.

Specifically, the CGMT shows that the PO given in (24) is asymptotically equivalent to the following formulation

$$\phi_p(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w} \in \mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u} \in \mathcal{S}_{\boldsymbol{u}}} \|\boldsymbol{u}\| \, \boldsymbol{g}^\top \boldsymbol{w} + \|\boldsymbol{w}\| \, \boldsymbol{h}^\top \boldsymbol{u} + \psi(\boldsymbol{w}, \boldsymbol{u}),$$

referred to as the auxiliary optimization (AO) problem. Before showing the equivalence between the PO and AO, the CGMT assumes that $\boldsymbol{G} \in \mathbb{R}^{n \times p}$, $\boldsymbol{g} \in \mathbb{R}^p$ and $\boldsymbol{h} \in \mathbb{R}^n$, all have i.i.d standard normal entries, the feasibility sets $\mathcal{S}_{\boldsymbol{w}} \subset \mathbb{R}^p$ and $\mathcal{S}_{\boldsymbol{u}} \subset \mathbb{R}^n$ are convex and compact, and the function $\psi(.,.) : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}$ is continuous *convex-concave* on $\mathcal{S}_{\boldsymbol{w}} \times \mathcal{S}_{\boldsymbol{u}}$. Moreover, the function $\psi(.,.)$ is independent of the matrix $\boldsymbol{G}$. Under these assumptions, the CGMT [17, Theorem 6.1] shows that for any $\chi \in \mathbb{R}$ and $\zeta > 0$, it holds

$$\mathbb{P}\left(\left|\Phi_p(\boldsymbol{B}) - \chi\right| > \zeta\right) \leq 2\mathbb{P}\left(\left|\phi_p(\boldsymbol{g}, \boldsymbol{h}) - \chi\right| > \zeta\right). \tag{25}$$

Additionally, the CGMT [17, Theorem 6.1] provides the following conditions under which the optimal solutions of the PO and AO concentrates around the same set.

Theorem 2 (CGMT Framework): Consider an open set $\mathcal{S}_{p,\epsilon}$. Moreover, define the set $\mathcal{S}^c_{p,\epsilon} = \mathcal{S}_{\boldsymbol{w}} \setminus \mathcal{S}_{p,\epsilon}$. Let ϕ_p and ϕ^c_p be the optimal cost values of the AO formulation in (25) with feasibility sets $\mathcal{S}_{\boldsymbol{w}}$ and $\mathcal{S}^c_{p,\epsilon}$, respectively. Assume that the following properties are all satisfied

(1) There exists a constant ϕ such that the optimal cost ϕ_p converges in probability to ϕ as p goes to $+\infty$.
(2) There exists a positive constant $\zeta > 0$ such that $\phi^c_p \geq \phi + \zeta$ with probability going to 1 as $p \to +\infty$, for any fixed $\epsilon > 0$.

Then, the following convergence in probability holds

$$\left|\Phi_p - \phi_p\right| \xrightarrow{p} 0, \text{ and } \mathbb{P}(\widehat{\boldsymbol{w}}_p \in \mathcal{S}_{p,\epsilon}) \xrightarrow{p \to \infty} 1,$$

for any fixed $\epsilon > 0$, where Φ_p and $\widehat{\boldsymbol{w}}_p$ are the optimal cost and the optimal solution of the PO formulation in (24). Theorem 2 allows us to analyze the generally easy AO problem to infer asymptotic properties of the generally hard PO problem. Next, we use the CGMT to rigorously prove the technical results presented in Theorem 1.

B. Precise Analysis of the Source Formulation

The source formulation defined in (3) is well–studied in recent literature [25]. Specifically, it has been rigorously proved that the performance of the source formulation can be fully characterized after solving the following scalar formulation

$$\min_{q_s, r_s \geq 0} \sup_{\sigma > 0} \alpha_s \mathbb{E}\Big[\mathcal{M}_{\ell(Y_s,.)}\Big(r_s H_s + q_s S_s; \frac{r_s}{\sigma}\Big)\Big] - \frac{r_s \sigma}{2} + \frac{\lambda}{2}(q_s^2 + r_s^2). \tag{26}$$

Here, $Y_s = \varphi(S_s)$, and H_s and S_s are two independent standard Gaussian random variables. The expectation in (26) is taken over the random variables H_s and S_s. Furthermore, the function $\mathcal{M}_{\ell(Y_s,.)}$ introduced in the scalar optimization problem (26) is the Moreau envelope function defined in (13).

C. Precise Analysis of the Soft Transfer Approach

In this part, we provide a precise asymptotic analysis of the generalized transfer formulation given in (7). Specifically, we focus on analyzing the following formulation

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{p} \sum_{i=1}^{n_t} \ell\left(y_i; \boldsymbol{a}_i^\top \boldsymbol{w}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\big\|\boldsymbol{\Sigma}(\boldsymbol{w} - \widehat{\boldsymbol{w}}_s)\big\|^2,$$

where $\widehat{\boldsymbol{w}}_s$ is the optimal solution of the source formulation given in (3). Note that the vector $\widehat{\boldsymbol{w}}_s$ is independent of the training data of the target task. For simplicity of notation, we denote by $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^{n_t}$, the training data of the target task. Here, we use the CGMT framework introduced in Section VI-A to precisely analyze the above formulation.

1) Formulating the Auxiliary Optimization Problem: Our first objective is to rewrite the generalized formulation in the form of the PO problem given in (24). To this end, we introduce additional optimization variables. Specifically, the generalized formulation can be equivalently formulated as follows

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \max_{\boldsymbol{u} \in \mathbb{R}^{n_t}} \frac{1}{p}\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{w} - \frac{1}{p}\sum_{i=1}^{n_t} \ell^\star\left(y_i; u_i\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\big\|\boldsymbol{\Sigma}(\boldsymbol{w} - \widehat{\boldsymbol{w}}_s)\big\|^2. \tag{27}$$

Here, the optimization vector $\boldsymbol{u} \in \mathbb{R}^{n_t}$ is formed as $\boldsymbol{u} = [u_1, \ldots, u_{n_t}]^\top$, the data matrix $\boldsymbol{A} \in \mathbb{R}^{n_t \times p}$ is given by $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^\top$. Additionally, the function $\ell^\star(y; .)$ denotes the convex conjugate function of the loss function $\ell(y; .)$. First, observe that the CGMT framework assumes that the feasibility sets of the minimization and maximization problems are compact. Then, our next step is to show that the formulation given in (27) satisfies this assumption.

Lemma 1 (Primal–Dual Compactness): Assume that $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{u}}$ are the optimal solutions of the optimization problem in (27). Then, there exist two constants $C_w > 0$ and $C_u > 0$ such that the following convergence in probability holds

$$\mathbb{P}(\|\widehat{\boldsymbol{w}}\| \leq C_w) \xrightarrow{p} 1, \ \mathbb{P}(\|\widehat{\boldsymbol{u}}\|/\sqrt{n_t} \leq C_u) \xrightarrow{p} 1. \tag{28}$$

The proof of Lemma 1 is omitted since it follows the same steps of the results presented in [20, Lemma 1] and [20, Lemma 2]. The proof of the above result follows using Assumption 3, Assumption 5 and the asymptotic results in [26, Theorem 2.1] to prove the compactness of the optimal solution $\widehat{\boldsymbol{w}}$. Then, use the result in [27, Proposition 11.3] and Assumption 3 to show the compactness of the optimal dual vector $\widehat{\boldsymbol{u}}$.

The theoretical result in Lemma 1 shows that the optimization problem in (27) can be equivalently formulated with compact feasibility sets on events with probability going to one. Then, it suffices to study the constrained version of (27). Note that the data labels $\{y_i\}_{i=1}^{n_t}$ depend on the data matrix $\boldsymbol{A}$. Then, one can decompose the matrix $\boldsymbol{A}$ as follows

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}_t} + \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp} = \boldsymbol{A}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top + \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}.$$

Here, the matrix $\boldsymbol{P}_{\boldsymbol{\xi}_t} \in \mathbb{R}^{p\times p}$ denotes the projection matrix onto the space spanned by the vector $\boldsymbol{\xi}_t$, and the matrix $\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp} = \boldsymbol{I}_p - \boldsymbol{\xi}_t\boldsymbol{\xi}_t^{\top}$ denotes the projection matrix onto the orthogonal complement of the space spanned by the vector $\boldsymbol{\xi}_t$. Note that we can express $\boldsymbol{A}$ as follows without changing its statistics

$$\boldsymbol{A} = \boldsymbol{s}_t\boldsymbol{\xi}_t^{\top} + \boldsymbol{G}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}, \tag{29}$$

where $\boldsymbol{s}_t \sim \mathcal{N}(0, \boldsymbol{I}_{n_t})$ and the components of the matrix $\boldsymbol{G} \in \mathbb{R}^{n_t\times p}$ are drawn independently from a standard Gaussian distribution and where $\boldsymbol{s}_t$ and $\boldsymbol{G}$ are independent. This means that the formulation in (27) can be expressed as follows

$$\begin{aligned}\min_{\|\boldsymbol{w}\|\leq C_w}\max_{\boldsymbol{u}\in\mathcal{C}_t} &\frac{1}{p}\boldsymbol{u}^{\top}\boldsymbol{G}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}\boldsymbol{w} + \frac{1}{p}\boldsymbol{u}^{\top}\boldsymbol{s}_t\boldsymbol{\xi}_t^{\top}\boldsymbol{w} + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\\ &- \frac{1}{p}\sum_{i=1}^{n_t}\ell^{\star}\left(y_i;u_i\right) + \frac{1}{2}\left\|\boldsymbol{\Sigma}(\boldsymbol{w}-\widehat{\boldsymbol{w}}_s)\right\|^2,\end{aligned} \tag{30}$$

where the set $\mathcal{C}_t = \{\boldsymbol{u} : \|\boldsymbol{u}\|/\sqrt{n_t} \leq C_u\}$. Note that the formulation in (30) is in the form of the primary formulation given in (24). Here, the function $\psi(.,.)$ is defined as follows

$$\begin{aligned}\psi(\boldsymbol{w},\boldsymbol{u}) = &\frac{1}{p}\boldsymbol{u}^{\top}\boldsymbol{s}_t\boldsymbol{\xi}_t^{\top}\boldsymbol{w} + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \frac{1}{p}\sum_{i=1}^{n_t}\ell^{\star}\left(y_i;u_i\right)\\ &+ \frac{1}{2}\left\|\boldsymbol{\Sigma}(\boldsymbol{w}-\widehat{\boldsymbol{w}}_s)\right\|^2.\end{aligned} \tag{31}$$

