entropy

MDPI

Article

Phase Transitions in Transfer Learning for High-Dimensional Perceptrons

Oussama Dhifallah * and Yue M. Lu

John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA; yuelu@seas.harvard.edu

* Correspondence: oussama_dhifallah@g.harvard.edu

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.

Keywords: transfer learning; statistics; phase transitions

Citation: Dhifallah, O.; Lu, Y. M. Phase Transitions in Transfer Learning for High-Dimensional Perceptrons. *Entropy* **2021**, *23*, 400. https://doi.org/10.3390/e23040400

Academic Editors: Nariman Farsad, Marco Mondelli and Morteza Mardani

Received: 4 January 2021
Accepted: 24 March 2021
Published: 27 March 2021

1. 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 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.

1.1. 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 assume that the labels $\{y_{s,i}\}_{i=1}^{n_s}$ are generated according to the following model:

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

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

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

where $\boldsymbol{\xi}_t \in \mathbb{R}^p$ is an unknown target teacher vector. We measure the similarity of the two tasks using

$$\rho \overset{\text{def}}{=} \frac{\boldsymbol{\xi}_t^\top \boldsymbol{\xi}_s}{\|\boldsymbol{\xi}_t\| \|\boldsymbol{\xi}_s\|}, \tag{3}$$

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:

$$\widehat{\boldsymbol{w}}_s = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{p} \sum_{i=1}^{n_s} \ell\left(y_{s,i}; \boldsymbol{a}_{s,i}^\top \boldsymbol{w}\right) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2, \tag{4}$$

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

$$\begin{cases} \ell(y; x) = \widehat{\ell}(y - x), & \text{for regression task} \\ \ell(y; x) = \widehat{\ell}(yx), & \text{for classification task,} \end{cases} \tag{5}$$

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:

$$\widehat{\boldsymbol{w}}_t = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{p} \sum_{i=1}^{n_t} \ell\left(y_{t,i}; \boldsymbol{a}_{t,i}^\top \boldsymbol{w}\right) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \tag{6}$$

$$\text{s.t. } \boldsymbol{Q}\boldsymbol{w} = \boldsymbol{Q}\widehat{\boldsymbol{w}}_s. \tag{7}$$

The vector $\widehat{\boldsymbol{w}}_s$ 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 $\delta = m/p \leq 1$. Here, m denotes the number of transferred components. Thus, on average, we retain δp number of entries from the source optimal vector $\widehat{\boldsymbol{w}}_s$. In addition to a possible improvement in 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 (6) 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

$$\widehat{\boldsymbol{w}}_t = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\operatorname{argmin}}\ \frac{1}{p} \sum_{i=1}^{n_t} \ell\left(y_{t,i}; \boldsymbol{a}_{t,i}^\top \boldsymbol{w}\right) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{2} \|\boldsymbol{\Sigma}(\boldsymbol{w} - \widehat{\boldsymbol{w}}_s)\|^2, \tag{8}$$

with $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ denoting some weighting matrix. In what follows, we refer to (8) as the *soft transfer* formulation, since it relaxes the strict equality in (6). In fact, the hard transfer in (6) is just a special case of the soft transfer formulation, if we set $\boldsymbol{\Sigma}$ to be a diagonal matrix in which the 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

$$\widehat{y}_{t,\text{new}} = \widehat{\varphi}[\widehat{\boldsymbol{w}}_t^\top \boldsymbol{a}_{t,\text{new}}], \tag{9}$$

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

$$\mathcal{E}_{\text{test}} = \frac{1}{4^v}\mathbb{E}\Big[\big(y_{t,\text{new}} - \widehat{\varphi}(\widehat{\boldsymbol{w}}_t^\top \boldsymbol{a}_{t,\text{new}})\big)^2\Big], \tag{10}$$

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

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

to quantify the performance of the training process. Here, we measure the training error on the training data without regularization.

1.2. Main Contributions

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

1.2.1. Precise Asymptotic Analysis

We present a precise asymptotic analysis of the transfer learning approaches introduced in (6) and (8) for Gaussian feature vectors and under regularity conditions on the eigenvalue distribution of the weighting matrix $\boldsymbol{\Sigma}$. 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 1a 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 (6) 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 (8) seems to achieve more favorable generalization errors compared to the hard formulation. In Figure 1b, 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).

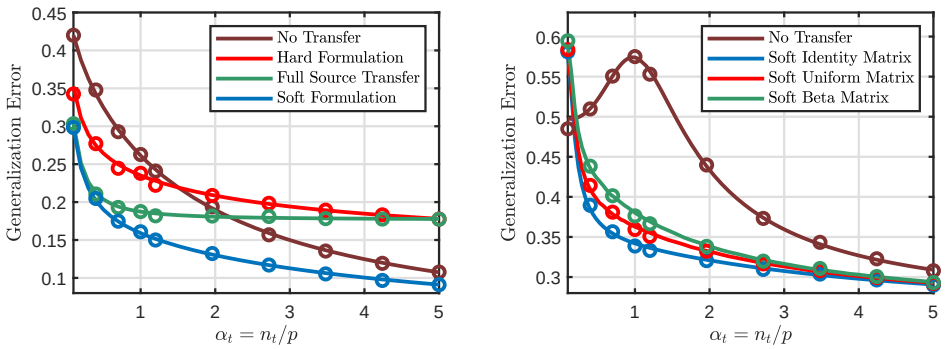

Figure 1. Theoretical predictions v.s. numerical simulations obtained by averaging over 100 independent Monte Carlo trials with dimension $p = 2500$. (**a**) Binary classification with logistic loss. We take $\alpha_s = 10\alpha_t$, $\lambda = 0.3$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p/\sqrt{5}$, 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 (8). Soft Identity Matrix: $\boldsymbol{\Sigma}$ is an identity matrix. Soft Uniform Matrix: $\boldsymbol{\Sigma}$ is a random matrix with diagonal elements drawn from the uniform distribution. Soft Beta Matrix: $\boldsymbol{\Sigma}$ is a random matrix with diagonal elements drawn from the beta distribution. 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 $\rho = 0.8$.

1.2.2. 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 2a 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 also appear in nonlinear regression, as shown in Figure 2b. 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*

$$\rho > \rho_c(\alpha_s, \alpha_t) = 1 - \frac{\mathbb{E}[\varphi^2(z)] - \mathbb{E}^2[z\varphi(z)]}{2\,\mathbb{E}^2[z\varphi(z)]}\left(\frac{1}{\alpha_t - 1} - \frac{1}{\alpha_s - 1}\right), \tag{11}$$

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

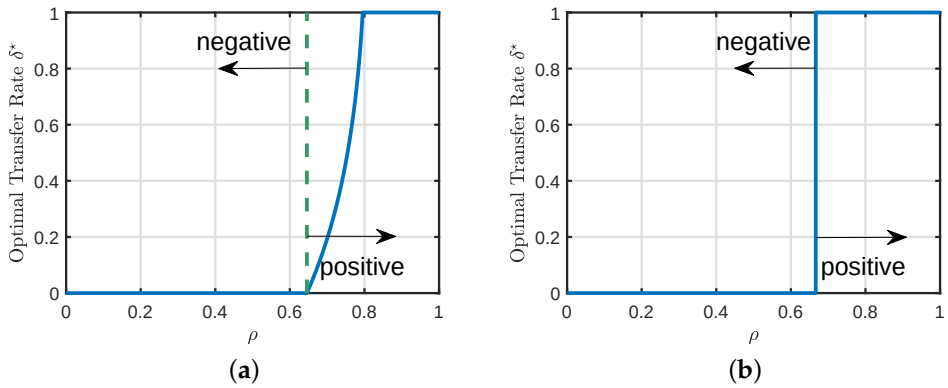

Figure 2. Phase transitions of the hard transfer formulation. When the similarity ρ between the two tasks is small, we are in the negative transfer regime, where we should not transfer the knowledge from the source task. However, as ρ moves past a critical threshold, we enter the positive transfer regime. (**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.

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 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 (11) 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.

1.3. 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 was 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].

1.4. Organization

The rest of this paper is organized as follows. Section 2 states the technical assumptions under which our results are obtained. Section 3 provides an asymptotic characterization of the soft transfer formulation. Precise analysis of the hard transfer formulation is presented in Section 4. We provide remarks about our approach in Section 5. Our theoretical predictions hold for general convex loss functions. We specialize these results to the settings of nonlinear regression and binary classification in Section 6, where we also provide additional numerical results to validate our predictions. Section 7 provides detailed proof of the technical statements introduced in Sections 3 and 4. Section 8 concludes the paper. The

Appendix provides additional technical details.

2. 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.*

Moreover, our results are valid in a 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* $p \to \infty$*.*

The CGMT framework makes specific assumptions about the loss function and the feasibility sets. To guarantee these assumptions, this paper considers a family of loss functions that satisfy the following conditions. Note that the assumption is stated for the target task, but we assume that it is also valid for the source task.

Assumption 3 (Loss Function)**.** *If* $\lambda > 0$*, the loss function* $\ell(y;.)$ *defined in* (5) *is a proper convex function in* $\mathbb{R}$*. If* $\lambda = 0$*, the loss function* $\ell(y;.)$ *defined in* (5) *is a proper strongly convex function in* $\mathbb{R}$*, where the constant* $S > 0$ *is a strong convexity parameter. In this case, we only consider the case when* $\alpha_t > 1$*. 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 and* $\sim$ *denoting equality in distribution. Denote by* $\partial\mathcal{L}$ *the sub-differential set of* $\mathcal{L}(\boldsymbol{x})$*. Then, for any constant* $C > 0$*, there exists a constant* $R > 0$ *such that*

$$\begin{cases} \mathbb{P}\Big(\sup_{\|\boldsymbol{v}\| \le C\sqrt{n_t}} \sup_{\boldsymbol{s} \in \partial\mathcal{L}(\boldsymbol{v})} \|\boldsymbol{s}\| \le R\sqrt{n_t}\Big) \xrightarrow{p\to\infty} 1. \\ \mathbb{P}\Big(\sup_{\|\boldsymbol{v}\| \le C\sqrt{n_t}} |\mathcal{L}(\boldsymbol{v})| \le Rn_t\Big) \xrightarrow{p\to\infty} 1. \end{cases} \tag{12}$$

Furthermore, we consider the following assumption to guarantee that the generalization error defined in (10) concentrates in the large system limit.

Assumption 4 (Regularity Conditions)**.** *The data-generating function* $\varphi(\cdot)$ *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*

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

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* $\sigma_{min,1}(\boldsymbol{\Lambda})$ *and* $\sigma_{min,2}(\boldsymbol{\Lambda})$ *denote its two smallest eigenvalues. There exists a constant* $\mu_{min} \ge 0$ *such that*

$$\begin{cases} \sigma_{min,1}(\boldsymbol{\Lambda}) \xrightarrow{p} \mu_{min} \\ |\sigma_{min,1}(\boldsymbol{\Lambda}) - \sigma_{min,2}(\boldsymbol{\Lambda})| \xrightarrow{p} 0. \end{cases} \tag{13}$$

Moreover, we assume that empirical distribution of the eigenvalues of the matrix $\mathbf{\Lambda}$ *converges weakly to a probability distribution* $\mathbb{P}_\mu(\cdot)$.

The above assumptions are essential to show that the soft formulation in (8) concentrates in the large system limit. We provide more details about these assumptions in Appendix A.

3. Sharp Asymptotic Analysis of Soft Transfer Formulation

In this section, we study the asymptotic properties of soft transfer formulation. Specifically, we provide a precise characterization of the training and generalization errors corresponding to (8).

The asymptotic performance of the source formulation defined in (4) has been studied in the literature [24]. In particular, it has been shown that the asymptotic limit of the source formulation in (4) can be quantified by the following deterministic optimization problem:

$$\min_{q_s, r_s \geq 0} \sup_{\sigma_s > 0} \alpha_s \mathbb{E}\Big[\mathcal{M}_{\ell(Y_s,.)}\Big(r_s H_s + q_s S_s; \frac{r_s}{\sigma_s}\Big)\Big] - \frac{r_s \sigma_s}{2} + \frac{\lambda}{2}(q_s^2 + r_s^2), \tag{14}$$

where $Y_s = \varphi(S_s)$ 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 (14) is the Moreau envelope function defined as

$$\mathcal{M}_{\ell(y,.)}(a; b) = \min_{c \in \mathbb{R}} \ell(y; c) + \frac{1}{2b}(c - a)^2. \tag{15}$$

The expectation in (14) is taken over the random variables H_s and S_s.

In our work, we focus on the target problem with soft transfer, as formulated in (8). It turns out that the asymptotic performance of the target problem can also be characterized by a deterministic optimization problem:

$$\begin{aligned} \min_{q_t, r_t \geq 0} \sup_{\sigma_t > -\mu_{\min}} & -\frac{\sigma_t r_t^2}{2} + \frac{1}{2}\Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big) T_2(\sigma_t) \\ & + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_t H_t + q_t S_t; T_1(\sigma_t)\Big)\Big] + \frac{\lambda}{2}(q_t^2 + r_t^2) \\ & - \frac{1}{2}(q_t - \rho q_s^\star)^2 (\sigma_t - 1/T_1(\sigma_t)), \end{aligned} \tag{16}$$

where $Y_t = \varphi(S_t)$, 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 (16), the constants $q_s^\star$ and $r_s^\star$ are optimal solutions of the asymptotic formulation given in (14). Moreover, the functions $T_1(.)$ and $T_2(.)$ are defined as follows:

$$T_1(\sigma_t) = \mathbb{E}_\mu[1/(\mu + \sigma_t)], \; T_2(\sigma_t) = \mathbb{E}_\mu[\mu \sigma_t/(\mu + \sigma_t)],$$

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 Assumptions 1–5 are satisfied. Then, the training error corresponding to the soft transfer formulation in* (8) *converges in probability as follows:*

$$\mathcal{E}_{train} \xrightarrow{p \to \infty} C_t^\star - \frac{\lambda}{2}\big((q_t^\star)^2 + (r_t^\star)^2\big), \tag{17}$$

where $C_t^\star$ denotes the minimum value achieved by the scalar formulation introduced in (16), and $q_t^\star$ and $r_t^\star$ are optimal solutions of the scalar formulation in (16). Moreover, the generalization error introduced in (10) corresponding to soft transfer formulation converges in probability as follows:

$$\mathcal{E}_{test} \xrightarrow{p\to\infty} \frac{1}{4^v}\mathbb{E}\Big[\big(\varphi(\nu_1)-\widehat{\varphi}(\nu_2)\big)^2\Big], \tag{18}$$

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). A detailed proof is provided in Section 7.3. 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 7.3 shows that the deterministic problems in (14) and (16) are the asymptotic limits of the source and target formulations given in (4) and (8), respectively. Moreover, it shows that the deterministic problems (14) and (16) are strictly convex in the minimization 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 soft transfer formulation can be fully characterized using the optimal solutions of scalar formulation in (16). Moreover, from its definition, (16) depends on the optimal solutions of the scalar formulation in (14) 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.*

4. Sharp Asymptotic Analysis of Hard Transfer Formulation

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

4.1. Asymptotic Predictions

As mentioned earlier, the hard transfer formulation can be recovered from (8) 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 Assumptions 1–4 are satisfied. Then, the asymptotic limit of the hard formulation defined in (6) is given by the following deterministic formulation:*

$$\begin{aligned}\min_{q_t,r_t\geq 0}\sup_{\sigma>0}\ &\frac{\lambda}{2}(q_t^2+r_t^2)+\frac{\sigma\delta}{2}\big[(1-\rho^2)(q_s^\star)^2+(r_s^\star)^2\big]\\ &+\alpha_t\mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_tH_t+q_tS_t;\frac{1-\delta}{\sigma}\Big)\Big]-\frac{\sigma r_t^2}{2}\\ &+\frac{\sigma\delta}{2(1-\delta)}(q_t-\rho q_s^\star)^2.\end{aligned} \tag{19}$$

Additionally, the training and generalization errors associated with the hard formulation converge in probability to the limits given in (17) and (18), respectively.

