The Annals of Statistics
2021, Vol. 49, No. 6, 3618–3649
https://doi.org/10.1214/21-AOS2102

ADAPTIVE TRANSFER LEARNING

BY HENRY W. J. REEVE[1], TIMOTHY I. CANNINGS[2] AND RICHARD J. SAMWORTH[3]

[1]*School of Mathematics, University of Bristol, henry.reeve@bristol.ac.uk*
[2]*School of Mathematics, University of Edinburgh, timothy.cannings@ed.ac.uk*
[3]*Statistical Laboratory, University of Cambridge, r.samworth@statslab.cam.ac.uk*

In transfer learning, we wish to make inference about a target population when we have access to data both from the distribution itself, and from a different but related source distribution. We introduce a flexible framework for transfer learning in the context of binary classification, allowing for covariate-dependent relationships between the source and target distributions that are not required to preserve the Bayes decision boundary. Our main contributions are to derive the minimax optimal rates of convergence (up to polylogarithmic factors) in this problem, and show that the optimal rate can be achieved by an algorithm that adapts to key aspects of the unknown transfer relationship, as well as the smoothness and tail parameters of our distributional classes. This optimal rate turns out to have several regimes, depending on the interplay between the relative sample sizes and the strength of the transfer relationship, and our algorithm achieves optimality by careful, decision tree-based calibration of local nearest-neighbour procedures.

1. Introduction. Transfer learning refers to statistical problems in which we wish to make inference about a test data population, but where some (typically, the large majority) of our training data come from a related but distinct distribution. Such problems arise in many natural, practical settings: for instance, we may wish to understand the effectiveness of a treatment on a particular subgroup of a population, but still wish to exploit information about its efficacy on the wider population under study. In medical applications, we may be interested in making predictions in a given experimental setting, or using a particular piece of equipment, but also have data obtained under different scenarios or measured with different devices. Closely related problems have recently been of interest to many communities, sometimes studied under the banner of label noise (Blanchard et al. (2016), Cannings, Fan and Samworth (2020), Frénay and Verleysen (2014)), multi-task learning (Caruana (1997), Maurer, Pontil and Romera-Paredes (2016)) or distributional robustness (Christiansen et al. (2020), Sinha, Namkoong and Duchi (2018), Weichwald and Peters (2021)). For recent survey papers on transfer learning, see Pan and Yang (2009), Storkey (2009) and Weiss, Khoshgoftaar and Wang (2016).

We focus here on transfer learning in the context of binary classification, both due to the latter's fundamental importance as a canonical problem in modern statistics and machine learning, and because, as we shall see, its structure is particularly amenable to algorithms that seek to exploit relationships between the training and test distributions. To set the scene for our contributions, let P and Q denote two distributions on $\mathbb{R}^d \times \{0, 1\}$, with corresponding generic random pairs (X^P, Y^P) and (X^Q, Y^Q), respectively. We think of P as a source distribution, from which most of our training data are generated, and Q as a target distribution, from which we may have some training data, and about which we wish to make inference. Let

Received July 2020; revised June 2021.
MSC2020 subject classifications. Primary 62G99; Secondary 62C20, 62G20.
Key words and phrases. Transfer learning, classification, decision trees, nearest neighbours, nonparametric, minimax.

$\eta_P, \eta_Q : \mathbb{R}^d \to [0, 1]$ denote the source and target regression functions, respectively, defined by

$$\eta_P(x) := \mathbb{P}(Y^P = 1 | X^P = x) \quad \text{and} \quad \eta_Q(x) := \mathbb{P}(Y^Q = 1 | X^Q = x). \tag{1}$$

Our main working assumption on the relationship between P and Q will be that our feature space $\mathbb{R}^d$ can be partitioned into finitely many cells $\mathcal{X}_\ell^*$, and for each cell there exists a *transfer function* $g_\ell : [0, 1] \to [0, 1]$ such that η_P can be approximated by $g_\ell \circ \eta_Q$ on $\mathcal{X}_\ell^*$. We will further assume that the cells arise from a decision tree partition (Breiman et al. (1984)), and that each transfer function satisfies

$$\frac{g_\ell(z) - g_\ell(1/2)}{z - 1/2} \geq \phi \tag{2}$$

for some $\phi > 0$, and all $z \in [0, 1/2) \cup (1/2, 1]$. Thus, in the simplest case where we have just a single cell, P and Q are connected via the fact that the propensity under the source distribution to have a Class 1 label at $x \in \mathbb{R}^d$ only depends on x through $\eta_Q(x)$, which reflects the propensity under the target distribution to have a Class 1 label at x. Our condition (2) is of course satisfied if each g_ℓ is differentiable with $g_\ell'(z) \geq \phi$ for $z \in [0, 1]$, and ensures in particular that when $\eta_P = g_\ell \circ \eta_Q$ holds exactly on $\mathcal{X}_\ell^*$, we have $\mathrm{sgn}\{\eta_P(x) - g_\ell(1/2)\} = \mathrm{sgn}\{\eta_Q(x) - 1/2\}$ on that cell. Importantly, though, condition (2) does not require that $g_\ell(1/2) = 1/2$.

To give an example where such a relationship between P and Q might be expected, suppose that we wish to predict At Risk individuals for a disease (e.g., breast cancer), on the basis of a set of covariates x. Due to the difficulties and expense of large-scale testing, only a small number of individuals in the general population are assessed (e.g., via a mammogram), but those displaying symptoms have a much greater propensity to be tested. In this example, we think of our (large) data set from P as being a set of individuals for whom we have recorded relevant covariates, and for whom we record a label $Y^P = 1$ if and only if the individual has both been tested, and has been assessed to be At Risk as a result. On the other hand, our main interest is in whether individuals are At Risk, regardless of whether or not they have been tested. Our (small) data set from Q then is obtained by testing a number of uniformly randomly-chosen individuals from the general population, and we record $Y^Q = 1$ if and only if the individual is assessed to be At Risk. We can think of our training data from both P and Q as being generated from independent and identically distributed triples (T, X, Y), where T is a binary indicator of whether or not a test has been conducted before the start of the study, where X encodes covariates, and where Y indicates whether or not an individual is At Risk. However, in our source sample, we only observe $X^P = X$ and $Y^P = TY$, while in our target sample, we see $X^Q = X$ and $Y^Q = Y$. Thus, in this example, the marginal distributions of X^P and X^Q are the same, while the regression functions η_P and η_Q satisfy $\eta_Q \geq \eta_P$. In fact, in this formulation, we have

$$\eta_P(x) = \mathbb{P}(TY = 1 | X = x) = \mathbb{P}(T = 1 | X = x, Y = 1)\eta_Q(x).$$

The relationship $\eta_P = g \circ \eta_Q$ then holds if T and (X, Y) are conditionally independent given $\eta_Q(X)$. More generally in this example, we might construct a decision tree partition based on geographical location and income, for instance, and ask only that this relationship hold approximately for each cell of the partition.

As another example, in tax fraud detection, most individuals can only be subjected to a simple screening procedure due to the administrative burden. Hence, in order to assess the reliability of their detection algorithms, a government agency might draw a separate, smaller sample of individuals, chosen uniformly at random from the population, for a more formal audit. Here, $Y^P = 1$ if the screening flags a potentially fraudulent return, $Y^Q = 1$ if the audit detects fraud, and $X^P = X^Q = X$ encodes covariates. Since

$$\eta_P(x) = \mathbb{P}(Y^P = 1 | X = x, Y^Q = 1)\eta_Q(x) + \mathbb{P}(Y^P = 1 | X = x, Y^Q = 0)\{1 - \eta_Q(x)\},$$

the modelling assumption $\eta_P = g \circ \eta_Q$ holds if the conditional probabilities above only depend on x through $\eta_Q(x)$. In practice, there may be additional dependencies, for example, based on profession, income bracket and domicile status, but the modelling relationship may still hold approximately on the cells of a suitable decision tree partition. Further examples may be found in computer vision, precision medicine, natural language processing and many other areas.

In line with the above examples, then we will assume a transfer learning setting where we have access to independent data $\mathcal{D}_P := ((X_1^P, Y_1^P), \dots, (X_{n_P}^P, Y_{n_P}^P))$ from P and $\mathcal{D}_Q := ((X_1^Q, Y_1^Q), \dots, (X_{n_Q}^Q, Y_{n_Q}^Q))$ from Q, and wish to classify a new observation $(X^Q, Y^Q) \sim Q$. Our first contribution is to formalise the new, decision tree-based transfer framework to incorporate the broad range of relationships between source and target distributions seen in practical applications such as those mentioned above. In particular, in contrast to most other work in this area, our highly flexible form of relationship between η_P and η_Q does not require that the Bayes decision boundaries agree for the two populations; we also allow the marginal distributions of X^P and X^Q to differ, and do not assume that these distributions have densities that are bounded away from zero on their respective supports. The classes of distributions we consider then combine local smoothness assumptions on η_P and η_Q with tail assumptions on the marginal distribution of X^Q and the marginal distribution of X^P. To understand the fundamental difficulty of the transfer learning problem, we derive a minimax lower bound that comprises several regimes, according to the relative sample sizes and the strength of the transfer relationship, as measured by the distributional parameters of our classes. The next challenge is to introduce a new method for the transfer learning task; our basic idea is to use $\mathcal{D}_P$ to construct a local nearest-neighbour based estimate of η_P, and then perform empirical risk minimisation with $\mathcal{D}_Q$ to estimate the underlying decision tree partition and the values of the transfer functions at $1/2$. We derive a high-probability upper bound for the excess test error of our procedure, which, together with our lower bound, reveals that our algorithm attains the minimax optimal rate, up to a poly-logarithmic factor. A notable feature of our methodology is that the only inputs required are $\mathcal{D}_P$ and $\mathcal{D}_Q$; in particular, it is adaptive to the unknown transfer relationship in the primary regime of interest, as well as the smoothness and tail parameters of our distributional classes, and the confidence with which the test error bound holds.

Interest in transfer learning has been growing considerably in recent years. One broad line of work considers the setting where the practitioner only has access to labelled data from P, possibly with some additional unlabelled data from Q. A popular approach in that context is to formulate a measure of discrepancy between the distributions P and Q and to give test error bounds in terms of this discrepancy (Ben-David et al. (2010a, 2010b), Germain et al. (2015), Mansour, Mohri and Rostamizadeh (2009), Mohri and Muñoz Medina (2012), Cortes, Mohri and Muñoz Medina (2019)). This strategy has been shown to yield distribution-free bounds with wide applicability, but whenever the discrepancy is nonzero, the excess error is not guaranteed to converge to zero with the sample size. In order to achieve consistent classification, we must impose additional structure (Ben-David et al. (2010b)), for example, by focusing on label shift, covariate shift or label noise, each of which may be viewed as a special case of transfer learning. In label shift (Lipton, Wang and Smola (2018), Zhang et al. (2013)), the marginal distributions of Y^P and Y^Q differ, but the class-conditional covariate distributions are the same for P and Q. Covariate shift (Candela et al. (2009), Gretton et al. (2009), Sugiyama, Suzuki and Kanamori (2012)) concerns scenarios where the regression functions η_P and η_Q are assumed to be equal, but the marginal distributions of P and Q may differ. In label noise (Blanchard et al. (2017), Reeve and Kabán (2019a), Scott (2019), Scott and Zhang (2019)), η_P and η_Q differ. This does not necessarily preclude consistent classification, even when $n_Q = 0$, provided that additional restrictions are met. For instance, the Bayes classifier may still be the same for P and Q, in which case one can sometimes proceed as if there were no label noise (Cannings, Fan and Samworth (2020), Menon, van Rooyen and Natarajan (2018)); alternatively, if the label noise only depends on the true class label, then the label noise parameters may be estimated under certain identifiability assumptions (Blanchard et al. (2016), Reeve and Kabán (2019b)).