One can easily see that the optimization problem in (30) has compact convex feasibility sets. Moreover, the function $\psi(.,.)$ is continuous, convex–concave and independent of the Gaussian matrix $\boldsymbol{G}$. This shows that the assumptions of the CGMT are all satisfied by the primary formulation in (30). Then, following the CGMT framework, the auxiliary formulation corresponding to our primary problem in (30) can be expressed as follows

$$\begin{aligned}\min_{\|\boldsymbol{w}\|\leq C_w}\max_{\boldsymbol{u}\in\mathcal{C}_t} &\frac{\|\boldsymbol{u}\|}{p}\boldsymbol{g}^{\top}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}\boldsymbol{w} + \frac{1}{p}\boldsymbol{u}^{\top}\boldsymbol{s}_t\boldsymbol{\xi}_t^{\top}\boldsymbol{w} + \frac{\boldsymbol{h}^{\top}\boldsymbol{u}}{p}\left\|\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}\boldsymbol{w}\right\|\\ &+ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \frac{1}{p}\sum_{i=1}^{n_t}\ell^{\star}\left(y_i;u_i\right) + \frac{1}{2}\left\|\boldsymbol{\Sigma}(\boldsymbol{w}-\widehat{\boldsymbol{w}}_s)\right\|^2,\end{aligned} \tag{32}$$

where $\boldsymbol{g} \in \mathbb{R}^p$ and $\boldsymbol{h} \in \mathbb{R}^{n_t}$ are two independent standard Gaussian vectors. The rest of the proof focuses on simplifying the obtained AO formulation and study its asymptotic properties.

2) Simplifying the AO Problem of the Target Task: Here, we focus on simplifying the auxiliary formulation corresponding to the target task. We start our analysis by decomposing the target optimization vector $\boldsymbol{w} \in \mathbb{R}^p$ as follows

$$\boldsymbol{w} = (\boldsymbol{\xi}_t^{\top}\boldsymbol{w})\boldsymbol{\xi}_t + \boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp}\boldsymbol{r}_t. \tag{33}$$

Here, $\boldsymbol{r}_t \in \mathbb{R}^{p-1}$ is a free vector, $\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp} \in \mathbb{R}^{p\times(p-1)}$ is formed by an orthonormal basis orthogonal to the vector $\boldsymbol{\xi}_t$. Now, define the variable q_t as follows $q_t = \boldsymbol{\xi}_t^{\top}\boldsymbol{w}$. Based on the result in Lemma 1 and the decomposition in (33), there exists $C_{q_t} > 0$, $C_r > 0$ and $C_u > 0$ such that our auxiliary formulation can be asymptotically expressed in terms of the variables q_t and $\boldsymbol{r}_t$ as follows

$$\begin{aligned}\min_{(q_t,\boldsymbol{r}_t)\in\mathcal{T}_1}\max_{\boldsymbol{u}\in\mathcal{C}_t} &\frac{\|\boldsymbol{u}\|}{p}\boldsymbol{g}^{\top}\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp}\boldsymbol{r}_t + \frac{\|\boldsymbol{r}_t\|}{p}\boldsymbol{h}^{\top}\boldsymbol{u} + \frac{q_t}{p}\boldsymbol{u}^{\top}\boldsymbol{s}_t + \frac{\lambda}{2}q_t^2\\ &+ \frac{\lambda}{2}\|\boldsymbol{r}_t\|^2 - \frac{1}{p}\sum_{i=1}^{n_t}\ell^{\star}\left(y_i;u_i\right) + \frac{1}{2}q_t^2V_{p,t} - q_tV_{p,ts}\\ &+ \frac{1}{2}\boldsymbol{r}_t^{\top}(\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp})^{\top}\boldsymbol{\Lambda}\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp}\boldsymbol{r}_t + q_t\boldsymbol{\xi}_t^{\top}\boldsymbol{\Lambda}\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp}\boldsymbol{r}_t - \boldsymbol{r}_t^{\top}(\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp})^{\top}\boldsymbol{\Lambda}\widehat{\boldsymbol{w}}_s,\end{aligned}$$

Here, we drop terms independent of the optimization variables and the matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ is defined as $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{\top}\boldsymbol{\Sigma}$. Additionally, the feasibility set $\mathcal{T}_1$ is defined as follows

$$\mathcal{T}_1 = \Big\{(q_t,\boldsymbol{r}_t) : |q_t| \leq C_{q_t}, \|\boldsymbol{r}_t\| \leq C_r\Big\}. \tag{34}$$

Here, the sequence of random variables $V_{t,n}$ and $V_{ts,n}$ are defined as follows

$$V_{p,t} = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{\xi}_t, \;\; V_{p,ts} = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \widehat{\boldsymbol{w}}_s. \tag{35}$$

Next, we focus on simplifying the obtained auxiliary formulation. Our strategy is to solve over the direction of the optimization vector $\boldsymbol{r} \in \mathbb{R}^{p-1}$. This step requires the interchange of a non–convex minimization and a non–concave maximization. We can easily justify the interchange using the theoretical result in [17, Lemma A.3]. The main argument is that the strong convexity of the primary formulation in (30) allows us to perform such interchange in the corresponding auxiliary formulation. The optimization problem over the vector $\boldsymbol{r}_t$ with fixed norm, i.e. $\|\boldsymbol{r}_t\| = r_t$, can be formulated as follows

$$C_p^\star = \min_{\boldsymbol{r}_t \in \mathbb{R}^{p-1}} \;\; \boldsymbol{b}_p^\top \boldsymbol{r}_t + \frac{1}{2} \boldsymbol{r}_t^\top \boldsymbol{\Lambda}^\perp \boldsymbol{r}_t, \quad \text{s.t.} \;\; \|\boldsymbol{r}_t\| = r_t, \tag{36}$$

Here, we ignore constant terms independent of $\boldsymbol{r}_t$, the matrix $\boldsymbol{\Lambda}^\perp \in \mathbb{R}^{(p-1)\times(p-1)}$ and the vector $\boldsymbol{b}_p \in \mathbb{R}^{p-1}$ can be expressed as follows

$$\begin{aligned} \boldsymbol{\Lambda}^\perp = (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp, \;\; \boldsymbol{b}_p = \frac{\|\boldsymbol{u}\|}{p} (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{g} + q_t (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \boldsymbol{\xi}_t^\top \\ - (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \widehat{\boldsymbol{w}}_s. \end{aligned}$$

The optimization problem in (36) is non–convex given the norm equality constraint. It is well–studied in the literature [28] and is known as the trust region subproblem. Using the same analysis in [20], the optimal cost value of the optimization problem (36) can be expressed in terms of a one-dimensional optimization problem as follows

$$C_p^\star = \sup_{\sigma > -\mu_p} \left\{ -\frac{1}{2} \boldsymbol{b}_p^\top [\boldsymbol{\Lambda}^\perp + \sigma \boldsymbol{I}_{p-1}]^{-1} \boldsymbol{b}_p - \frac{\sigma r_t^2}{2} \right\}, \tag{37}$$

where μ_p is the minimum eigenvalue of the matrix $\boldsymbol{\Lambda}^\perp$, denoted by $\sigma_{\min}(\boldsymbol{\Lambda}^\perp)$. This result can be seen by equivalently formulating the non–convex problem in (36) as follows

$$C_p^\star = \min_{\boldsymbol{r}_t \in \mathbb{R}^{p-1}} \max_{\sigma \in \mathbb{R}} \;\; \boldsymbol{b}_p^\top \boldsymbol{r}_t + \frac{1}{2} \boldsymbol{r}_t^\top \boldsymbol{\Lambda}^\perp \boldsymbol{r}_t + \frac{\sigma}{2} \left(\|\boldsymbol{r}_t\|^2 - r_t^2 \right).$$