4.2. 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 (19).

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 (4) and (6) 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 (4) and (6).
(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 where the function $\widehat{\varphi}(\cdot)$ in (9) is the identity function.

Proposition 1 (Regression Phase Transition)**.** *In addition to conditions (a)–(c) introduced above, assume that the predefined function $\widehat{\varphi}(\cdot)$ in (9) is the identity function. Let $\delta^\star$ be the optimal transfer rate that leads to the lowest generalization error in the hard formulation (6). Then,*

$$\delta^\star=\begin{cases}0 & \text{if } \rho<\rho_c(\alpha_s,\alpha_t)\\ 1 & \text{if } \rho>\rho_c(\alpha_s,\alpha_t),\end{cases} \tag{20}$$

where $\rho_c(\alpha_s,\alpha_t)$ is defined in (11).

The result of Proposition 1, for which the proof can be found in Section 7.4, 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 is $\rho<\rho_c(\alpha_s,\alpha_t)$, the optimal transfer ratio is $\delta^\star=0$, 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.

4.3. Sufficient Condition

Next, we consider a binary classification task, where the nonlinear functions $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are both the sign function. In this part, we provide a *sufficient* condition for when the hard transfer is beneficial. Before stating our predictions, we need a few definitions related to the Moreau envelope function defined in (15). For simplicity of notation, we refer to the Moreau envelope function as $\mathcal{M}_\ell(\cdot,\cdot)$. Based on [25], $\mathcal{M}_\ell(\cdot,\cdot)$ is differentiable in $\mathbb{R}\times\mathbb{R}^+$. We refer to its derivatives with respect to the first and second arguments as $\mathcal{M}'_{\ell,1}(\cdot,\cdot)$ and $\mathcal{M}'_{\ell,2}(\cdot,\cdot)$, respectively. If $\mathcal{M}_\ell(\cdot,\cdot)$ is twice differentiable, we refer to its second derivative with respect to the first and second arguments as $\mathcal{M}''_{\ell,1}(\cdot,\cdot)$ and $\mathcal{M}''_{\ell,2}(\cdot,\cdot)$, respectively. Additionally, we refer to its second derivative with respect to the first then the second arguments as $\mathcal{M}''_{\ell,12}(\cdot,\cdot)$.

We define q_0, r_0, and σ_0 as the optimal solutions of the standard learning formulation (i.e., $\delta=0$ in (19)). Moreover, we define the constants β_1 and β_2 as follows:

$$\beta_1=(1-\rho^2)(q_s^\star)^2+(r_s^\star)^2,\ \beta_2=\rho q_s^\star, \tag{21}$$

where $q_s^\star$ and $r_s^\star$ are optimal solutions of the deterministic source formulation given in (14). Define the constants I_{11}, I_{12}, I_{13}, and I_{14} as follows:

$$\begin{cases} I_{11}=\alpha_t\mathbb{E}\Big(SH\mathcal{M}''_{\ell,1}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big),\ I_{12}=\alpha_t\mathbb{E}\Big(S^2\mathcal{M}''_{\ell,1}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big)+\lambda\\ I_{13}=-\frac{\alpha_t}{\sigma_0^2}\mathbb{E}\Big(S\mathcal{M}''_{\ell,12}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big);\ I_{14}=-\frac{\alpha_t}{\sigma_0}\mathbb{E}\Big(S\mathcal{M}''_{\ell,12}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big)+\sigma_0(q_0-\beta_2),\end{cases}$$

where $Y=\varphi(S)$, and H and S are two independent standard Gaussian random variables. Now, define the constants I_{21}, I_{22}, I_{23}, and I_{24} as follows:

$$\begin{cases} I_{21}=\alpha_t\mathbb{E}\Big(H^2\mathcal{M}''_{\ell,1}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big)-\sigma_0+\lambda,\ I_{22}=\alpha_t\mathbb{E}\Big(HS\mathcal{M}''_{\ell,1}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big)\\ I_{23}=-\frac{\alpha_t}{\sigma_0^2}\mathbb{E}\Big(H\mathcal{M}''_{\ell,12}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big)-r_0,\ I_{24}=-\frac{\alpha_t}{\sigma_0}\mathbb{E}\Big(H\mathcal{M}''_{\ell,12}[r_0H+q_0S;\frac{1}{\sigma_0}]\Big).\end{cases}$$

Finally, define the constants I_{31}, I_{32}, I_{33}, and I_{34} as follows:

$$\begin{cases} I_{31} = -\frac{\alpha_t}{\sigma_0^2}\mathbb{E}\Big(H\mathcal{M}''_{\ell,21}[r_0H + q_0S; \frac{1}{\sigma_0}]\Big) - r_0, \; I_{32} = -\frac{\alpha_t}{\sigma_0^2}\mathbb{E}\Big(S\mathcal{M}''_{\ell,12}[r_0H + q_0S; \frac{1}{\sigma_0}]\Big) \\ I_{33} = \frac{2\alpha_t}{\sigma_0^3}\mathbb{E}\Big(\mathcal{M}'_{\ell,2}[r_0H + q_0S; \frac{1}{\sigma_0}]\Big) + \frac{\alpha_t}{\sigma_0^4}\mathbb{E}\Big(\mathcal{M}''_{\ell,2}[r_0H + q_0S; \frac{1}{\sigma_0}]\Big) \\ I_{34} = \frac{\alpha_t}{\sigma_0^2}\mathbb{E}\Big(\mathcal{M}'_{\ell,2}[r_0H + q_0S; \frac{1}{\sigma_0}]\Big) + \frac{\alpha_t}{\sigma_0^3}\mathbb{E}\Big(\mathcal{M}''_{\ell,2}[r_0H + q_0S; \frac{1}{\sigma_0}]\Big) + \frac{1}{2}(q_0 - \beta_2)^2 + \frac{\beta_1}{2}. \end{cases}$$

Now, we are ready to state our sufficient condition.

Proposition 2 (Classification)**.** *Assume that the Moreau envelope function is twice continuously differentiable almost everywhere in $\mathbb{R} \times \mathbb{R}^+$ and that the above expectations are all well-defined. Moreover, assume that both $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are the sign function. Then,*

$$\delta^\star > 0 \quad \textit{if} \quad q_0' r_0 - q_0 r_0' > 0, \tag{22}$$

where q_0' and r_0' are solutions to the following linear system of equations:

$$\begin{cases} I_{11}r_0' + I_{12}q_0' + I_{13}\sigma_0' + I_{14} = 0 \\ I_{21}r_0' + I_{22}q_0' + I_{23}\sigma_0' + I_{24} = 0 \\ I_{31}r_0' + I_{32}q_0' + I_{33}\sigma_0' + I_{34} = 0. \end{cases} \tag{23}$$

We prove this result at the end of Section 7. Note that Proposition 2 is valid for a general family of loss functions and general regularization strength $\lambda \geq 0$. For instance, we can see that the results stated in Proposition 2 are valid for the squared loss and the least absolute deviation (LAD) loss, i.e.,

$$\begin{cases} \ell(y; x) = \frac{1}{2}(1 - yx)^2 \\ \ell(y; x) = |1 - yx|. \end{cases} \tag{24}$$

Unlike (20), the result in (22) only provides a *sufficient* condition for when the hard transfer is beneficial. Nevertheless, our numerical simulations show that the sufficient condition in (22) provides a good prediction of the phase transition boundary for the majority of parameter settings.

5. Remarks

5.1. Learning Formulations

Given that the target task predicts the new label with $\widehat{\varphi}(\cdot)$, it is more natural to consider loss functions satisfying the following form:

$$\ell(y_i; \widehat{\varphi}(\boldsymbol{a}_i^\top \boldsymbol{w})).$$

In this case, the convexity assumption is not necessarily satisfied since the loss function can be viewed as the composition of a convex function with a nonlinear function. To guarantee the convexity, we need additional assumptions on the function $\widehat{\varphi}(\cdot)$. Moreover, note that, once the convexity is guaranteed, the function $\widehat{\varphi}(\cdot)$ can be absorbed by the loss function $\ell(\cdot; \cdot)$.

5.2. Transition from Negative to Positive Transfer

Our first simulation example in Figure 2 shows that the optimal transfer rate $\delta^\star$ can be 1 while the similarity ρ is still less than 1. Here, we provide an intuitive explanation of this behavior.

Given that the source and target feature vectors are generated from the same distribution, one can see that the source labels can be equivalently expressed as follows:

$$y_{s,i} = \varphi(\rho \boldsymbol{a}_{t,i}^\top \boldsymbol{\xi}_t + z_i), \; \forall i \in \{1, \ldots, n_s\}, \tag{25}$$

where $\{z_i\}_{i=1}^{n_s}$ is an additive noise caused by the mismatch between the source and target hidden vectors. Moreover, note that the noise strength depends on the similarity measure ρ.

First, consider the case when the number of source samples is bigger than the number of target samples (i.e., $\alpha_s > \alpha_t$). We can see that a large value of ρ means that the source and target models are very closely related. Then, one can expect that the additional available data in the source task will be capable of defeating the effects of noise in (25) for large values of ρ. Specifically, it is expected in this regime that the source model will perform better than the standard learning formulation for values of ρ close to 1. However, as we decrease the similarity ρ, the source model will have a small information about the target data. Then, the performance of the hard formulation is expected to be lower than the standard formulation for small values of ρ. In this regime, the source information may hurt the generalization performance of the target task. Then, we need to only transfer a portion of the source information (see Figure 2a). In some settings, the transition is sharp, which means that the source information is irrelevant for the target task when ρ is smaller than a threshold (see Figure 2b).

Second, consider the case when the number of source samples is smaller than the number of target samples (i.e., $\alpha_s < \alpha_t$). Given the observation in (25), the performance of the standard method is expected to be better than the hard formulation for all possible values of ρ in this regime (see Figure 7).

6. 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.

6.1. Model Assumptions

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

$$y_i = \max(\boldsymbol{a}_i^\top \boldsymbol{\xi}, 0), \ \forall i \in \{1, \dots, n\}. \tag{26}$$

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

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

where 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:

$$y_i = \text{sign}(\boldsymbol{a}_i^\top \boldsymbol{\xi}), \ \forall i \in \{1, \dots, n\}, \tag{27}$$

where 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 (8) 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 optimal solutions of the target scalar formulation given in (16). The generalization error corresponding to the hard formulation given in (6) can be expressed in a similar fashion.

6.2. Phase Transitions in the Hard Formulation

In Section 4, we 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.

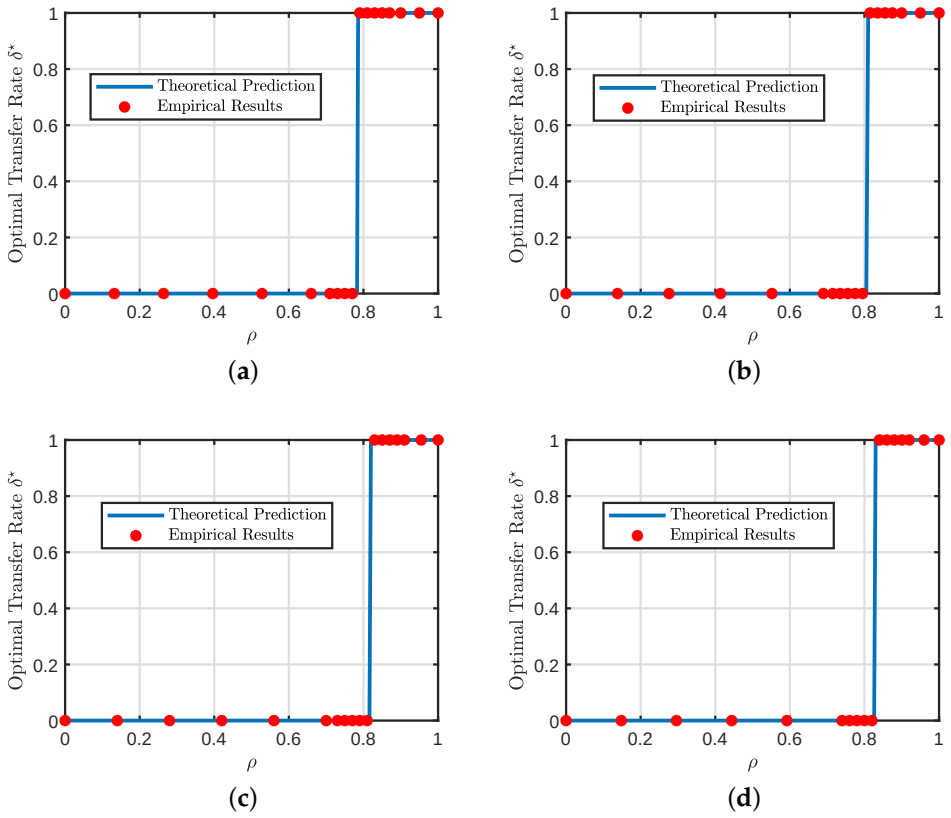

Figure 3. Additional illustrations of the phase transition phenomenon. (**a**) Regression (squared loss, $\alpha_t = 0.5$, and $\alpha_s = 3\alpha_t$) (**b**) Regression (squared loss, $\alpha_t = 2$, and $\alpha_s = 2\alpha_t$) (**c**) Binary classification (squared loss, $\alpha_t = 1.5$, and $\alpha_s = 3\alpha_t$) (**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 4 for multiple values of δ. The empirical results are averaged over 100 independent Monte Carlo trials with $p = 2500$.

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, Figure 3c,d show that the loss function has a small effect on the phase transition boundary.

6.3. Sufficient Condition for the Hard Formulation

In Section 4, we presented a sufficient condition for positive transfer. This sufficient condition is valid for a general family of loss functions and a general regularization strength. The main purpose of this experiment, shown in Figure 4, is to illustrate the precision of the sufficient condition for two particular loss functions, i.e., the squared loss and LAD loss.

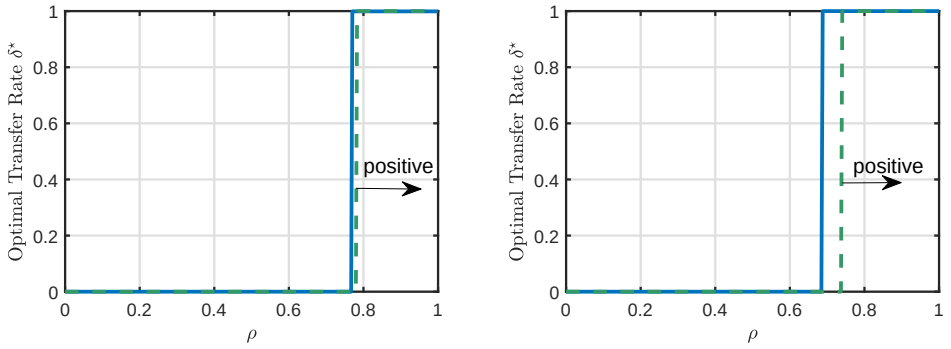

Figure 4. Illustrations of the sufficient condition in Proposition 2. (**a**) Classification (squared loss, $\alpha_t = 1.5$, and $\alpha_s = 8\alpha_t$) (**b**) Classification (LAD loss, $\alpha_t = 1.5$, and $\alpha_s = 8\alpha_t$). In all the experiments, we set the regularization strength to be $\lambda = 0.1$. The blue line represents our theoretical predictions of the optimal transfer rate obtained by solving our asymptotic results in Section 4 for multiple values of δ. The green line represents our sufficient condition for positive transfer stated in Proposition 2.

In all the cases shown in Figure 4, we can see that the transition from negative to positive transfer is a discontinuous jump from standard learning to full source transfer. Additionally, Figure 4a,b show that the sufficient condition summarized in Proposition 2 provides a good prediction of the phase transition boundary for the considered setting.