Other related work that considers the current setting where the statistician has access to labelled data from both source and target distributions includes Kpotufe and Martinet (2018) for the covariate shift problem, and Hanneke and Kpotufe (2019) and Cai and Wei (2021) for general transfer learning. The frameworks of these papers ensure that $(\eta_P(x) - 1/2)(\eta_Q(x) - 1/2) > 0$ whenever $\eta_Q(x) \neq 1/2$, and hence the Bayes classifiers for P and Q are equal. In our terminology, this corresponds to the special case where $g_\ell(1/2) = 1/2$ for all ℓ. In each of these works, the authors obtain minimax rates of convergence for the excess error in their respective problems, which in particular reveal that consistent classification is possible with $\mathcal{D}_P$ alone, and the effect of $\mathcal{D}_Q$ is to improve the rates. The only work in this context of which we are aware that allows the Bayes classifier for the two distributions to differ is the very recent contribution of Maity, Sun and Banerjee (2020). These authors consider the label shift problem, so the differences between P and Q are captured through a single parameter governing the similarity of $\mathbb{P}(Y^P = 1)$ and $\mathbb{P}(Y^Q = 1)$. Maity, Sun and Banerjee (2020) show how this parameter can be efficiently estimated from the data (which can even be unlabelled), and are therefore also able to obtain minimax rates of convergence for their problem.

One of our main goals in this work is to allow more flexible forms of transfer, to make our framework applicable to the examples discussed above. The price we pay for this generality is that our rates of convergence are necessarily slower than those of Kpotufe and Martinet (2018), Cai and Wei (2021), Hanneke and Kpotufe (2019) and Maity, Sun and Banerjee (2020). Nonetheless, our minimax rates conclusively demonstrate the benefits of transfer learning in a highly flexible setting.

The remainder of this paper is organised as follows: in Section 2, we introduce our general transfer learning framework, and state our main minimax optimality result (Theorem 1). Section 3 gives a formal description of our algorithm, as well as a high-probability upper bound for its excess test error (Theorem 2), while a conclusion is provided in Section 4. The proofs of Theorem 2 and the upper bound in Theorem 1 are given in Section 5, and the proof of the lower bound in Theorem 1 is provided in Section 6. Auxiliary results and illustrative examples are deferred to the online Supplementary Material (Reeve, Cannings and Samworth (2021)) and are prefaced by the letter 'S'; there we also present the results of a brief simulation study.

We conclude this Introduction with some notation used throughout the paper. Given a set A, we write $|A|$ for its cardinality, and $\mathrm{Par}(A)$ for the set of all finite partitions of A, that is, the set consisting of elements of the form $\{A_1, \dots, A_m\}$, with $A_1, \dots, A_m$ pairwise disjoint and $\bigcup_{\ell=1}^m A_\ell = A$. We let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, and for $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$. For $x \in \mathbb{R}^d$, we write $\|x\|$ for the Euclidean norm of x, and, given $r > 0$, we write $B_r(x) := \{y \in \mathbb{R}^d : \|y - x\| < r\}$ for the open Euclidean ball of radius r about x. We let $\mathcal{L}_d$ denote Lebesgue measure on $\mathbb{R}^d$, and let $V_d := \mathcal{L}_d(B_1(0)) = \pi^{d/2}/\Gamma(1 + d/2)$. For $x \geq 0$, we let $\log_+(x) := \log x$ if $x \geq e$, and $\log_+(x) := 1$ otherwise. If μ, ν are probability measures on $(\mathcal{X}, \mathcal{A})$, then we write $\mathrm{TV}(\mu, \nu) := \sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|$ for their total variation distance, and if μ is absolutely continuous with respect to ν with Radon–Nikodym derivative $d\mu/d\nu$, we write $\mathrm{KL}(\mu, \nu) := \int_{\mathcal{X}} \log(\frac{d\mu}{d\nu})\, d\mu$ for the Kullback–Leibler divergence from ν to μ. Finally, the support of a probability measure μ on $\mathbb{R}^d$, denoted $\mathrm{supp}(\mu)$, is defined to be the intersection of all closed sets $C \subseteq \mathbb{R}^d$ with $\mu(C) = 1$.

2. Statistical setting and main result. Let P, Q be distributions on $\mathbb{R}^d \times \{0,1\}$ and let $(X^P, Y^P) \sim P$ and $(X^Q, Y^Q) \sim Q$. We recall the definitions of the regression functions η_P and η_Q from (1), and write μ_P and μ_Q for the marginal distributions of X^P and X^Q, respectively.

A *classifier* is a Borel measurable function $f : \mathbb{R}^d \to \{0,1\}$. In practice, classifiers are constructed on the basis of training data, and we will assume that for some $n_P, n_Q \in \mathbb{N}_0$, we have access to independent pairs $(X_1^P, Y_1^P), \ldots, (X_{n_P}^P, Y_{n_P}^P) \sim P$ and $(X_1^Q, Y_1^Q), \ldots, (X_{n_Q}^Q, Y_{n_Q}^Q) \sim Q$. Recall that as shorthand, we denote $\mathcal{D}_P = ((X_1^P, Y_1^P), \ldots, (X_{n_P}^P, Y_{n_P}^P))$ and $\mathcal{D}_Q = ((X_1^Q, Y_1^Q), \ldots, (X_{n_Q}^Q, Y_{n_Q}^Q))$. A *data-dependent classifier* $\hat{f}$ is a measurable function from $(\mathbb{R}^d \times \{0,1\})^{n_P} \times (\mathbb{R}^d \times \{0,1\})^{n_Q} \times \mathbb{R}^d$ to $\{0,1\}$, and we let $\hat{\mathcal{F}}_{n_P,n_Q}$ denote the set of all such data-dependent classifiers. In this work, the first arguments of $\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}$ will always be $\mathcal{D}_P$ and $\mathcal{D}_Q$, so we will often suppress all but the final argument of $\hat{f}$, noting also that the mapping $x \mapsto \hat{f}(x)$ is a classifier. Conversely, any classifier may be regarded as a data-dependent classifier that is constant in all but its final argument. The *test error* of $\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}$ is given by

$$\mathcal{R}(\hat{f}) := \mathbb{P}(\hat{f}(X^Q) \neq Y^Q | \mathcal{D}_P, \mathcal{D}_Q), \tag{3}$$

where $(X^Q, Y^Q) \sim Q$ is independent of our training data, and is minimised for every $\mathcal{D}_P$ and $\mathcal{D}_Q$ by the *Bayes classifier* f_Q^*, where $f_Q^*(x) := \mathbb{1}_{\{\eta_Q(x) \geq 1/2\}}$. The *excess test error* of $\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}$ is given by

$$\mathcal{E}(\hat{f}) := \mathcal{R}(\hat{f}) - \mathcal{R}(f_Q^*) = \int_{\{x : \hat{f}(x) \neq f_Q^*(x)\}} |2\eta_Q(x) - 1| \, d\mu_Q(x). \tag{4}$$

In order to provide a formal statement of our key transfer assumption, we first define the notion of a decision tree partition.

DEFINITION 1 (Decision tree partitions). Let $\mathbb{T}_1 := \{\{\mathbb{R}^d\}\} \subseteq \mathrm{Par}(\mathbb{R}^d)$, and for $L \geq 2$, define the subset of $\mathrm{Par}(\mathbb{R}^d)$ given by

$$\mathbb{T}_L := \bigl\{\{\mathcal{X}_1, \ldots, \mathcal{X}_{L-1} \cap H_{j,s}, \mathcal{X}_{L-1} \setminus H_{j,s}\} : j \in [d], s \in \mathbb{R}, \{\mathcal{X}_1, \ldots, \mathcal{X}_{L-1}\} \in \mathbb{T}_{L-1}\bigr\},$$

where $H_{j,s} := \{(x_t)_{t \in [d]} \in \mathbb{R}^d : x_j \geq s\}$ for $j \in [d]$ and $s \in \mathbb{R}$. The set of all decision tree partitions is $\bigcup_{L \in \mathbb{N}} \mathbb{T}_L$.

We illustrate some elements of $\mathbb{T}_1, \mathbb{T}_2, \mathbb{T}_3$ and $\mathbb{T}_4$ when $d = 2$ in Figure 1.

ASSUMPTION 1 (Transfer). There exist $\{\mathcal{X}_1^*, \ldots, \mathcal{X}_{L^*}^*\} \in \mathbb{T}_{L^*}$, as well as $\Delta \in [0,1)$, $\phi \in (0,1)$ and transfer functions $g_1, \ldots, g_{L^*} : [0,1] \to [0,1]$ such that $|\eta_P(x) - g_\ell(\eta_Q(x))| \leq \Delta$ for every $\ell \in [L^*]$ and $x \in \mathcal{X}_\ell^*$; moreover,

$$\frac{g_\ell(z) - g_\ell(1/2)}{z - 1/2} \geq \phi \tag{5}$$

for every $\ell \in [L^*]$ and $z \in [0, 1/2) \cup (1/2, 1]$.

To understand this assumption, first consider the case where $\Delta = 0$. Then our condition states that for the $\mathcal{X}_\ell^*$ cell of our decision tree partition, we have the relationship $\eta_P = g_\ell \circ \eta_Q$, so that within this cell, $\eta_P(x)$ only depends on x through $\eta_Q(x)$. Moreover, (5) asks that each g_ℓ is strictly increasing at 1/2, and is of course satisfied if each g_ℓ is differentiable with $g_\ell'(z) \geq \phi$ for $z \in [0,1]$. More generally, for $\Delta > 0$, Assumption 1 only requires that

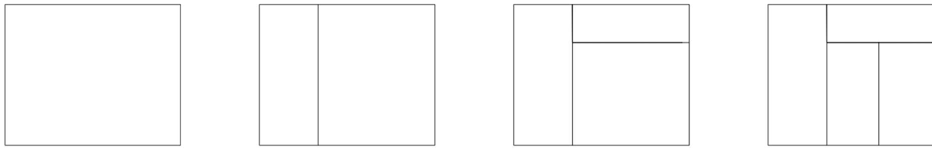

FIG. 1. *Illustration of elements of* $\mathbb{T}_1, \mathbb{T}_2, \mathbb{T}_3$ *and* $\mathbb{T}_4$.

the relationship $\eta_P = g_\ell \circ \eta_Q$ holds to within an error of Δ on each cell of our decision tree partition.