Then, show that the optimal σ satisfies a constraint that preserves the convexity over $\boldsymbol{w}$. This allows us to interchange the maximization and minimization and solve over the vector $\boldsymbol{w}$. The above analysis shows that the AO formulation corresponding to our primary problem can be expressed as follows

$$\begin{aligned} &\min_{(q_t, r_t) \in \mathcal{T}_2} \max_{\boldsymbol{u} \in \mathcal{C}_t} \sup_{\sigma > -\mu_p} \frac{r_t}{p} \boldsymbol{h}^\top \boldsymbol{u} + \frac{q_t}{p} \boldsymbol{u}^\top \boldsymbol{s}_t + \frac{\lambda}{2} q_t^2 + \frac{\lambda}{2} r_t^2 \\ &- \frac{1}{p} \sum_{i=1}^{n_t} \ell^\star \left(y_i; u_i \right) + \frac{1}{2} q_t^2 V_{p,t} - q_t V_{p,ts} - \frac{\|\boldsymbol{u}\|^2}{2p} T_{p,g}(\sigma) \\ &- \frac{\sigma r_t^2}{2} - \frac{1}{2} q_t^2 T_{p,t}(\sigma) - \frac{1}{2} T_{p,s}(\sigma) + q_t T_{p,ts}(\sigma). \end{aligned} \tag{38}$$

Here, the set $\mathcal{T}_2$ has the same definition as the set $\mathcal{T}_1$ except that we replace $\|\boldsymbol{r}_t\|$ by r_t. Here, the sequence of random functions $T_{p,g}(.)$, $T_{p,t}(.)$, $T_{p,s}(.)$ and $T_{p,ts}(.)$ can be expressed as follows

$$\begin{cases} T_{p,g}(\sigma) = \frac{1}{p} \boldsymbol{g}^\top \boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp [\boldsymbol{\Lambda}^\perp + \sigma \boldsymbol{I}_{p-1}]^{-1} (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{g} \\ T_{p,t}(\sigma) = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp [\boldsymbol{\Lambda}^\perp + \sigma \boldsymbol{I}_{p-1}]^{-1} (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \boldsymbol{\xi}_t \\ T_{p,s}(\sigma) = \widehat{\boldsymbol{w}}_s^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp [\boldsymbol{\Lambda}^\perp + \sigma \boldsymbol{I}_{p-1}]^{-1} (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \widehat{\boldsymbol{w}}_s \\ T_{p,ts}(\sigma) = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp [\boldsymbol{\Lambda}^\perp + \sigma \boldsymbol{I}_{p-1}]^{-1} (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \widehat{\boldsymbol{w}}_s. \end{cases}$$

Note that the formulation in (38) is obtained after dropping terms that converge in probability to zero. This simplification can be justified using a similar analysis as in [20, Lemma 3]. The main idea is to show that both loss functions converge uniformly to the same limit.

Next, the objective is to simplify the obtained AO formulation over the optimization vector $\boldsymbol{u} \in \mathbb{R}^{n_t}$. Based on the property stated in [20, Lemma 4], the optimization over the vector $\boldsymbol{u}$ can be expressed as follows

$$\begin{aligned} I_p^\star &= \max_{\boldsymbol{u}\in\mathcal{C}_t} r_t \boldsymbol{h}^\top \boldsymbol{u} + q_t \boldsymbol{u}^\top \boldsymbol{s}_t - \sum_{i=1}^{n_t} \ell^\star\left(y_i; u_i\right) - \frac{\|\boldsymbol{u}\|^2}{2} T_{p,g}(\sigma) \\ &= \sum_{i=1}^{n_t} \mathcal{M}_{\ell(y_i,.)}\Big(r_t h_i + q_t s_{t,i}; T_{p,g}(\sigma)\Big). \end{aligned}$$

This result is valid on events with probability going to one as p goes to $+\infty$. Here, the function $\mathcal{M}_{\ell(y_i,.)}$ is the Moreau envelope function defined in (13). The proof of this property is omitted since it follows the same ideas of [20, Lemma 4]. The main idea is to use Assumption 3 to show that the optimal solution of the unconstrained version of the maximization problem is bounded asymptotically. Then, use the property introduced in [27, Example 11.26] to complete the proof. Now, our auxiliary formulation can be asymptotically simplified to a scalar optimization problem as follows

$$\begin{aligned} &\min_{(q_t,r_t)\in\mathcal{T}_2} \sup_{\sigma>-\mu_p} \frac{\lambda}{2}(q_t^2 + r_t^2) + \frac{1}{2} q_t^2 V_{p,t} - q_t V_{p,ts} - \frac{\sigma r_t^2}{2} \\ &+ \frac{1}{p}\sum_{i=1}^{n_t} \mathcal{M}_{\ell(y_i,.)}\Big(r_t h_i + q_t s_{t,i}; T_{p,g}(\sigma)\Big) - \frac{1}{2} q_t^2 T_{p,t}(\sigma) \\ &- \frac{1}{2} T_{p,s}(\sigma) + q_t T_{p,ts}(\sigma). \end{aligned} \tag{39}$$

Note that the auxiliary formulation in (39) has now scalar optimization variables. Then, it remains to study its asymptotic properties. We refer to this problem as the target scalar formulation.

3) Asymptotic Analysis of the Target Scalar Formulation: In this part, we study the asymptotic properties of the scalar formulation expressed in (39). We start our analysis by studying the asymptotic properties of the sequence of random variables $V_{p,t}$ and $V_{p,ts}$ and the sequence of random functions $T_{p,g}(.)$, $T_{p,t}(.)$, $T_{p,s}(.)$ and $T_{p,ts}(.)$ as given in the following Lemma.

Lemma 2 (Asymptotic Properties): Define V as follows $V = \mathbb{E}_\mu[\mu]$, where the expectation is over the probability distribution $\mathbb{P}_\mu(.)$ defined in Assumption 5. First, the random variable μ_p converges in probability to $\mu_{\min}$, where $\mu_{\min}$ is defined in Assumption 5. For any fixed $\sigma > 0$, the following convergence in probability holds true

$$\begin{cases} V_{p,t} \xrightarrow{p} V, \;\; V_{p,ts} \xrightarrow{p} V_{ts} = \rho q_s^\star V \\ T_{p,t}(\sigma - \mu_p) \xrightarrow{p} T_t(\sigma - \mu_{\min}) \\ T_{p,ts}(\sigma - \mu_p) \xrightarrow{p} T_{ts}(\sigma - \mu_{\min}) \\ T_{p,s}(\sigma - \mu_p) \xrightarrow{p} T_s(\sigma - \mu_{\min}) \\ T_{p,g}(\sigma - \mu_p) \xrightarrow{p} T_1(\sigma - \mu_{\min}). \end{cases}$$

Here, the deterministic functions $T_t(.)$, $T_{ts}(.)$, $T_s(.)$, $T_1(.)$ and $T_3(.)$ are defined as follows

$$\begin{cases} T_t(\sigma) = V + \sigma - 1/T_1(\sigma), \;\; T_{ts}(\sigma) = \rho q_s^\star T_t(\sigma) \\ T_s(\sigma) = ((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2) T_3(\sigma) + (\rho q_s^\star)^2 T_t(\sigma) \\ T_1(\sigma) = \mathbb{E}_\mu\left[1/(\mu+\sigma)\right], \;\; T_3(\sigma) = \mathbb{E}_\mu\left[\mu^2/(\mu+\sigma)\right]. \end{cases}$$

Moreover, the constants $q_s^\star$ and $r_s^\star$ are the optimal solutions of the source asymptotic formulation defined in (26). The detailed proof of Lemma 2 is provided in Appendix VIII. Now that we obtained the asymptotic properties of the sequence of random variables, it remains to study the asymptotic properties of the optimal cost and optimal

solution set of the scalar formulation in (39). To state our first asymptotic result, we define the following deterministic optimization problem

$$\begin{aligned}\min_{(q_t,r_t)\in\mathcal{T}_2}\sup_{\sigma>-\mu_{\min}}&\frac{\lambda}{2}(q_t^2+r_t^2)+\frac{1}{2}q_t^2V-q_tV_{ts}-\frac{\sigma r_t^2}{2}\\&+\alpha_t\mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_tH_t+q_tS_t;T_g(\sigma)\Big)\Big]-\frac{1}{2}T_s(\sigma)\\&+q_tT_{ts}(\sigma)-\frac{1}{2}q_t^2T_t(\sigma),\end{aligned}\tag{40}$$

where H_t and S_t are two independent standard Gaussian random variables and $Y_t=\varphi(S_t)$. Here, the function $\mathcal{M}_{\ell(Y_t,.)}$ denotes the Moreau envelope function defined in (13) and the expectation is take over the random variables H_t, S_t and the possibly random function $\varphi(.)$. Now, we are ready to state our asymptotic property of the cost function of (39).

Lemma 3 (Cost Function of the Traget AO Formulation): Define $\mathcal{O}_{p,t}(.)$ as the loss function of the target scalar optimization problem given in (39). Additionally, define $\mathcal{O}_t(.)$ as the cost function of the deterministic formulation in (40). Then, the following convergence in probability holds true.

$$\mathcal{O}_{p,t}(q_t,r_t,\sigma-\mu_p)\ \xrightarrow{p}\ \mathcal{O}_t(q_t,r_t,\sigma-\mu_{\min}),\tag{41}$$

for any fixed feasible q_t, r_t and $\sigma>0$.

