Sample Complexity of Transfer Learning: an Optimal Transport Approach

Haoyang Cao, Xin Guo, Wenpin Tang, Guan Wang

***Abstract*—Transfer learning is an essential technique for many machine learning/AI models of complex structures such as large language models and generative AI. The essence of transfer learning is to leverage knowledge from resolved source tasks for a new target task, especially when the sample size m of the training data for the latter is low. In this work, we rigorously analyze the potential benefit of transfer learning in terms of sample efficiency. Specifically, taking an optimal transport viewpoint of transfer learning, we find that when the data dimension d is higher than 3, the sample complexity for transfer learning is $O(m^{-(\alpha+1)/d})$, with α indicating the smoothness of the data distribution, as opposed to the $O(m^{-p/d})$ sample complexity for direct learning with p indicating the smoothness of the optimal target model. Our finding theoretically supports a better sample efficiency for transfer learning, when the target task is optimizing over a family of not-so-smooth models (i.e., highly complex networks with the possible use of non-smooth activation functions). Using image classification as an example, we numerically demonstrate the sample efficiency for transfer learning, that is, in the data hungry regime, the model performance can be significantly improved by transfer learning.**

***Index Terms*—Transfer learning, sample complexity, optimal transport, nonparametric estimation, image classification.**

I. Introduction

TRANSFER learning is a machine learning technique where a model developed for a resolved task is incorporated into the training for a model on a new and related task. The key idea is to leverage knowledge from the resolved learning problem for which the data is abundant to improve the learning outcome in the new learning task where the data are scarce. For example, it has now become a common practice in image classification to reuse models such as `ResNet-50` pretrained over a large dataset of generic images (e.g., ImageNet). Transfer learning has demonstrated empirical success across different fields and is particularly useful when the target domain has limited labeled data, and when source and target tasks share underlying structure (e.g., image features, language patterns), including medical image analysis [1]–[3], natural language processing [4]–[6], large multimodal models [7]–[9], and biomedicine [10], [11].

a) Transfer learning and sample complexity.: Sample complexity refers to the number of training examples needed for a model to learn a task to a desired level of accuracy. Mathematically, this can be framed using PAC learning with prior knowledge [12], [13], Bayesian approaches [14], or domain adaptation bounds [15]. The empirical success has demonstrated that transfer learning can reduce sample complexity if the source task helps constrain the hypothesis space for the target task. In particular, if the source and target tasks are well-aligned, fewer samples are needed to achieve generalization. Theoretically, there has been a line of works analyzing the sample complexity for various transfer learning scenarios. In distributional transfer, the authors of [16] introduce transfer exponents to measure the target-relevant value of source samples, while [17] and [18] derive minimax and adaptive rates for classification under source–target drift. These works formalize transfer as an effective sample-size improvement where the source contribution depends on problem-specific relatedness. Complementary lower-bound results show that such gains are not automatic: without appropriate distributional information, multitask or transfer procedures may fail to improve over target-only learning, and adaptive transfer can be substantially harder than oracle transfer [19], [20]. Recent robust methods therefore aim to identify useful sources or avoid negative transfer. For example, the authors of [21] study unreliable source data through an "ambiguity level" between source and target regression functions and propose Transfer Around Boundary, which selectively uses source information near the decision boundary while retaining risk-improvement guarantees. The authors of [22] provide statistical guarantees for transfer learning from a representation learning perspective. Taking a similar representation learning perspective, the authors of [23] establish few-shot bounds that reduce target dependence from the ambient dimension d to an intrinsic representation dimension k once the representation is learned. The authors of [24] provide related upper and lower bounds for meta-learning linear representations. More recent work relaxes exact sharing assumptions by allowing similar but non-identical representations or outlier tasks [25], [26]. Casting multi-task bandit problems into representation learning, the authors of [27] derive instance-specific lower bounds for sample complexity. The authors of [28] show that the sample complexity can be significantly reduced by applying transfer learning across multiple reinforcement learning tasks with unknown Markov decision processes; such advantage was also shown in [29]. The authors of [30] analyze representational transfer in reinforcement learning and showed that source tasks can help learn features that allow target-task algorithms to approach the sample complexity achievable with known ground-truth representations. In contextual bandits, the authors of [31] derive minimax regret rates under covariate shift and quantified how auxiliary source data contribute to target-domain learning. The authors of [32] establish rigorous bounds for the

H. Cao is with Department of Applied Mathematics and Statistics, John Hopkins University, Baltimore, MD, USA.

X. Guo is with Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA, USA.

W. Tang is with Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA.

G. Wang is with Tsinghua-Berkeley Shenzhen Institute, Shenzhen, Guangdong, China.

sample efficiency advantage by employing transfer learning for linear reinforcement learning problems. Transfer has also been studied in structured high-dimensional models, where source data can estimate a shared component and target data need only estimate a sparse or low-dimensional correction [33]–[36]. In modern pretraining settings, recent theory shows that unlabeled or source-task data can reduce downstream labeled sample complexity when they recover reusable features [37]–[39]. Related sample-efficiency guarantees have also been developed for conditional diffusion models [40]. The authors of [41] adopt an entropic optimal transport approach to solve domain adaptation problems and established the corresponding sample complexity. The authors of [42] derive the sample complexity for domain adaptation for neural networks training. The authors of [43] provide guarantees on error rates and sample requirements in the presence of domain adaptation via importance sampling-based corrections.

b) Statistical optimal transport (OT).: It has emerged as a powerful framework to compare, align, and analyze probability distributions. Building on the foundational work of [44], which formalized OT theory and its geometric structure, a growing body of research has focused on statistical aspects such as estimation, convergence rates, and computational efficiency. Especially, to address the computational challenges of OT, regularized approaches such as entropic OT [45] and sliced OT [46] have been developed, yielding scalable algorithms with statistical guarantees. Applications span a wide range of domains, including generative modeling [47], domain adaptation [48], and causal inference [49], where OT provides both a metric for distributional comparison and a tool for constructing statistically principled estimators. Recent directions explore robust OT, and connections with kernel methods [50], further cementing OT's role as a unifying statistical framework for modern data analysis. Refer to [51] for a textbook account.

c) Our work.: We take an optimal transport viewpoint of transfer learning to analyze its impact on sample complexity. We interpret transfer learning as properly transporting the feature extraction mechanism from an resolved learning task to a new one. By casting direct learning as nonparametric estimation, we are able to provide the sample efficiencies for both approaches. In particular, we find that when the dimension of the input data is fixed, what influences the sample complexity the most is the smoothness of the optimal target model for direct learning, or the smoothness of input data distribution for transfer learning; a better smoothness condition implies a better sample complexity.