Our next assumption concerns the mass of the source and target distributions in the tails. Given a probability distribution μ on $\mathbb{R}^d$ and $d_0 \in [0, d]$, we define the *lower density* $\omega_{\mu,d_0} : \mathbb{R}^d \to [0, 1]$ of μ by

$$\omega_{\mu,d_0}(x) := \inf_{r \in (0,1)} \frac{\mu(B_r(x))}{r^{d_0}}. \tag{6}$$

For intuition, when μ is absolutely continuous with respect to the volume form on a d_0-dimensional, orientable manifold in $\mathbb{R}^d$ with density (Radon–Nikodym derivative) f_μ, and if the infimum in (6) is replaced with a $\liminf$ as $r \searrow 0$, then ω_{μ,d_0} is almost everywhere equal to a constant multiple of f_μ (Ledrappier and Young (1985), Lemma 4.1.2). This explains our lower density terminology. Further properties of this lower density, which can in fact be defined on general separable metric spaces, can be inferred from common assumptions in the classification literature, including an assumption of *regular support* (Audibert and Tsybakov (2007)) and a *strong minimal mass assumption* (Gadat, Klein and Marteau (2016)); see Lemmas S1 and S2 for details. We also note that the definition in (6) has some similarities with that of a Hardy–Littlewood operator (Hardy and Littlewood (1930)), though one important difference with the standard definition is that here an infimum replaces a supremum.

ASSUMPTION 2 (Marginals). There exist $d_Q \in [1, d]$, $\gamma_Q > 0$ and $C_{P,Q} > 1$ such that

$$\mu_Q\big(\{x \in \mathbb{R}^d : \omega_{\mu_Q,d_Q}(x) < \xi\}\big) \le C_{P,Q} \cdot \xi^{\gamma_Q} \tag{7}$$

for all $\xi > 0$. Moreover, there exist $d_P \in [d_Q, d]$ and $\gamma_P > 0$ such that

$$\mu_Q\big(\{x \in \mathbb{R}^d : \omega_{\mu_P,d_P}(x) < \xi\}\big) \le C_{P,Q} \cdot \xi^{\gamma_P} \tag{8}$$

for all $\xi > 0$.

To understand the first part of Assumption 2, first consider the case where μ_Q is absolutely continuous with respect to $\mathcal{L}_d$. In that case, condition (7) can be viewed as similar to other tail conditions in the classification literature that control the μ_Q measure of the set on which this density is small (e.g., Gadat, Klein and Marteau (2016), Assumption A4). Thus, (7) is a generalisation of such a tail condition, because we do not require μ_Q to be absolutely continuous with respect to $\mathcal{L}_d$, and instead work with its lower density ω_{μ_Q,d_Q}. The great advantage of this formulation in (7) is that it allows us to avoid assuming that this lower density is bounded away from zero on the support of μ_Q; Example S2 provides a simple, univariate parametric family of densities $\{f_\gamma : \gamma > 0\}$ for which $\gamma_Q = \gamma$ is the optimal choice.

Further intuition about the first part of Assumption 2 can be gained from several results in the Supplementary Material that we now summarise. In Lemma S5, we show that if μ_Q has a finite ρth moment for some $\rho > 0$, then (7) holds with $d_Q = d$ and $\gamma_Q = \rho/(\rho + d)$. The proof relies on Vitali's covering lemma (e.g., Evans and Gariepy (2015), Theorem 1), and we believe the result may find application elsewhere; see Remark S1 after Lemma S5. As a consequence of a general result about Weibull-type tails (Lemma S6), Proposition S12 shows that when μ_Q has a log-concave density on $\mathbb{R}^d$ with d_0-dimensional support, (7) holds with $d_Q = d_0$ and any $\gamma_Q < 1$; in fact, when $d_Q = 1$, we may even take $\gamma_Q = 1$ (Proposition S11). Moreover, Proposition S13 extends these results to finite mixtures of log-concave distributions, with $C_{P,Q}$ depending linearly on the number of mixture components (and not depending on the mixing proportions). In fact, more generally, Propositions S7 and S8 provide simple stability results for the property (7) under finite mixtures and products, respectively. As additional important examples, whenever μ_Q has bounded, d_0-dimensional support, we may take $d_Q = d_0$ and $\gamma_Q = 1$ (by Lemma S4); moreover, if μ_Q has a density that is bounded away from zero on a d_0-dimensional, regular support, then we may take $d_Q = d_0$ and γ_Q to be arbitrarily large (by Lemma S1).

The second part of Assumption 2 relates μ_P and μ_Q together: it controls the μ_Q measure of the set on which the lower density of μ_P is small, thereby capturing the extent to which the source measure covers the target measure. For instance, Example S4 reveals that when $\mu_Q = N(0, 1)$ and $\mu_P = N(0, \sigma^2)$, we may take $\gamma_P = \sigma^2$, while from Example S5, we see that when $\mu_Q = N(0, 1)$ and $\mu_P = N(a, 1)$ for some $a \neq 0$, we may take any $\gamma_P < 1$. We remark that if (7) holds and if, in the terminology of Kpotufe and Martinet (2018), (P, Q) have *transfer-exponent* $\kappa \in [0, \infty]$, then (8) holds for any $d_P \geq d_Q + \kappa$ and with $\gamma_P = \gamma_Q$; see Lemma S14. Moreover, Example S6 provides a prototypical setting where working with the condition (8) allows us to obtain faster rates of convergence than would be the case if we instead deduced this rate from the corresponding transfer-exponent.

Our next two assumptions are standard margin (e.g., Mammen and Tsybakov (1999), Polonik (1995)) and smoothness assumptions. We emphasise that these are only imposed on the distribution Q, and we require no corresponding properties for P.

ASSUMPTION 3 (Margin). There exist $\alpha > 0$ and $C_{\mathrm{M}} \geq 1$ such that for all $\zeta > 0$ we have $\mu_Q(\{x \in \mathbb{R}^d : |\eta_Q(x) - 1/2| < \zeta\}) \leq C_{\mathrm{M}} \cdot \zeta^\alpha$.

ASSUMPTION 4 (Smoothness). There exist $\beta \in (0, 1]$ and $C_{\mathrm{S}} \geq 1$ such that $|\eta_Q(x_0) - \eta_Q(x_1)| \leq C_{\mathrm{S}} \cdot \|x_0 - x_1\|^\beta$ for all $x_0, x_1 \in \mathbb{R}^d$.

It will be convenient to write θ for the vector of parameters that appear in Assumptions 2–4, namely $(d_Q, \gamma_Q, d_P, \gamma_P, C_{P,Q}, \alpha, C_{\mathrm{M}}, \beta, C_{\mathrm{S}})$, and to write Θ for the corresponding parameter space. We will also make use of an augmented parameter vector that incorporates the additional parameters that appear in Assumption 1, by letting $\theta^\sharp := (\Delta, \phi, L^*, \theta)$, with corresponding parameter space $\Theta^\sharp$. For $\theta^\sharp \in \Theta^\sharp$, we write $\mathcal{P}_{\theta^\sharp}$ for the set of pairs (P, Q) of distributions satisfying Assumptions 1–4 with parameter $\theta^\sharp$.

We are now in a position to state our main result.

THEOREM 1. *Fix* $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$ *with* $\beta/(2\beta + d_Q) < \gamma_Q$, $\beta/(2\beta + d_P) < \gamma_P$, $\alpha\beta \leq d_Q$, $\gamma_P(1 - \gamma_Q) \leq \gamma_Q$ *and* $C_{\mathrm{M}} \geq 1 + 2^{2d_Q/\beta} d_Q^{d_Q/2} V_{d_Q}$. *For* $j \in \{\mathrm{L}, \mathrm{U}\}$, *let*

$$A^j_{n_P,n_Q} := \left(\frac{a^j_0}{\phi^2 \cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\gamma_P(2\beta+d_P)+\alpha\beta}} + \min\left\{\left(\frac{L^* a^j_1}{n_Q}\right)^{\frac{1+\alpha}{2+\alpha}}, (1-\phi)^{1+\alpha}\right\} + \left(\frac{\Delta}{\phi}\right)^{1+\alpha},$$

$$B^j_{n_Q} := \left(\frac{b^j}{n_Q}\right)^{\frac{\beta\gamma_Q(1+\alpha)}{\gamma_Q(2\beta+d_Q)+\alpha\beta}},$$

where $a^{\mathrm{L}}_0 = a^{\mathrm{L}}_1 = b^{\mathrm{L}} := 1$, $a^{\mathrm{U}}_0 := \log_+(n_P)$, $a^{\mathrm{U}}_1 := \log_+(L^* d(n_P + n_Q))$ *and* $b^{\mathrm{U}} := \log_+(n_Q)$. *Then there exist* $c_\theta, C_\theta > 0$, *depending only on* θ, *such that*

$$c_\theta\big(A^{\mathrm{L}}_{n_P,n_Q} \wedge B^{\mathrm{L}}_{n_Q} \wedge 1\big) \leq \inf_{\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}} \sup_{(P,Q) \in \mathcal{P}_{\theta^\sharp}} \mathbb{E}\{\mathcal{E}(\hat{f})\} \leq C_\theta\big(A^{\mathrm{U}}_{n_P,n_Q} \wedge B^{\mathrm{U}}_{n_Q} \wedge 1\big). \tag{9}$$

Theorem 1 establishes the optimal rates of convergence for the excess risk over our classes, up to logarithmic factors. It is important to note that c_θ and C_θ do not depend on (Δ, ϕ, L^*) (and nor on n_P or n_Q); thus the theorem reveals the optimal dependence of the worst-case excess risk on these parameters, too. Moreover, as we will show in Theorem 2, the minimax rate can be achieved up to a poly-logarithmic factor when $\phi \le 1 - n_Q^{-1/(2+\alpha)}$ by a procedure that is completely adaptive, in the sense that it only takes $\mathcal{D}_P$ and $\mathcal{D}_Q$ as inputs (and not any component of $\theta^\sharp$).

The restrictions on the parameters in Theorem 1 are mild. For instance, by Lemma S5, the conditions $\beta/(2\beta + d_Q) < \gamma_Q$ and $\gamma_P(1 - \gamma_Q) \le \gamma_Q$ hold whenever $\sup_{(P,Q)\in\mathcal{P}_{\theta^\sharp}} \mathbb{E}(\|X^Q\|^{1\vee d_Q\gamma_P}) < \infty$. The condition $\alpha\beta \le d_Q$ rules out 'super-fast rates' (in the terminology of Audibert and Tsybakov (2007)) and is guaranteed to hold whenever there exist $(P, Q) \in \mathcal{P}_{\theta^\sharp}$ and $x_0 \in \mathbb{R}^d$ such that $\eta_Q(x_0) = 1/2$ and $\omega_{\mu_Q,d_Q}(x_0) > 0$ (see Lemma S15). The first two parameter restrictions in Theorem 1 are only required for the upper bound, while the other three are only needed for the lower bound. We also remark that Theorem 1 holds even when n_P or n_Q are zero. In the former case, the problem reduces to a standard classification problem, while in the latter case, Theorem 1 provides results for relaxations of the covariate shift model in which η_P and η_Q are close.