The proof of the asymptotic property stated in Lemma 3 uses the asymptotic results stated in Lemma 2. Moreover, it uses the weak law of large numbers to show that the empirical mean of the Moreau envelope concentrates around its expected value. Based on Assumption 3, one can see that the following pointwise convergence is valid

$$\frac{1}{p}\sum_{i=1}^{n_t}\mathcal{M}_{\ell(y_i,.)}\big(r_th_i+q_ts_{t,i};x\big)\xrightarrow{p}\mathbb{E}\big[\mathcal{M}_{\ell(Y,.)}\big(r_tH+q_tS;x\big)\big].$$

Here H and S are independent standard Gaussian random variables and $Y=\varphi(S)$. The above property is valid for any $x>0$, $r_t\geq 0$ and q_t. Based on [27, Theorem 2.26], the Moreau envelope function is convex and continuously differentiable with respect $x>0$. Combining this with [29, Theorem 7.46], the above asymptotic function is continuous in $x>0$. Then, using Lemma 2, the uniform convergence and the continuity property, we conclude that the empirical average of the Moreau envelope converges in probability to the following function

$$\mathbb{E}\big[\mathcal{M}_{\ell(Y,.)}\big(r_tH+q_tS;T_g(\sigma-\mu_{\min})\big)\big],\tag{42}$$

for any fixed feasible q_t, r_t and $\sigma>0$. This completes the proof of Lemma 3. The analysis in [20, Lemma 6] can also be applied here to show that the formulation in (40) is strictly concave in the maximization variable σ for fixed feasible (q_t,r_t). Define the following function

$$(q_t,r_t)\rightarrow\sup_{\sigma>-\mu_{\min}}\mathcal{O}_t(q_t,r_t,\sigma),\tag{43}$$

where $\mathcal{O}_t(.)$ denotes the cost function of the deterministic formulation in (40). The analysis in [20, Lemma 6] can also be used here to show that the function defined in (43) is strongly convex in (q_t,r_t) with a strong convexity parameter λ.

Now, we use these properties to show that the optimal solution set of the formulation in (39) converges in probability to the optimal solution set of the formulation in (40).

Lemma 4 (Consistency of the Target AO Formulation): Define $\mathcal{P}_{p,t}$ and $\mathcal{P}_t$ as the optimal set of (q_t,r_t) of the optimization problems formulated in (39) and (40), respectively. Moreover, define $\mathcal{O}_{p,t}^\star$ and $\mathcal{O}_t^\star$ as the optimal cost values of the optimization problems formulated in (39) and (40), respectively. Then, the following converges in probability holds true

$$\mathcal{O}_{p,t}^\star\xrightarrow{p}\mathcal{O}_t^\star,\ \mathbb{D}(\mathcal{P}_{p,t},\mathcal{P}_t)\xrightarrow{p}0,\tag{44}$$

where $\mathbb{D}(\mathcal{A},\mathcal{B})$ denotes the deviation between the sets $\mathcal{A}$ and $\mathcal{B}$ and is defined as $\mathbb{D}(\mathcal{A},\mathcal{B})=\sup_{\boldsymbol{c}_1\in\mathcal{A}}\inf_{\boldsymbol{c}_2\in\mathcal{B}}\|\boldsymbol{c}_1-\boldsymbol{c}_2\|$.

The stated result can be proved by first observing that the loss function $\mathcal{O}_t(.)$ corresponding to the deterministic formulation in (40) satisfies the following

$$\lim_{\sigma \to +\infty} \mathcal{O}_t(q_t, r_t, \sigma - \mu_{\min}) = -\infty. \tag{45}$$

For any $r_t > 0$ and any fixed q_t. Combining this with the convergence result in Lemma 3, [17, Lemma B.1] and [17, Lemma B.2], we obtain the following asymptotic result

$$\sup_{\sigma>0} \mathcal{O}_{p,t}(q_t, r_t, \sigma - \mu_p) \xrightarrow{p} \sup_{\sigma>0} \mathcal{O}_t(q_t, r_t, \sigma - \mu_{\min}).$$

Note that if $r_t = 0$, the supremum in the above convergence result occurs at $\sigma \to +\infty$. However, it can be checked that the above convergence result still hold. Based on [20, Lemma 6], the cost function of the minimization problem in (40) is strongly convex in (q_t, r_t). Then, based on [30, Theorem II.1] and [31, Theorem 2.1], we obtain the convergence result in Lemma 4. Now that we obtained the asymptotic problem, it remains to study the asymptotic properties of the training and generalization errors corresponding to the target formulation in (7).

4) Specialization to the Hard Formulation: Before starting the analysis of the generalization error, we specialize our general analysis to the hard transfer formulation. To obtain the asymptotic limit of the hard formulation, we specialize the general results in (40) to the following probability distribution

$$\mathbb{P}_p(\mu) = (1-\delta)d(\mu) + \delta d(\mu - p), \tag{46}$$

where the function $x \to d(x-a)$ is the dirac delta function defined at a. Then, we study the obtained asymptotic results when p goes to $+\infty$. Note that the probability distribution in (46) satisfies Assumption 5. Then, the asymptotic limit of the soft formulation corresponding to the probability distribution $\mathbb{P}_\mu(.)$, defined in (46), can be expressed as follows

$$\begin{aligned}\min_{(q_t,r_t)\in\mathcal{T}_2} \sup_{\sigma>0} \; & \frac{\lambda}{2}(q_t^2 + r_t^2) + \frac{T_2(\sigma)}{2}\big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\big) \\ & + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_t H_t + q_t S_t; T_1(\sigma)\Big)\Big] - \frac{\sigma r_t^2}{2} \\ & - \frac{1}{2}\left(q_t - \rho q_s^\star\right)^2 \left(\sigma - 1/T_1(\sigma)\right),\end{aligned} \tag{47}$$

where the functions $T_1(.)$ and $T_2(.)$ are defined as follows

$$\begin{cases} T_1(\sigma) = (1-\delta)/\sigma + \delta/(p+\sigma) \\ T_2(\sigma) = \delta p \sigma/(p+\sigma). \end{cases} \tag{48}$$

First, one can see that the loss function of (47), denoted by $h_p(.)$, converges as follows

$$\lim_{p \to +\infty} h_p(q_t, r_t, \sigma) = h(q_t, r_t, \sigma), \tag{49}$$

for any fixed q_t, r_t and σ in the feasibility set of the formulation (47). Here, the function $h(.)$ is defined as follows

$$\begin{aligned} h(q_t, r_t, \sigma) = & \frac{\lambda}{2}(q_t^2 + r_t^2) + \frac{\sigma\delta}{2}\Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big) \\ & + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_t H_t + q_t S_t; \frac{1-\delta}{\sigma}\Big)\Big] - \frac{\sigma r_t^2}{2} \\ & + \frac{\sigma\delta}{2(1-\delta)}\left(q_t - \rho q_s^\star\right)^2, \end{aligned} \tag{50}$$

for any fixed q_t, r_t and σ in the feasibility set of the formulation (47). Based on the analysis in [20, Lemma 6], one can see that the formulation in (47) and the function in (50) are strictly convex in the minimization variables

and strictly concave in the maximization variable. Then, based on the analysis in Section VI-C2 and [31, Theorem 2.1], the asymptotic limit of the soft formulation defined in (47) simplifies to the following formulation

$$\begin{aligned}\min_{(q_t,r_t)\in\mathcal{T}_2}\sup_{\sigma>0}\ &\frac{\lambda}{2}(q_t^2+r_t^2)+\frac{\sigma\delta}{2}\left((1-\rho^2)(q_s^\star)^2+(r_s^\star)^2\right)\\&+\alpha_t\mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_tH_t+q_tS_t;\frac{1-\delta}{\sigma}\Big)\Big]-\frac{\sigma r_t^2}{2}\\&+\frac{\sigma\delta}{2(1-\delta)}\left(q_t-\rho q_s^\star\right)^2.\end{aligned}\tag{51}$$

This shows that the asymptotic limit of the hard formulation is the deterministic problem (51).

5) Asymptotic Analysis of the Training and Generalization Errors: First, the generalization error corresponding to the target task is given by

$$\mathcal{E}_{\text{test}}=\frac{1}{4^v}\mathbb{E}\left[\left(\varphi(\boldsymbol{a}_{t,\text{new}}^\top\boldsymbol{\xi}_t)-\widehat{\varphi}(\widehat{\boldsymbol{w}}_t^\top\boldsymbol{a}_{t,\text{new}})\right)^2\right],\tag{52}$$

where $\boldsymbol{a}_{t,\text{new}}$ is an unseen target feature vector. Now, consider the following two random variables

$$\nu_1=\boldsymbol{a}_{t,\text{new}}^\top\boldsymbol{\xi}_t,\ \text{and}\ \nu_2=\widehat{\boldsymbol{w}}_t^\top\boldsymbol{a}_{t,\text{new}}.$$

Given $\widehat{\boldsymbol{w}}_t$ and $\boldsymbol{\xi}_t$, the random variables ν_1 and ν_2 have a bivaraite Gaussian distribution with zero mean vector and covariance matrix given as follows

$$\boldsymbol{C}_p=\begin{bmatrix}\|\boldsymbol{\xi}_t\|^2 & \boldsymbol{\xi}_t^\top\widehat{\boldsymbol{w}}_t\\ \boldsymbol{\xi}_t^\top\widehat{\boldsymbol{w}}_t & \|\widehat{\boldsymbol{w}}_t\|^2\end{bmatrix}.\tag{53}$$

To precisely analyze the asymptotic behavior of the generalization error, it suffices to analyze the properties of the covariance matrix $\boldsymbol{C}_p$. Define the random variables $\widehat{q}_{p,t}^\star$ and $\widehat{r}_{p,t}^\star$ for the target task as follows

$$\widehat{q}_{p,t}^\star=\boldsymbol{\xi}_t^\top\widehat{\boldsymbol{w}}_t,\ \text{and}\ \widehat{r}_{p,t}^\star=\left\|(\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top\widehat{\boldsymbol{w}}_t\right\|,\tag{54}$$

where $\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp$ is defined in Section VI-C2. Then, the covariance matrix $\boldsymbol{C}_p$ given in (53) can be expressed as follows

$$\begin{bmatrix}1 & \widehat{q}_{p,t}^\star\\ \widehat{q}_{p,t}^\star & (\widehat{q}_{p,t}^\star)^2+(\widehat{r}_{p,t}^\star)^2\end{bmatrix}.$$

Hence, to study the asymptotic properties of the generalization error, it suffices to study the asymptotic properties of the random quantities $\widehat{q}_{p,t}^\star$ and $\widehat{r}_{p,t}^\star$.