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

In this experiment, we empirically explore the impact of the weighting matrix $\mathbf{\Sigma}$ on the generalization error corresponding to the soft formulation. We focus on the binary classification problem with logistic loss. The weighting matrix in (8) takes the following form:

$$\mathbf{\Sigma} = \sqrt{\beta_t}\boldsymbol{V}, \tag{28}$$

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; and (3): *Soft Beta*: similar to (2), but with the diagonal entries drawn from the beta distribution, followed by rescaling to the unit mean.

Figure 5a shows that the considered weighting matrix choices have similar generalization performances, with the identity matrix being slightly better than the other alternatives. Moreover, Figure 5b illustrates the effects of the parameter β_t in (28) on the generalization performance. It points to the interesting possibility of "designing" the optimal weight matrix to minimize the generalization error.

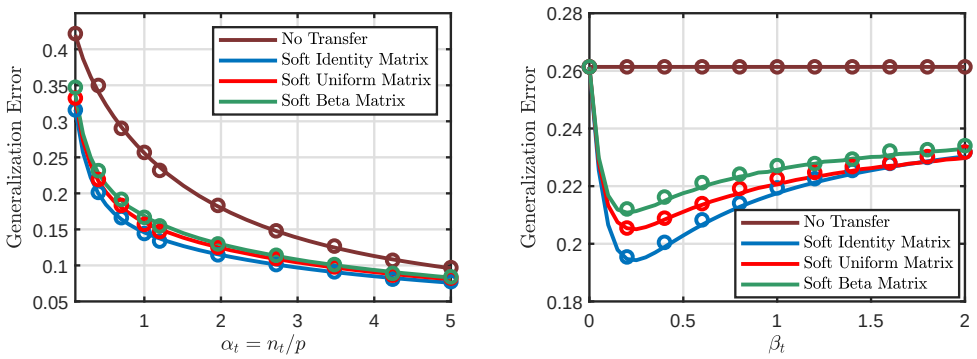

Figure 5. Continuous line: theoretical predictions. Circles: numerical simulations. (**a**) $\alpha_s = 6\alpha_t$, $\lambda = 0.1$, $\beta_t = 1/10$, and $\rho = 0.9$. (**b**) $\alpha_t = 1$, $\alpha_s = 5\alpha_t$, $\lambda = 0.3$, and $\rho = 0.75$. In all the experiments, we consider the binary classification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

6.5. Soft and Hard Transfer Comparison

In this simulation example, we consider the regression model and compare the performances of the hard and soft transfer formulations as functions of α_t and ρ.

Figure 6a 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 6b 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.

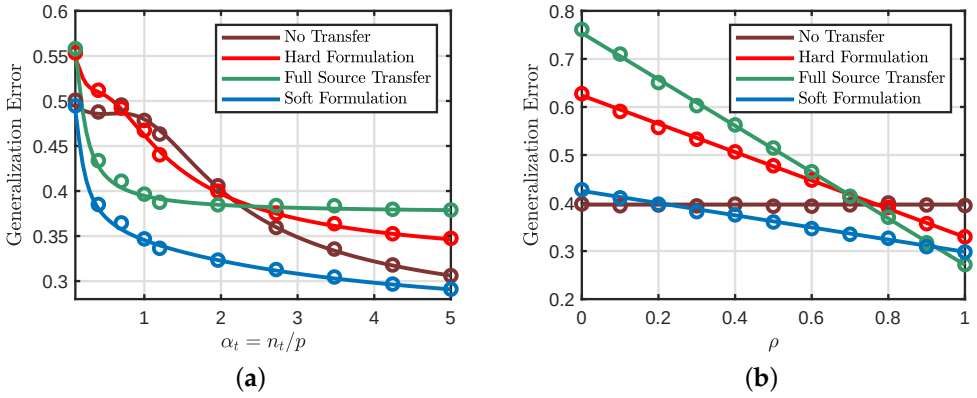

Figure 6. Continuous line: theoretical predictions. Circles: numerical simulations. (**a**) $\alpha_s = 12\alpha_t$, $\lambda = 0.2$, and $\rho = 0.75$. (**b**) $\alpha_t = 1.5$, $\alpha_s = 8\alpha_t$, and $\lambda = 0.4$. In all the experiments, we consider the regression setting with a squared loss. 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 $p = 1000$.

6.6. Effects of the Source Parameters

In the last simulation example, we consider the regression and classification models. We study the performance of the hard and soft transfer formulations when $\alpha_s < \alpha_t$.

Figure 7a considers the regression model. It first shows that the soft transfer formulation provides a slightly better generalization performance compared to the standard method. This behavior can be explained by the fact that the soft formulation requires the target weight vector to be close and not necessarily equal to the source weight vector. Additionally, the source model carries some information about the target task.

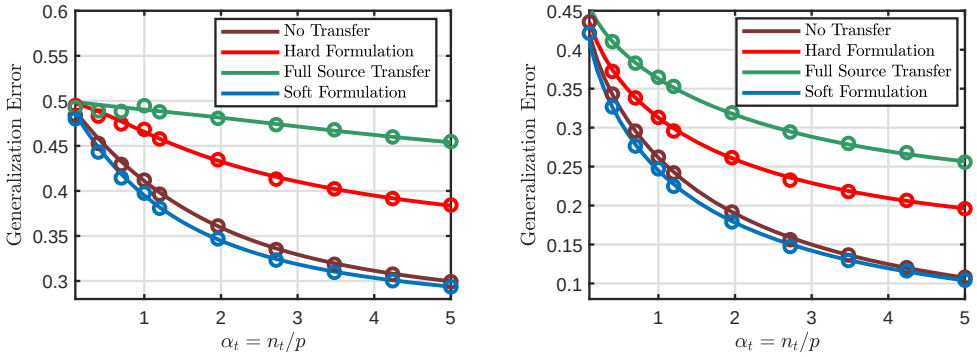

Figure 7. Continuous line: theoretical predictions. Circles: numerical simulations. (**a**) $\alpha_s = 0.5\alpha_t$, $\lambda = 0.6$, and $\rho = 0.7$. We consider the regression setting with a squared loss. (**b**) $\alpha_s = 0.5\alpha_t$, $\lambda = 0.3$, and $\rho = 0.8$. We consider the classification setting with a logistic 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 60 independent Monte Carlo trials, and we set $p = 1000$.

We can also see that the hard transfer approach is not beneficial when the number of source samples is smaller than the number of target samples. This result can be explained by the fact that the hard formulation restricts some entries in the target weight vector to be exactly equal to the corresponding entries in the source weight vector. Moreover, the source model is not perfectly aligned with the target model and has smaller data than the target model (see Section 5.2).

The same behavior can be observed in Figure 7b, which considers the classification model.

7. Technical Details

In this section, we provide a detailed proof of Theorem 1, and Proportions 1 and 2. Specifically, we focus on analyzing the generalized formulation in (8) using the CGMT framework introduced in the following part.

7.1. Technical Tool: Convex Gaussian Min–Max Theorem

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

$$\Phi_p(\boldsymbol{G}) = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \boldsymbol{u}^\top \boldsymbol{G}\boldsymbol{w} + \psi(\boldsymbol{w},\boldsymbol{u}). \tag{29}$$

Specifically, the CGMT shows that the PO given in (29) is asymptotically equivalent to the following formulation:

$$\phi_p(\boldsymbol{g},\boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \|\boldsymbol{u}\|\boldsymbol{g}^\top\boldsymbol{w} + \|\boldsymbol{w}\|\boldsymbol{h}^\top\boldsymbol{u} + \psi(\boldsymbol{w},\boldsymbol{u}), \tag{30}$$

referred to as the auxiliary optimization (AO) problem. Before showing the equivalence between PO and AO, the CGMT assumes that $\boldsymbol{G} \in \mathbb{R}^{n\times p}$, $\boldsymbol{g} \in \mathbb{R}^p$, and $\boldsymbol{h} \in \mathbb{R}^n$; that all have independent and identically distributed standard normal entries; that 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 that 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 $\zeta > 0$, the following holds:

$$\mathbb{P}\big(|\Phi_p(\boldsymbol{G}) - \chi| > \zeta\big) \leq 2\mathbb{P}\big(|\phi_p(\boldsymbol{g}, \boldsymbol{h}) - \chi| > \zeta\big). \tag{31}$$

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 $\mathcal{S}_p$. Moreover, define the set $\mathcal{S}_p^c = \mathcal{S}_{\boldsymbol{w}} \setminus \mathcal{S}_p$. Let ϕ_p and ϕ_p^c be the optimal cost values of AO formulation in (30) with feasibility sets $\mathcal{S}_{\boldsymbol{w}}$ and $\mathcal{S}_p^c$, 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 $\phi_p^c \geq \phi + \zeta$ with probability going to 1 as $p \to +\infty$.

Then, the following convergence in probability holds:

$$|\Phi_p - \phi_p| \overset{p\to+\infty}{\longrightarrow} 0, \text{ and } \mathbb{P}(\widehat{\boldsymbol{w}}_p \in \mathcal{S}_p) \overset{p\to+\infty}{\longrightarrow} 1,,$$

where Φ_p and $\widehat{\boldsymbol{w}}_p$ are the optimal cost and the optimal solution of the PO formulation in (29).

Theorem 2 allows us to analyze the generally easy AO problem to infer the asymptotic properties of the generally hard PO problem. Next, we use the CGMT to rigorously prove the technical results presented in Theorem 1.

7.2. Precise Analysis of the Source Formulation

The source formulation defined in (4) is well-studied in recent literature [26]. Specifically, it has been rigorously proven that the performance of the source formulation can be fully characterized after solving the following scalar formulation:

$$\min_{q_s, r_s \geq 0} \sup_{\sigma_s > 0} \alpha_s \mathbb{E}\Big[\mathcal{M}_{\ell(Y_s,.)}\Big(r_s H_s + q_s S_s; \frac{r_s}{\sigma_s}\Big)\Big] - \frac{r_s \sigma_s}{2} + \frac{\lambda}{2}(q_s^2 + r_s^2), \tag{32}$$

where $Y_s = \varphi(S_s)$, and H_s and S_s are two independent standard Gaussian random variables. The expectation in (32) 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 (32) is the Moreau envelope function defined in (15).

7.3. Precise Analysis of the Soft Transfer Approach

In this part, we provide a precise asymptotic analysis of the generalized transfer formulation given in (8). Specifically, we focus on analyzing the following formulation:

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{p} \sum_{i=1}^{n_t} \ell\Big(y_i; \boldsymbol{a}_i^\top \boldsymbol{w}\Big) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\|\boldsymbol{\Sigma}(\boldsymbol{w} - \widehat{\boldsymbol{w}}_s)\|^2, \tag{33}$$

where $\widehat{\boldsymbol{w}}_s$ is the optimal solution of the source formulation given in (4). 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 7.1 to precisely analyze the above formulation.

7.3.1. Formulating the Auxiliary Optimization Problem

Our first objective is to rewrite the generalized formulation in the form of the PO problem given in (29). To this end, we introduce additional optimization variables. Specifically, the generalized formulation can be equivalently formulated as follows:

$$\min_{\boldsymbol{w}\in\mathbb{R}^p}\max_{\boldsymbol{u}\in\mathbb{R}^{n_t}} \frac{1}{p}\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{w} - \frac{1}{p}\sum_{i=1}^{n_t}\ell^\star(y_i;u_i) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\|\boldsymbol{\Sigma}(\boldsymbol{w}-\widehat{\boldsymbol{w}}_s)\|^2, \tag{34}$$

where the optimization vector $\boldsymbol{u}\in\mathbb{R}^{n_t}$ is formed as $\boldsymbol{u}=[u_1,\ldots,u_{n_t}]^\top$ and the data matrix $\boldsymbol{A}\in\mathbb{R}^{n_t\times p}$ is given by $\boldsymbol{A}=[\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{n_t}]^\top$. 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 (34) satisfies this assumption.

Lemma 1 (Primal-Dual Compactness)**.** *Assume that $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{u}}$ are optimal solutions of the optimization problem in (34). Then, there exist two constants $C_w>0$ and $C_u>0$ such that the following convergence in probability holds:*

$$\mathbb{P}(\|\widehat{\boldsymbol{w}}\|\leq C_w)\xrightarrow{p\to+\infty}1,\ \mathbb{P}(\|\widehat{\boldsymbol{u}}\|/\sqrt{n_t}\leq C_u)\xrightarrow{p\to+\infty}1. \tag{35}$$

A detailed proof of Lemma 1 is provided in Appendix B. The proof of the above result follows using Assumption 3 to prove the compactness of the optimal solution $\widehat{\boldsymbol{w}}$. Moreover, it uses the asymptotic results in [27] (Theorem 2.1), which provides the concentration properties of the minimum and maximum eigenvalues of random matrices. To show the compactness of the optimal dual vector $\widehat{\boldsymbol{u}}$, we use Assumption 3 and the result in [25] (Proposition 11.3), which provides the inversion rules for subgradient relations.

The theoretical result in Lemma 1 shows that the optimization problem in (34) can be equivalently formulated with compact feasibility sets on events with probability going to one. Then, it suffices to study the constrained version of (34). 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:

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

where 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:

$$\boldsymbol{A}=\boldsymbol{s}_t\boldsymbol{\xi}_t^\top+\boldsymbol{G}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}, \tag{36}$$

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. Here, (36) represents an equality in distribution. This means that the formulation in (34) can be expressed as follows:

$$\min_{\|\boldsymbol{w}\|\leq C_w}\max_{\boldsymbol{u}\in\mathcal{C}_t}\frac{1}{p}\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}\boldsymbol{w}+\frac{1}{p}\boldsymbol{u}^\top\boldsymbol{s}_t\boldsymbol{\xi}_t^\top\boldsymbol{w}+\frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \frac{1}{p}\sum_{i=1}^{n_t}\ell^\star(y_i;u_i)+\frac{1}{2}\|\boldsymbol{\Sigma}(\boldsymbol{w}-\widehat{\boldsymbol{w}}_s)\|^2, \tag{37}$$

where the set $\mathcal{C}_t$ is defined as $\mathcal{C}_t = \{\boldsymbol{u} : \|\boldsymbol{u}\|/\sqrt{n_t} \leq C_u\}$. Note that the formulation in (37) is in the form of the primary formulation given in (29). Here, the function $\psi(.,.)$ is defined as follows:

$$\psi(\boldsymbol{w},\boldsymbol{u}) = \frac{1}{p}\boldsymbol{u}^\top \boldsymbol{s}_t \boldsymbol{\xi}_t^\top \boldsymbol{w} + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \frac{1}{p}\sum_{i=1}^{n_t} \ell^\star(y_i;u_i) + \frac{1}{2}\|\boldsymbol{\Sigma}(\boldsymbol{w}-\widehat{\boldsymbol{w}}_s)\|^2. \tag{38}$$

One can easily see that the optimization problem in (37) 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 (37). Then, following the CGMT framework, the auxiliary formulation corresponding to our primary problem in (37) can be expressed as follows:

$$\min_{\|\boldsymbol{w}\|\leq C_w} \max_{\boldsymbol{u}\in\mathcal{C}_t} \frac{\|\boldsymbol{u}\|}{p}\boldsymbol{g}^\top \boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}\boldsymbol{w} + \frac{1}{p}\boldsymbol{u}^\top \boldsymbol{s}_t\boldsymbol{\xi}_t^\top \boldsymbol{w} + \frac{\boldsymbol{h}^\top \boldsymbol{u}}{p}\|\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}\boldsymbol{w}\| + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \frac{1}{p}\sum_{i=1}^{n_t}\ell^\star(y_i;u_i) + \frac{1}{2}\|\boldsymbol{\Sigma}(\boldsymbol{w}-\widehat{\boldsymbol{w}}_s)\|^2, \tag{39}$$

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 on studying its asymptotic properties.