By careful inspection of the proof of Theorem 1, we see that the first terms $A^{\mathrm{L}}_{n_P,n_Q}$ and $A^{\mathrm{U}}_{n_P,n_Q}$ in the bounds are due to the transfer learning error, and comprise three separate contributions. The first term arises from the error incurred in estimating η_P. The second represents the difficulty of identifying the correct decision tree partition $\{\mathcal{X}_1^*, \ldots, \mathcal{X}_{L^*}^*\}$, as well as learning $g_1(1/2), \ldots, g_{L^*}(1/2)$; this term is negligible if ϕ is sufficiently close to 1. In fact, as we will see from the proof of Theorem 2 below, it is not necessary to carry out this step when ϕ is close to 1. Finally, the third term reflects the extent to which η_P can be approximated by $g_\ell \circ \eta_Q$ on $\mathcal{X}_\ell^*$. The $B^{\mathrm{L}}_{n_Q}$ and $B^{\mathrm{U}}_{n_Q}$ terms represent the rate of convergence achievable by ignoring $\mathcal{D}_P$ and performing standard classification using $\mathcal{D}_Q$; in the context of transfer learning, our primary interest is in the setting where $n_P \gg n_Q$, and where the minima in (9) are attained by $A^{\mathrm{L}}_{n_P,n_Q}$ and $A^{\mathrm{U}}_{n_P,n_Q}$, respectively.

To set the rates $B^{\mathrm{L}}_{n_Q}$ and $B^{\mathrm{U}}_{n_Q}$ in context, it may be helpful to consider the case where μ_Q is absolutely continuous with respect to $\mathcal{L}_d$ with a density that is bounded away from zero on its regular support; see Definition S1. In that case, we may take γ_Q to be arbitrarily large and $d_Q = d$; notice that setting $\gamma_Q = \infty$ and $d_Q = d$ in $B^{\mathrm{L}}_{n_Q}$ recovers the rate $n_Q^{-\frac{\beta(1+\alpha)}{2\beta+d}}$ for the standard classification problem (with no source data) in Audibert and Tsybakov (2007) under this regular support hypothesis. Returning to the more general transfer learning setting, if we take $\gamma_P = \gamma_Q = \infty$ and $d_P = d_Q = d$, then the minimum of the first term in $A^{\mathrm{L}}_{n_P,n_Q}$ and $B^{\mathrm{L}}_{n_Q}$ matches the rate obtained by Cai and Wei (2021). The second and third terms in $A^{\mathrm{L}}_{n_P,n_Q}$ represent the necessary additional price for the generality of our framework.

To illustrate Theorem 1, and ignoring logarithmic factors for simplicity, consider the special case where $\gamma_P = \gamma_Q$ and $d_P = d_Q$ (which would in particular be the case if the marginal distributions μ_P and μ_Q coincide). Then Theorem 1 reveals that in order for transfer learning to be effective (as opposed to simply constructing a classifier based on $\mathcal{D}_Q$), we require $\phi^2 \cdot n_P \gg n_Q$. If we further assume that $\gamma_P = \gamma_Q = 1$, that $d_P = d_Q = d$, that $\Delta = 0$ and that $\alpha = \beta = 1$, then we benefit from transfer learning provided that $\phi^2 \cdot n_P \gg n_Q$ and $L^* \ll n_Q^{d/(d+3)}$. In general, the scope for transfer learning to have an impact increases as γ_P and ϕ increase, and as d_P, L^* and Δ decrease.

3. Methodology and upper bound. In this section, we introduce our adaptive algorithm for transfer learning and provide a high-probability bound for its excess risk. To understand the main idea, consider the case where $\Delta = 0$ in Assumption 1, and where we are told the correct decision tree partition and transfer functions. In this setting, when $x \in \mathcal{X}_\ell^*$, the sign of $\eta_P(x) - g_\ell(1/2)$ agrees with the sign of $\eta_Q(x) - 1/2$, so we aim to construct a nearest-neighbour based estimate of the former quantity using $\mathcal{D}_P$. In practice, this estimate will depend on a choice of decision tree, but this can be calibrated using a subsample from $\mathcal{D}_Q$. Separately, we also construct a standard k nearest-neighbour estimate of $\eta_Q(x) - 1/2$ via the same subsample from $\mathcal{D}_Q$, and make our final choice between the two data-dependent classifiers using empirical risk minimisation over the held-out data from $\mathcal{D}_Q$. The independence of the two subsamples from $\mathcal{D}_Q$ allows us to work conditionally on the first subsample at this final step to obtain our final performance guarantees.

In giving a formal description of our algorithm, we will assume that $n_Q \geq 2$ (when $n_Q \leq 1$, the upper bound in Theorem 1 is attained by applying a nearest-neighbour method to $\mathcal{D}_P$), and it will also be convenient initially to assume that $n_P > 0$. For $x \in \mathbb{R}^d$ and $k \in [n_P]$, we let $X_{(k)}^P \equiv X_{(k)}^P(x)$ denote the kth nearest neighbour of x in $\mathcal{D}_P$ in Euclidean norm (where for definiteness, in the case of ties, we preserve the original ordering of the indices), and let $Y_{(k)}^P \equiv Y_{(k)}^P(x)$ denote the concomitant label. We then split $\mathcal{D}_Q$ into two subsamples $\mathcal{D}_Q^0 := ((X_1^Q, Y_1^Q), \dots, (X_{\lfloor n_Q/2 \rfloor}^Q, Y_{\lfloor n_Q/2 \rfloor}^Q))$ and $\mathcal{D}_Q^1 := ((X_{\lfloor n_Q/2 \rfloor+1}^Q, Y_{\lfloor n_Q/2 \rfloor+1}^Q), \dots, (X_{n_Q}^Q, Y_{n_Q}^Q))$. For $k \in [\lfloor n_Q/2 \rfloor]$, we let $X_{(k)}^Q \equiv X_{(k)}^Q(x)$ denote the kth nearest neighbour of x in $\mathcal{D}_Q^0$, and similarly let $Y_{(k)}^Q \equiv Y_{(k)}^Q(x)$ denote the concomitant label.

Given $L \in \mathbb{N}$ and a decision tree partition $\{\mathcal{X}_1, \dots, \mathcal{X}_L\} \in \mathbb{T}_L$, we define the *leaf function* $\ell : \mathbb{R}^d \to [L]$ by $\ell(x) := j$ whenever $x \in \mathcal{X}_j$. Let $\mathcal{H}_L$ denote the set of decision tree functions $h : \mathbb{R}^d \to (0, 1)$ of the form $x \mapsto \tau_{\ell(x)}$ for some $\{\mathcal{X}_1, \dots, \mathcal{X}_L\} \in \mathbb{T}_L$ with leaf function ℓ, and some $(\tau_1, \dots, \tau_L) \in \{0, 1/n_P, 2/n_P, \dots, 1\}^L$. It is also convenient to define $\mathcal{H}_0$ to consist of the single (constant) function that maps $\mathbb{R}^d$ to $1/2$ (this will handle the case when ϕ is very close to 1). Given $k \in [n_P]$, $L \in \mathbb{N}_0$ and $h \in \mathcal{H}_L$, we let

$$\hat{m}_{k,h}^P(x) := \frac{1}{k} \sum_{i=1}^{k} \{Y_{(i)}^P(x) - h(X_{(i)}^P(x))\} \tag{10}$$

denote an empirical estimate of $\eta_P(x) - g_{\ell(x)}(1/2)$. To choose k, we fix a robustness parameter $\sigma \in [n_P^2]/n_P = \{1/n_P, 2/n_P, \dots, n_P\}$, and use a Lepski-type procedure to define

$$\hat{k} \equiv \hat{k}_{\sigma,h}^P(x) := \max\left\{k \in [n_P - 1] : |\hat{m}_{r,h}^P(x)| \leq \frac{\sigma}{\sqrt{r}} \text{ for all } r \in [k]\right\} + 1. \tag{11}$$

Fixing a confidence level $\delta \in (0, 1)$, we will see in Proposition 3 that the choice $\sigma^* = \min\{\lceil 3 \log_+^{1/2}(n_P/\delta) \rceil, n_P\}$ yields classifiers that perform well with probability at least $1 - \delta$. However, we seek a procedure with simultaneous guarantees across all levels δ, so we will provide a data-dependent choice below. We now choose h by applying empirical risk minimisation over $\mathcal{D}_Q^0$, so that

$$\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}_L} \sum_{i=1}^{\lfloor n_Q/2 \rfloor} \{Y_i^Q \mathbb{1}_{\{\hat{m}_{\hat{k},h}^P(X_i^Q) < 0\}} + (1 - Y_i^Q) \mathbb{1}_{\{\hat{m}_{\hat{k},h}^P(X_i^Q) \geq 0\}}\}. \tag{12}$$

As defined, $\hat{h}$ involves a minimisation over an infinite set of decision tree functions; however, by Lemma 13, a minimiser can be found by restricting the class $\mathcal{H}_L$ to a finite set that may in principle be computed from the data. See Section S2 for a discussion of implementational aspects. Having determined $\hat{h}$, we can now define a family $\hat{\mathcal{F}}^P := \{\hat{f}_{\sigma,L}^P : \sigma \in [n_P^2]/n_P, L \in \{0\} \cup [n_Q]\} \subseteq \hat{\mathcal{F}}_{n_P,n_Q}$, where $\hat{f}_{\sigma,L}^P(x) := \mathbb{1}_{\{\hat{m}_{\hat{k},\hat{h}}^P(x) \geq 0\}}$. If $n_P = 0$, then we set $\hat{\mathcal{F}}^P := \varnothing$.

The second part of our procedure involves applying a k nearest-neighbour classifier to $\mathcal{D}_Q^0$. More precisely, for $k \in [\lfloor n_Q/2 \rfloor]$, we first define

$$\hat{m}_k^Q(x) := \frac{1}{k}\sum_{i=1}^{k}\left\{Y_{(i)}^Q(x) - \frac{1}{2}\right\}. \tag{13}$$

Given $\sigma \in [n_Q^2]/n_Q$, we select a number of neighbours

$$\tilde{k} \equiv \tilde{k}_\sigma^Q(x) := \max\left\{k \in [\lfloor n_Q/2 \rfloor - 1] : |\hat{m}_r^Q(x)| \le \frac{\sigma}{\sqrt{r}} \text{ for all } r \in [k]\right\} + 1, \tag{14}$$

and define another family $\hat{\mathcal{F}}^Q := \{\hat{f}_\sigma^Q : \sigma \in [n_Q^2]/n_Q\} \subseteq \hat{\mathcal{F}}_{n_P,n_Q}$ by $\hat{f}_\sigma^Q(x) := \mathbb{1}_{\{\hat{m}_{\tilde{k}}^Q(x) \ge 0\}}$.