Lemma 5 (Consistency of the Target Formulation): The random quantities $\widehat{q}_{p,t}^\star$ and $\widehat{r}_{p,t}^\star$ satisfy the following asymptotic properties

$$\widehat{q}_{p,t}^\star\xrightarrow{p}q_t^\star,\ \text{and}\ \widehat{r}_{p,t}^\star\xrightarrow{p}r_t^\star,$$

where $q_t^\star$ and $r_t^\star$ are the optimal solutions of the deterministic formulation stated in (40).
To prove the above asymptotic result, we define $\widetilde{q}_{p,t}^\star$ and $\widetilde{r}_{p,t}^\star$ as follows

$$\widetilde{q}_{p,t}^\star=\boldsymbol{\xi}_t^\top\widetilde{\boldsymbol{w}}_t,\ \text{and}\ \widetilde{r}_{p,t}^\star=\left\|(\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top\widetilde{\boldsymbol{w}}_t\right\|,\tag{55}$$

where $\widetilde{\boldsymbol{w}}_t$ is the optimal solution of the auxiliary formulation in (32). Given the result in Lemma 4 and the analysis in Sections VI-C2 and VI-C3, the convergence result in Lemma 4 is also satisfied by our auxiliary formulation in (32), i.e.

$$\widetilde{q}_{p,t}^\star\xrightarrow{p}q_t^\star,\ \text{and}\ \widetilde{r}_{p,t}^\star\xrightarrow{p}r_t^\star.$$

The rest of the proof of the convergence result stated in Lemma 5 is based on the CGMT framework, i.e. Theorem 2. Specifically, it follows after showing that the assumptions in Theorem 2 are all satisfied. Note that the cost function of the problem (40) is strongly convex in the minimization variables. Then, based on [30, Theorem II.1], the cost function of the optimization problem in (39) converges uniformly to the cost function of (40). Combine

this with the compactness of the feasibility sets to see that the conditions in Theorem 2 are all satisfied. Then, the convergence result in Lemma 5 follows.

Note that the CGMT framework applied to prove Lemma 5 also shows that the optimal cost value of the soft target formulation in (7) converges in probability to the optimal cost value of the deterministic formulation given in (40). Combining this with the result in Lemma 5 shows the convergence property of the training error stated in (15). Now, it remains to show the convergence of the generalization error. It suffices to show that the generalization error defined in (52) is continuous in the quantities $\widehat{q}^{\star}_{p,t}$ and $\widehat{r}^{\star}_{p,t}$. This follows based on Assumption 4 and the continuity under integral sign property [32]. This shows the convergence result in (16) which completes the proof of Theorem 1 and Corollary 1. Note that the above analysis of the soft target formulation in (7) is valid for any choice of C_{q_t} and C_r that satisfy the result in Lemma 1. One can ignore these bounds given the convexity properties of the deterministic formulation in (40). This leads to the scalar formulations introduced in (14) and (17).

D. Phase Transitions in the Hard formulation

In this part, we provide a rigorous proof of Proposition 1 and Proposition 2. Here, we consider the squared–loss function. In this case, the deterministic source formulation given in (12) can be simplified as follows

$$\min_{q_s,r_s\geq 0} \frac{1}{2}\max\Big\{-r_s+\sqrt{\alpha_s}(q_s^2+r_s^2+v_s-2q_sc_s)^{\frac{1}{2}},0\Big\}^2 + \frac{\lambda}{2}(q_s^2+r_s^2). \tag{56}$$

The constants v_s and c_s are defined as $v_s=\mathbb{E}[Y_s^2]$ and $c_s=\mathbb{E}[S_sY_s]$, where $Y_s=\varphi(S_s)$ and S_s is a standard Gaussian random variable. Additionally, the target scalar formulation given in (14) can be simplified as follows

$$\min_{q_t,r_t\geq 0}\sup_{\sigma>0} \frac{\lambda}{2}(q_t^2+r_t^2)+\frac{\sigma\delta}{2}\Big((1-\rho^2)(q_s^{\star})^2+(r_s^{\star})^2\Big) + \frac{\alpha_t\sigma}{2(1-\delta)+2\sigma}(r_t^2+q_t^2+v_t-2q_tc_t)-\frac{\sigma r_t^2}{2} + \frac{\sigma\delta}{2(1-\delta)}\left(q_t-\rho q_s^{\star}\right)^2. \tag{57}$$

Here, the constants v_t and c_t are defined as $v_t=\mathbb{E}[Y_t^2]$ and $c_t=\mathbb{E}[Y_tS_t]$, where $Y_t=\varphi(S_t)$ and S_t is a standard Gaussian random variable. Under the conditions stated in Proposition 1 and Proposition 2, the source deterministic formulation given in (56) can be simplified as follows

$$\min_{q_s,r_s\geq 0} -r_s+\sqrt{\alpha_s}(q_s^2+r_s^2+v_s-2q_sc_s)^{\frac{1}{2}}. \tag{58}$$

Note that one can easily solve over the variables q_s and r_s. Specifically, the optimal solutions of (58) can be expressed as follows

$$q_s^{\star}=c_s,\ \text{and}\ r_s^{\star}=\sqrt{v_s-c_s^2}/\sqrt{\alpha_s-1}. \tag{59}$$

Moreover, the target deterministic formulation given in (57) can be expressed as follows

$$\min_{q_t,r_t\geq 0}\sup_{\sigma>0} \frac{\sigma\delta}{2}\beta_2+\frac{\alpha_t\sigma}{2(1-\delta)+2\sigma}(r_t^2+q_t^2+v_t-2q_tc_t) - \frac{\sigma r_t^2}{2}+\frac{\sigma\delta}{2(1-\delta)}\left(q_t-\beta_1\right)^2, \tag{60}$$

where β_1 and β_2 are given by

$$\beta_1=\rho q_s^{\star},\ \beta_2=\Big((1-\rho^2)(q_s^{\star})^2+(r_s^{\star})^2\Big). \tag{61}$$

Before solving the optimization problem in (60), we consider the following change of variable

$$x_t^2=r_t^2-\delta\beta_2-\frac{\delta}{1-\delta}(q_t-\beta_1)^2. \tag{62}$$

Note that the above change of variable is valid since the formulation in (60) requires the right hand side of (62) to be positive. Therefore, the formulation in (60) can be expressed in terms of x_t instead of r_t as follows

$$\min_{q_t,x_t\geq 0}\sup_{\sigma>0}\frac{\alpha_t\sigma}{2(1-\delta)+2\sigma}\big(x_t^2+\delta\beta_2+\frac{\delta}{1-\delta}(q_t-\beta_1)^2 + q_t^2+v_t-2q_tc_t\big)-\frac{\sigma x_t^2}{2}. \tag{63}$$

Now, it can be easily checked that the above optimization problem can be solved over the variable σ to give the following formulation

$$\min_{q_t,x_t\geq 0}\frac{1}{2}\max\Big\{-x_t\sqrt{1-\delta}+\sqrt{\alpha_t}\big(x_t^2+\delta\beta_2 + \frac{\delta}{1-\delta}(q_t-\beta_1)^2+q_t^2+v_t-2q_tc_t\big)^{\frac{1}{2}},0\Big\}^2.$$

It is now clear that one can solve the problem in (63) in closed form. Moreover, it can be easily checked that the optimal solutions of the optimization problem (60) can be expressed as follows

$$\begin{cases} q_t^\star=(1-\delta)c_t+\delta\beta_1 \\ (r_t^\star)^2=\frac{1-\delta}{\alpha_t+\delta-1}\left((\delta-1)c_t^2+\delta\beta_1^2+\delta\beta_2+v_t-2\delta\beta_1c_t\right) \\ \qquad +\delta\beta_2+\delta(1-\delta)(c_t-\beta_1)^2. \end{cases}$$

Then, the asymptotic limit of the generalization error corresponding to the hard formulation can be determined in closed–form. Since the source and target models given in (1) and (2) use the same data generating function, the constants v_t, c_t, v_s and c_s are all equal. We express them as v and c in the rest of the proof.

1) Regression Model: In this part, we assume that the function $\widehat{\varphi}(.)$ is the identity function. Based on the asymptotic result stated in Corollary 1, the asymptotic limit of the generalization error corresponding to the hard formulation can be expressed as follows

$$\mathcal{E}_{\text{test}}=v-2cq_t^\star+(q_t^\star)^2+(r_t^\star)^2.$$

It can be easily checked that the generalization error can be express as follows

$$\mathcal{E}_{\text{test}}=\frac{\alpha_t}{\alpha_t+\delta-1}\left(\delta\{(c-\beta_1)^2+\beta_2\}+(v-c^2)\right). \tag{64}$$

Note that the the generalization error obtained above depends explicitly on δ. Now, it suffices to study the derivative of $\mathcal{E}_{\text{test}}$ to find the properties of the optimal transfer rate δ that minimizes the generalization error. Note that the derivative can be expressed as follows

$$\mathcal{E}'_{\text{test}}(\delta)=\frac{(\alpha_t-1)\{(c-\beta_1)^2+\beta_2\}-(v-c^2)}{(\alpha_t+\delta-1)^2}. \tag{65}$$

This shows that the derivative of the generalization error has the same sign as the numerator. This means that the optimal transfer rate satisfies the following

$$\delta^\star=\begin{cases}1 & \text{if } Z_t<0\\ 0 & \text{if } Z_t>0\\ [0\ 1] & \text{otherwise,}\end{cases} \tag{66}$$

where Z_t is given by

$$Z_t=(\alpha_t-1)\{(c-\beta_1)^2+\beta_2\}-(v-c^2). \tag{67}$$

It can be easily shown that the condition in (66) can be expressed as the one given in (18). This completes the proof of Proposition 1.