7.3.2. 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:

$$\boldsymbol{w} = (\boldsymbol{\xi}_t^\top \boldsymbol{w})\boldsymbol{\xi}_t + \boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp}\boldsymbol{r}_t, \tag{40}$$

where $\boldsymbol{r}_t \in \mathbb{R}^{p-1}$ is a free vector and $\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp} \in \mathbb{R}^{p\times(p-1)}$ 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 (40), there exist $C_{q_t} > 0$, $C_r > 0$, 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:

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

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:

$$\mathcal{T}_1 = \Big\{(q_t,\boldsymbol{r}_t) : |q_t| \leq C_{q_t}, \|\boldsymbol{r}_t\| \leq C_r\Big\}. \tag{41}$$

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

$$V_{p,t} = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda}\boldsymbol{\xi}_t,\ V_{p,ts} = \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda}\widehat{\boldsymbol{w}}_s. \tag{42}$$

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 an interchange

between non-convex minimization and non-concave maximization. We can justify the interchange using the theoretical result in [17] (Lemma A.3). The main argument in [17] (Lemma A.3) is that the strong convexity of the primary formulation in (37) allows us to perform such an 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:

$$C_p^\star = \min_{\boldsymbol{r}_t \in \mathbb{R}^{p-1}} \boldsymbol{b}_p^\top \boldsymbol{r}_t + \frac{1}{2} \boldsymbol{r}_t^\top \boldsymbol{\Lambda}^\perp \boldsymbol{r}_t, \text{ s.t. } \|\boldsymbol{r}_t\| = r_t. \tag{43}$$

Here, we ignore constant terms independent of $\boldsymbol{r}_t$, and 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:

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

The optimization problem in (43) 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 as in [20], the optimal cost value of the optimization problem (43) can be expressed in terms of a one-dimensional optimization problem as follows:

$$C_p^\star = \sup_{\sigma_t > -\mu_p} \left\{ -\frac{1}{2} \boldsymbol{b}_p^\top [\boldsymbol{\Lambda}^\perp + \sigma_t \boldsymbol{I}_{p-1}]^{-1} \boldsymbol{b}_p - \frac{\sigma_t r_t^2}{2} \right\}, \tag{44}$$

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 (43) as follows:

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

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

$$\begin{aligned} \min_{(q_t, r_t) \in \mathcal{T}_2} \max_{\boldsymbol{u} \in \mathcal{C}_t} \sup_{\sigma_t > -\mu_p} & \frac{r_t}{p} \boldsymbol{h}^\top \boldsymbol{u} + \frac{q_t}{p} \boldsymbol{u}^\top \boldsymbol{s}_t + \frac{\lambda}{2} q_t^2 + \frac{\lambda}{2} r_t^2 \\ & - \frac{1}{p} \sum_{i=1}^{n_t} \ell^\star(y_i; u_i) + \frac{1}{2} q_t^2 V_{p,t} - q_t V_{p,ts} - \frac{\|\boldsymbol{u}\|^2}{2p} T_{p,g}(\sigma_t) \\ & - \frac{\sigma_t r_t^2}{2} - \frac{1}{2} q_t^2 T_{p,t}(\sigma_t) - \frac{1}{2} T_{p,s}(\sigma_t) + q_t T_{p,ts}(\sigma_t), \end{aligned} \tag{45}$$

where the set $\mathcal{T}_2$ has the same definition as the set $\mathcal{T}_1$ except that we replace $\|\boldsymbol{r}_t\|$ with 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:

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

Note that the formulation in (45) is obtained after dropping terms that converge in probability to zero. This simplification can be justified using a similar analysis to that in [20]

(Lemma 3). The main idea in [20] (Lemma 3) 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:

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

This result is valid on events with probability going to one as p goes to $+\infty$. Here, the function $\mathcal{M}_{\ell(y_i,\cdot)}$ is the Moreau envelope function defined in (15). The proof of this property is omitted since it follows the same ideas as [20] (Lemma 4). The main idea in [20] (Lemma 4) is to use Assumption 3 to show that the optimal solution of the unconstrained version of the maximization problem is bounded asymptotically and then to use the property introduced in [25] (Example 11.26) to complete the proof. Now, our auxiliary formulation can be asymptotically simplified to a scalar optimization problem as follows:

$$\begin{aligned} &\min_{(q_t, r_t) \in \mathcal{T}_2} \sup_{\sigma_t > -\mu_p} \frac{\lambda}{2}(q_t^2 + r_t^2) - \frac{\sigma_t r_t^2}{2} - \frac{1}{2} q_t^2 Z_{p,t}(\sigma_t) - \frac{1}{2} Z_{p,s}(\sigma_t) \\ &+ \frac{1}{p} \sum_{i=1}^{n_t} \mathcal{M}_{\ell(y_i,\cdot)}\Big(r_t h_i + q_t s_{t,i}; T_{p,g}(\sigma_t)\Big) + q_t Z_{p,ts}(\sigma_t), \end{aligned} \tag{46}$$

where the functions $Z_{p,t}(\cdot)$, $Z_{p,ts}(\cdot)$, and $Z_{p,s}(\cdot)$ are defined as follows:

$$Z_{p,t}(\sigma_t) = T_{p,t}(\sigma_t) - V_{p,t}, \; Z_{p,ts}(\sigma_t) = T_{p,ts}(\sigma_t) - V_{p,ts} \tag{47}$$

$$Z_{p,s}(\sigma_t) = T_{p,s}(\sigma_t) - V_{p,s}, \text{ where } V_{p,s} = \widehat{\boldsymbol{w}}_s^\top \boldsymbol{\Lambda} \widehat{\boldsymbol{w}}_s. \tag{48}$$

Note that the auxiliary formulation in (46) now has scalar optimization variables. Then, it remains to study its asymptotic properties. We refer to this problem as the target scalar formulation.

7.3.3. Asymptotic Analysis of the Target Scalar Formulation

In this part, we study the asymptotic properties of the target scalar formulation expressed in (46). We start our analysis by studying the asymptotic properties of the sequence of random functions $T_{p,g}(.)$, $Z_{p,t}(.)$, $Z_{p,s}(.)$, and $Z_{p,ts}(.)$ as given in the following lemma.

Lemma 2 (Asymptotic Properties)**.** *First, the random variable μ_p converges in probability to μ_{min}, where μ_{min} is defined in Assumption 5. For any fixed $\sigma > 0$, the following convergence in probability holds true:*

$$\begin{cases} Z_{p,t}(\sigma - \mu_p) \overset{p \to +\infty}{\to} Z_t(\sigma - \mu_{min}) \\ Z_{p,ts}(\sigma - \mu_p) \overset{p \to +\infty}{\to} Z_{ts}(\sigma - \mu_{min}) \\ Z_{p,s}(\sigma - \mu_p) \overset{p \to +\infty}{\to} Z_s(\sigma - \mu_{min}) \\ T_{p,g}(\sigma - \mu_p) \overset{p \to +\infty}{\to} T_g(\sigma - \mu_{min}) = T_1(\sigma - \mu_{min}). \end{cases}$$

Here, the deterministic functions $Z_t(.)$, $Z_{ts}(.)$, $Z_s(.)$, $T_1(.)$, and $T_3(.)$ are defined as follows:

$$\begin{cases} Z_t(\sigma) = \sigma - 1/T_1(\sigma), \; Z_{ts}(\sigma) = \rho q_s^\star Z_t(\sigma) \\ Z_s(\sigma) = ((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2) T_3(\sigma) + (\rho q_s^\star)^2 Z_t(\sigma) \\ T_1(\sigma) = \mathbb{E}_\mu[1/(\mu + \sigma)], \; T_3(\sigma) = -\mathbb{E}_\mu[\mu\sigma/(\mu + \sigma)]. \end{cases}$$

Moreover, the constants $q_s^\star$ and $r_s^\star$ are optimal solutions of the source asymptotic formulation defined in (32).

A detailed proof of Lemma 2 is provided in Appendix C. 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 (46). To state our first asymptotic result, we define the following deterministic optimization problem:

$$\min_{(q_t,r_t)\in\mathcal{T}_2} \sup_{\sigma_t>-\mu_{\min}} \frac{\lambda}{2}(q_t^2+r_t^2)-\frac{\sigma_t r_t^2}{2}-\frac{1}{2}Z_s(\sigma_t)-\frac{1}{2}q_t^2 Z_t(\sigma_t) + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_t H_t+q_t S_t; T_g(\sigma_t)\Big)\Big]+q_t Z_{ts}(\sigma_t), \tag{49}$$

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 (15) and the expectation is taken over the random variables H_t and S_t, and the possibly random function $\varphi(.)$. Now, we are ready to state our asymptotic property of the cost function of (46).

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 (46). Additionally, define $\mathcal{O}_t(.)$ as the cost function of the deterministic formulation in (49). Then, the following convergence in probability holds true:*

$$\mathcal{O}_{p,t}(q_t,r_t,\sigma_t-\mu_p) \overset{p\to+\infty}{\to} \mathcal{O}_t(q_t,r_t,\sigma_t-\mu_{min}), \tag{50}$$

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

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:

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{M}_{\ell(y_{i,.})}\big(r_t h_i+q_t s_{t,i};x\big) \overset{p\to+\infty}{\to} \mathbb{E}\big[\mathcal{M}_{\ell(Y,.)}\big(r_t H+q_t S;x\big)\big],$$

where 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 [25] (Theorem 2.26), the Moreau envelope function is convex and continuously differentiable with respect to $x>0$. 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:

$$\mathbb{E}\big[\mathcal{M}_{\ell(Y,.)}\big(r_t H+q_t S;T_g(\sigma_t-\mu_{\min})\big)\big], \tag{51}$$

for any fixed feasible q_t, r_t, and $\sigma_t>0$. This completes the proof of Lemma 3.

Before continuing our analysis, we provide the convexity properties of the cost function of the deterministic problem in (49) in the following lemma.

Lemma 4 (Strong Conexity)**.** *Define $\mathcal{O}_t(\cdot,\cdot,\cdot)$ as the cost function of the optimization problem in (49). Then, $\mathcal{O}_t(\cdot,\cdot,\cdot)$ is concave in the maximization variable σ_t for any fixed feasible (q_t,r_t). Moreover, define the function $\mathcal{O}_t(\cdot,\cdot)$ as follows:*

$$\mathcal{O}_t(q_t,r_t)=\sup_{\sigma_t>-\mu_{min}} f(q_t,r_t,\sigma_t). \tag{52}$$

Then, the function $\mathcal{O}_t(\cdot,\cdot)$ is strongly convex in the minimization variables (q_t,r_t).

The proof of Lemma 4 is provided in Appendix D. Now, we use these properties to show that the optimal solution set of the formulation in (46) converges in probability to the optimal solution set of the formulation in (49).

Lemma 5 (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 (46) and (49). Moreover, define $\mathcal{O}^\star_{p,t}$ and $\mathcal{O}^\star_t$ as the optimal cost values of the optimization problems formulated in (46) and (49). Then, the following converges in probability holds true:*

$$\mathcal{O}^\star_{p,t} \overset{p\to+\infty}{\to} \mathcal{O}^\star_t, \ \mathbb{D}(\mathcal{P}_{p,t}, \mathcal{P}_t) \overset{p\to+\infty}{\to} 0, \tag{53}$$

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 proven by first observing that the loss function $\mathcal{O}_t(.)$ corresponding to the deterministic formulation in (49) satisfies the following:

$$\lim_{\sigma_t \to +\infty} \mathcal{O}_t(q_t, r_t, \sigma_t - \mu_{\min}) = -\infty \tag{54}$$

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

$$\sup_{\sigma_t > 0} \mathcal{O}_{p,t}(q_t, r_t, \sigma_t - \mu_p) \overset{p\to+\infty}{\to} \sup_{\sigma_t > 0} \mathcal{O}_t(q_t, r_t, \sigma_t - \mu_{\min}).$$

Here, the results in [17] (Lemma B.1) and [17] (Lemma B.2) provide convergence properties of minimization problems over open sets. Note that, if $r_t = 0$, the supremum in the above convergence result occurs at $\sigma_t \to +\infty$. However, it can be checked that the above convergence result still holds. Based on Lemma 4, the cost function of the minimization problem in (49) is strongly convex in (q_t, r_t). Moreover, the feasibility set of the minimization problem is convex and compact. Additionally, the cost function of the minimization problem in (49) is continuous in the feasibility set. Then, using the results in [30] (Theorem II.1) and [31] (Theorem 2.1), we obtain the convergence properties stated in Lemma 5. Here, the results in [30] (Theorem II.1) and [31] (Theorem 2.1) provide uniform convergence and consistency properties of convex optimization problems.

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 (8).

7.3.4. Specialization to Hard Formulation

Before starting the analysis of the generalization error, we specialize our general analysis to the hard transfer formulation. First, note that $\delta = 1$ implies that the hard transfer formulation is equivalent to the source formulation. Next, we assume that $\delta < 1$. To obtain the asymptotic limit of the hard formulation, we specialize the general results in (49) to the following probability distribution:

$$\mathbb{P}_p(\mu) = \begin{cases} 0 & \text{with probability } (1-\delta) \\ +\infty & \text{with probability } \delta. \end{cases} \tag{55}$$

Note that the probability distribution in (55) satisfies Assumption 5. Then, the asymptotic

limit of the soft formulation corresponding to the probability distribution $\mathbb{P}_{\mu}(.)$, defined in (55), can be expressed as follows:

$$\begin{aligned}\min_{(q_t,r_t)\in\mathcal{T}_2}\sup_{\sigma_t>0}\ &\frac{\lambda}{2}(q_t^2+r_t^2)+\frac{\sigma_t\delta}{2}\Big((1-\rho^2)(q_s^\star)^2+(r_s^\star)^2\Big)\\&+\alpha_t\mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,.)}\Big(r_tH_t+q_tS_t;\frac{1-\delta}{\sigma_t}\Big)\Big]-\frac{\sigma_t r_t^2}{2}\\&+\frac{\sigma_t\delta}{2(1-\delta)}(q_t-\rho q_s^\star)^2.\end{aligned}\tag{56}$$

This shows that the asymptotic limit of the hard formulation is the deterministic problem (56).

7.3.5. Asymptotic Analysis of the Training and Generalization Errors

First, the generalization error corresponding to the target task is given by

$$\mathcal{E}_{\text{test}}=\frac{1}{4^v}\mathbb{E}\Big[\Big(\varphi(\boldsymbol{a}_{t,\text{new}}^\top\boldsymbol{\xi}_t)-\widehat{\varphi}(\widehat{\boldsymbol{w}}_t^\top\boldsymbol{a}_{t,\text{new}})\Big)^2\Big],\tag{57}$$

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:

$$\boldsymbol{C}_p=\begin{bmatrix}\|\boldsymbol{\xi}_t\|^2 & \boldsymbol{\xi}_t^\top\widehat{\boldsymbol{w}}_t\\ \boldsymbol{\xi}_t^\top\widehat{\boldsymbol{w}}_t & \|\widehat{\boldsymbol{w}}_t\|^2\end{bmatrix}.\tag{58}$$

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:

$$\widehat{q}_{p,t}^\star=\boldsymbol{\xi}_t^\top\widehat{\boldsymbol{w}}_t,\ \text{and}\ \widehat{r}_{p,t}^\star=\|(\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top\widehat{\boldsymbol{w}}_t\|,\tag{59}$$

where $\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp$ is defined in Section 7.3.2. Then, the covariance matrix $\boldsymbol{C}_p$ given in (58) 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 6 (Consistency of the Target Formulation)**.** *The random quantities $\widehat{q}_{p,t}^\star$ and $\widehat{r}_{p,t}^\star$ satisfy the following asymptotic properties:*

$$\widehat{q}_{p,t}^\star\xrightarrow{p\to+\infty}q_t^\star,\ \textit{and}\ \widehat{r}_{p,t}^\star\xrightarrow{p\to+\infty}r_t^\star,$$

where $q_t^\star$ and $r_t^\star$ are the optimal solutions of the deterministic formulation stated in (49).