We conduct numerical experiments on image classification in two settings: the standard Office-31 dataset and the diagnosis of retinopathy of prematurity (ROP). The results show that transfer learning consistently outperforms direct learning across all levels of data availability and under all evaluation metrics, aligning with our theoretical analysis on sample complexity. The advantage of transfer learning becomes particularly pronounced as the amount of available data decreases.

d) Organization of the paper.: We present the transfer learning problem setting and preliminary analysis in Section II, where the general setup of nonparametric estimation can be found in Section II-A and the corresponding transfer learning formulation is in Section II-B. The theoretical result is in Section III and the numerical experiments in Section IV.

e) Notations.: Throughout the paper, we specify the following notations.

- For a random variable X, P_X denotes the probability distribution of X, and $\mathbb{E}X$ denotes its expectation.
- For f a function on U, and μ a probability distribution on U, $f_{\#}\mu$ denotes the pushforward of μ by f.
- For a function $f : \mathbb{R}^d \to \mathbb{R}$, $\nabla f := (\partial_1 f \ldots \partial_d f)$ denotes the gradient of f.
- For $\alpha > 0$, let $[\alpha]$ the integer part of α. We say that a function f is of class $\mathcal{C}^\alpha$ if f is $[\alpha]$-times continuous differentiable, and the $[\alpha]^{th}$ derivatives of f is $(\alpha-[\alpha])$-Hölder continuous.

II. Problem Setting and Preliminary Analysis

In this section, we fix a problem setting for transfer learning, as in [52]. For simplicity and without much loss of generality, we focus on a supervised learning framework.

We denote (X_T, Y_T) for the pair of input and output random variables taking values in $\mathcal{X}_T \times \mathcal{Y}_T$ for a target task T, where $\mathcal{X}_T$ is assumed to be compact subset of $\mathbb{R}^d$ with $d \in \mathbb{N}^+$ and $\mathcal{Y}_T = \mathbb{R}$. The loss for any admissible target model f_T is denoted as $\mathcal{L}_T(f_T) = \mathbb{E}_{X_T,Y_T}[L_T(f_T(X_T), Y_T)]$. Here the loss function $L_T : \mathcal{Y}_T \times \mathcal{Y}_T \to [0, +\infty)$ with $L_T(y, y) = 0$ for any $y \in \mathcal{Y}_T$. Note that the target loss $\mathcal{L}_T$ is subject to the joint distribution of (X_T, Y_T), and we assume this joint distribution to admit an absolutely continuous density.

Let f_T^* be an optimal target model among the set of admissible target models $A_T \subset \{f | f : \mathcal{X}_T \to \mathcal{Y}_T\}$ that minimizes the loss $\mathcal{L}_T$, then one may assume that the optimal target model is given by

$$f_T^*(x) = \mathbb{E}(Y_T \,|\, X_T = x), \quad x \in \mathcal{X}_T. \tag{II.1}$$

This holds under a general class of loss functions $L_T : \mathcal{Y}_T \times \mathcal{Y}_T \to \mathbb{R}$ called Bregman loss function, including mean squared error and (generalized) Kullback-Leibler divergence [53].

Under a direct learning approach, f_T^* is estimated using $m \in \mathbb{N}^+$ i.i.d. samples of (X_T, Y_T) available in the dataset $\mathcal{D}_T = \{(X_T^i, Y_T^i)\}_{i=1}^m$. In contrast, transfer learning is to construct an estimate of f_T^* by exploiting some already resolved problem, called the source task. Let $(X_S, Y_S) \in \mathcal{X}_S \times \mathcal{Y}_S$ be the pair of input and output random variables for S with $\mathcal{X}_S = \mathcal{X}_T$ and $\mathcal{Y}_S = \mathcal{Y}_T$, let $f_S^* : \mathcal{X}_S \to \mathcal{Y}_S$ be the optimal source model, called the pretrained model, that minimizes some loss $\mathcal{L}_S$ which is subject to the joint distribution of (X_S, Y_S). We also assume that the joint distribution of (X_S, Y_S) admits an absolutely continuous density function.

When the source task is sufficiently similar to the target task, the pretrained model f_S^* is believed to contain a feature extractor unit that is suitable for both tasks, in the following sense:

$$f_T^* \approx \mathcal{T}_{S\to T}^{\mathcal{Y}*} \circ f_S^* \circ \mathcal{T}_{T\to S}^{\mathcal{X}*}; \tag{II.2}$$

here, $\mathcal{T}_{T\to S}^{\mathcal{X}*} : \mathcal{X}_T \to \mathcal{X}_S$ aims to transport $Law(X_T)$ to $Law(X_S)$, and $\mathcal{T}_{S\to T}^{\mathcal{Y}*}$ aims to transport $Law(f_S^*(X_S))$ to

$Law(f_T^*(X_T))$. It is thus natural to adopt an optimal transport viewpoint to analysis sample efficiency of such a transfer learning approach. (See [44], [54] for more details). We refer to (II.3) as an illustration of this transfer learning framework.

$$\begin{tikzcd}
\mathcal{X}_S \ni X_S \arrow[rr, Rightarrow, "\text{Pretrained model } f_S^*"] & & f_S^*(X_S) \in \mathcal{Y}_S \arrow[d, Rightarrow, "\mathcal{T}_{S\to T}^{\mathcal{Y}*}"] \\
\mathcal{X}_T \ni X_T \arrow[u, Rightarrow, "\mathcal{T}_{T\to S}^{\mathcal{X}*}"] \arrow[rr, dashed, "\text{Direct learning}", "f_T^* \in \arg\min_{f\in A_T} \mathcal{L}_T(f_T)"'] & & f_T^*(X_T) \in \mathcal{Y}_T
\end{tikzcd} \tag{II.3}$$

A. Nonparametric Estimation in Direct Learning

In direct learning, it is to characterize the optimal relation $f_T^* : \mathcal{X}_T \to \mathcal{Y}_T$ prescribed in (II.1) between X_T and Y_T using the sample data $\mathcal{D}_T = \{(X_T^i, Y_T^i)\}_{i=1}^m$. Such an estimation can be done by parametric or nonparametric methods. A classical result of Stone [55], [56] shows how well the regression function f_T^* can be estimated with respect to the number of sample data m.