2) Classification Model: In this part, we assume that the function $\widehat{\varphi}(.)$ is the sign function. Based on the asymptotic result stated in Corollary 1, the asymptotic limit of the generalization error corresponding to the hard formulation can be expressed as follows

$$\mathcal{E}_{\text{test}} = \frac{1}{\pi}\text{acos}\Big(\frac{q_t^\star}{\sqrt{(q_t^\star)^2+(r_t^\star)^2}}\Big). \tag{68}$$

Given the closed-form expressions of the solutions, the generalization error can be expressed in terms of the transfer rate δ as follows

$$\mathcal{E}_{\text{test}}(\delta) = \frac{1}{\pi}\text{acos}\Big(\frac{(a\delta+c)\sqrt{\delta+\alpha_t-1}}{\sqrt{T_1\delta^2+T_2\delta+T_3}}\Big).$$

Here, the constant terms a, T_1, T_2 and T_3 are independent of the transfer rate δ and are given by

$$\begin{aligned}
&a = \rho c - c,\ T_1 = -2c^2 + 2c^2\rho \\
&T_2 = \alpha_t(v-c^2)/(\alpha_s-1) + 4c^2 - 2c^2\rho - v \\
&T_3 = (\alpha_t-2)c^2 + v.
\end{aligned}$$

Given that the acos(.) function is strictly decreasing, it suffices to find the maximum of the following function

$$g(\delta) = \frac{(a\delta+c)\sqrt{\delta+\alpha_t-1}}{\sqrt{T_1\delta^2+T_2\delta+T_3}}, \tag{69}$$

to determine the properties of the optimal transfer rate $\delta^\star$ that gives the lowest generalization error. It can be easily checked that the derivative of the function $g(.)$ with respect to δ can be fully characterized by analyzing the following third degree polynomial

$$h(\delta) = Z_1\delta^3 + Z_2\delta^2 + Z_3\delta + Z_4. \tag{70}$$

Here, Z_1, Z_2, Z_3 and Z_4 are independent of δ and can be expressed as follows

$$\begin{aligned}
&Z_1 = aT_1,\ Z_2 = 2aT_2 - cT_1, \\
&Z_3 = 3aT_3 + a(\alpha_t-1)T_2 - 2c(\alpha_t-1)T_1 \\
&Z_4 = (2(\alpha_t-1)a + c)T_3 - c(\alpha_t-1)T_2.
\end{aligned}$$

We can see that the function $g(.)$ is increasing at $\delta = 0$ when $Z_4 > 0$. This means that the function $\mathcal{E}_{\text{test}}(.)$ is decreasing at $\delta = 0$ when $Z_4 > 0$. This means that there exists $\delta_p > 0$ such that the hard transfer with δ_p is better than the standard transfer in this case. It can be easily checked that this is equivalent to the condition provided in Proposition 2. This completes the proof of the theoretical statement in Proposition 2.

VII. Conclusion

In this paper, we presented a precise characterization of the asymptotic properties of two simple transfer learning formulations. Specifically, our results show that the training and generalization errors corresponding to the considered transfer formulations converge to deterministic functions. These functions can be explicitly found by combining the solutions of two deterministic scalar optimization problems. Our simulation results validate our theoretical predictions and reveal the existence of a phase transition phenomenon in the hard transfer formulation. Specifically, it shows that the hard transfer formulation moves from negative transfer to positive transfer when the similarity of the source and target tasks move past a well-defined critical threshold.

VIII. Appendix: Proof of Lemma 2

To prove the convergence properties stated in Lemma 2, we show first that they are valid for the auxiliary formulation corresponding to the source problem.

A. Auxiliary Convergence

Note that the analysis present in Section VI is also valid for the source problem. This is because the formulation in (7) is equivalent to the source problem in (3) if $\boldsymbol{\Sigma}$ is the all zero matrix and we use the source training data. Then, we can see that the optimal solution of the auxiliary formulation corresponding to the source problem, denoted by $\widetilde{\boldsymbol{w}}_s$, can be expressed as follows

$$\widetilde{\boldsymbol{w}}_s = q_{p,s}^{\star}\boldsymbol{\xi}_s - \frac{r_{p,s}^{\star}}{\|\widetilde{\boldsymbol{g}}_s\|}\boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp}\widetilde{\boldsymbol{g}}_s, \tag{71}$$

where $\widetilde{\boldsymbol{g}}_s = (\boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp})^{\top}\boldsymbol{g}_s$ and $\boldsymbol{g}_s$ has independent standard Gaussian components. Here, $\boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp} \in \mathbb{R}^{p\times(p-1)}$ is formed by an orthonormal basis orthogonal to the vector $\boldsymbol{\xi}_s$. Additionally, our analysis in Section VI shows that the following convergence in probability holds

$$q_{p,s} \xrightarrow{p} q_s^{\star} \text{ and } r_{p,s}^{\star} \xrightarrow{p} r_s^{\star}. \tag{72}$$

Here, $q_s^{\star}$ and $r_s^{\star}$ are the optimal solutions of asymptotic limit of the source formulation defined in (12).

Based on Assumption 5, the random variable μ_p converges in probability to $\mu_{\min}$, where $\mu_{\min}$ is defined in Assumption 5. Using [33, Proposition 3], Assumptions 1 and 5, the sequence of random variables $V_{p,t}$ converges pointwisely in probability to the constant $V = \mathbb{E}_{\mu}[\mu]$, where the expectation is taken over the probability distribution $\mathbb{P}_{\mu}(.)$ defined in Assumption 5. Now, we study the properties of the remaining functions using the optimal solution of the auxiliary formulation defined in (71), i.e. $\widetilde{\boldsymbol{w}}_s$, instead of $\widehat{\boldsymbol{w}}_s$. For instance, we first study the random sequence $\widetilde{V}_{p,ts} = \boldsymbol{\xi}_t^{\top}\boldsymbol{\Lambda}\widetilde{\boldsymbol{w}}_s$ to infer the asymptotic properties of $V_{p,ts}$.

Exploiting the predictions stated in (71) and (72), the sequence of random variables $\widetilde{V}_{p,ts}$ converges in probability to the following constant

$$V_{ts} = q_s^{\star}\rho V, \tag{73}$$

First, fix $\sigma > -\mu_{\min}$. Then, based on the convergence of μ_p and [33, Proposition 3], the sequence of random functions $T_{p,g}(.)$ converges in probability as follows

$$T_{p,g}(\sigma) \xrightarrow{p} T_g(\sigma) = \mathbb{E}_{\mu}\left[1/(\mu+\sigma)\right]. \tag{74}$$

Now, we express σ as $\sigma = \sigma' - x$, where $\sigma' > 0$. This means that the following convergence in probability holds true

$$T_{p,g}(\sigma' - x) \xrightarrow{p} T_g(\sigma' - x), \tag{75}$$

for any $x < \sigma' + \mu_{\min}$. Note that the functions $T_{p,g}(.)$ and $T_g(.)$ are both convex and continuous in the variable x in the set $[0,\ \sigma' + \mu_{\min}[$. Then, based on [30, Theorem II.1], the convergence in (75) is uniform in the variable x in the compact set $[0,\ \sigma'/2 + \mu_{\min}]$. Now, note that μ_p converges in probability to $\mu_{\min}$. Therefore, we obtain the following convergence in probability

$$T_{p,g}(\sigma' - \mu_p) \xrightarrow{p} T_g(\sigma' - \mu_{\min}), \tag{76}$$

valid for any fixed $\sigma' > 0$. Using the block matrix inversion lemma, the function $T_{p,t}(.)$ can be expressed as follows

$$\begin{aligned} T_{p,t}(\sigma) &= \boldsymbol{\xi}_t^{\top}\boldsymbol{\Lambda}\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp}[(\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp})^{\top}\boldsymbol{\Lambda}\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp} + \sigma\boldsymbol{I}_{p-1}]^{-1}(\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp})^{\top}\boldsymbol{\Lambda}\boldsymbol{\xi}_t \\ &= \boldsymbol{\xi}_t^{\top}\boldsymbol{\Lambda}\boldsymbol{\xi}_t + \sigma - \frac{1}{\boldsymbol{\xi}_t^{\top}[\boldsymbol{\Lambda} + \sigma\boldsymbol{I}_p]^{-1}\boldsymbol{\xi}_t}. \end{aligned} \tag{77}$$

Then, using the theoretical results stated in [33, Proposition 3], the functions $T_{p,t}(.)$ converges in probability as follows

$$T_{p,t}(\sigma) \xrightarrow{p} T_t(\sigma) = V + \sigma - \frac{1}{\mathbb{E}_{\mu}\left[1/(\mu+\sigma)\right]}. \tag{78}$$