To prove the above asymptotic result, we define $\widetilde{q}_{p,t}^\star$ and $\widetilde{r}_{p,t}^\star$ as follows:

$$\widetilde{q}_{p,t}^\star=\boldsymbol{\xi}_t^\top\widetilde{\boldsymbol{w}}_t,\ \text{and}\ \widetilde{r}_{p,t}^\star=\|(\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top\widetilde{\boldsymbol{w}}_t\|,\tag{60}$$

where $\widetilde{\boldsymbol{w}}_t$ is the optimal solution of the auxiliary formulation in (39). Given the result in Lemma 5 and the analysis in Sections 7.3.2 and 7.3.3, the convergence result in Lemma 5 is also satisfied by our auxiliary formulation in (39), i.e.,

$$\widetilde{q}^{\star}_{p,t} \xrightarrow{p\to+\infty} q^{\star}_t, \text{ and } \widetilde{r}^{\star}_{p,t} \xrightarrow{p\to+\infty} r^{\star}_t.$$

The rest of the proof of the convergence result stated in Lemma 6 is based on the CGMT framework, i.e., Theorem 2. Specifically, it follows after showing that the assumptions in Theorem 2 are all satisfied. First, we define the set $\mathcal{S}_p$ in Theorem 2 as follows:

$$\mathcal{S}_p = \{\boldsymbol{w} \in \mathbb{R}^p : |\boldsymbol{\xi}_t^{\top}\boldsymbol{w} - q_t^{\star}| < \epsilon\} \cup \{\boldsymbol{w} \in \mathbb{R}^p : |\|(\boldsymbol{B}_{\boldsymbol{\xi}_t}^{\perp})^{\top}\boldsymbol{w}\| - r_t^{\star}| < \epsilon\}, \tag{61}$$

where $q_t^{\star}$ and $r_t^{\star}$ are the optimal solutions of the deterministic formulation stated in (49). Note that the cost function of the problem (49) is strongly convex in the minimization variables. Based on the analysis in the previous sections, note that the feasibility sets of the problems defined in Theorem 2 are compact asymptotically. Moreover, the analysis in the previous sections shows that there exists a constant ϕ such that the optimal cost ϕ_p defined in Theorem 2 converges in probability to ϕ as p goes to $+\infty$. Additionally, the same analysis in the previous sections shows that there exists a constant ϕ^c such that the optimal cost ϕ_p^c defined in Theorem 2 converges in probability to ϕ^c as p goes to $+\infty$. The strong convexity property of the cost function of the optimization problem in (49) can then be used to show that there exists $\zeta > 0$ such that $\phi^c > \phi + \zeta$. This implies that the second assumption in Theorem 2 is satisfied for the considered set $\mathcal{S}_p$ and any fixed $\epsilon > 0$. This then shows that the convergence results in Lemma 6 are all satisfied.

Note that the CGMT framework applied to prove Lemma 6 also shows that the optimal cost value of the soft target formulation in (8) converges in probability to the optimal cost value of the deterministic formulation given in (49). Combining this with the result in Lemma 6 shows the convergence property of the training error stated in (17). Now, it remains to show the convergence of the generalization error. It suffices to show that the generalization error defined in (57) 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 (18), which completes the proof of Theorem 1. Note that the above analysis of the soft target formulation in (8) 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 (49). This leads to the scalar formulations introduced in (16) and (19).

7.4. Phase Transitions in Hard Formulation

In this part, we provide a rigorous proof of Proposition 1. Here, we consider the squared loss function. In this case, the deterministic source formulation given in (14) can be simplified as follows:

$$\min_{q_s, r_s \geq 0} \frac{1}{2} \max\Big\{ -r_s + \sqrt{\alpha_s}(q_s^2 + r_s^2 + v_s - 2q_s c_s)^{\frac{1}{2}}, 0 \Big\}^2 + \frac{\lambda}{2}(q_s^2 + r_s^2), \tag{62}$$

where the constants v_s and c_s are defined as $v_s = \mathbb{E}[Y_s^2]$ and $c_s = \mathbb{E}[S_s Y_s]$, $Y_s = \varphi(S_s)$, and S_s is a standard Gaussian random variable. Additionally, the target scalar formulation given in (16) can be simplified as follows:

$$\begin{aligned}\min_{q_t, r_t \geq 0} \sup_{\sigma_t > 0} & \frac{\lambda}{2}(q_t^2 + r_t^2) + \frac{\sigma_t \delta}{2}\Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big) \\ & + \frac{\alpha_t \sigma_t}{2(1-\delta) + 2\sigma_t}(r_t^2 + q_t^2 + v_t - 2q_t c_t) - \frac{\sigma_t r_t^2}{2} \\ & + \frac{\sigma_t \delta}{2(1-\delta)}(q_t - \rho q_s^\star)^2, \end{aligned} \tag{63}$$

where the constants v_t and c_t are defined as $v_t = \mathbb{E}[Y_t^2]$ and $c_t = \mathbb{E}[Y_t S_t]$, $Y_t = \varphi(S_t)$, and S_t is a standard Gaussian random variable. Under the conditions stated in Proposition 1, the source deterministic formulation given in (62) can be simplified as follows:

$$\min_{q_s, r_s \geq 0} \; -r_s + \sqrt{\alpha_s}(q_s^2 + r_s^2 + v_s - 2q_s c_s)^{\frac{1}{2}}. \tag{64}$$

Note that one can easily solve the variables q_s and r_s. Specifically, the optimal solutions of (64) can be expressed as follows:

$$q_s^\star = c_s, \text{ and } r_s^\star = \sqrt{v_s - c_s^2}/\sqrt{\alpha_s - 1}. \tag{65}$$

Moreover, the target deterministic formulation given in (63) can be expressed as follows:

$$\begin{aligned}\min_{q_t, r_t \geq 0} \sup_{\sigma_t > 0} & \frac{\sigma_t \delta}{2}\beta_2 + \frac{\alpha_t \sigma_t}{2(1-\delta) + 2\sigma_t}(r_t^2 + q_t^2 + v_t - 2q_t c_t) - \frac{\sigma_t r_t^2}{2} \\ & + \frac{\sigma_t \delta}{2(1-\delta)}(q_t - \beta_1)^2, \end{aligned} \tag{66}$$

where β_1 and β_2 are given by

$$\beta_1 = \rho q_s^\star, \; \beta_2 = \Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big). \tag{67}$$

Before solving the optimization problem in (66), we consider the following change in variable:

$$x_t^2 + r_t^2 - \delta\beta_2 - \frac{\delta}{1-\delta}(q_t - \beta_1)^2. \tag{68}$$

Note that the above change in variable is valid since the formulation in (66) requires the left-hand side of (68) to be positive. Therefore, the formulation in (66) can be expressed in terms of x_t instead of r_t as follows:

$$\min_{q_t, x_t \geq 0} \sup_{\sigma_t > 0} \frac{\alpha_t \sigma_t}{2(1-\delta) + 2\sigma_t}\left(x_t^2 + \delta\beta_2 + \frac{\delta}{1-\delta}(q_t - \beta_1)^2 + q_t^2 + v_t - 2q_t c_t\right) - \frac{\sigma_t x_t^2}{2}. \tag{69}$$

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

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

It is now clear that one can solve the problem in (69) in closed form. Moreover, it can be easily checked that the optimal solutions of the optimization problem (66) can be expressed as follows:

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

Then, the asymptotic limit of the generalization error corresponding to the hard formulation can be determined in closed-form. Given that 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.

Next, 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:

$$\mathcal{E}_{\text{test}} = \frac{\alpha_t}{\alpha_t+\delta-1}\Big(\delta\{(c-\beta_1)^2+\beta_2\} + (v-c^2)\Big). \tag{70}$$

Note that 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:

$$\mathcal{E}'_{\text{test}}(\delta) = \frac{(\alpha_t-1)\{(c-\beta_1)^2+\beta_2\} - (v-c^2)}{(\alpha_t+\delta-1)^2}. \tag{71}$$

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:

$$\delta^\star = \begin{cases} 1 & \text{if } Z_t < 0 \\ 0 & \text{if } Z_t > 0 \\ [0\ 1] & \text{otherwise,} \end{cases} \tag{72}$$

where Z_t is given by

$$Z_t = (\alpha_t-1)\{(c-\beta_1)^2+\beta_2\} - (v-c^2). \tag{73}$$

It can be easily shown that the condition in (72) can be expressed as the one given in (20). This completes the proof of Proposition 1.

7.5. Sufficient Condition for the Hard Formulation

In this part, we provide a rigorous proof to Proposition 2. Suppose that the assumptions in Proposition 2 are all satisfied. Additionally, 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:

$$\mathcal{E}_{\text{test}}(\delta) = \frac{1}{\pi}\text{acos}\Big(\frac{q_t(\delta)}{\sqrt{(q_t(\delta))^2 + (r_t(\delta))^2}}\Big), \tag{74}$$

where $q_t(\delta)$ and $r_t(\delta)$ are optimal solutions to the deterministic problem in (19) for fixed δ. A simple sufficient condition for positive transfer is when $\mathcal{E}_{\text{test}}(\delta)$ is decreasing at $\delta = 0$.

This means that there exists some $\delta > 0$ such that the transfer learning method introduced in (6) is better than the standard method when the following function increases at $\delta = 0$:

$$g(\delta) = \frac{q_t(\delta)}{\sqrt{(q_t(\delta))^2 + (r_t(\delta))^2}}. \tag{75}$$

After computing the derivative of the function $g(\cdot)$ at zero, one can see that the transfer learning method introduced in (6) is better than the standard method when the following condition is true:

$$q_t'(0) r_t(0) - q_t(0) r_t'(0) > 0, \tag{76}$$

where $q_t(0)$ and $r_t(0)$ denote the optimal solutions of the standard learning formulation (i.e., $\delta = 0$ in (19)). Additionally, $q_t'(0)$ and $r_t'(0)$ denote the derivative of the functions $q_t(\delta)$ and $r_t(\delta)$ at $\delta = 0$. The above analysis shows that it suffices to find the values of $q_t'(0)$ and $r_t'(0)$ to fully characterize the sufficient condition in (76). Before stating our analysis, we define β_1 and β_2 as follows:

$$\beta_1 = \Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big), \ \beta_2 = \rho q_s^\star, \tag{77}$$

where $q_s^\star$ and $r_s^\star$ are the optimal solutions of the deterministic source formulation given in (14).

Note that the optimal solution of the deterministic formulation in (19) satisfy the following system of equations:

$$\begin{cases} \alpha_t \mathbb{E}\Big(S\mathcal{M}_{\ell,1}'[r_t(\delta)H + q_t(\delta)S; \frac{1-\delta}{\sigma_t(\delta)}]\Big) + \frac{\delta\sigma_t(\delta)}{1-\delta}(q_t(\delta) - \beta_2) + \lambda q_t(\delta) = 0 \\ \alpha_t \mathbb{E}\Big(H\mathcal{M}_{\ell,1}'[r_t(\delta)H + q_t(\delta)S; \frac{1-\delta}{\sigma_t(\delta)}]\Big) - \sigma_t(\delta) r_t(\delta) + \lambda r_t(\delta) = 0 \\ \frac{\delta}{2}\beta_1 - \frac{\alpha_t(1-\delta)}{\sigma_t(\delta)^2}\mathbb{E}\Big(\mathcal{M}_{\ell,2}'[r_t(\delta)H + q_t(\delta)S; \frac{1-\delta}{\sigma_t(\delta)}]\Big) - \frac{r_t(\delta)^2}{2} + \frac{\delta}{2(1-\delta)}(q_t(\delta) - \beta_2)^2 = 0. \end{cases}$$

The derivative of the first equation at $\delta = 0$ can be expressed as follows:

$$\begin{aligned} &\alpha_t \mathbb{E}\Big((SHr'(0) + S^2 q'(0))\mathcal{M}_{\ell,1}''[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \sigma_t(0)(q_t(0) - \beta_2) \\ &- \frac{\alpha_t}{\sigma_t(0)^2}(\sigma_t(0) + \sigma_t'(0))\mathbb{E}\Big(S\mathcal{M}_{\ell,12}''[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \lambda q_t'(0) = 0, \end{aligned} \tag{78}$$

where $q_t(0)$, $r_t(0)$, and $\sigma_t(0)$ denote optimal solutions of the standard learning formulation (i.e., $\delta = 0$ in (19)). This means that they are known. Moreover, $q_t'(0)$, $r_t'(0)$ and $\sigma_t'(0)$ are unknown and denote the derivative of the functions $q_t(\delta)$, $r_t(\delta)$, and $\sigma_t(\delta)$ at $\delta = 0$. Now, define the constants I_{11}, I_{12}, I_{13}, and I_{14} as follows:

$$\begin{cases} I_{11} = \alpha_t \mathbb{E}\Big(SH\mathcal{M}_{\ell,1}''[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) \\ I_{12} = \alpha_t \mathbb{E}\Big(S^2\mathcal{M}_{\ell,1}''[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \lambda \\ I_{13} = -\frac{\alpha_t}{\sigma_t(0)^2}\mathbb{E}\Big(S\mathcal{M}_{\ell,12}''[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) \\ I_{14} = -\frac{\alpha_t}{\sigma_t(0)}\mathbb{E}\Big(S\mathcal{M}_{\ell,12}''[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \sigma_t(0)(q_t(0) - \beta_2). \end{cases} \tag{79}$$

This means that the equation in (78) can be expressed as follows:

$$I_{11} r_t'(0) + I_{12} q_t'(0) + I_{13}\sigma_t'(0) + I_{14} = 0. \tag{80}$$

Similarly, the derivative of the second equation at $\delta = 0$ can be expressed as follows:

$$\alpha_t \mathbb{E}\Big((H^2 r'(0) + HSq'(0))\mathcal{M}''_{\ell,1}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) - \sigma'_t(0)r_t(0) - \sigma_t(0)r'_t(0)$$
$$- \frac{\alpha_t}{\sigma_t(0)^2}(\sigma_t(0) + \sigma'_t(0))\mathbb{E}\Big(H\mathcal{M}''_{\ell,12}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \lambda r'_t(0) = 0. \tag{81}$$

Now, define the constants I_{21}, I_{22}, I_{23}, and I_{24} as follows:

$$\begin{cases} I_{21} = \alpha_t \mathbb{E}\Big(H^2\mathcal{M}''_{\ell,1}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) - \sigma_t(0) + \lambda \\ I_{22} = \alpha_t \mathbb{E}\Big(HS\mathcal{M}''_{\ell,1}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) \\ I_{23} = -\frac{\alpha_t}{\sigma_t(0)^2}\mathbb{E}\Big(H\mathcal{M}''_{\ell,12}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) - r_t(0) \\ I_{24} = -\frac{\alpha_t}{\sigma_t(0)}\mathbb{E}\Big(H\mathcal{M}''_{\ell,12}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big). \end{cases} \tag{82}$$

This means that the equation in (81) can be expressed as follows:

$$I_{21}r'_t(0) + I_{22}q'_t(0) + I_{23}\sigma'_t(0) + I_{24} = 0. \tag{83}$$

Moreover, the derivative of the third equation at $\delta = 0$ can be expressed as follows:

$$\frac{\beta_1}{2} + \frac{\alpha_t}{\sigma_t(0)^3}(\sigma_t(0) + 2\sigma'_t(0))\mathbb{E}\Big(\mathcal{M}'_{\ell,2}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) - r'_t(0)r_t(0)$$
$$- \frac{\alpha_t}{\sigma_t(0)^2}\mathbb{E}\Big((Hr'(0) + Sq'(0))\mathcal{M}''_{\ell,21}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \frac{1}{2}(q_t(0) - \beta_2)^2$$
$$+ \frac{\alpha_t}{\sigma_t(0)^4}(\sigma_t(0) + \sigma'_t(0))\mathbb{E}\Big(\mathcal{M}''_{\ell,2}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) = 0. \tag{84}$$