Our final data-dependent classifier then is obtained by empirical risk minimisation over $\mathcal{D}_Q^1$: we pick

$$\hat{f}_{\mathrm{ATL}} \in \operatorname*{argmin}_{f \in \hat{\mathcal{F}}^P \cup \hat{\mathcal{F}}^Q} \sum_{i=\lfloor n_Q/2 \rfloor + 1}^{n_Q} \mathbb{1}_{\{f(X_i^Q) \neq Y_i^Q\}}.$$

The following theorem provides a high-probability bound on the performance of $\hat{f}_{\mathrm{ATL}}$ over $\mathcal{P}_{\theta^\sharp}$:

THEOREM 2. *Fix $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$ with $\beta/(2\beta + d_P) < \gamma_P$ and $\beta/(2\beta + d_Q) < \gamma_Q$. Given $n_P \in \mathbb{N}_0$, $n_Q \ge 2$ and $\delta \in (0, 1)$, we let*

$$A_{n_P,n_Q,\delta} := \left(\frac{a_{0,\delta}}{\phi^2 \cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\gamma_P(2\beta+d_P)+\alpha\beta}} + \min\left\{\left(\frac{L^* a_{1,\delta}}{n_Q}\right)^{\frac{1+\alpha}{2+\alpha}}, (1-\phi)^{1+\alpha}\right\} + \left(\frac{\Delta}{\phi}\right)^{1+\alpha},$$

$$B_{n_Q,\delta} := \left(\frac{b_\delta}{n_Q}\right)^{\frac{\beta\gamma_Q(1+\alpha)}{\gamma_Q(2\beta+d_Q)+\alpha\beta}}, \qquad D_{n_P,n_Q,\delta} := \left(\frac{d_\delta}{n_Q}\right)^{\frac{1+\alpha}{2+\alpha}},$$

where $a_{0,\delta} := \log_+(n_P/\delta)$, $a_{1,\delta} := \log_+(L^* d n_P/\delta)$, $b_\delta := \log_+(n_Q/\delta)$ and $d_\delta := \log_+((n_P + n_Q)/\delta)$. Then there exists $C_\theta > 0$, depending only on θ, such that

$$\sup_{(P,Q) \in \mathcal{P}_{\theta^\sharp}} \mathbb{P}\{\mathcal{E}(\hat{f}_{\mathrm{ATL}}) > C_\theta \cdot (\min(A_{n_P,n_Q,\delta}, B_{n_Q,\delta}) + D_{n_P,n_Q,\delta})\} \le \delta. \tag{15}$$

An important point to note is that the definition of $\hat{f}_{\mathrm{ATL}}$ does not depend on the confidence level δ, yet the probabilistic guarantee in (15) holds simultaneously over all such levels. The terms $A_{n_P,n_Q,\delta}$ and $B_{n_Q,\delta}$ are very closely related to A_{n_P,n_Q}^{U} and $B_{n_Q}^{\mathrm{U}}$ in Theorem 1; indeed, the only changes are in the logarithmic factor. Integrating the tail probability bound (15) over $\delta \in (0, 1)$ therefore reveals that in the primary regimes of interest, the upper bound in Theorem 1 can be attained using an algorithm that is agnostic to Δ, ϕ and L^*. Comparing Theorem 2 with the upper bound in Theorem 1, we see that there is an additional term $D_{n_P,n_Q,\delta}$. This term only contributes when ϕ is extremely close to 1 (i.e., when $1 - \phi \ll n_Q^{-1/(2+\alpha)}$ up to a logarithmic factor) and $A_{n_P,n_Q,\delta} \ll B_{n_Q,\delta}$. In this case, η_P is very close to η_Q, and the upper bound in Theorem 1 can be attained by applying a standard nearest- neighbour method to $\mathcal{D}_P$.

We recall that the second term $B_{n_Q,\delta}$ in (15) arises from ignoring $\mathcal{D}_P$ and performing classification using $\mathcal{D}_Q$. Our analysis here builds on prior work on error rates in k nearest-neighbour classification (e.g., Kulkarni and Posner (1995), Hall, Park and Samworth (2008), Samworth (2012), Chaudhuri and Dasgupta (2014), Biau and Devroye (2015), Gadat, Klein and Marteau (2016), Reeve and Brown (2017), Cannings, Berrett and Samworth (2020)); see also the seminal early work by Fix and Hodges (1951), Cover and Hart (1967) and Stone (1977). The main novelty in our arguments, however, is in obtaining the $A_{n_P,n_Q,\delta}$ term, which quantifies the extent to which our algorithm can exploit $\mathcal{D}_P$ to classify data from Q more accurately than can be done with $\mathcal{D}_Q$ alone. Here, we combine analyses of nearest-neighbour classification (but using $\mathcal{D}_P$ instead of $\mathcal{D}_Q$) with a covering number argument for the number of possible decision trees on $\mathcal{D}_P$ (Biau and Devroye (2013), Scott and Nowak (2006), Wager and Walther (2015)), allowing for an approximation error.

4. Conclusion. In this paper, we have argued that transfer learning has great potential for practitioners in the modern data-rich era. Frequently, there is an abundance of data that, while not arising from the target population, are still able to provide useful information about inferential questions of interest. We have introduced a general framework to study this phenomenon in the context of binary classification, and have derived the optimal rates of convergence in this setting. Moreover, we have shown that these optimal rates are attainable by a fully adaptive algorithm that takes only our source and target data as inputs.

The scope of transfer learning is very wide indeed, encompassing not only other forms of transfer relationship and data acquisition mechanisms, but also alternative learning tasks such as regression, density estimation and clustering. We therefore look forward to future developments in this field.

5. Proofs of Theorem 2 and upper bound in Theorem 1. The proof of Theorem 2 is split into two subsections: the first controls the contribution to the excess test error of a data-dependent classifier calibrated via a given decision tree, while the second handles the additional error incurred in choosing the decision tree and other tuning parameters via empirical risk minimisation. Both subsections require several intermediate results.

5.1. *Excess test error of decision tree-calibrated nearest-neighbour classifiers*. We introduce some additional terminology. Given $\sigma > 0$, $L \in \mathbb{N}_0$ and $h \in \mathcal{H}_L$, define $\hat{f}^P_{\sigma,h} \in \hat{\mathcal{F}}_{n_P,n_Q}$ by

$$\hat{f}^P_{\sigma,h}(x) := \mathbb{1}_{\{\hat{m}^P_{\hat{k},h}(x) \geq 0\}}, \tag{16}$$

where $\hat{m}^P_{k,h}(\cdot)$ and $\hat{k} \equiv \hat{k}^P_{\sigma,h}(\cdot)$ are defined in (10) and (11), respectively. Note that $\hat{f}^P_{\sigma,h}(x)$ is measurable with respect to the sigma algebra generated by $\mathcal{D}_P$, for every $x \in \mathbb{R}^d$. Proposition 3 below is the main result of this subsection, and provides a high-probability bound for the excess test error of $\hat{f}^P_{\sigma,h}$ for a particular choice of σ and a general decision tree h. It will be applied three times in the proof of Theorem 2.

PROPOSITION 3. *Let $n_P \in \mathbb{N}$. Fix $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$, where $\theta = (d_Q, \gamma_Q, d_P, \gamma_P, C_{P,Q}, \alpha, C_{\mathrm{M}}, \beta, C_{\mathrm{S}})$, with $\beta/(2\beta + d_P) < \gamma_P$, and $(P, Q) \in \mathcal{P}_{\theta^\sharp}$. For $h \in \mathcal{H}_{L^*} \cup \mathcal{H}_0$, let*

$$\Delta_h := \Delta + \max_{\ell \in [L^*]} \sup_{x \in \mathcal{X}^*_\ell} \bigl|h(x) - g_\ell(1/2)\bigr|.$$

Then there exists $\tilde{C}_\theta > 0$, depending only on θ, such that for every $\delta \in (0, 1)$, if we set $\sigma^* = \min\{\lceil 3 \log_+^{1/2}(n_P/\delta)\rceil, n_P\}$, then

$$\mathbb{P}\biggl[\mathcal{E}(\hat{f}^P_{\sigma^*,h}) > \tilde{C}_\theta \biggl\{\biggl(\frac{\log_+(n_P/\delta)}{\phi^2 \cdot n_P}\biggr)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta+\gamma_P(2\beta+d_P)}} + \biggl(\frac{\Delta_h}{\phi}\biggr)^{1+\alpha}\biggr\}\biggr] \leq \delta. \tag{17}$$

The first term in the probability bound in (17) corresponds to the difficulty of estimating η_P, while, in the second term, Δ_h quantifies the approximation error of the decision tree function h. The proof of Proposition 3 is given after several preliminary lemmas.

For $\delta \in (0,1)$ and $x \in \mathbb{R}^d$ with $\omega_{\mu_P,d_P}(x) > 0$, we define the event

$$E_1^\delta(x) := \bigcap_{\substack{k \in [n_P] \\ 4\log_+(n_P/\delta) \le k < n_P \cdot \omega_{\mu_P,d_P}(x)/2}} \biggl\{ \| X_{(k)}^P(x) - x \| \le \biggl(\frac{2k}{n_P \cdot \omega_{\mu_P,d_P}(x)} \biggr)^{1/d_P} \biggr\}.$$

LEMMA 4. *Let $n_P \in \mathbb{N}$ and $(P,Q) \in \mathcal{P}_{\theta^\sharp}$. For $x \in \mathbb{R}^d$ with $\omega_{\mu_P,d_P}(x) > 0$, we have $\mathbb{P}(E_1^\delta(x)^c) \le \delta$.*

PROOF. Suppose that $k \in [n_P]$ satisfies $4\log_+(n_P/\delta) \le k < n_P \cdot \omega_{\mu_P,d_P}(x)/2$, and let $r \equiv r_k := \{2k/(n_P \cdot \omega_{\mu_P,d_P}(x))\}^{1/d_P}$. Since $r < 1$, we have

$$\mu_P\bigl(B_r(x)\bigr) \ge \omega_{\mu_P,d_P}(x) \cdot r^{d_P} = \frac{2k}{n_P}.$$

Hence, by the multiplicative Chernoff bound (McDiarmid (1998), Theorem 2.3(c)), we have

$$\begin{aligned}
\mathbb{P}\biggl\{ \| X_{(k)}^P(x) - x \| > \biggl(\frac{2k}{n_P \cdot \omega_{\mu_P,d_P}(x)} \biggr)^{1/d_P} \biggr\} &\le \mathbb{P}\biggl\{ \sum_{i=1}^{n_P} \mathbb{1}_{\{X_i^P \in B_r(x)\}} < k \biggr\} \\
&\le \mathbb{P}\biggl\{ \sum_{i=1}^{n_P} \mathbb{1}_{\{X_i^P \in B_r(x)\}} < \frac{n_P}{2} \cdot \mu_P\bigl(B_r(x)\bigr) \biggr\} \\
&\le e^{-n_P \cdot \mu_P(B_r(x))/8} \le e^{-k/4} \le \frac{\delta}{n_P}.
\end{aligned}$$

The conclusion of the lemma now follows by a union bound. □

LEMMA 5. *Let $n_P \in \mathbb{N}$, let $(P,Q) \in \mathcal{P}_{\theta^\sharp}$ and let $x \in \mathbb{R}^d$ be such that $\omega_{\mu_P,d_P}(x) > 0$. On the event $E_1^\delta(x)$, we have that*

$$\max_{i \in [k]} \bigl| \eta_Q\bigl(X_{(i)}^P(x)\bigr) - \eta_Q(x) \bigr| < C_S \cdot \biggl(\frac{2 \cdot \max\{k, \lceil 4\log_+(n_P/\delta) \rceil\}}{n_P \cdot \omega_{\mu_P,d_P}(x)} \biggr)^{\beta/d_P} \tag{18}$$

for all $k \in [n_P]$.