Theorem II.1. *[55], [56] Assume that $f_T^*(\cdot)$ in* (II.1) *is p-times differentiable. Then there exists an estimate $\hat{f}_m(\cdot)$ (based on m samples) such that*

$$\mathbb{E}|\hat{f}_m - f_T^*|_{L^2} \le C m^{-\frac{p}{2p+d}}, \tag{II.4}$$

for some $C > 0$ (independent of m). Moreover, the rate $m^{-\frac{p}{2p+d}}$ on the right side of (II.4) *is optimal in some asymptotic sense.*

Refer to [56, p.1042] for the precise definition of optimality stated in Theorem II.1. Note that the sample complexity (II.4) of nonparametric estimation depends on:

- the dimension d of the target space $\mathcal{X}_T$,
- the smoothness p of the regression function $f_T^*(x) := \mathbb{E}(Y_T \mid X_T = x)$.

B. Transfer Learning via Optimal Transport

When data is scarce, adopting a direct learning approach becomes challenging, especially when the training sample size m is too small relative to the model's complexity. In such cases, transfer learning offers a powerful alternative, by leveraging a pretrained model that has already been trained on a much larger dataset.

To formalize the transfer learning idea in (II.2), consider the following optimization problem,

$$\min_{\mathcal{T}_{T\to S}^{\mathcal{X}}, \mathcal{T}_{S\to T}^{\mathcal{Y}}} \mathbb{E}\left(Y_T - \mathcal{T}_{S\to T}^{\mathcal{Y}} \circ f_S^* \circ \mathcal{T}_{T\to S}^{\mathcal{X}}(X_T)\right)^2, \tag{II.5}$$

where $\mathcal{T}_{T\to S}^{\mathcal{X}}$ (resp. $\mathcal{T}_{S\to T}^{\mathcal{Y}}$) is a transfer map from $\mathcal{X}_T$ to $\mathcal{X}_S$ (resp. from $\mathcal{Y}_S$ to $\mathcal{Y}_T$). The transfer learning (II.5) can be viewed as a constrained optimization problem, with the transfer map $\mathcal{X}_S \to \mathcal{Y}_S$ set to be f_S^*. This transfer learning is inherently different from optimal transport between different dimensions [57], [58]. Second, it is possible to use a general loss function $L_T(\cdot,\cdot)$, which yields the objective:

$$\min_{\mathcal{T}_{T\to S}^{\mathcal{X}}, \mathcal{T}_{S\to T}^{\mathcal{Y}}} \mathbb{E} L_T\left(\mathcal{T}_{S\to T}^{\mathcal{Y}} \circ f_S^* \circ \mathcal{T}_{T\to S}^{\mathcal{X}}(X_T),\, Y_T\right).$$

The problem (II.5) corresponds to the quadratic loss $L_T(y, y') = (y - y')^2$ for any $y, y' \in \mathcal{Y}_T$, and other popular choices of the loss includes $L_T(y, y') = |y - y'|^q$ for $q > 1$. For ease of presentation, we focus on the quadratic loss, and the learning problem (II.5) in the sequel. Finally, for implementation we need to replace the expectation in (II.5) with the empirical average:

$$\min_{\mathcal{T}_{T\to S}^{\mathcal{X}}, \mathcal{T}_{S\to T}^{\mathcal{Y}}} \frac{1}{m} \sum_{i=1}^m \left(Y_T^i - \mathcal{T}_{S\to T}^{\mathcal{Y}} \circ f_S^* \circ \mathcal{T}_{T\to S}^{\mathcal{X}}(X_T^i)\right)^2. \tag{II.6}$$

The following lemma gives an easy sufficient condition for the problem (II.5) to have a solution.

Lemma II.2. *Let $f_T^*(\cdot)$ be defined by* (II.1)*, and assume that there exist transfer maps $\mathcal{T}_{T\to S}^{\mathcal{X}*}$, $\mathcal{T}_{S\to T}^{\mathcal{Y}*}$ such that*

$$f_T^* = \mathcal{T}_{S\to T}^{\mathcal{Y}*} \circ f_S^* \circ \mathcal{T}_{T\to S}^{\mathcal{X}*}. \tag{II.7}$$

Then $(\mathcal{T}_{T\to S}^{\mathcal{X}*}, \mathcal{T}_{S\to T}^{\mathcal{Y}*})$ is a solution to the problem (II.5).

Proof. The result follows from the fact that $f_T^*(X_T) := \mathbb{E}(Y_T \mid X_T)$ is the orthogonal projection of Y_T on the σ-field generated by X_T. $\square$

Lemma II.2 provides the existence of a solution to the learning problem (II.5). In fact, there may be many (possibly infinitely many) solutions to the problem (II.5), all of which are equally "good" because they lead to the same optimal prediction $f_T^*(\cdot)$. However, the non-uniqueness of solutions makes the analysis subtle: we need to choose a "canonical" solution as the ground truth for analysis. Here our idea is to use optimal transport theory, by choosing $(\mathcal{T}_{T\to S}^{\mathcal{X}*}, \mathcal{T}_{S\to T}^{\mathcal{Y}*})$ that are the (unique) solutions to some transportation problems. To this end, we make the following assumption.

Assumption II.3. *There exist transfer maps $\mathcal{T}_{T\to S}^{\mathcal{X}*}$, $\mathcal{T}_{S\to T}^{\mathcal{Y}*}$, and a random variable X_S on $\mathcal{X}_S$ such that:*

1) *The relation* (II.7) *holds.*
2) *The map $\mathcal{T}_{T\to S}^{\mathcal{X}*}$ solves the Monge problem:*

$$\min_{P_{X_S} = \mathcal{T}_\# P_{X_T}} \mathbb{E}(\mathcal{T} X_T - X_S)^2. \tag{II.8}$$

3) *Denote $\bar{Y}_S := f_S^*(X_S)$ and $\bar{Y}_T := f_T^*(X_T^*)$, so that $P_{\bar{Y}_S} := f_{S\#}^* P_{X_S}$ and $P_{\bar{Y}_T} := f_{T\#}^* P_{X_T}$. The map $\mathcal{T}_{S\to T}^{\mathcal{Y}*}$ solves the Monge problem:*

$$\min_{P_{\bar{Y}_T} = \mathcal{T}_\# P_{\bar{Y}_S}} \mathbb{E}(\mathcal{T}\bar{Y}_S - \bar{Y}_T)^2. \tag{II.9}$$

With the above assumptions, the following proposition resolves the non-uniqueness issue.