We define the constants I_{31}, I_{32}, I_{33}, and I_{34} as follows:

$$\begin{cases} I_{31} = -\frac{\alpha_t}{\sigma_t(0)^2}\mathbb{E}\Big(H\mathcal{M}''_{\ell,21}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) - r_t(0) \\ I_{32} = -\frac{\alpha_t}{\sigma_t(0)^2}\mathbb{E}\Big(S\mathcal{M}''_{\ell,12}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) \\ I_{33} = \frac{2\alpha_t}{\sigma_t(0)^3}\mathbb{E}\Big(\mathcal{M}'_{\ell,2}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \frac{\alpha_t}{\sigma_t(0)^4}\mathbb{E}\Big(\mathcal{M}''_{\ell,2}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) \\ I_{34} = \frac{\alpha_t}{\sigma_t(0)^2}\mathbb{E}\Big(\mathcal{M}'_{\ell,2}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) + \frac{\alpha_t}{\sigma_t(0)^3}\mathbb{E}\Big(\mathcal{M}''_{\ell,2}[r_t(0)H + q_t(0)S; \frac{1}{\sigma_t(0)}]\Big) \\ \quad + \frac{1}{2}(q_t(0) - \beta_2)^2 + \frac{\beta_1}{2}. \end{cases}$$

Therefore, the equation in (84) can be expressed as follows:

$$I_{31}r'_t(0) + I_{32}q'_t(0) + I_{33}\sigma'_t(0) + I_{34} = 0. \tag{85}$$

The above analysis shows that the values of $q'_t(0)$ and $r'_t(0)$ can be determined after solving the following system of linear equations:

$$\begin{cases} I_{11}r'_t(0) + I_{12}q'_t(0) + I_{13}\sigma'_t(0) + I_{14} = 0 \\ I_{21}r'_t(0) + I_{22}q'_t(0) + I_{23}\sigma'_t(0) + I_{24} = 0 \\ I_{31}r'_t(0) + I_{32}q'_t(0) + I_{33}\sigma'_t(0) + I_{34} = 0, \end{cases} \tag{86}$$

over the three unknowns $q'_t(0)$, $r'_t(0)$, and $\sigma'_t(0)$. This completes the proof of Proposition 2.

8. Conclusions

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.

Author Contributions: Conceptualization, O.D. and Y.M.L.; methodology, O.D. and Y.M.L.; software, O.D.; validation, O.D. and Y.M.L.; formal analysis, O.D. and Y.M.L.; investigation, O.D. and Y.M.L.; resources, O.D. and Y.M.L.; data curation, O.D.; writing—original draft preparation, O.D.; writing—review and editing, O.D. and Y.M.L.; visualization, O.D.; supervision, Y.M.L.; project administration, Y.M.L.; funding acquisition, Y.M.L. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Technical Assumptions

Note that Assumption 1 is essential to show that the soft formulation in (4) concentrates in the large system limit. It also guarantees that the vectors $\boldsymbol{\xi}_t \in \mathbb{R}^p$ and $\boldsymbol{\xi}_s \in \mathbb{R}^p$ have correlations equal to ρ, asymptotically. This is aligned with the definition in (3). Assumption 4 is also introduced to guarantee that the generalization error concentrates in the large system limit. It is satisfied by popular regression and classification models. For instance, observe that the conditions in Assumption 4 are all satisfied by the regression model considering $\varphi : x \to \max(x, 0)$. Moreover, they are satisfied by the binary classification model considering $\varphi : x \to \text{sign}(x)$.

The analysis presented in this paper mostly focuses on regularized transfer learning formulations (i.e., $\lambda > 0$). The convexity properties in Assumption 3 are essential to apply the CGMT framework. Moreover, the properties in (12) are used to guarantee the compactness assumptions in the CGMT framework (see Theorem 2). In this appendix, we check the validity of Assumption 3 using popular loss functions, i.e., squared loss for regression tasks and logistic and hinge losses for binary classification tasks. To this end, assume that C is an arbitrary fixed positive constant.

1. Squared loss: It is easy to see that the squared loss is a proper strongly convex function in $\mathbb{R}$, where 1 is a strong convexity parameter. Moreover, $\mathcal{L}(\cdot)$ and its sub-differential set $\partial\mathcal{L}(\cdot)$ can be expressed as follows:

$$\mathcal{L}(\boldsymbol{v}) = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{y}\|^2, \;\; \partial\mathcal{L}(\boldsymbol{v}) = \{\boldsymbol{v} - \boldsymbol{y}\}, \tag{A1}$$

where the vector $\boldsymbol{y}$ is formed by the concatenation of $\{y_i\}_{i=1}^{n_t}$. Then, there exists $R > 0$ such that

$$\sup_{\|\boldsymbol{v}\| \le C\sqrt{n_t}} |\mathcal{L}(\boldsymbol{v})| \le Rn_t, \quad \sup_{\|\boldsymbol{v}\| \le C\sqrt{n_t}} \sup_{\boldsymbol{s} \in \partial\mathcal{L}(\boldsymbol{v})} \|\boldsymbol{s}\| = \sup_{\|\boldsymbol{v}\| \le C\sqrt{n_t}} \|\boldsymbol{v} - \boldsymbol{y}\| \le R\sqrt{n_t}, \tag{A2}$$

with probability going to 1 as p grow to $+\infty$. The inequality follows using the regularity condition in Assumption 4 and the weak law of large numbers. Then, the squared loss satisfies Assumption 3 for any $\lambda \ge 0$.

2. Logistic loss: Now, we consider the logistic loss applied to a binary classification model (i.e., $y_i \in \{-1, 1\}$). Note that the logistic loss is a proper convex function in $\mathbb{R}$. Moreover, $\mathcal{L}(\cdot)$ and its sub-differential set $\partial\mathcal{L}(\cdot)$ are given by

$$\mathcal{L}(\boldsymbol{v}) = \sum_{i=1}^{n_t} \log(1 + e^{-y_i v_i}), \;\; \partial\mathcal{L}(\boldsymbol{v}) = \{\boldsymbol{x}\}, \text{ where } x_i = \frac{-y_i e^{-y_i v_i}}{1 + e^{-y_i v_i}}, \; \forall i \in \{1 \dots n_t\}. \tag{A3}$$

First, observe that the loss $\mathcal{L}(\cdot)$ satisfies the following inequality:

$$|\mathcal{L}(\boldsymbol{v})| \leq n_t + \|\boldsymbol{v}\|_1. \tag{A4}$$

This means that there exists $R_1 > 0$ such that the following inequality is valid:

$$\sup_{\|\boldsymbol{v}\| \leq C\sqrt{n_t}} |\mathcal{L}(\boldsymbol{v})| \leq R_1 n_t. \tag{A5}$$

Additionally, the following results hold true:

$$\sup_{\|\boldsymbol{v}\| \leq C\sqrt{n_t}} \; \sup_{\boldsymbol{s} \in \partial\mathcal{L}(\boldsymbol{v})} \|\boldsymbol{s}\| = \sup_{\|\boldsymbol{v}\| \leq C\sqrt{n_t}} \|\boldsymbol{x}\| \leq \Big(\sum_{i=1}^{n_t} y_i^2\Big)^{\frac{1}{2}}. \tag{A6}$$

This means that there exists $R_2 > 0$ such that the following inequality is valid:

$$\sup_{\|\boldsymbol{v}\| \leq C\sqrt{n_t}} \; \sup_{\boldsymbol{s} \in \partial\mathcal{L}(\boldsymbol{v})} \|\boldsymbol{s}\| \leq R_2\sqrt{n_t}. \tag{A7}$$

Then, there exists a universal constant $R > 0$ such that Assumption 3 is satisfied for the logistic loss for any $\lambda > 0$.

3. Hinge loss: Finally, we consider the hinge loss applied to a binary classification model (i.e., $y_i \in \{-1, 1\}$). It is clear that the hinge loss is a proper convex function in $\mathbb{R}$. Moreover, $\mathcal{L}(\cdot)$ is given by $\mathcal{L}(\cdot) = \sum_{i=1}^{n_t} \max(1 - y_i v_i, 0)$. Following [33], the sub-differential set $\partial\mathcal{L}(\cdot)$ can be expressed as follows:

$$\partial\mathcal{L}(\boldsymbol{v}) = \Big\{\frac{1}{2}\boldsymbol{D}(\boldsymbol{1} + \boldsymbol{g}) : \|\boldsymbol{g}\|_\infty \leq 1, \boldsymbol{g}^\top(\boldsymbol{D}\boldsymbol{v} + \boldsymbol{1}) = \|\boldsymbol{D}\boldsymbol{v} + \boldsymbol{1}\|_1\Big\}, \tag{A8}$$

where $\boldsymbol{D} \in \mathbb{R}^{n_t \times n_t}$ is a diagonal matrix with diagonal entries $\{y_i\}_{i=1}^{n_t}$. Note that the loss function $\mathcal{L}(\cdot)$ satisfies the following inequality:

$$|\mathcal{L}(\boldsymbol{v})| \leq n_t + \|\boldsymbol{v}\|_1. \tag{A9}$$

This means that there exists $R_1 > 0$ such that the following inequality is valid:

$$\sup_{\|\boldsymbol{v}\| \leq C\sqrt{n_t}} |\mathcal{L}(\boldsymbol{v})| \leq R_1 n_t. \tag{A10}$$

Moreover, the result in (A8) shows that any element $\boldsymbol{s}$ in the sub-differential set $\partial\mathcal{L}(\boldsymbol{v})$ satisfies the following :

$$\|\boldsymbol{s}\| \leq \frac{1}{2}\sqrt{n_t} + \frac{1}{2}\|\boldsymbol{g}\| \leq \frac{1}{2}\sqrt{n_t} + \frac{1}{2}\sqrt{n_t}\|\boldsymbol{g}\|_\infty \leq \sqrt{n_t}. \tag{A11}$$

This means that there exists $R_2 > 0$ such that the following inequality is valid:

$$\sup_{\|\boldsymbol{v}\| \leq C\sqrt{n_t}} \; \sup_{\boldsymbol{s} \in \partial\mathcal{L}(\boldsymbol{v})} \|\boldsymbol{s}\| \leq R_2\sqrt{n_t}. \tag{A12}$$

Then, there exists a universal constant $R > 0$ such that Assumption 3 is satisfied for the hinge loss for any $\lambda > 0$.

Appendix B. Proof of Lemma 1

Appendix B.1. Primal Compactness

We start our analysis by assuming that $\lambda > 0$. We first consider the compactness of the source problem given in (4). Note that the formulation in (4) has a unique optimal solution. Assume that $\widehat{\boldsymbol{w}}_{s,p} \in \mathbb{R}^p$ is the unique optimal solution of the optimization problem given in (4). The analysis in [20] (Lemma 1) can be used to prove that there exists $C_1 > 0$ such that the following inequality is valid:

$$\|\widehat{\boldsymbol{w}}_{s,p}\|^2 \leq C_1, \tag{A13}$$

with probability going to one as $p \to \infty$. Moreover, observe that the formulation in (33) has a unique optimal solution. Assume that $\widehat{\boldsymbol{w}}_{t,p} \in \mathbb{R}^p$ is the unique optimal solution of the optimization problem given in (33). Assumption 3 supposes that the loss function is proper. Then, we can conclude that there exists $C_2 > 0$ such that

$$\ell(y, z) \geq -C_2, \ \forall z \in \mathbb{R}. \tag{A14}$$

Now, we define $O^{\star}_{t,p}$ as the optimal objective value of the formulation in (33). Then, we can see that there exists $C_3 > 0$ such that

$$\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}_{t,p}\|^2 \leq O^{\star}_{t,p} + C_3. \tag{A15}$$

Given that $\widehat{\boldsymbol{w}}_{s,p}$ is a feasible solution in the formulation given in (33), we obtain the following inequality:

$$\frac{\lambda}{2}\|\widehat{\boldsymbol{w}}_{t,p}\|^2 \leq \frac{1}{p}\sum_{i=1}^{p} \ell(y_i; \boldsymbol{a}_i^{\top}\widehat{\boldsymbol{w}}_{s,p}) + \frac{\lambda}{2}\|\widehat{\boldsymbol{w}}_{s,p}\|^2 + C_3. \tag{A16}$$

Based on [27] (Theorem 2.1), the following convergence in probability holds:

$$\frac{\|\boldsymbol{A}\|}{\sqrt{n_t}} \xrightarrow{p \to +\infty} \frac{\sqrt{\alpha_t} + 1}{\sqrt{\alpha_t}}, \tag{A17}$$

where the matrix $\boldsymbol{A} \in \mathbb{R}^{n_t \times p}$ is formed by the concatenation of vectors $\{\boldsymbol{a}_i\}_{i=1}^{n_t}$. Then, there exists $C_4 > 0$ such that the following inequality is valid:

$$\|\boldsymbol{A}\widehat{\boldsymbol{w}}_{s,p}\| \leq \|\boldsymbol{A}\|\|\widehat{\boldsymbol{w}}_{s,p}\| \leq C_4\sqrt{n_t}, \tag{A18}$$

with probability going to one as $p \to \infty$. Combining this with the assumption in (12), we see that there exists $C_5 > 0$ such that the following inequality is valid:

$$\frac{1}{n_t}|\sum_{i=1}^{p} \ell(y_i; \boldsymbol{a}_i^{\top}\widehat{\boldsymbol{w}}_{s,p})| \leq C_5. \tag{A19}$$

Given that $\lambda > 0$ and the result in (A13), we conclude that there exists $C_6 > 0$ such that the following holds:

$$\|\widehat{\boldsymbol{w}}_{t,p}\|^2 \leq C_6, \tag{A20}$$

with probability going to one as $p \to \infty$.