PROOF. First, if $n_P \cdot \omega_{\mu_P,d_P}(x)/2 \le k \le n_P$ or $\lceil 4\log_+(n_P/\delta) \rceil > n_P$, then the result follows from the fact that the right-hand side of (18) is at least 1. Second, if $4\log_+(n_P/\delta) \le k < n_P \cdot \omega_{\mu_P,d_P}(x)/2$, then (18) follows from the definition of $E_1^\delta(x)$ combined with Assumption 4. Finally, if $k < \lceil 4\log_+(n_P/\delta) \rceil \le n_P$, then on $E_1^\delta(x)$,

$$\begin{aligned}
\max_{i \in [k]} \bigl| \eta_Q\bigl(X_{(i)}^P(x)\bigr) - \eta_Q(x) \bigr| &\le \max_{i \in [\min\{\lceil 4\log_+(n_P/\delta) \rceil, n_P\}]} \bigl| \eta_Q\bigl(X_{(i)}^P(x)\bigr) - \eta_Q(x) \bigr| \\
&\le C_S \cdot \biggl(\frac{2 \cdot \lceil 4\log_+(n_P/\delta) \rceil}{n_P \cdot \omega_{\mu_P,d_P}(x)} \biggr)^{\beta/d_P} \\
&= C_S \cdot \biggl(\frac{2 \cdot \max\{k, \lceil 4\log_+(n_P/\delta) \rceil\}}{n_P \cdot \omega_{\mu_P,d_P}(x)} \biggr)^{\beta/d_P},
\end{aligned}$$

where the second inequality follows from the first two cases applied to $\lceil 4\log_+(n_P/\delta) \rceil$. □

For $\delta \in (0,1)$ and $x \in \mathbb{R}^d$, we now define another event

$$E_2^\delta(x) := \bigcap_{k=1}^{n_P} \left\{ \frac{1}{k} \sum_{i=1}^{k} \bigl[Y_{(i)}^P(x) - \eta_P\bigl(X_{(i)}^P(x)\bigr)\bigr] \le \sqrt{\frac{\log_+(n_P/\delta)}{2k}} \right\}.$$

LEMMA 6. *Let $n_P \in \mathbb{N}$ and $(P, Q) \in \mathcal{P}_{\theta^\sharp}$. For every $\delta \in (0,1)$ and $x \in \mathbb{R}^d$, we have $\mathbb{P}(E_2^\delta(x)^c) \le \delta$.*

PROOF. First, note that conditional on $(X_i^P)_{i \in [n_P]}$, the labels $Y_{(1)}^P(x), \dots, Y_{(n_P)}^P(x)$ are independent Bernoulli random variables with respective means $\eta_P(X_{(1)}^P(x)), \dots, \eta_P(X_{(n_P)}^P(x))$. Hence, by Hoeffding's inequality, for each $k \in [n_P]$,

$$\mathbb{P}\left\{ \frac{1}{k} \sum_{i=1}^{k} \bigl[Y_{(i)}^P(x) - \eta_P\bigl(X_{(i)}^P(x)\bigr)\bigr] > \sqrt{\frac{\log_+(n_P/\delta)}{2k}} \,|\, (X_i^P)_{i \in [n_P]} \right\} \le \frac{\delta}{n_P}.$$

The conclusion of the lemma follows by taking expectations over $(X_i^P)_{i \in [n_P]}$, and then a union bound over $k \in [n_P]$. □

LEMMA 7. *Let $n_P \in \mathbb{N}$ and $(P, Q) \in \mathcal{P}_{\theta^\sharp}$. Suppose that $x \in \mathbb{R}^d$ and $h \in \mathcal{H}_{L^*} \cup \mathcal{H}_0$ satisfy*

$$\left| \eta_Q(x) - \frac{1}{2} \right| \ge 50 \cdot C_{\mathrm{S}} \left(\frac{\log_+(n_P/\delta)}{\phi^2 \cdot n_P \cdot \omega_{\mu_P, d_P}(x)} \right)^{\frac{\beta}{2\beta + d_P}} + \frac{2\Delta_h}{\phi},$$

and let $\sigma^* = \min\{\lceil 3 \log_+^{1/2}(n_P/\delta) \rceil, n_P\}$. Then, on the event $E_1^\delta(x) \cap E_2^\delta(x)$, we have $\hat{f}_{\sigma^*, h}^P(x) = f_Q^*(x)$.

PROOF. For the purpose of the proof, we let

$$\epsilon := 50 \cdot C_{\mathrm{S}} \left(\frac{\log_+(n_P/\delta)}{\phi^2 \cdot n_P \cdot \omega_{\mu_P, d_P}(x)} \right)^{\frac{\beta}{2\beta + d_P}} + \frac{2\Delta_h}{\phi},$$

so that $|\eta_Q(x) - 1/2| \ge \epsilon$ (in particular, this means that $\epsilon \le 1/2$). We consider only the case where $\eta_Q(x) - 1/2 \ge \epsilon$, since the case where $\eta_Q(x) - 1/2 \le -\epsilon$ follows by symmetry. We further suppose throughout the proof that the event $E_1^\delta(x) \cap E_2^\delta(x)$ holds. Let ℓ^* denote the leaf function corresponding to $\{\mathcal{X}_1^*, \dots, \mathcal{X}_{L^*}^*\} \in \mathbb{T}_{L^*}$ in Assumption 1. Fixing $k \in [n_P]$, it then follows from Assumption 1 and Lemma 5 that for all $i \in [k]$, and $h \in \mathcal{H}_{L^*} \cup \mathcal{H}_0$,

$$\begin{aligned}
\eta_P\bigl(X_{(i)}^P\bigr) - h\bigl(X_{(i)}^P\bigr) &= \eta_P\bigl(X_{(i)}^P\bigr) - g_{\ell^*(X_{(i)}^P)}\bigl(\eta_Q\bigl(X_{(i)}^P\bigr)\bigr) \\
&\quad + g_{\ell^*(X_{(i)}^P)}\bigl(\eta_Q\bigl(X_{(i)}^P\bigr)\bigr) - g_{\ell^*(X_{(i)}^P)}(1/2) + g_{\ell^*(X_{(i)}^P)}(1/2) - h\bigl(X_{(i)}^P\bigr) \\
&\ge \phi \cdot \bigl\{\eta_Q\bigl(X_{(i)}^P\bigr) - 1/2\bigr\} - \Delta - \max_{\ell \in [L^*]} \sup_{x \in \mathcal{X}_\ell^*} \bigl|h(x) - g_\ell(1/2)\bigr| \\
&\ge \phi \cdot \bigl\{\eta_Q\bigl(X_{(i)}^P\bigr) - 1/2\bigr\} - \frac{\phi \cdot \epsilon}{2} \\
&\ge \phi \cdot \left\{ \frac{\epsilon}{2} - C_{\mathrm{S}} \cdot \left(\frac{2 \cdot \max\{k, \lceil 4 \log_+(n_P/\delta) \rceil\}}{n_P \cdot \omega_{\mu_P, d_P}(x)} \right)^{\beta/d_P} \right\}.
\end{aligned}$$

We deduce that

$$
\begin{aligned}
\hat{m}^P_{k,h}(x) &\equiv \frac{1}{k}\sum_{i=1}^{k}\{Y^P_{(i)}(x) - h(X^P_{(i)})\} \\
&\geq \phi \cdot \left\{\frac{\epsilon}{2} - C_{\mathrm{S}} \cdot \left(\frac{2 \cdot \max\{k, \lceil 4\log_+(n_P/\delta)\rceil\}}{n_P \cdot \omega_{\mu_P,d_P}(x)}\right)^{\beta/d_P}\right\} - \sqrt{\frac{\log_+(n_P/\delta)}{2k}}
\end{aligned} \tag{19}
$$

for all $k \in [n_P]$. Now define

$$
k^* := \min\left\{k \in \mathbb{N} : \phi \cdot C_{\mathrm{S}} \cdot \left(\frac{2k}{n_P \cdot \omega_{\mu_P,d_P}(x)}\right)^{\beta/d_P} \geq \sqrt{\frac{\log_+(n_P/\delta)}{k}}\right\}.
$$

Since $\epsilon \leq 1/2$, we have $n_P \geq n_P \cdot \omega_{\mu_P,d_P}(x) \geq (100C_S)^{(2\beta+d_P)/\beta}\phi^{-2}\log_+(n_P/\delta) \geq (100C_S)^{(2\beta+d_P)/\beta}\log_+(n_P/\delta)$, so

$$
\lceil 4\log_+(n_P/\delta)\rceil < k^* \leq \left\lceil \left(\frac{\log_+(n_P/\delta)}{\phi^2}\right)^{\frac{d_P}{2\beta+d_P}} \cdot \left(n_P \cdot \omega_{\mu_P,d_P}(x)\right)^{\frac{2\beta}{2\beta+d_P}}\right\rceil \leq n_P.
$$

Moreover, we also see that $\sigma^* = \min\{\lceil 3\log_+^{1/2}(n_P/\delta)\rceil, n_P\} = \lceil 3\log_+^{1/2}(n_P/\delta)\rceil$. Hence by (19), we have

$$
\begin{aligned}
\hat{m}^P_{k^*,h}(x) - \frac{\sigma^*}{\sqrt{k^*}} &> \phi \cdot \left\{\frac{\epsilon}{2} - C_{\mathrm{S}} \cdot \left(\frac{2k^*}{n_P \cdot \omega_{\mu_P,d_P}(x)}\right)^{\beta/d_P}\right\} - 5\sqrt{\frac{\log_+(n_P/\delta)}{k^*}} \\
&\geq \phi \cdot \left\{\frac{\epsilon}{2} - 6C_{\mathrm{S}} \cdot \left(\frac{2k^*}{n_P \cdot \omega_{\mu_P,d_P}(x)}\right)^{\beta/d_P}\right\} \\
&= \phi \cdot \left\{\frac{\epsilon}{2} - 24C_{\mathrm{S}} \cdot \left(\frac{\log_+(n_P/\delta)}{\phi^2 \cdot n_P \cdot \omega_{\mu_P,d_P}(x)}\right)^{\frac{\beta}{2\beta+d_P}}\right\} > 0.
\end{aligned} \tag{20}
$$

We conclude that $\hat{k} \equiv \hat{k}^P_{\sigma^*,h}(x) \leq k^*$. Moreover, by applying (19) once again, we have for $k < k^*$ that

$$
\begin{aligned}
&\hat{m}^P_{k,h}(x) + \frac{\sigma^*}{\sqrt{k}} \\
&\geq \phi \cdot \left\{\frac{\epsilon}{2} - C_{\mathrm{S}} \cdot \left(\frac{2 \cdot \max\{k, \lceil 4\log_+(n_P/\delta)\rceil\}}{n_P \cdot \omega_{\mu_P,d_P}(x)}\right)^{\beta/d_P}\right\} - \sqrt{\frac{\log_+(n_P/\delta)}{k}} + \frac{\sigma^*}{\sqrt{k}} \\
&> -\sqrt{\frac{\log_+(n_P/\delta)}{\max\{k, \lceil 4\log_+(n_P/\delta)\rceil\}}} - \sqrt{\frac{\log_+(n_P/\delta)}{k}} + \frac{\sigma^*}{\sqrt{k}} \geq 0.
\end{aligned} \tag{21}
$$

But by definition of $\hat{k}$, we have $|\hat{m}^P_{\hat{k},h}(x)| > \sigma^*/\hat{k}^{1/2}$, so from (20) and (21), we deduce that $\hat{m}^P_{\hat{k},h}(x) > 0$, and hence that $\hat{f}^P_{\sigma^*,h}(x) = \mathbb{1}_{\{\hat{m}^P_{\hat{k},h}(x) \geq 0\}} = 1 = f^*_Q(x)$, as required. □

We now define a random subset $A_\delta(\mathcal{D}_P)$ of $\mathbb{R}^d$ by

$$
A_\delta(\mathcal{D}_P) := \left\{x \in \mathbb{R}^d : E_1^{\delta/(2n_P^{1+\alpha})}(x) \cap E_2^{\delta/(2n_P^{1+\alpha})}(x) \text{ holds}\right\}.
$$

LEMMA 8. *Let $n_P \in \mathbb{N}$ and $(P, Q) \in \mathcal{P}_{\theta^\sharp}$. We have* $\mathbb{P}\{\mu_Q(A_\delta(\mathcal{D}_P)^c) \geq 1/n_P^{1+\alpha}\} \leq \delta$.