Proposition II.4. *The transfer maps $(\mathcal{T}_{T\to S}^{\mathcal{X}*}, \mathcal{T}_{S\to T}^{\mathcal{Y}*})$ in Assumption II.3 solve the problem* (II.5)*, and are unique almost everywhere.*

Proof. The fact that $(\mathcal{T}_{T\to S}^{\mathcal{X}*}, \mathcal{T}_{S\to T}^{\mathcal{Y}*})$ solve the problem (II.5) follow from Lemma II.2. By Brenier's theorem [54, Chapter 3], $\mathcal{T}_{T\to S}^{\mathcal{X}*}$ (resp. $\mathcal{T}_{S\to T}^{\mathcal{Y}*}$) is the unique (in the sense of almost everywhere) optimal transport map to the Monge problem in (2) (resp. in (3)). $\square$

Proposition II.4 shows that Assumption II.3 provides a canonical optimal solution to the transfer learning problem

(II.5), with additional optimal transport constraints $P_{X_S} = \mathcal{T}^{\mathcal{X}}_{T\to S\#}P_{X_T}$ and $P_{\bar{Y}_T} = \mathcal{T}^{\mathcal{Y}}_{S\to T\#}P_{\bar{Y}_S}$. Note that the quadratic costs in Assumption II.3 (2)-(3) can be replaced with any strictly convex costs (see [59]). Moreover, each of $(\mathcal{T}^{\mathcal{X}*}_{T\to S}, \mathcal{T}^{\mathcal{Y}*}_{S\to T})$, as optimal transport maps, can be expressed as the gradient of a convex function that solves a Monge-Ampère equation [60], [61]. See also [62]–[64] for the regularity theory of optimal transport maps.

III. Sample Complexity of Transfer Learning

In this section, we consider the sample complexity of the transfer learning problem. To establish the sample complexity result, we need some further assumptions.

Assumption III.1. *Let $(\mathcal{T}^{\mathcal{X}*}_{T\to S}, \mathcal{T}^{\mathcal{Y}*}_{S\to T})$ be specified by Assumption II.3.*

1) *There exist $C > c > 0$ such that $c \le |\nabla f^*_S| \le C$.*
2) *$\mathcal{T}^{\mathcal{Y}*}_{S\to T}$ is Lipschitz.*

Generally, the optimal transport maps are not Lipschitz. A sufficient condition for $\mathcal{T}^{\mathcal{Y}*}_{S\to T}$ to be Lipschitz is that both P_{Y_T} and P_{Y_S} are log-concave – this result is known as Caffarelli's contraction theorem [65], (see also [66], [67] for extensions).

Our main result is stated as follows. The proof relies on the recently developed statistical theory of optimal transport, see e.g., [50], [51].

Theorem III.2. *Let Assumption II.3–III.1 hold. Suppose that the densities of P_{X_T} and P_{X_S} are of class $\mathcal{C}^\alpha$ for some $\alpha > 0$. Then there exist transfer maps $(\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T}, \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S})$ (based on m samples) such that*

$$\mathbb{E}|\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - f^*_T|_{L^2} \le C(\kappa_m + m^{-\frac{1}{2}}), \quad \text{(III.1)}$$

for some $C > 0$ (independent of m), where

$$\kappa_m := \begin{cases} m^{-\frac{1}{2}} & \text{for} \quad d = 1, \\ \sqrt{\log m/m} & \text{for} \quad d = 2, \\ m^{-\frac{\alpha+1}{2\alpha+d}} & \text{for} \quad d \ge 3. \end{cases} \quad \text{(III.2)}$$

Proof. By Assumption II.3, we have:

$$f^*_T = \mathcal{T}^{\mathcal{Y}*}_{S\to T} \circ f^*_S \circ \mathcal{T}^{\mathcal{X}*}_{T\to S},$$

where $\mathcal{T}^{\mathcal{X}*}_{T\to S}$ (resp. $\mathcal{T}^{\mathcal{Y}*}_{S\to T}$) is the optimal transport map of the problem (II.8) (resp. (II.9)). For the moment, let $\hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S}$ (resp. $\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T}$) be any empirical optimal transport map to the problem (II.8) (resp. (II.9)), based on m i.i.d. samples from the distribution $P_{(X_T,X_S)}$ (resp. $P_{(Y_S,Y_T)}$). Denote by

$$\hat{f}_m := \hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S}, \quad \text{(III.3)}$$

the prediction via transfer learning. By the triangle inequality, we have:

$$\mathbb{E}|\hat{f}_m - f^*_T|_{L^2} \le (a) + (b), \quad \text{(III.4)}$$

where

$$\begin{aligned} (a) &:= \mathbb{E}|\mathcal{T}^{\mathcal{Y}*}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - \mathcal{T}^{\mathcal{Y}*}_{S\to T} \circ f^*_S \circ \mathcal{T}^{\mathcal{X}*}_{T\to S}|_{L^2}, \\ (b) &:= \mathbb{E}|\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - \mathcal{T}^{\mathcal{Y}*}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S}|_{L^2}. \end{aligned}$$

We proceed to estimating the terms (a) and (b).

$\boxed{\text{Term (a)}}$ By the Lipschitz assumption on $\mathcal{T}^{\mathcal{Y}*}_{S\to T}$ and f^*_S, we get:

$$|\mathcal{T}^{\mathcal{Y}*}_{S\to T}\circ f^*_S\circ\hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - \mathcal{T}^{\mathcal{Y}*}_{S\to T}\circ f^*_S\circ\mathcal{T}^{\mathcal{X}*}_{T\to S}|_{L^2} \le C_1|\hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - \mathcal{T}^{\mathcal{X}*}_{T\to S}|_{L^2},$$

for some $C_1 > 0$. By [50, Theorem 18], there exists a (kernel) transport map $\hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S}$ such that $\mathbb{E}|\hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - \mathcal{T}^{\mathcal{X}*}_{T\to S}|_{L^2} \le C_2\kappa_m$ for some $C_2 > 0$, where κ_m is defined by (III.2). This yields the estimate:

$$(a) \le C_3\kappa_m, \quad \text{for some } C_3 > 0 \text{ (independent of } m). \quad \text{(III.5)}$$

$\boxed{\text{Term (b)}}$ Denote by $g_m := f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S}$. We can check that $\inf_x |\partial_1 g_m| > C_4$ for some $C_4 > 0$ (independent of m). Next we complete $g_m(x)$ with $(g_m(x), x_2, \dots, x_n)$ to form a basis, and obtain:

$$\begin{aligned} &|\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - \mathcal{T}^{\mathcal{Y}*}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S}|_{L^2} \\ &= |\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} \circ g_m - \mathcal{T}^{\mathcal{Y}*}_{S\to T} \circ g_m|_{L^2} \\ &\le C_5|\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} - \mathcal{T}^{\mathcal{Y}*}_{S\to T}|_{L^2}, \end{aligned}$$

where the inequality follows from a change of variables $x \leftrightarrow (g_m(x), x_2, \dots, x_n)$, with the Jacobian matrix of form:

$$\begin{pmatrix} (\partial_1 g_m)^{-1} & 0 & \cdots & 0 \\ * & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & 0 & \cdots & 1 \end{pmatrix},$$

and the fact that $\mathcal{X}_T$ is a compact subset of $\mathbb{R}^d$. Again by [50, Theorem 18] (specializing to $d = 1$), there exists a (kernel) transport map $\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T}$ such that $\mathbb{E}|\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} - \mathcal{T}^{\mathcal{Y}*}_{S\to T}|_{L^2} \le C_6/\sqrt{m}$ for some $C_6 > 0$. As a result,

$$(b) \le \frac{C_7}{\sqrt{m}}, \quad \text{for some } C_7 > 0 \text{ (independent of } m). \quad \text{(III.6)}$$