Combine this with the above analysis to obtain the following convergence in probability

$$T_{p,t}(\sigma' - \mu_p) \xrightarrow{p} T_t(\sigma' - \mu_{\min}), \tag{79}$$

valid for any $\sigma' > 0$. Based on the result in (71), the sequence of random functions $\widetilde{T}_{p,ts}(.)$ converges in probability to the following function

$$T_{ts}(\sigma) = q_s^\star \rho T_t(\sigma). \tag{80}$$

Combine this with the above analysis to obtain the following convergence in probability

$$\widetilde{T}_{p,ts}(\sigma' - \mu_p) \xrightarrow{p} T_{ts}(\sigma' - \mu_{\min}), \tag{81}$$

valid for any $\sigma' > 0$. Using the same analysis and based on (71) and (72), one can see that the sequence of random functions $\widetilde{T}_{p,s}(.)$ converges in probability to the following function

$$\begin{aligned}\widetilde{T}_{p,s}(\sigma) \xrightarrow{p} T_s(\sigma) &= (\rho q_s^\star)^2 T_t(\sigma) \\ &+ \left((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\right) \mathbb{E}_\mu \left[\mu^2/(\mu+\sigma)\right].\end{aligned} \tag{82}$$

Combine this with the above analysis to obtain the following convergence in probability

$$\widetilde{T}_{p,s}(\sigma' - \mu_p) \xrightarrow{p} T_s(\sigma' - \mu_{\min}), \tag{83}$$

valid for any $\sigma' > 0$. The above analysis shows that the asymptotic properties stated in Lemma 2 are valid for the AO formulation corresponding to the source problem. Now, it remains to show that these properties also hold for the primary formulation.

B. Primary Convergence

Now, we show that the convergence properties proved above are also valid for the primary problem. To this end, we show that all the assumptions in Theorem 2 are satisfied. We start our proof by defining the following open set

$$\mathcal{T}_\epsilon = \{\boldsymbol{w} \in \mathbb{R}^p : \left|\boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{w} - V_{ts}\right| < \epsilon\}.$$

Now, we consider the feasibility set $\mathcal{D}_\epsilon = \mathcal{T}_1/\mathcal{S}_\epsilon$, where $\mathcal{T}_1$ is defined in (34). Based on the analysis of the generalized target formulation in Section VI-C2, one can see that the AO formulation corresponding to the source formulation with the set $\mathcal{D}_\epsilon$ can be asymptotically expressed as follows

$$\begin{aligned}\mathfrak{V}_p : \min_{(q_s,r_s)\in\mathcal{T}_2} \min_{\boldsymbol{r}_s \in \widetilde{\mathcal{D}}_\epsilon} \max_{\boldsymbol{u}\in\mathcal{C}_s} &\frac{\|\boldsymbol{u}\|}{p} \boldsymbol{g}_s^\top \boldsymbol{B}_{\boldsymbol{\xi}_s}^\perp \boldsymbol{r}_s + \frac{q_s}{p} \boldsymbol{u}^\top \boldsymbol{s}_s \\ &+ \frac{\lambda}{2}(q_s^2 + \|\boldsymbol{r}_s\|^2) + \frac{1}{p}\|\boldsymbol{r}_s\| \, \boldsymbol{h}_s^\top \boldsymbol{u} - \frac{1}{p}\sum_{i=1}^{n_s} \ell^\star\left(y_{s,i}; u_i\right).\end{aligned}$$

Here, the feasibility set $\mathcal{T}_2$ is defined in Section VI-C2 and the feasibility set $\widetilde{\mathcal{D}}_\epsilon$ is given by

$$\begin{aligned}\Big\{\boldsymbol{r}_s : &\left|q_s \rho V_{p,t} + q_s\sqrt{1-\rho^2} V_{p,r} + \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_s}^\perp \boldsymbol{r}_s - V_{ts}\right| \geq \epsilon \\ &, \|\boldsymbol{r}_s\| = r_s\Big\}.\end{aligned}$$

This follows based on the decomposition in (33) and where $V_{p,t}$ is defined in Section VI-C2 and $V_{p,r} = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{\xi}_r$. Note that the optimization problem given in $\mathfrak{V}_p$ can be equivalently formulated as follows

$$\begin{aligned}\mathfrak{V}_p : \min_{(q_s,r_s)\in\widehat{\mathcal{S}}_\epsilon} \min_{\boldsymbol{r}_s \in \widetilde{\mathcal{D}}_\epsilon} \max_{\boldsymbol{u}\in\mathcal{C}_s} &\frac{\|\boldsymbol{u}\|}{p} \boldsymbol{g}_s^\top \boldsymbol{B}_{\boldsymbol{\xi}}^\perp \boldsymbol{r}_s + \frac{q_s}{p} \boldsymbol{u}^\top \boldsymbol{s}_s \\ &+ \frac{\lambda}{2}(q_s^2 + \|\boldsymbol{r}_s\|^2) + \frac{1}{p}\|\boldsymbol{r}_s\| \, \boldsymbol{h}_s^\top \boldsymbol{u} - \frac{1}{p}\sum_{i=1}^{n_s} \ell^\star\left(y_{s,i}; u_i\right).\end{aligned}$$

Here, we replace the feasibility set $\mathcal{T}_2$ by the feasibility set $\widehat{\mathcal{S}}_\epsilon$ defined as follows

$$\begin{aligned}\Big\{\Big|q_s \rho V_{p,t} + q_s\sqrt{1-\rho^2} V_{p,r} - r_s \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_s}^\perp \frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|} - V_{ts}\Big| \\ \geq \epsilon\Big\} \cap \mathcal{T}_2,\end{aligned}$$

where $\widetilde{\boldsymbol{g}}_s = (\boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp})^\top \boldsymbol{g}_s$. This follows since the first set in $\widehat{\mathcal{S}}_\epsilon$ satisfies the condition in the set $\widetilde{\mathcal{D}}_\epsilon$. Now, assume that $\widehat{\phi}_p^\star$ is the optimal cost value of the optimization problem $\mathfrak{V}_p$ and define the function $\widehat{h}_p(.)$ as follows

$$\widehat{h}_p(q_s, r_s) = \min_{\boldsymbol{r}_s \in \widetilde{\mathcal{D}}_\epsilon} \max_{\boldsymbol{u} \in \mathcal{C}_s} \frac{\|\boldsymbol{u}\|}{p} \boldsymbol{g}_s^\top \boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp} \boldsymbol{r}_s + \frac{q_s \boldsymbol{u}^\top \boldsymbol{s}_s}{p} + \frac{\lambda}{2}(q_s^2 + r_s^2) + \frac{r_s}{p} \boldsymbol{h}_s^\top \boldsymbol{u} - \frac{1}{p} \sum_{i=1}^{n_s} \ell^\star \left(y_{s,i}; u_i\right),$$

in the set $\widehat{\mathcal{S}}_\epsilon$. Based on the max–min inequality [34], the function $\widehat{h}_p(.)$ can be lower bounded by the following function

$$\widetilde{h}_p(q_s, r_s) = \max_{\boldsymbol{u} \in \mathcal{C}_s} \min_{\boldsymbol{r}_s \in \widetilde{\mathcal{D}}_\epsilon} \frac{\|\boldsymbol{u}\|}{p} \boldsymbol{g}_s^\top \boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp} \boldsymbol{r}_s + \frac{q_s \boldsymbol{u}^\top \boldsymbol{s}_s}{p} + \frac{\lambda}{2}(q_s^2 + r_s^2) + \frac{r_s}{p} \boldsymbol{h}_s^\top \boldsymbol{u} - \frac{1}{p} \sum_{i=1}^{n_s} \ell^\star \left(y_{s,i}; u_i\right).$$

This is valid for any $(q_s, r_s) \in \widehat{\mathcal{S}}_\epsilon$. Moreover, note that the following inequality holds true

$$\min_{\boldsymbol{r}_s \in \widetilde{\mathcal{D}}_\epsilon} \frac{\|\boldsymbol{u}\|}{p} \boldsymbol{g}_s^\top \boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp} \boldsymbol{r}_s \geq -\frac{\|\boldsymbol{u}\|}{p} \left\| (\boldsymbol{B}_{\boldsymbol{\xi}_s}^{\perp})^\top \boldsymbol{g}_s \right\| r_s, \tag{84}$$

for any $(q_s, r_s) \in \widehat{\mathcal{S}}_\epsilon$. Following the generalized analysis in Section VI-C2, one can see that the auxiliary problem corresponding to the source formulation can be expressed as follows

$$\min_{(q_s, r_s) \in \mathcal{T}_2} \sup_{\sigma > 0} \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{M}_{\ell(y_{s,i}, .)} \left(r_s h_{s,i} + q_s s_{s,i}; \frac{r_s \|\widetilde{\boldsymbol{g}}_s\|}{\sqrt{n_s}\sigma} \right) - \frac{r_s \sigma}{2} \frac{\|\widetilde{\boldsymbol{g}}_s\|}{\sqrt{n_s}} + \frac{\lambda}{2}(q_s^2 + x_s^2), \tag{85}$$

This means that the function $\widetilde{h}_p(.)$ can be lower bounded by the cost function of the minimization problem formulated in (85) denoted by $\widehat{g}_p(.)$, i.e.