Now, we consider the case when $\lambda = 0$. Define $g_{s,p}(\cdot)$, $O^{\star}_{s,p}$, and $\widehat{\boldsymbol{w}}_{s,p}$ as the cost function, the optimal cost value, and the optimal solution of the formulation in (4). Moreover, define $g_{t,p}(\cdot)$, $O^{\star}_{t,p}$, and $\widehat{\boldsymbol{w}}_{t,p}$ as the cost function, the optimal cost value, and the optimal solution of the formulation in (33). Note that the loss function $\ell(y, \cdot)$ is strongly convex

with a strong convexity parameter $S > 0$. Then, for any $x_1, x_2 \in \mathbb{R}$, the following property is valid:

$$\ell\Big(y, \frac{x_1 + x_2}{2}\Big) \leq \frac{1}{2}\ell(y, x_1) + \frac{1}{2}\ell(y, x_2) - \frac{S}{8}|x_1 - x_2|^2. \tag{A21}$$

This means that, for any $i \in \{1, \ldots, n\}$, the following property is valid:

$$\ell\Big(y_i, \frac{\boldsymbol{a}_i^\top \boldsymbol{w}_1 + \boldsymbol{a}_i^\top \boldsymbol{w}_2}{2}\Big) \leq \frac{1}{2}\ell(y_i, \boldsymbol{a}_i^\top \boldsymbol{w}_1) + \frac{1}{2}\ell(y_i, \boldsymbol{a}_i^\top \boldsymbol{w}_2) - \frac{S}{8}|\boldsymbol{a}_i^\top \boldsymbol{w}_1 - \boldsymbol{a}_i^\top \boldsymbol{w}_2|^2, \tag{A22}$$

where n can be the number of samples of the source task or target task and $\{y_i\}_{i=1}^n$ are the labels of the source task or target task. Given the convexity of the norm, we obtain the following inequality:

$$g_p\Big(\frac{\boldsymbol{w}_1 + \boldsymbol{w}_2}{2}\Big) \leq \frac{1}{2}g_p(\boldsymbol{w}_1) + \frac{1}{2}g_p(\boldsymbol{w}_2) - \frac{S}{8p}\|\boldsymbol{A}(\boldsymbol{w}_1 - \boldsymbol{w}_2)\|^2, \tag{A23}$$

where $g_p(\cdot)$ can be the cost value of the source task or target task formulations. Now, we focus on the source formulation. Take $\boldsymbol{w}_1 = \widehat{\boldsymbol{w}}_{s,p}$ and $\boldsymbol{w}_2 = 0$. Moreover, see that the loss function is proper. Then, there exists $C_7 > 0$ such that

$$\frac{S}{8p}\|\boldsymbol{A}\widehat{\boldsymbol{w}}_{s,p}\|^2 \leq C_7 + \frac{1}{2}g_{s,p}(0). \tag{A24}$$

Given the assumption in (12), $S > 0$, $\alpha_s > 1$, and the analysis in [27] (Theorem 2.1), there exists $C_8 > 0$ such that

$$\|\widehat{\boldsymbol{w}}_{s,p}\|^2 \leq C_8. \tag{A25}$$

Now, we focus on the target task. Take $\boldsymbol{w}_1 = \widehat{\boldsymbol{w}}_{t,p}$ and $\boldsymbol{w}_2 = \widehat{\boldsymbol{w}}_{s,p}$. Moreover, see that the loss function is proper. Then, there exists $C_9 > 0$ such that

$$\frac{S}{8p}\|\boldsymbol{A}(\widehat{\boldsymbol{w}}_{t,p} - \widehat{\boldsymbol{w}}_{s,p})\|^2 \leq C_9 + \frac{1}{2}g_{t,p}(\widehat{\boldsymbol{w}}_{s,p}). \tag{A26}$$

Given the assumption in (12), the result in (A25), $S > 0$, $\alpha_t > 1$, and the analysis in [27] (Theorem 2.1), there exists $C_{10} > 0$ such that

$$\|\widehat{\boldsymbol{w}}_{t,p}\|^2 \leq C_{10}. \tag{A27}$$

This completes the first part of the proof of Lemma 1.

Appendix B.2. Dual Compactness

The analysis in Appendix B.1 shows that the formulation in (34) can be equivalently formulated, where the primal feasibility set is given by

$$\|\boldsymbol{w}\|^2 \leq C, \tag{A28}$$

where $C > 0$ is a sufficiently large constant that satisfies the analysis in Appendix B.1. Now, define $\widehat{\boldsymbol{u}}_p$ as the optimal solution of the formulation in (34). Additionally, define the function $\mathcal{L}^\star(\cdot)$ as $\mathcal{L}^\star(\boldsymbol{u}) = \sum_{i=1}^{n_t} \ell^\star(y_i; u_i)$. We can see that the optimal vector $\widehat{\boldsymbol{u}}_p$ solves the following maximization problem:

$$\widehat{\boldsymbol{u}}_p = \underset{\boldsymbol{u} \in \mathbb{R}^{n_t}}{\operatorname{argmax}} \ \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{w} - \mathcal{L}^\star(\boldsymbol{u}),$$

where the data matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n_t}]^\top \in \mathbb{R}^{n_t \times p}$. Now, we denote by $\partial \mathcal{L}^\star(\boldsymbol{u})$ the sub-differential set of the function $\mathcal{L}^\star(\cdot)$ evaluated at $\boldsymbol{u}$. Therefore, the solution of the above maximization problem satisfies the following condition:

$$\boldsymbol{A}\boldsymbol{w} \in \partial \mathcal{L}^\star(\widehat{\boldsymbol{u}}_p). \tag{A29}$$

Now, we use the result in [25] (Proposition 11.3) to show that the condition in (A29) can be equivalently expressed as follows:

$$\widehat{\boldsymbol{u}}_p \in \partial \mathcal{L}(\boldsymbol{A}\boldsymbol{w}), \tag{A30}$$

where the loss function $\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n_t} \ell(y_i; w_i)$ based on [25] (Proposition 11.22). Note that the introduced constraint in (A28) is satisfied. Moreover, the analysis presented in (A17) shows that there exists $C_1 > 0$ such that the following inequality holds:

$$\|\boldsymbol{A}\boldsymbol{w}\|^2 \leq C_1 n_t, \tag{A31}$$

with probability going to one as p goes to $+\infty$. Now, we use the assumption in (12) to conclude that there exists $C_2 > 0$ such that the following inequality holds:

$$\|\widehat{\boldsymbol{u}}_p\|^2 \leq C_2 n_t, \tag{A32}$$

with probability going to one as p goes to $+\infty$. This completes the proof of Lemma 1.

Appendix C. 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.

Appendix C.1. Auxiliary Convergence

Note that the analysis present in Section 7 is also valid for the source problem. This is because the formulation in (8) is equivalent to the source problem in (4) 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:

$$\widetilde{\boldsymbol{w}}_s = q^\star_{p,s}\boldsymbol{\xi}_s - \frac{r^\star_{p,s}}{\|\widetilde{\boldsymbol{g}}_s\|}\boldsymbol{B}^\perp_{\boldsymbol{\xi}_s}\widetilde{\boldsymbol{g}}_s, \tag{A33}$$

where $\widetilde{\boldsymbol{g}}_s = (\boldsymbol{B}^\perp_{\boldsymbol{\xi}_s})^\top \boldsymbol{g}_s$ and $\boldsymbol{g}_s$ has independent standard Gaussian components. Here, $\boldsymbol{B}^\perp_{\boldsymbol{\xi}_s} \in \mathbb{R}^{p\times(p-1)}$ is formed by an orthonormal basis orthogonal to the vector $\boldsymbol{\xi}_s$. Additionally, our analysis in Section 7 shows that the following convergence in probability holds:

$$q_{p,s} \xrightarrow{p\to+\infty} q^\star_s \text{ and } r^\star_{p,s} \xrightarrow{p\to+\infty} r^\star_s. \tag{A34}$$

Here, $q^\star_s$ and $r^\star_s$ are the optimal solutions of the asymptotic limit of the source formulation defined in (14).

Note that μ_p can be expressed as follows:

$$\mu_p = \sigma_{\min}((\boldsymbol{B}^\perp_{\boldsymbol{\xi}_t})^\top \boldsymbol{\Lambda} \boldsymbol{B}^\perp_{\boldsymbol{\xi}_t}). \tag{A35}$$

Using the eigenvalue interlacing theorem, one can see that

$$\sigma_{\min,1}(\boldsymbol{\Lambda}) \leq \mu_p \leq \sigma_{\min,2}(\boldsymbol{\Lambda}). \tag{A36}$$

Then, using the assumption in (13), we can see that the random variable μ_p converges in probability to $\mu_{\min}$, where $\mu_{\min}$ is defined in Assumption 5. Now, we study the properties

of the remaining functions using the optimal solution of the auxiliary formulation defined in (A33), 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 $V_{p,ts}$.

First, fix $\sigma > -\mu_{\min}$. Then, based on the convergence of μ_p and [34] (Proposition 3), the sequence of random functions $T_{p,g}(.)$ converges in probability as follows:

$$T_{p,g}(\sigma) \overset{p\to+\infty}{\to} T_g(\sigma) = \mathbb{E}_\mu[1/(\mu+\sigma)]. \tag{A37}$$

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

$$T_{p,g}(\sigma' - x) \overset{p\to+\infty}{\to} T_g(\sigma' - x), \tag{A38}$$

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 (A38) 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:

$$T_{p,g}(\sigma' - \mu_p) \overset{p\to+\infty}{\to} T_g(\sigma' - \mu_{\min}), \tag{A39}$$

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

$$\begin{aligned} T_{p,t}(\sigma) &= \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp [(\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp + \sigma \boldsymbol{I}_{p-1}]^{-1} (\boldsymbol{B}_{\boldsymbol{\xi}_t}^\perp)^\top \boldsymbol{\Lambda} \boldsymbol{\xi}_t \\ &= \boldsymbol{\xi}_t^\top \boldsymbol{\Lambda} \boldsymbol{\xi}_t + \sigma - \frac{1}{\boldsymbol{\xi}_t^\top [\boldsymbol{\Lambda} + \sigma \boldsymbol{I}_p]^{-1} \boldsymbol{\xi}_t}. \end{aligned} \tag{A40}$$

Therefore, we obtain the following expression:

$$Z_{p,t}(\sigma) = \sigma - \frac{1}{\boldsymbol{\xi}_t^\top [\boldsymbol{\Lambda} + \sigma \boldsymbol{I}_p]^{-1} \boldsymbol{\xi}_t}. \tag{A41}$$

Then, using the theoretical results stated in [34] (Proposition 3), the functions $Z_{p,t}(.)$ converges in probability as follows:

$$Z_{p,t}(\sigma) \overset{p\to+\infty}{\to} Z_t(\sigma) = \sigma - \frac{1}{\mathbb{E}_\mu[1/(\mu+\sigma)]}. \tag{A42}$$

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

$$Z_{p,t}(\sigma' - \mu_p) \overset{p\to+\infty}{\to} Z_t(\sigma' - \mu_{\min}), \tag{A43}$$

valid for any $\sigma' > 0$. Based on the result in (A33), the sequence of random functions $\widetilde{Z}_{p,ts}(.)$ converges in probability to the following function:

$$Z_{ts}(\sigma) = q_s^\star \rho Z_t(\sigma). \tag{A44}$$

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

$$\widetilde{Z}_{p,ts}(\sigma' - \mu_p) \overset{p\to+\infty}{\to} Z_{ts}(\sigma' - \mu_{\min}), \tag{A45}$$

valid for any $\sigma' > 0$. Using the same analysis and based on (A33) and (A34), one can see that the sequence of random functions $\widetilde{Z}_{p,s}(.)$ converges in probability to the following function:

$$\widetilde{Z}_{p,s}(\sigma) \overset{p\to+\infty}{\to} Z_s(\sigma) = (\rho q_s^\star)^2 Z_t(\sigma) - \Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big)\mathbb{E}_\mu[\mu\sigma/(\mu+\sigma)]. \tag{A46}$$

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

$$\widetilde{Z}_{p,s}(\sigma' - \mu_p) \overset{p\to+\infty}{\to} Z_s(\sigma' - \mu_{\min}), \tag{A47}$$

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.

Appendix C.2. Primary Convergence

Here, we assume that $\lambda > 0$. The case when $\lambda = 0$ can be conducted similarly. 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:

$$\mathcal{T}_\epsilon = \{\boldsymbol{w} \in \mathbb{R}^p : |\boldsymbol{\xi}_t^\top [\boldsymbol{\Lambda} + \sigma \boldsymbol{I}_p]^{-1} \boldsymbol{w} - \rho q_s^\star K| < \epsilon\},$$

where K is defined as follows:

$$K = \mathbb{E}_\mu[1/(\mu+\sigma)]. \tag{A48}$$

Now, we consider the feasibility set $\mathcal{D}_\epsilon = \mathcal{T}_1/\mathcal{S}_\epsilon$, where $\mathcal{T}_1$ is defined in (41). Based on the analysis of the generalized target formulation in Section 7.3.2, one can see that the AO formulation corresponding to the source formulation with the set $\mathcal{D}_\epsilon$ can be asymptotically expressed as follows:

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

Here, the feasibility set $\mathcal{T}_2$ is defined in Section 7.3.2 and the feasibility set $\widetilde{\mathcal{D}}_\epsilon$ is given by

$$\Big\{\boldsymbol{r}_s : |q_s \rho K_{p,t} + q_s\sqrt{1-\rho^2}K_{p,r} + \boldsymbol{\xi}_t^\top [\boldsymbol{\Lambda} + \sigma \boldsymbol{I}_p]^{-1} \boldsymbol{B}_{\boldsymbol{\xi}_s}^\perp \boldsymbol{r}_s - \rho q_s^\star K| \geq \epsilon , \|\boldsymbol{r}_s\| = r_s\Big\}.$$

This follows based on the decomposition in (40) and where $K_{p,t} = \boldsymbol{\xi}_t^\top [\boldsymbol{\Lambda} + \sigma \boldsymbol{I}_p]^{-1} \boldsymbol{\xi}_t$ and $K_{p,r} = \boldsymbol{\xi}_t^\top [\boldsymbol{\Lambda} + \sigma \boldsymbol{I}_p]^{-1} \boldsymbol{\xi}_r$. Note that the optimization problem given in $\mathfrak{V}_p$ can be equivalently formulated as follows:

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

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

$$\Big\{|q_s\rho K_{p,t}+q_s\sqrt{1-\rho^2}K_{p,r}-r_s\boldsymbol{\xi}_t^\top[\boldsymbol{\Lambda}+\sigma\boldsymbol{I}_p]^{-1}\boldsymbol{B}_{\widetilde{\boldsymbol{\xi}}_s}^{\perp}\frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|}-\rho q_s^\star K| \geq \epsilon\Big\}\cap\mathcal{T}_2,$$

where $\widetilde{\boldsymbol{g}}_s=(\boldsymbol{B}_{\widetilde{\boldsymbol{\xi}}_s}^{\perp})^\top\boldsymbol{g}_s$. 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:

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

in the set $\widehat{\mathcal{S}}_\epsilon$. Based on the max–min inequality [35], the function $\widehat{h}_p(.)$ can be lower bounded by the following function:

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

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

$$\min_{\boldsymbol{r}_s\in\widetilde{\mathcal{D}}_\epsilon}\frac{\|\boldsymbol{u}\|}{p}\boldsymbol{g}_s^\top\boldsymbol{B}_{\widetilde{\boldsymbol{\xi}}_s}^{\perp}\boldsymbol{r}_s\geq-\frac{\|\boldsymbol{u}\|}{p}\|(\boldsymbol{B}_{\widetilde{\boldsymbol{\xi}}_s}^{\perp})^\top\boldsymbol{g}_s\|r_s, \tag{A49}$$

for any $(q_s,r_s)\in\widehat{\mathcal{S}}_\epsilon$. Following the generalized analysis in Section 7.3.2, one can see that the auxiliary problem corresponding to the source formulation can be expressed as follows:

$$\min_{(q_s,r_s)\in\mathcal{T}_2}\sup_{\sigma_s>0}\frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{M}_{\ell(y_{s,i},.)}\Big(r_sh_{s,i}+q_ss_{s,i};\frac{r_s\|\widetilde{\boldsymbol{g}}_s\|}{\sqrt{n_s}\sigma_s}\Big)-\frac{r_s\sigma_s}{2}\frac{\|\widetilde{\boldsymbol{g}}_s\|}{\sqrt{n_s}}+\frac{\lambda}{2}(q_s^2+r_s^2). \tag{A50}$$

This means that the function $\widetilde{h}_p(.)$ can be lower bounded by the cost function of the minimization problem formulated in (A50) denoted by $\widehat{g}_p(.)$, i.e.,

$$\widehat{g}_p(q_s,r_s)\leq\widetilde{h}_p(q_s,r_s). \tag{A51}$$

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 7.3.1. 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:

$$\widehat{g}_p(\beta\boldsymbol{v}_1+(1-\beta)\boldsymbol{v}_2)\leq\beta\widehat{g}_p(\boldsymbol{v}_1)+(1-\beta)\widehat{g}_p(\boldsymbol{v}_2)-\frac{\lambda}{2}\beta(1-\beta)\|\boldsymbol{v}_1-\boldsymbol{v}_2\|^2, \tag{A52}$$

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 (A50). Then, the inequality in (A52) implies the following inequality:

$$\phi_p^\star \leq \widehat{g}_p(\boldsymbol{v}_2) - \frac{\lambda}{2}\beta\|\boldsymbol{v}_p^\star - \boldsymbol{v}_2\|^2. \tag{A53}$$

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:

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

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

$$\min_{\boldsymbol{v}\in\widehat{\mathcal{S}}_\epsilon}\widehat{g}_p(\boldsymbol{v}) \leq \widehat{\phi}_p^\star. \tag{A54}$$

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:

$$\frac{\lambda}{4}\min_{\boldsymbol{v}\in\widehat{\mathcal{S}}_\epsilon}\|\boldsymbol{v}_p^\star - \boldsymbol{v}\|^2 \geq \epsilon', \tag{A55}$$

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

$$\epsilon \leq |q_s\rho K_{p,t} + q_s\sqrt{1-\rho^2}K_{p,r} - r_s\boldsymbol{\xi}_t^\top[\boldsymbol{\Lambda} + \sigma\boldsymbol{I}_p]^{-1}\boldsymbol{B}_{\boldsymbol{\xi}}^\perp\frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|} - \rho q_s^\star K| \leq$$
$$|q_s\rho K_{p,t} - \rho q_s^\star K| + |q_s\sqrt{1-\rho^2}||K_{p,r}| + |r_s||\boldsymbol{\xi}_t^\top[\boldsymbol{\Lambda} + \sigma\boldsymbol{I}_p]^{-1}\boldsymbol{B}_{\boldsymbol{\xi}}^\perp\frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|}|.$$

Based on the analysis in Appendix C.1, we have the following convergence in probability:

$$|q_s\rho K_{p,t} - \rho q_s^\star K| \overset{p\to+\infty}{\to} |q_s - q_s^\star|\rho K$$
$$|q_s|\sqrt{1-\rho^2}|K_{p,r}| \overset{p\to+\infty}{\to} 0,\ |r_s||\boldsymbol{\xi}_t^\top[\boldsymbol{\Lambda} + \sigma\boldsymbol{I}_p]^{-1}\boldsymbol{B}_{\boldsymbol{\xi}}^\perp\frac{\widetilde{\boldsymbol{g}}_s}{\|\widetilde{\boldsymbol{g}}_s\|}| \overset{p\to+\infty}{\to} 0. \tag{A56}$$

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

$$|q_s - q_s^\star|\rho K \geq \epsilon'', \tag{A57}$$

with probability going to 1 as $p \to \infty$. Combining this with Assumption 5 and the consistency result stated in (A34) shows that there exists $\epsilon' > 0$ such that the following inequality holds:

$$\frac{\lambda}{4}\min_{\boldsymbol{v}\in\widehat{\mathcal{D}}_\epsilon}\|\boldsymbol{v}_p^\star - \boldsymbol{v}\|^2 \geq \epsilon', \tag{A58}$$

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

$$\widehat{\phi}_p^\star \geq \phi_p^\star + \epsilon', \tag{A59}$$

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:

$$\boldsymbol{\xi}_t^\top [\boldsymbol{\Lambda} + \sigma \boldsymbol{I}_p]^{-1} \widehat{\boldsymbol{w}}_s \overset{p \to +\infty}{\to} q_s^\star \rho K, \tag{A60}$$

where $\widehat{\boldsymbol{w}}_s$ is the optimal solution of the source problem (4). Following the same analysis, one can also show the convergence properties stated in Lemma 2.

Appendix D. Proof of Lemma 4

Here, we assume that $\lambda > 0$. The case when $\lambda = 0$ can be conducted similarly. The cost function of the optimization problem (49) can be expressed as follows:

$$\begin{aligned}\mathcal{O}_t(q_t, r_t, \sigma_t) = \frac{\lambda}{2}(q_t^2 + r_t^2) - \frac{\sigma_t r_t^2}{2} - \frac{1}{2} Z_s(\sigma_t) - \frac{1}{2} q_t^2 Z_t(\sigma_t) \\ + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_{t,\cdot})}\Big(r_t H_t + q_t S_t; T_g(\sigma_t)\Big)\Big] + q_t Z_{ts}(\sigma_t).\end{aligned} \tag{A61}$$

Note that the function $\mathcal{O}_t(\cdot, \cdot, \cdot)$ can be expressed as follows:

$$\begin{aligned}\mathcal{O}_t(q_t, r_t, \sigma_t) = -\frac{\sigma_t r_t^2}{2} + \frac{1}{2}\Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big) T_2(\sigma_t) \\ + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_{t,\cdot})}\Big(r_t H_t + q_t S_t; T_1(\sigma_t)\Big)\Big] + \frac{\lambda}{2}(q_t^2 + r_t^2) \\ - \frac{1}{2}(q_t - \rho q_s^\star)^2 (\sigma_t - 1/T_1(\sigma_t)).\end{aligned} \tag{A62}$$

Here, the functions $T_1(.)$ and $T_2(.)$ are defined as follows:

$$T_1(\sigma_t) = \mathbb{E}_\mu[1/(\mu + \sigma_t)], \; T_2(\sigma_t) = \mathbb{E}_\mu[\mu \sigma_t / (\mu + \sigma_t)].$$

Based on Assumption 5, the functions $T_1(\cdot)$ and $T_2(\cdot)$ are twice continuously differentiable in the feasibility set. We start our analysis by showing that the function $\mathcal{O}_t(\cdot, \cdot, \cdot)$ is concave in the variable σ_t for fixed feasible (q_t, r_t). First, note that the function $T_2(\cdot)$ is concave in the feasibility set. Now, define the function $g(\cdot)$ as follows:

$$g(\sigma_t) = \frac{1}{T_1(\sigma_t)}. \tag{A63}$$

Then, we can see that the second derivative of the function $g(\cdot)$ can be expressed as follows:

$$g''(\sigma_t) = -\frac{T_1''(\sigma_t) T_1(\sigma_t) - 2 T_1'(\sigma_t)^2}{T_1(\sigma_t)^3}. \tag{A64}$$

Here, the first and second derivatives of the function $T_1(\cdot)$ can be expressed as follows:

$$T_1'(\sigma_t) = -\mathbb{E}_\mu[1/(\mu + \sigma_t)^2], \; T_1''(\sigma_t) = 2\mathbb{E}_\mu[1/(\mu + \sigma_t)^3]. \tag{A65}$$

Then, using the Cauchy–Schwarz inequality, one can see that the second derivative of the function $g(\cdot)$ is negative. This implies the concavity of the function $g(\cdot)$. Therefore, using the properties in [35] (Section 3.2), the function $\mathcal{O}_t(\cdot, \cdot, \cdot)$ is concave in the variable σ_t.

Now, we focus on proving the strong convexity properties. Define the function $\mathcal{O}_t(\cdot, \cdot)$ as follows:

$$\mathcal{O}_t(q_t, r_t) = \sup_{\sigma_t > -\mu_{\min}} \mathcal{O}_t(q_t, r_t, \sigma_t). \tag{A66}$$

Note that the term $\frac{\lambda}{2}(q_t^2 + r_t^2)$ is strongly convex in the variables (q_t, r_t). Then, to prove our property it suffices to show that the following function is jointly convex in the variables (q_t, r_t) in the feasibility set:

$$h(q_t, r_t) = \sup_{\sigma_t > -\mu_{\min}} -\frac{\sigma_t r_t^2}{2} + \frac{1}{2}\Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big)T_2(\sigma_t) - \frac{1}{2}(q_t - \rho q_s^\star)^2(\sigma_t - 1/T_1(\sigma_t)) + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,\cdot)}\Big(r_t H_t + q_t S_t; T_1(\sigma_t)\Big)\Big], \tag{A67}$$

Note that the function $h(\cdot,\cdot)$ can also be expressed as follows:

$$h(q_t, r_t) = \sup_{\sigma_t > -\mu_{\min}} \min_{0 \le \tau \le C_\tau} \frac{\sigma_t \tau^2}{2} - \tau r_t \sigma_t + \frac{1}{2}\Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big)T_2(\sigma_t) + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,\cdot)}\Big(r_t H_t + q_t S_t; T_1(\sigma_t)\Big)\Big] - \frac{1}{2}(q_t - \rho q_s^\star)^2(\sigma_t - 1/T_1(\sigma_t)). \tag{A68}$$

Here, the feasibility set of the variable τ is bounded given that the optimal τ satisfies $\tau^\star = r_t$. It can be easily seen that the cost function of the optimization problem in (A68) is convex in τ and concave in σ_t. Then, using the result in [36], the function $h(\cdot,\cdot)$ can also be expressed as follows:

$$h(q_t, r_t) = \inf_{0 < \tau \le C_\tau} \sup_{\sigma_t > -\mu_{\min}} \frac{\sigma_t \tau}{2} - r_t \sigma_t + \frac{1}{2}\Big((1-\rho^2)(q_s^\star)^2 + (r_s^\star)^2\Big)T_2(\sigma_t/\tau) + \alpha_t \mathbb{E}\Big[\mathcal{M}_{\ell(Y_t,\cdot)}\Big(r_t H_t + q_t S_t; T_1(\sigma_t/\tau)\Big)\Big] - \frac{1}{2}(q_t - \rho q_s^\star)^2(\sigma_t/\tau - 1/T_1(\sigma_t/\tau)). \tag{A69}$$

Then, to prove our property, it suffices to show that the cost function of the above problem is jointly convex in the variables (q_t, r_t, τ). Using the positivity of the second derivative, it is easy to see that the function $\tau \to T_2(\sigma_t/\tau)$ is convex. Now, using the analysis below equation (161) in [20] (Appendix H), we can see that the remaining functions are jointly convex in the variables (q_t, r_t, τ). We omit these steps since they are similar to the approach employed in [20] (Appendix H). This shows that the function $\mathcal{O}_t(\cdot,\cdot)$ is strongly convex in the variables (q_t, r_t).

References

1. Pratt, L.Y.; Mostow, J.; Kamm, C.A. Direct Transfer of Learned Information among Neural Networks. In Proceedings of the Ninth National Conference on Artificial Intelligence—Volume 2, AAAI'91, Anaheim, CA, USA, 14–19 July 1991; pp. 584–589.
2. Pratt, L.Y. Discriminability-Based Transfer between Neural Networks. In *Advances in Neural Information Processing Systems*; Hanson, S., Cowan, J., Giles, C., Eds.; Morgan-Kaufmann: Burlington, MA, USA, 1993; Volume 5, pp. 204–211.
3. Perkins, D.; Salomon, G. *Transfer of Learning*; Pergamon: Oxford, UK, 1992.
4. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.
5. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. *arXiv* **2018**, arXiv:1808.01974.
6. Rosenstein, M.T.; Marx, Z.; Kaelbling, P.K.; Dietterich, T.G. To transfer or not to transfer. In *NIPS Workshop on Transfer Learning*; NIPS: Vancouver, BC, Canada, 2005.
7. Bakker, B.; Heskes, T. Task Clustering and Gating for Bayesian Multitask Learning. *J. Mach. Learn. Res.* **2003**, *4*, 83–99.
8. Ben-David, S.; Schuller, R. Exploiting Task Relatedness for Multiple Task Learning. In *Learning Theory and Kernel Machines*; Schölkopf, B., Warmuth, M.K., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 567–580.
9. Kornblith, S.; Shlens, J.; Le, Q.V. Do Better ImageNet Models Transfer Better? *arXiv* **2019**, arXiv:1805.08974.
10. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable Are Features in Deep Neural Networks? *arXiv* **2014**, arXiv:1411.1792.
11. Tommasi, T.; Orabona, F.; Caputo, B. Learning Categories from Few Examples with Multi Model Knowledge Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 928–941.
12. Yang, J.; Yan, R.; Hauptmann, A.G. Adapting SVM Classifiers to Data with Shifted Distributions. In Proceedings of the Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), Omaha, NE, USA, 28–31 October 2007; pp. 69–76.
13. Lampinen, A.K.; Ganguli, S. An Analytic Theory of Generalization Dynamics and Transfer Learning in Deep Linear Networks. *arXiv* **2019**, arXiv:1809.10374.
14. Dar, Y.; Baraniuk, R.G. Double Double Descent: On Generalization Errors in Transfer Learning between Linear Regression Tasks. *arXiv* **2021**, arXiv:2006.07002.

15. Saglietti, L.; Zdeborová, L. Solvable Model for Inheriting the Regularization through Knowledge Distillation. *arXiv* **2020**, arXiv:2012.00194.
16. Stojnic, M. A Framework to Characterize Performance of LASSO Algorithms. *arXiv* **2013**, arXiv:1303.7291.
17. Thrampoulidis, C.; Abbasi, E.; Hassibi, B. Precise high-dimensional error analysis of regularized M-estimators. In Proceedings of the 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 29 September–2 October 2015; pp. 410–417.
18. Gordon, Y. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*; Lindenstrauss, J., Milman, V.D., Eds.; Springer: Berlin/Heidelberg, Germany, 1988; pp. 84–106.
19. Dhifallah, O.; Thrampoulidis, C.; Lu, Y.M. Phase Retrieval via Polytope Optimization: Geometry, Phase Transitions, and New Algorithms. *arXiv* **2018**, arXiv:1805.09555.
20. Dhifallah, O.; Lu, Y.M. A Precise Performance Analysis of Learning with Random Features. *arXiv* **2020**, arXiv:2008.11904.
21. Salehi, F.; Abbasi, E.; Hassibi, B. The Impact of Regularization on High-dimensional Logistic Regression. *arXiv* **2019**, arXiv:1906.03761.
22. Kammoun, A.; Alouini, M.S. On the Precise Error Analysis of Support Vector Machines. *arXiv* **2020**, arXiv:2003.12972.
23. Mignacco, F.; Krzakala, F.; Lu, Y.M.; Zdeborová, L. The Role of Regularization in Classification of High-Dimensional Noisy Gaussian Mixture. *arXiv* **2020**, arXiv:2002.11544.
24. Aubin, B.; Krzakala, F.; Lu, Y.M.; Zdeborová, L. Generalization Error in High-Dimensional Perceptrons: Approaching Bayes Error with Convex Optimization. *arXiv* **2020**, arXiv:2006.06560.
25. Rockafellar, R.T.; Wets, R.J.B. *Variational Analysis*; Springer: Berlin/Heidelberg, Germany, 1998.
26. Thrampoulidis, C.; Oymak, S.; Hassibi, B. Regularized Linear Regression: A Precise Analysis of the Estimation Error. In *Proceedings of the 28th Conference on Learning Theory*; Grünwald, P., Hazan, E., Kale, S., Eds.; PMLR: Paris, France, 2015; Volume 40, Proceedings of Machine Learning Research, pp. 1683–1709.
27. Rudelson, M.; Vershynin, R. Non-Asymptotic Theory of Random Matrices: Extreme Singular Values. *arXiv* **2010**, arXiv:1003.2990.
28. Adachi, S.; Iwata, S.; Nakatsukasa, Y.; Takeda, A. Solving the Trust-Region Subproblem By a Generalized Eigenvalue Problem. *SIAM J. Optim.* **2017**, *27*, 269–291,
29. Shapiro, A.; Dentcheva, D.; Ruszczyński, A. *Lectures on Stochastic Programming: Modeling and Theory* , 2nd ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2014.
30. Andersen, P.K.; Gill, R.D. Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann. Statist.* **1982**, *10*, 1100–1120.
31. Newey, W.K.; Mcfadden, D. Chapter 36 Large sample estimation and hypothesis testing. In *Handbook of Econometrics*; Elsevier: Amsterdam, The Netherlands, 1994; p. 2111.
32. Schilling, R.L. *Measures, Integrals and Martingales*; Cambridge University Press: Cambridge, UK, 2005.
33. Shor, N. *Minimization Methods for Non-Differentiable Functions*; Springer: Berlin/Heidelberg, Germany, 1985.
34. Debbah, M.; Hachem, W.; Loubaton, P.; de Courville, M. MMSE analysis of certain large isometric random precoded systems. *IEEE Trans. Inf. Theory* **2003**, *49*, 1293–1311.
35. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
36. Sion, M. On general minimax theorems. *Pac. J. Math.* **1958**, *8*, 171–176.