Combining (III.4), (III.5) and (III.6) yields the desired result. □

Remark on Theorem III.2. Of particular interest is the high-dimensional case ($d \to \infty$), where the bound (III.1) leads to:

$$\mathbb{E}|\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S} - f^*_T|_{L^2} \le Cm^{-\frac{\alpha+1}{2\alpha+d}}. \quad \text{(III.7)}$$

Note that the convergence rate depends on:

- the dimension d of the target space $\mathcal{X}_T$,
- the smoothness α of the target distribution (not on the smoothness p of the regressor $\mathbb{E}(Y_T \mid X_T = x)$).

Comparing with Theorem II.1, L^2 errors for the learned models are upper bounded by

$$\begin{cases} O\left(m^{-\frac{p}{d}}\right) & \text{for direct approach,} \\ O\left(m^{-\frac{\alpha+1}{d}}\right) & \text{for transfer learning approach.} \end{cases} \quad \text{(III.8)}$$

So in the direct approach, the exponent is proportional to $\frac{p}{d}$, where p is the smoothness of the regression kernel $\mathbb{E}(Y_T \mid X_T = x)$; while in the transfer learning case, only the smoothness α of the source/target distribution matters.

A typical scenario where the transfer learning approach has a better sample complexity is that the regressor $\mathbb{E}(Y_T \mid X_T = x)$ is lesser smooth, e.g., $p = 1$ that yields the rate $m^{-\frac{1}{d}}$; while the source/target distributions are much smoother ($\alpha \to +\infty$). In particular, if

$$\alpha + 1 > p, \tag{III.9}$$

then the transfer learning approach is advantageous over the direct approach in terms of the sample complexity. In practice, the above situation can occur when the goal of the target task is to learn a highly complex network containing non-smooth activation functions, while the input data distributions are approximately of exponential family.

IV. Numerical Experiments

In this section, we use image classification to study sample complexity between transfer learning and direct learning. We analyze numerically two cases. The first is the classification problem using the benchmark dataset, Office-31. The second is the classification problem in diagosis for one of the eye diseases for pre-mature infants called retinopathy of prematurity (ROP). In both cases, we vary the amount of the corresponding target data used for the training data, and compare the model performance between direct learning and transfer learning. Our experiments results are shown to be consistent with the theoretical results presented in Sections III. In particular, when data is scarce, the advantage of transfer learning over direct learning becomes most pronounced.

a) Evaluation metrics and neural network setup.: For both direct and transfer learning, we adopt the following four performance metrics for the ROP diagnosis results:

1) AUROC evaluates classification models' ability to discriminate between positive and negative classes across all thresholds,
2) Accuracy is the ratio of correct predictions out of all predictions, i.e., $\frac{\text{TruePositives}+\text{TrueNegatives}}{\text{All predictions}}$,
3) Precision is the ratio of true positives over all reported positives, i.e., $\frac{\text{TruePositives}}{\text{TruePositives}+\text{FalsePositives}}$, and
4) Sensitivity is the number of true positives detected over the total number of positive cases in the test dataset, i.e., $\frac{\text{TruePositives}}{\text{TruePositives}+\text{FalseNegatives}}$.

For each metric, the numerical value ranges in $[0, 1]$, with higher values indicating better learning outcome. An AUROC score of 1.0 indicates a perfect model, and 0.5 indicating random guessing. We also report the average relative performance differences between transfer learning and direct learning, where

$$\Delta\text{metric} = \frac{\text{metric}_{\text{TL}} - \text{metric}_{\text{direct learning}}}{\text{metric}_{\text{direct learning}}} \times 100\%.$$

The neural network for classification consists of three units: the resizing unit, the feature extractor unit implemented via `ResNet50`, and a classifier layer.

A. Image Classification on the Office-31 Dataset

To study the sample complexity for transfer learning versus direct learning, we start with a standard image classification problem with a benchmark dataset Office-31 [68].

In this dataset, there are three domains: Amazon (A), Webcam (W), and DSLR (D), containing 4114 images of 31 categories of objects in an office environment. In this experiment, images from domains D and W are seen as the target data, with a total number of $\text{total} = 1296$. When applying transfer learning, illustrated in Figure **??**, the feature extractor, with input dimension $3 \times 244 \times 244$ and output dimension 2048, is pretrained using either ImageNet (in which case the weights are publicly available in [69]) or domain A as the source data. The resizing unit and the classifier layer are retrained as the input and output transport mappings, respectively.

By the direct learning approach, the neural network is trained with a random initialization. We vary the percentage of the target data (i.e. D+W Data $\%$) being used for both the training in direct learning and the fine-tuning in transfer learning, so that $m = \text{total} \times$ D+W Data $\%$.

a) Results. : The average performance evaluation for both direct and transfer learning is summarized in Table I. It is clear that transfer learning, either using ImageNet or domain A as source data, outperforms direct learning across all metrics, and at any given level of target data availability. At the data-scarce regime, i.e., D+W Data % $\leq 30\%$, transfer learning significantly dominates direct learning. In the extreme case with 10% of the target data in the training set, direct learning achieves only 0.3 in Precision, while the transfer learning achieves over 0.6.

To further demonstrated the benefit of transfer learning, Table III reports the average relative performance differences between transfer learning (ImageNet version) and direct learning. Here the benefit of transfer learning become increasingly more prominent as the target data becomes increasingly scarce. Our numerical analysis also shows that the choice of different source task and data may affect the quality of transfer learning.

The numerical results align with the theoretical findings in Section III. From Figure **??**, we see that the target classification model employs a `ResNet-50`-based feature extractor, containing multiple layers with `ReLU` as the activation functions; such an architecture reduces the smoothness p of the optimal target model, i.e., $p \leq 1$. The distribution of images can be approximated as a Gaussian mixture model [70] which is smooth, i.e., $\alpha = +\infty$. These settings make (III.9) hold.