$$\widehat{g}_p(q_s, r_s) \leq \widetilde{h}_p(q_s, r_s). \tag{86}$$

Here, both functions are defined in the feasibility set $\widehat{\mathcal{S}}_\epsilon$. Now, define $\phi_p^\star$ as the optimal cost value of the auxiliary optimization problem corresponding to the source formulation defined in Section VI-C1. Note that the loss function $\widehat{g}_p(.)$ is strongly convex in the variables (q_s, r_s) with strong convexity parameter $\lambda > 0$. This means that for any $\beta \in [0, 1]$, $(q_{s,1}, r_{s,1}) \in \mathcal{T}_2$ and $(q_{s,2}, r_{s,2}) \in \mathcal{T}_2$, we have the following inequality

$$\widehat{g}_p(\beta \boldsymbol{v}_1 + (1-\beta)\boldsymbol{v}_2) \leq \beta \widehat{g}_p(\boldsymbol{v}_1) + (1-\beta)\widehat{g}_p(\boldsymbol{v}_2) - \frac{\lambda}{2}\beta(1-\beta)\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2, \tag{87}$$

where $\boldsymbol{v}_1 = [q_{s,1}, r_{s,1}]$ and $\boldsymbol{v}_2 = [q_{s,2}, r_{s,2}]$. Take $\boldsymbol{v}_1$ as $\boldsymbol{v}_p^\star$ which represents the optimal solution of the optimization problem (85). Then, the inequality in (87) implies the following inequality

$$\phi_p^\star \leq \widehat{g}_p(\boldsymbol{v}_2) - \frac{\lambda}{2}\beta \left\| \boldsymbol{v}_p^\star - \boldsymbol{v}_2 \right\|^2. \tag{88}$$

This is valid for any $\boldsymbol{v}_2$ in the set $\mathcal{T}_2$. Now, taking $\beta = 1/2$ and the minimum over $\boldsymbol{v}_2$ in the set $\widehat{\mathcal{S}}_\epsilon$ in both sides, we obtain the following inequality

$$\phi_p^\star + \frac{\lambda}{4} \min_{\boldsymbol{v} \in \widehat{\mathcal{S}}_\epsilon} \left\| \boldsymbol{v}_p^\star - \boldsymbol{v} \right\|^2 \leq \min_{\boldsymbol{v} \in \widehat{\mathcal{S}}_\epsilon} \widehat{g}_p(\boldsymbol{v}).$$

Based on the above analysis, note that the following inequality also holds true

$$\min_{\boldsymbol{v} \in \widehat{\mathcal{S}}_\epsilon} \widehat{g}_p(\boldsymbol{v}) \leq \widehat{\phi}_p^\star. \tag{89}$$

Then, to verify the assumption of [17, Theorem 6.1], it remains to show that there exists $\epsilon' > 0$ such that, the following inequality holds

$$\frac{\lambda}{4} \min_{\boldsymbol{v} \in \widehat{\mathcal{S}}_\epsilon} \left\| \boldsymbol{v}_p^\star - \boldsymbol{v} \right\|^2 \geq \epsilon', \tag{90}$$

with probability going to 1 as $p \to \infty$. Note that any element in the set $\widehat{\mathcal{S}}_\epsilon$ satisfies the following inequality

$$\epsilon \leq \left| q_s \rho V_{p,t} + q_s \sqrt{1-\rho^2} V_{p,r} - r_s \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}}^\perp \frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|} - V_{ts} \right| \leq$$
$$\left| q_s \rho V_{p,t} - V_{ts} \right| + \left| q_s \sqrt{1-\rho^2} \right| \left| V_{p,r} \right| + |r_s| \left| \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}}^\perp \frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|} \right|.$$

Based on the analysis in Section VIII-A, we have the following convergence in probability

$$\left| q_s \rho V_{p,t} - V_{ts} \right| \xrightarrow{p} \left| q_s - q_s^\star \right| \rho V$$
$$|q_s| \sqrt{1-\rho^2} \left| V_{p,r} \right| \xrightarrow{p} 0, \;\; |r_s| \left| \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}}^\perp \frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|} \right| \xrightarrow{p} 0. \tag{91}$$

This means that there exists $\epsilon'' > 0$ such that any elements in the set $\widehat{\mathcal{S}}_\epsilon$ satisfies the following inequality

$$\left| q_s - q_s^\star \right| \rho V \geq \epsilon'', \tag{92}$$

with probability going to 1 as $p \to \infty$. Combining this with Assumption 5 and the consistency result stated in (72) shows that there exists $\epsilon' > 0$ such that the following inequality holds

$$\frac{\lambda}{4} \min_{\boldsymbol{v} \in \widehat{\mathcal{D}}_\epsilon} \left\| \boldsymbol{v}_p^\star - \boldsymbol{v} \right\|^2 \geq \epsilon', \tag{93}$$

with probability going to 1 as $p \to \infty$. This also proves that there exists $\epsilon' > 0$ such that the following inequality holds

$$\widehat{\phi}_p^\star \geq \phi_p^\star + \epsilon', \tag{94}$$

with probability going to 1 as $p \to \infty$. This completes the verification of the assumptions in Theorem 2. This means that the optimal solution of the primary problem belongs to the set $\mathcal{S}_\epsilon$ on events with probability going to 1 as $p \to \infty$. Since the choice of ϵ is arbitrary, we obtain the following asymptotic result

$$V_{p,ts} = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \widehat{\boldsymbol{w}}_s \xrightarrow{p} q_s^\star \rho V, \tag{95}$$

where $\widehat{\boldsymbol{w}}_s$ is the optimal solution of the source problem (3). Following the same analysis, one can also show the remaining convergence properties stated in Lemma 2.

REFERENCES

[1] L. Y. Pratt, J. Mostow, and C. A. Kamm, "Direct transfer of learned information among neural networks," in *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'91. AAAI Press, 1991, p. 584–589.

[2] L. Y. Pratt, "Discriminability-based transfer between neural networks," in *Advances in Neural Information Processing Systems*, S. Hanson, J. Cowan, and C. Giles, Eds., vol. 5. Morgan-Kaufmann, 1993, pp. 204–211. [Online]. Available: https://proceedings.neurips.cc/paper/1992/file/67e103b0761e60683e83c559be18d40c-Paper.pdf

[3] D. Perkins and G. Salomon, "Transfer of learning," *Oxford, England:Pergamon*, 1992.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," 2018.

[6] L. P. K. M. T. Rosenstein, Z. Marx and T. G. Dietterich, "To transfer or not to transfer," in *NIPS workshop on transfer learning*, 2005.

[7] B. Bakker and T. Heskes, "Task clustering and gating for bayesian multitask learning," *J. Mach. Learn. Res.*, vol. 4, no. null, p. 83–99, Dec. 2003.

[8] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Learning Theory and Kernel Machines*, B. Schölkopf and M. K. Warmuth, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 567–580.

[9] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" 2019.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" 2014.

[11] T. Tommasi, F. Orabona, and B. Caputo, "Learning categories from few examples with multi model knowledge transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 928–941, 2014.

[12] J. Yang, R. Yan, and A. G. Hauptmann, "Adapting svm classifiers to data with shifted distributions," in *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, 2007, pp. 69–76.

[13] A. K. Lampinen and S. Ganguli, "An analytic theory of generalization dynamics and transfer learning in deep linear networks," 2019.

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

[15] L. Saglietti and L. Zdeborová, "Solvable model for inheriting the regularization through knowledge distillation," 2020.

[16] M. Stojnic, "A framework to characterize performance of lasso algorithms," 2013.

[17] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Precise error analysis of regularized m-estimators in high-dimensions," *CoRR*, vol. abs/1601.06233, 2016. [Online]. Available: http://arxiv.org/abs/1601.06233

[18] Y. Gordon, "On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$," in *Geometric Aspects of Functional Analysis*, J. Lindenstrauss and V. D. Milman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1988, pp. 84–106.

[19] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, "Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms," *CoRR*, vol. abs/1805.09555, 2018.

[20] O. Dhifallah and Y. M. Lu, "A precise performance analysis of learning with random features," 2020.

[21] F. Salehi, E. Abbasi, and B. Hassibi, "The impact of regularization on high-dimensional logistic regression," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 12 005–12 015.

[22] A. Kammoun and M.-S. Alouini, "On the precise error analysis of support vector machines," 2020.

[23] F. Mignacco, F. Krzakala, Y. M. Lu, and L. Zdeborová, "The role of regularization in classification of high-dimensional noisy gaussian mixture," 2020.

[24] B. Aubin, F. Krzakala, Y. M. Lu, and L. Zdeborová, "Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization," 2020.

[25] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *Proceedings of The 28th Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, P. Grünwald, E. Hazan, and S. Kale, Eds., vol. 40. Paris, France: PMLR, 03–06 Jul 2015, pp. 1683–1709.

[26] M. Rudelson and R. Vershynin, "Non-asymptotic theory of random matrices: extreme singular values," 2010.

[27] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. SpringerVerlag Berlin Heidelberg, 1998.

[28] S. Adachi, S. Iwata, Y. Nakatsukasa, and A. Takeda, "Solving the trust-region subproblem by a generalized eigenvalue problem," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 269–291, 2017. [Online]. Available: https://doi.org/10.1137/16M1058200

[29] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2014.

[30] P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: A large sample study," *Ann. Statist.*, vol. 10, no. 4, pp. 1100–1120, 12 1982. [Online]. Available: https://doi.org/10.1214/aos/1176345976

[31] W. K. Newey and D. Mcfadden, "Chapter 36 large sample estimation and hypothesis testing," in *of Handbook of Econometrics*, 1994, p. 2111.

[32] R. L. Schilling, *Measures, Integrals and Martingales*. Cambridge University Press, 2005.

[33] M. Debbah, W. Hachem, P. Loubaton, and M. de Courville, "Mmse analysis of certain large isometric random precoded systems," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1293–1311, 2003.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.