PROOF. By Markov's inequality and Fubini's theorem, as well as Lemmas 4 and 6, we have

$$\begin{aligned}\mathbb{P}\{\mu_Q(A_\delta(\mathcal{D}_P)^c) \geq 1/n_P^{1+\alpha}\} &\leq n_P^{1+\alpha} \cdot \mathbb{E}\{\mu_Q(A_\delta(\mathcal{D}_P)^c)\} \\ &= n_P^{1+\alpha} \cdot \mathbb{E}\int_{\mathbb{R}^d} \mathbb{1}_{\{x \in A_\delta(\mathcal{D}_P)^c\}}\, d\mu_Q(x) \\ &\leq n_P^{1+\alpha} \int_{\mathbb{R}^d} \bigl[\mathbb{P}\{E_1^{\delta/(2n_P^{1+\alpha})}(x)^c\} + \mathbb{P}\{E_2^{\delta/(2n_P^{1+\alpha})}(x)^c\}\bigr]\, d\mu_Q(x) \leq \delta,\end{aligned}$$

as required. □

We are now ready to provide the proof of Proposition 3.

PROOF OF PROPOSITION 3. We begin by introducing some further notation for the proof. Let

$$\Lambda_0 := 100 \cdot C_{\mathrm{S}} \cdot \left(\frac{\log_+(2n_P^{2+\alpha}/\delta)}{\phi^2 \cdot n_P}\right)^{\frac{\beta}{2\beta+d_P}},$$

$$t_0 := \min\left\{\left(\frac{\Lambda_0\phi}{4\Delta_h}\right)^{\frac{2\beta+d_P}{\beta}}, \Lambda_0^{\frac{\alpha(2\beta+d_P)}{\alpha\beta+\gamma_P(2\beta+d_P)}}\right\}.$$

We then generate a countable partition $(\mathcal{T}_j)_{j\in\mathbb{N}_0}$ of $A_\delta(\mathcal{D}_P)$ by

$$\mathcal{T}_0 := \{x \in A_\delta(\mathcal{D}_P) : \omega_{\mu_P,d_P}(x) \geq t_0\},$$

$$\mathcal{T}_j := \{x \in A_\delta(\mathcal{D}_P) : 2^{-j} \cdot t_0 \leq \omega_{\mu_P,d_P}(x) < 2^{-(j-1)} \cdot t_0\},$$

for $j \in \mathbb{N}$. By Lemma 7 for each $j \in \mathbb{N}_0$, we have

$$\begin{aligned}|2\eta_Q(x) - 1| \cdot \mathbb{1}_{\{x\in\mathcal{T}_j : \hat{f}^P_{\sigma^*,h}(x)\neq f^*_Q(x)\}} &< \Lambda_0 \cdot (2^{-j}\cdot t_0)^{-\frac{\beta}{2\beta+d_P}} + \frac{4\Delta_h}{\phi} \\ &\leq 2\Lambda_0 \cdot (2^{-j}\cdot t_0)^{-\frac{\beta}{2\beta+d_P}},\end{aligned}$$

where the second inequality uses $t_0 \leq (\frac{\Lambda_0\phi}{4\Delta_h})^{\frac{2\beta+d_P}{\beta}}$. By Assumption 3, we have

$$\begin{aligned}\int_{\{x\in\mathcal{T}_0 : \hat{f}^P_{\sigma^*,h}(x)\neq f^*_Q(x)\}} |2\eta_Q(x)-1|\, d\mu_Q(x) &\leq 2^{1+\alpha} C_{\mathrm{M}} \cdot {\Lambda_0}^{1+\alpha} \cdot t_0^{-\frac{\beta(1+\alpha)}{2\beta+d_P}} \\ &\leq 2^{1+\alpha} C_{\mathrm{M}} \cdot \max\left\{\left(\frac{4\Delta_h}{\phi}\right)^{1+\alpha}, \Lambda_0^{\frac{\gamma_P(1+\alpha)(2\beta+d_P)}{\alpha\beta+\gamma_P(2\beta+d_P)}}\right\}.\end{aligned}$$

On the other hand, by Assumption 2 and the assumption that $\gamma_P - \frac{\beta}{2\beta+d_P} > 0$, for $j \in \mathbb{N}$ we have

$$\begin{aligned}\int_{\{x\in\mathcal{T}_j : \hat{f}^P_{\sigma^*,h}(x)\neq f^*_Q(x)\}} |2\eta_Q(x)-1|\, d\mu_Q(x) &\leq (2^{1+\gamma_P} C_{P,Q}) \cdot \Lambda_0 \cdot (2^{-j}\cdot t_0)^{\gamma_P - \frac{\beta}{2\beta+d_P}} \\ &\leq (2^{1+\gamma_P} C_{P,Q}) \cdot \Lambda_0^{\frac{\gamma_P(1+\alpha)(2\beta+d_P)}{\alpha\beta+\gamma_P(2\beta+d_P)}} \cdot 2^{-j(\gamma_P - \frac{\beta}{2\beta+d_P})}.\end{aligned}$$

Putting the above together, we have

$$\mathcal{E}(\hat{f}^P_{\sigma^*,h}) = \int_{\{x\in\mathbb{R}^d : \hat{f}^P_{\sigma^*,h}(x)\neq f^*_Q(x)\}} |2\eta_Q(x)-1|\, d\mu_Q(x)$$

$$
\begin{aligned}
&\le \mu_Q\big(A_\delta(\mathcal{D}_P)^c\big) + \sum_{j=0}^{\infty} \int_{\{x\in\mathcal{T}_j : \hat{f}^P_{\sigma^*,h}(x)\neq f^*_Q(x)\}} |2\eta_Q(x)-1|\, d\mu_Q(x) \\
&\le \mu_Q\big(A_\delta(\mathcal{D}_P)^c\big) + 2^{1+\alpha} C_{\mathrm{M}} \cdot \left(\frac{4\Delta_h}{\phi}\right)^{1+\alpha} \\
&\quad + \left\{ 2^{1+\alpha} C_{\mathrm{M}} + (2^{1+\gamma_P} C_{P,Q}) \cdot \sum_{j=1}^{\infty} 2^{-j(\gamma_P - \frac{\beta}{2\beta+d_P})} \right\} \cdot \Lambda_0^{\frac{\gamma_P(1+\alpha)(2\beta+d_P)}{\alpha\beta+\gamma_P(2\beta+d_P)}} \\
&\le \mu_Q\big(A_\delta(\mathcal{D}_P)^c\big) + \frac{\tilde{C}_\theta}{2}\left\{ \left(\frac{\log_+(n_P/\delta)}{\phi^2\cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta+\gamma_P(2\beta+d_P)}} + \left(\frac{\Delta_h}{\phi}\right)^{1+\alpha} \right\},
\end{aligned}
$$

where $\tilde{C}_\theta \ge 2$ depends only on θ. Note that the final inequality again uses the hypothesis that $\gamma_P > \beta/(2\beta+d_P)$. By Lemma 8, it now follows that

$$
\mathbb{P}\left[\mathcal{E}(\hat{f}^P_{\sigma^*,h}) > \frac{1}{n_P^{1+\alpha}} + \frac{\tilde{C}_\theta}{2}\left\{ \left(\frac{\log_+(n_P/\delta)}{\phi^2\cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta+\gamma_P(2\beta+d_P)}} + \left(\frac{\Delta_h}{\phi}\right)^{1+\alpha} \right\} \right] \le \delta.
$$

Since $\frac{\beta\gamma_P}{\alpha\beta+\gamma_P(2\beta+d_P)} \le 1$, the conclusion follows. □

We now give the three different instantiations of Proposition 3 mentioned above.

COROLLARY 9. *Let $n_P \in \mathbb{N}$. Fix $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$, where $\theta = (d_Q, \gamma_Q, d_P, \gamma_P, C_{P,Q}, \alpha, C_{\mathrm{M}}, \beta, C_{\mathrm{S}})$, with $\beta/(2\beta+d_P) < \gamma_P$, and $(P,Q) \in \mathcal{P}_{\theta^\sharp}$. Taking $\tilde{C}_\theta \ge 2$ from Proposition 3, there exists $h^* \in \mathcal{H}_{L^*}$ such that for every $\delta \in (0,1)$, if we set $\sigma^* = \min\{\lceil 3\log_+^{1/2}(n_P/\delta)\rceil, n_P\}$, then*

$$
\mathbb{P}\left[\mathcal{E}(\hat{f}^P_{\sigma^*,h^*}) > (2^\alpha+1)\tilde{C}_\theta \left\{ \left(\frac{\log_+(n_P/\delta)}{\phi^2\cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta+\gamma_P(2\beta+d_P)}} + \left(\frac{\Delta}{\phi}\right)^{1+\alpha} \right\} \right] \le \delta.
$$

PROOF. The decision tree function $h^* \in \mathcal{H}_{L^*}$ is constructed as follows. Recall that $\{\mathcal{X}^*_1, \dots, \mathcal{X}^*_{L^*}\} \in \mathbb{T}_{L^*}$ denotes the decision tree partition given by Assumption 1, with corresponding transfer functions $g_1, \dots, g_{L^*} : [0,1] \to [0,1]$ and leaf function $\ell^* : \mathbb{R}^d \to [L^*]$. Fix $(\tau^*_1, \dots, \tau^*_{L^*}) \in \{0, 1/n_P, 2/n_P, \dots, 1\}^{L^*}$ such that $|\tau^*_\ell - g_\ell(1/2)| \le 1/n_P$ for each $\ell \in [L^*]$, and define $h^* : \mathbb{R}^d \to (0,1)$ by $h^*(x) := \tau^*_{\ell^*(x)}$. When $\phi^2 \cdot n_P \ge 1$, the claim now follows from Proposition 3, noting that

$$
\begin{aligned}
\left(\frac{\Delta + 1/n_P}{\phi}\right)^{1+\alpha} &\le 2^\alpha \left\{ \left(\frac{\Delta}{\phi}\right)^{1+\alpha} + \left(\frac{1}{\phi^2\cdot n_P}\right)^{(1+\alpha)/2} \right\} \\
&\le 2^\alpha \left(\frac{\Delta}{\phi}\right)^{1+\alpha} + 2^\alpha \left(\frac{\log_+(n_P/\delta)}{\phi^2\cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta+\gamma_P(2\beta+d_P)}}.
\end{aligned}
$$

But if $\phi^2 \cdot n_P < 1$, then the claim follows from the fact that $\mathcal{E}(\hat{f}^P_{\sigma^*,h^*}) \le 1$. □