B. Detecting ROP from Detections of DR

We now study the problem of detecting an eye disease called retinopathy of prematurity (ROP). To analyze the sample efficiency of transfer learning compared with direct learning, we compare the diagnosis performances of $\hat{\mathcal{T}}^{\mathcal{Y},m}_{S\to T} \circ f^*_S \circ \hat{\mathcal{T}}^{\mathcal{X},m}_{T\to S}$ and $\hat{f}_m$ as we vary the sample sizes m of ROP training data.

a) Classification problem for ROP.: ROP is a common retinal vascular disease primarily affecting prematurely-born infants or infants with low birth weight [71], which is one of the leading causes of infant blindness globally [72]–[74]. To build an efficient, e.g., deep-learning based, ROP detector, the major obstacles are the lack of sufficient data *and* the high accuracy requirement due to the specific case of ROP involving infants.

Technically, the ROP detection problem is a binary classification problem, with f^*_T being the optimal classifier. By direct learning, the classifier is trained entirely on the ROP dataset which contains 9727 anonymized images in total,

TABLE I: Office-31 Classification Performance Evaluation

D+W Data % (m)	AUROC			Accuracy			Precision			Sensitivity		
	Direct	Transfer-A	Transfer-ImageNet	Direct	Transfer-A	Transfer-ImageNet	Direct	Transfer-A	Transfer-ImageNet	Direct	Transfer-A	Transfer-ImageNet
100% (1296)	1.00	1.00	1.00	0.97	0.99	1.00	0.98	0.99	1.00	0.97	0.99	1.00
50% (648)	1.00	1.00	1.00	0.90	0.94	1.00	0.91	0.95	1.00	0.90	0.95	1.00
20% (259)	0.95	0.98	0.99	0.57	0.71	0.85	0.56	0.68	0.88	0.56	0.70	0.86
10% (130)	0.87	0.94	0.96	0.35	0.60	0.72	0.30	0.60	0.72	0.34	0.59	0.73

TABLE II: ROP Detection Performance Evaluation

ROP Data % (m)	AUROC		Accuracy		Precision		Sensitivity	
	Direct	Transfer	Direct	Transfer	Direct	Transfer	Direct	Transfer
100% (5838)	0.88	0.97	0.81	0.96	0.57	0.90	0.78	0.94
50% (2919)	0.85	0.97	0.82	0.95	0.60	0.88	0.71	0.90
10% (584)	0.81	0.95	0.80	0.91	0.56	0.80	0.67	0.84
1% (58)	0.72	0.84	0.73	0.83	0.44	0.63	0.49	0.67

TABLE III: Relative Improvement of Transfer Learning over Direct Learning for Office-31 Classification

D+W Data %	ΔAUROC	ΔAccuracy	ΔPrecision	ΔSensitivity
100%	0.02%	3.25%	2.07%	2.82%
50%	0.26%	11.40%	9.99%	10.98%
20%	4.56%	47.95%	55.62%	54.99%
10%	10.18%	109.09%	142.69%	111.77%

with 2310 positive samples and 7417 negative samples. The corresponding classifier after direct learning is $\hat{f}_m$.

b) Transfer learning from DR to ROP.: By the transfer learning approach, we leverage on the successes of the deep learning approach for diagnosis of diabetic retinopathy (DR), another retinal vascular disease for people with diabetes. The pre-trained models for DP is our choice of source task for the transfer learning study.

The DR dataset is much larger and contains 36126 anonymized images with 26548 positive samples and 9578 negative samples. Among all the 9727 ROP data, 1945 images (462 positive samples and 1483 negative samples) are used as the validation data and 1944 images (461 positive samples and 1483 negative samples) are reserved to be the test data. The maximum number of training samples available is $\mathrm{total} = 5838$ (1387 positive samples and 4451 negative samples).

With transfer learning, as shown in Figure 2, the feature extractor is first pretrained on DR dataset, denoted by f_S^*. Then, using the ROP dataset, this feature extractor is fine-tuned with the resizing unit as the input transport mapping $\hat{\mathcal{T}}_{T\to S}^{\mathcal{X},m}$ and the classifier layer as the output transport mapping $\hat{\mathcal{T}}_{S\to T}^{\mathcal{Y},m}$.

In the experiment, we vary the percentage of ROP data being used for both the training in direct learning and the fine-tuning in transfer learning, so that $m = \mathrm{total} \times$ ROP Data %.

c) Results. : Similar to the Office-31 example in Section IV-A, the numerical findings for ROP problem detailed in Tables II–IV are (again) consistent with the theoretical results in Sections III.

As shown in Table II, transfer learning demonstrates consistent performance improvements over direct learning across all levels of ROP data availability and under all evaluation metrics. Despite the relatively limited ROP dataset, transfer learning achieves and surpasses the critical threshold of 90 percent in every category. Remarkably, it requires only 973 images, i.e, less than 10 percent of the available ROP data, to attain AUROC and Accuracy values exceeding 0.9. In particular, with limited ROP data, transfer learning achieves impressive scores of 0.97 for AUROC and 0.96 for Accuracy. By contrast, direct learning fails to reach the 0.9 threshold in any category, even when nearly the full ROP dataset is utilized. In the extreme data scarce regime with less than 10 percent of ROP data, direct learning never reaches 0.5 under either Precision or Sensitivity.

Table IV further compares the mean relative performance difference between transfer learning and direct learning Δmetric. Here the advantage of transfer learning in the data-scarce regime is even more striking across the board. When the proportion of available ROP data is reduced to the extreme case of one percent (i.e., fewer than 100 images), transfer learning yields the most significant gains in both sensitivity and AUROC. These observations suggest that when the training data becomes increasingly limited, the benefit of transfer learning remains robust.

References

[1] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang, "Automatic ICD-9 coding via deep transfer learning," *Neurocomputing*, vol. 324, pp. 43–50, 2019.

[2] G. Wang, Y. Kikuchi, J. Yi, Q. Zou, R. Zhou, and X. Guo, "Transfer learning for retinal vascular disease detection: A pilot study with diabetic retinopathy and retinopathy of prematurity," *arXiv preprint arXiv:2201.01250*, 2022.

[3] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC Medical Imaging*, vol. 22, no. 1, p. 69, 2022.

[4] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *NACCL*, 2019, pp. 15–18.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NACCL*, vol. 1, 2019, pp. 4171–4186.

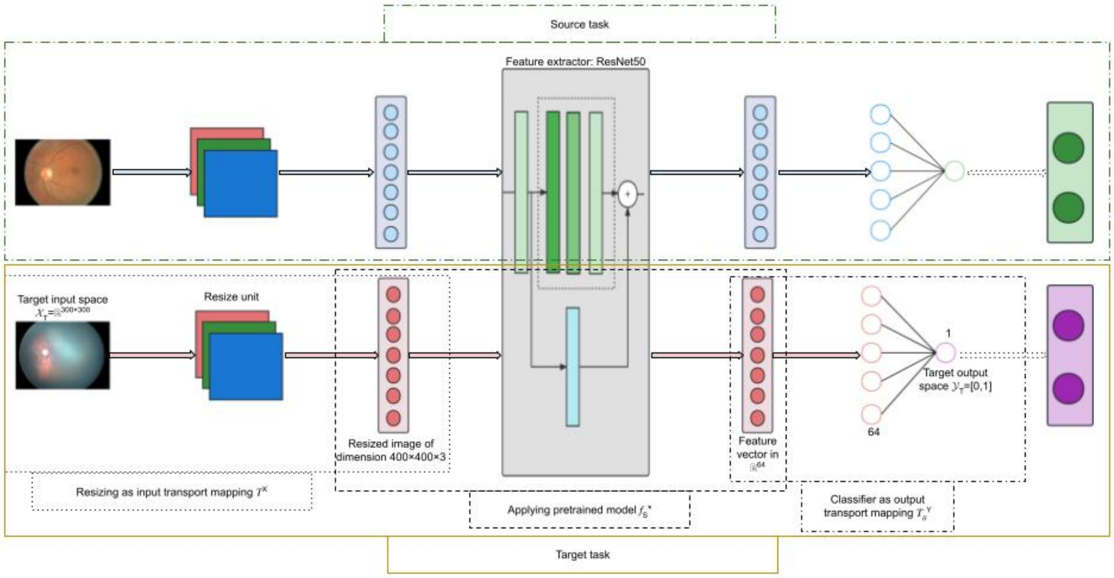

Fig. 2: Transfer learning

TABLE IV: Relative Improvement of Transfer Learning over Direct Learning for ROP Detection

ROP Data %	ΔAUROC	ΔAccuracy	ΔPrecision	ΔSensitivity
100%	10.05%	18.75%	58.01%	19.71%
50%	14.09%	15.37%	45.83%	27.19%
10%	16.58%	14.65%	43.68%	26.03%
1%	16.08%	13.04%	46.67%	49.46%

[6] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *CVPR*, 2022, pp. 5227–5237.

[7] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[8] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[9] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[10] D. Xu, J. X. Yao, S. Song, Z. Zhu, and J. Ji, "Ibex: Information-bottleneck-explored coarse-to-fine molecular generation under limited data," *arXiv preprint arXiv:2508.10775*, 2025.

[11] P. A. Apellaniz, B. A. Galende, and A. Jim, "Advancing cancer research with synthetic data generation in low-data scenarios," *IEEE Journal of Biomedical Informatics*, 2025.

[12] B. Eisenberg and R. L. Rivest, "On the sample complexity of pac-learning using random and chosen examples." in *COLT*, vol. 90, 1990, pp. 154–162.

[13] S. Hanneke, "The optimal sample complexity of pac learning," *J. Mach. Learn. Res.*, vol. 17, no. 38, pp. 1–15, 2016.

[14] D. Haussler, M. Kearns, and R. E. Schapire, "Bounds on the sample complexity of bayesian learning using information theory and the vc dimension," *Mach. Learn.*, vol. 14, no. 1, pp. 83–113, 1994.

[15] S. Ben-David and R. Urner, "On the hardness of domain adaptation and the utility of unlabeled target samples," in *ALT*, 2012, pp. 139–153.

[16] S. Hanneke and S. Kpotufe, "On the value of target data in transfer learning," in *Neurips*, vol. 32, 2019, pp. 9867–9877.

[17] T. T. Cai and H. Wei, "Transfer learning for nonparametric classification: Minimax rate and adaptive classifier," *Ann. Stat.*, vol. 49, no. 1, pp. 100–128, 2021.

[18] H. W. J. Reeve, T. I. Cannings, and R. J. Samworth, "Adaptive transfer learning," *Ann. Stat.*, vol. 49, no. 6, pp. 3618–3649, 2021.

[19] S. Hanneke and S. Kpotufe, "A no-free-lunch theorem for multitask learning," *Ann. Stat.*, vol. 50, no. 6, pp. 3119–3143, 2022.

[20] S. Hanneke, S. Kpotufe, and Y. Mahdaviyeh, "Limits of model selection under transfer learning," in *COLT*, vol. 36, 2023, pp. 5781–5812.

[21] J. Fan, C. Gao, and J. M. Klusowski, "Robust transfer learning with unreliable source data," *Ann. Stat.*, vol. 53, no. 4, pp. 1728–1752, 2025.

[22] N. Tripuraneni, M. Jordan, and C. Jin, "On the theory of transfer learning: The importance of task diversity," in *Neurips*, vol. 33, 2020, pp. 7852–7862.

[23] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," in *ICLR*, 2021.

[24] N. Tripuraneni, C. Jin, and M. I. Jordan, "Provable meta-learning of linear representations," in *ICML*, vol. 38, 2021, pp. 10 434–10 443.

[25] Y. Duan and K. Wang, "Adaptive and robust multi-task learning," *Ann. Stat.*, vol. 51, no. 5, pp. 2015–2039, 2023.

[26] Y. Tian, Y. Gu, and Y. Feng, "Learning from similar linear representations: Adaptivity, minimaxity, and robustness," *J. Mach. Learn. Res.*, vol. 26, no. 187, pp. 1–125, 2025.

[27] A. Russo and A. Proutiere, "On the sample complexity of representation learning in multi-task bandits with global and local structure," in *AAAI*, vol. 37, 2023, pp. 9658–9667.

[28] E. Brunskill and L. Li, "Sample complexity of multi-task reinforcement learning," in *UAI*, vol. 29, 2013, p. 122–131.

[29] K. Oguni, K. Narisawa, and A. Shinohara, "Reducing sample complexity in reinforcement learning by transferring transition and reward probabilities," in *ICAART*, vol. 2, 2014, pp. 632–638.

[30] A. Agarwal, Y. Song, W. Sun, K. Wang, M. Wang, and X. Zhang, "Prov-

able benefits of representational transfer in reinforcement learning," in *COLT*, vol. 36, 2023, pp. 2114–2187.

[31] C. Cai, T. T. Cai, and H. Li, "Transfer learning for contextual multi-armed bandits," *Ann. Stat.*, vol. 52, no. 1, pp. 207–232, 2024.

[32] M. Mahajan, A. Pacchiano, and X. Zhang, "Learning to undo: Transfer reinforcement learning under linear state space transformations," in *OpenReview*, 2025. [Online]. Available: https://openreview.net/forum?id=jI8a3s5xUz

[33] S. Li, T. T. Cai, and H. Li, "Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 84, no. 1, pp. 149–173, 2022.

[34] Y. Tian and Y. Feng, "Transfer learning under high-dimensional generalized linear models," *J. Am. Stat. Assoc.*, vol. 118, no. 544, pp. 2684–2697, 2023.

[35] S. S. Liu, "Unified transfer learning in high-dimensional linear regression," in *AISTATS*, vol. 27, 2024, pp. 1036–1044.

[36] Z. He, Y. Sun, and R. Li, "TransFusion: Covariate-shift robust transfer learning for high-dimensional regression," in *AISTATS*, vol. 27, 2024, pp. 703–711.

[37] J. Ge, S. Tang, J. Fan, and C. Jin, "On the provable advantage of unsupervised pretraining," in *ICLR*, 2024.

[38] T. Jones-McCormick, A. Jagannath, and S. Sen, "Provable benefits of unsupervised pre-training and transfer learning via single-index models," in *ICLM*, vol. 42, 2025, pp. 28 350–28 376.

[39] J. Tahir, S. Ganguli, and G. M. Rotskoff, "Features are fate: A theory of transfer learning in high-dimensional regression," in *ICML*, vol. 42, 2025, pp. 58 142–58 168.

[40] Z. Cheng, T. Xie, S. Zhang, and C. Zhang, "Provable sample-efficient transfer learning conditional diffusion models via representation learning," in *Neurips*, vol. 38, 2025.

[41] P. Rigollet and A. J. Stromme, "On the sample complexity of entropic optimal transport," *Ann. Stat.*, vol. 53, no. 1, pp. 61 – 90, 2025.

[42] E. Vural and H. Karaca, "A unified analysis of generalization and sample complexity for semi-supervised domain adaptation," *arXiv preprint arXiv:2507.22632*, 2025.

[43] C. X. Ren, Y. W. Luo, C. J. Guo, and X. L. Xu, "Partial domain adaptation via importance sampling-based shift correction," *IEEE Trans. Neural Netw.*, 2025.

[44] C. Villani, *Optimal transport*, ser. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2009, vol. 338, old and new.

[45] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *NIPS*, vol. 26, 2013, p. 2292–2300.

[46] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and Radon Wasserstein barycenters of measures," *J. Math. Imaging Vision*, vol. 51, pp. 22–45, 2015.

[47] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet, "Diffusion schrödinger bridge with applications to score-based generative modeling," in *Neurips*, vol. 34, 2021, pp. 17 695–17 709.

[48] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1853–1865, 2016.

[49] W. Torous, F. Gunsilius, and P. Rigollet, "An optimal transport approach to estimating causal effects via nonlinear difference-in-differences," *arXiv preprint arXiv:2108.05858*, 2021.

[50] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman, "Plugin estimation of smooth optimal transport maps," *Ann. Statist.*, vol. 52, no. 3, pp. 966–998, 2024.

[51] S. Chewi, J. Niles-Weed, and P. Rigollet, "Statistical optimal transport," *arXiv preprint arXiv:2407.18163*, 2024.

[52] H. Cao, H. Gu, X. Guo, and M. Rosenbaum, "Risk of transfer learning and its applications in finance," *arXiv preprint arXiv:2311.03283*, 2023.

[53] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a bregman predictor," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2664–2669, 2005.

[54] C. Villani, *Topics in optimal transportation*, ser. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2003, vol. 58.

[55] C. J. Stone, "Optimal rates of convergence for nonparametric estimators," *Ann. Statist.*, vol. 8, no. 6, pp. 1348–1360, 1980.

[56] ——, "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, vol. 10, no. 4, pp. 1040–1053, 1982.

[57] R. J. McCann and B. Pass, "Optimal transportation between unequal dimensions," *Arch. Ration. Mech. Anal.*, vol. 238, no. 3, pp. 1475–1520, 2020.

[58] B. Pass, "Regularity of optimal transportation between spaces with different dimensions," *Math. Res. Lett.*, vol. 19, no. 2, pp. 291–307, 2012.

[59] W. Gangbo and R. J. McCann, "The geometry of optimal transportation," *Acta Math.*, vol. 177, no. 2, pp. 113–161, 1996.

[60] Y. Brenier, "Décomposition polaire et réarrangement monotone des champs de vecteurs," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 305, no. 19, pp. 805–808, 1987.

[61] ——, "Polar factorization and monotone rearrangement of vector-valued functions," *Comm. Pure Appl. Math.*, vol. 44, no. 4, pp. 375–417, 1991.

[62] L. A. Caffarelli, "Boundary regularity of maps with convex potentials," *Comm. Pure Appl. Math.*, vol. 45, no. 9, pp. 1141–1151, 1992.

[63] ——, "The regularity of mappings with a convex potential," *J. Amer. Math. Soc.*, vol. 5, no. 1, pp. 99–104, 1992.

[64] ——, "Boundary regularity of maps with convex potentials. II," *Ann. of Math. (2)*, vol. 144, no. 3, pp. 453–496, 1996.

[65] ——, "Monotonicity properties of optimal transportation and the FKG and related inequalities," *Comm. Math. Phys.*, vol. 214, no. 3, pp. 547–563, 2000.

[66] G. Carlier, A. Figalli, and F. Santambrogio, "On optimal transport maps between 1/d-concave densities," *arXiv preprint arXiv:2404.05456*, 2024.

[67] M. Colombo, A. Figalli, and Y. Jhaveri, "Lipschitz changes of variables between perturbations of log-concave measures," *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, vol. 17, no. 4, pp. 1491–1519, 2017.

[68] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proceedings of the 11th European Conference on Computer Vision*. Springer, 2010, pp. 213–226.

[69] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[70] H. Permuter, J. Francos, and I. Jermyn, "A study of gaussian mixture models of color and texture features for image classification and segmentation," *Pattern recognition*, vol. 39, no. 4, pp. 695–706, 2006.

[71] W. M. Fierson, "Screening examination of premature infants for retinopathy of prematurity," *Pediatrics*, vol. 142, no. 6, 2018.

[72] H. Blencowe, J. Lawn, T. Vazquez, A. Fielder, and C. Gilbert, "Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010," *Pediatr. Res.*, vol. 74 Suppl 1, pp. 35–49, 12 2013.

[73] C. Gilbert, J. Rahi, M. Eckstein, J. O'sullivan, and A. Foster, "Retinopathy of prematurity in middle-income countries," *The Lancet*, vol. 350, no. 9070, p. 12–14, 1997.

[74] Early Treatment For Retinopathy Of Prematurity Cooperative Group, "Revised Indications for the Treatment of Retinopathy of Prematurity: Results of the Early Treatment for Retinopathy of Prematurity Randomized Trial," *Archives of Ophthalmology*, vol. 121, no. 12, pp. 1684–1694, 2003.