COROLLARY 10. *Let $n_P \in \mathbb{N}$. Fix $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$, where $\theta = (d_Q, \gamma_Q, d_P, \gamma_P, C_{P,Q}, \alpha, C_{\mathrm{M}}, \beta, C_{\mathrm{S}})$, with $\beta/(2\beta+d_P) < \gamma_P$, and $(P,Q) \in \mathcal{P}_{\theta^\sharp}$. Taking $\tilde{C}_\theta \ge 2$ from Proposition 3, and writing h_0 for the unique element of $\mathcal{H}_0$, for every $\delta \in (0,1)$, if we set $\sigma^* = \min\{\lceil 3\log_+^{1/2}(n_P/\delta)\rceil, n_P\}$, then*

$$
\mathbb{P}\left[\mathcal{E}(\hat{f}^P_{\sigma^*,h_0}) > 2^\alpha \tilde{C}_\theta \left\{ \left(\frac{\log_+(n_P/\delta)}{\phi^2\cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta+\gamma_P(2\beta+d_P)}} + \left(\frac{\Delta}{\phi}\right)^{1+\alpha} + (1-\phi)^{1+\alpha} \right\} \right] \le \delta.
$$

PROOF. Observe that, by Assumption 1, for any $\ell \in [L^*]$,

$$g_\ell(1/2) - 1/2 \leq -\{g_\ell(1) - g_\ell(1/2)\} + 1/2 \leq (1-\phi)/2,$$
$$1/2 - g_\ell(1/2) \leq -\{g_\ell(1/2) - g_\ell(0)\} + 1/2 \leq (1-\phi)/2.$$

Hence

$$\Delta_{h_0} = \Delta + \max_{\ell \in [L^*]} |g_\ell(1/2) - 1/2| \leq \Delta + \frac{1-\phi}{2}.$$

We assume without loss of generality that $\phi \geq 1/2$, since otherwise the conclusion follows from the facts that $\mathcal{E}(\hat{f}^P_{\sigma^*,h_0}) \leq 1$ and $\tilde{C}_\theta \geq 2$. The result then follows by Proposition 3. □

Now, for $\theta^\flat = (d_Q, \gamma_Q, C_{P,Q}, \alpha, C_{\mathrm{M}}, \beta, C_{\mathrm{S}})$, let $\mathcal{Q}_{\theta^\flat}$ denote the set of distributions Q on $\mathbb{R}^d \times \{0,1\}$ that satisfy Assumptions 3, 4 and the first part of Assumption 2 with parameter $\theta^\flat$.

COROLLARY 11. *Let $n_Q \in \mathbb{N}$. Fix $\theta^\flat = (d_Q, \gamma_Q, C_{P,Q}, \alpha, C_{\mathrm{M}}, \beta, C_{\mathrm{S}})$ with $\beta/(2\beta + d_Q) < \gamma_Q$, and $Q \in \mathcal{Q}_{\theta^\flat}$. There exists $\tilde{C}_{\theta^\flat} \geq 2$, depending only on $\theta^\flat$, such that for every $\delta \in (0,1)$, if we set $\tilde{\sigma} = \min\{\lceil 3 \log_+^{1/2}(n_Q/\delta)\rceil, n_Q\}$, then*

$$\mathbb{P}\biggl\{\mathcal{E}(\hat{f}^Q_{\tilde{\sigma}}) > \tilde{C}_{\theta^\flat}\biggl(\frac{\log_+(n_Q/\delta)}{n_Q}\biggr)^{\frac{\beta\gamma_Q(1+\alpha)}{\alpha\beta+\gamma_Q(2\beta+d_Q)}}\biggr\} \leq \delta.$$

PROOF. The result follows from Corollary 10, with Q in place of P, with $\mathcal{D}^0_Q$ in place of $\mathcal{D}_P$, and with $\Delta = 0$ and $\phi = 1$. □

5.2. *Empirical risk minimisation.* In this section, we control the additional error incurred by selecting our decision tree h and robustness parameter σ by empirical risk minimisation. Our analysis will make use of the following result, similar versions of which are well known (e.g., Tsybakov (2004)), but we include a proof for completeness.

PROPOSITION 12. *Suppose that Assumption 3 holds and let $\mathcal{F}$ be a nonempty, finite set of classifiers and let $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent pairs with distribution Q, and let $\hat{f}_n \in \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n \mathbb{1}_{\{f(X_i) \neq Y_i\}}$. Then, for any $\delta \in (0,1]$,*

$$\mathbb{P}\biggl\{\mathcal{E}(\hat{f}_n) > 2\mathcal{E}(f^*) + 64 C_{\mathrm{M}}^{\frac{1}{2+\alpha}}\biggl(\frac{\log(2|\mathcal{F}|/\delta)}{n}\biggr)^{\frac{1+\alpha}{2+\alpha}}\biggr\} \leq \delta.$$

PROOF. By Assumption 3, for every $\epsilon > 0$ and $f \in \mathcal{F}$,

$$\begin{aligned}
\mathcal{E}(f) &= \int_{\{x \in \mathbb{R}^d : f(x) \neq f^*_Q(x)\}} |2\eta_Q(x) - 1| \, d\mu_Q(x) \\
&\geq 2\epsilon \cdot \mu_Q(\{x \in \mathbb{R}^d : f(x) \neq f^*_Q(x) \text{ and } |\eta_Q(x) - 1/2| \geq \epsilon\}) \\
&\geq 2\epsilon \cdot \{\mu_Q(\{x \in \mathbb{R}^d : f(x) \neq f^*_Q(x)\}) - \mu_Q(\{x \in \mathbb{R}^d : |\eta_Q(x) - 1/2| < \epsilon\})\} \\
&\geq 2\epsilon \cdot \{\mu_Q(\{x \in \mathbb{R}^d : f(x) \neq f^*_Q(x)\}) - C_{\mathrm{M}} \cdot \epsilon^\alpha\}.
\end{aligned}$$

In particular, taking $\epsilon = \{\mu_Q(\{x \in \mathbb{R}^d : f(x) \neq f^*_Q(x)\})/(2C_{\mathrm{M}})\}^{1/\alpha}$, we deduce that

$$\mu_Q(\{x \in \mathbb{R}^d : f(x) \neq f^*_Q(x)\}) \leq (2C_{\mathrm{M}})^{\frac{1}{1+\alpha}} \cdot \mathcal{E}(f)^{\frac{\alpha}{1+\alpha}}. \tag{22}$$

Now, for each $f \in \mathcal{F}$, let $Z_i^f := \mathbb{1}_{\{f(X_i) \neq Y_i\}} - \mathbb{1}_{\{f_Q^*(X_i) \neq Y_i\}}$ for $i \in [n]$, noting that $\mathbb{E}(Z_i^f) = \mathcal{E}(f)$ and $\mathbb{E}\{(Z_i^f)^2\} \leq (2C_{\mathrm{M}})^{\frac{1}{1+\alpha}} \cdot \mathcal{E}(f)^{\frac{\alpha}{1+\alpha}}$ by (22). Define the event

$$E_3^\delta := \bigcap_{f \in \mathcal{F}} \left\{ \left| \frac{1}{n} \sum_{i=1}^n Z_i^f - \mathcal{E}(f) \right| \leq \sqrt{\frac{2(2C_{\mathrm{M}})^{\frac{1}{1+\alpha}} \mathcal{E}(f)^{\frac{\alpha}{1+\alpha}} \log(2|\mathcal{F}|/\delta)}{n}} + \frac{2\log(2|\mathcal{F}|/\delta)}{3n} \right\}.$$

Note that $|Z_i^f| \leq 1$, so by Bernstein's inequality (Bernstein (1924)) combined with a union bound, we have $\mathbb{P}\{(E_3^\delta)^c\} \leq \delta$. Hence, on the event E_3^δ, we have

$$\begin{aligned}
&\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n Z_i^{\hat{f}_n} - \frac{1}{n}\sum_{i=1}^n Z_i^{f^*} + \left| \frac{1}{n}\sum_{i=1}^n Z_i^{\hat{f}_n} - \mathcal{E}(\hat{f}_n) \right| + \left| \frac{1}{n}\sum_{i=1}^n Z_i^{f^*} - \mathcal{E}(f^*) \right| \\
&\quad \leq 2\sqrt{\frac{C_{\mathrm{M}}^{\frac{1}{1+\alpha}} \mathcal{E}(\hat{f}_n)^{\frac{\alpha}{1+\alpha}} \log(2|\mathcal{F}|/\delta)}{n}} + 2\sqrt{\frac{C_{\mathrm{M}}^{\frac{1}{1+\alpha}} \mathcal{E}(f^*)^{\frac{\alpha}{1+\alpha}} \log(2|\mathcal{F}|/\delta)}{n}} + \frac{4\log(2|\mathcal{F}|/\delta)}{3n} \\
&\quad \leq 4\sqrt{\frac{C_{\mathrm{M}}^{\frac{1}{1+\alpha}} \mathcal{E}(\hat{f}_n)^{\frac{\alpha}{1+\alpha}} \log(2|\mathcal{F}|/\delta)}{n}} + \frac{4\log(2|\mathcal{F}|/\delta)}{3n}.
\end{aligned}$$

Thus, by considering separately the cases $\mathcal{E}(\hat{f}_n) \leq 2\mathcal{E}(f^*)$ and $\mathcal{E}(\hat{f}_n) > 2\mathcal{E}(f^*)$, we see that on the event E_3^δ,

$$\mathcal{E}(\hat{f}_n) \leq 2\mathcal{E}(f^*) + 64 C_{\mathrm{M}}^{\frac{1}{2+\alpha}} \left(\frac{\log(2|\mathcal{F}|/\delta)}{n} \right)^{\frac{1+\alpha}{2+\alpha}},$$

as required. □

In order to apply Proposition 12, we will first derive a bound on the number of possible decision tree functions over $\mathcal{D}_P$. Recall for $L \in \mathbb{N}$ that $\mathcal{H}_L$ denotes the set of decision tree functions $h : \mathbb{R}^d \to (0, 1)$ to be those of the form $x \mapsto \tau_{\ell(x)}$ for some $\{\mathcal{X}_1, \ldots, \mathcal{X}_L\} \in \mathbb{T}_L$ with leaf function ℓ, and some $(\tau_1, \ldots, \tau_L) \in \{0, 1/n_P, 2/n_P, \ldots, 1\}^L$. Given a set $\mathcal{S} \subseteq \mathbb{R}^d$, we let $h|_{\mathcal{S}} : \mathcal{S} \to (0, 1)$ denote the restriction of h to $\mathcal{S}$.

LEMMA 13. *Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a set of cardinality at most n_P, and let $L \in \mathbb{N}$. Then the set $\{h|_{\mathcal{S}} : h \in \mathcal{H}_L\}$ has cardinality at most $\{Ld(n_P + 1)\}^{2L}$.*

PROOF. For the proof, let $\mathbb{L}_L$ denote the set of leaf functions $\ell : \mathbb{R}^d \to \{1, \ldots, L\}$ corresponding to decision tree partitions $\{\mathcal{X}_1, \ldots, \mathcal{X}_L\} \in \mathbb{T}_L$. We begin by bounding the cardinality of the set of restricted leaf functions $\{\ell|_{\mathcal{S}} : \ell \in \mathbb{L}_L\}$. Observe that each restricted leaf function $\ell_{\mathcal{S}}$ may be constructed recursively by a sequence of $L - 1$ splits. Each split point may be specified by choosing:

(a) one of at most $L - 1$ existing leaf nodes;
(b) one of d dimensions to split along;
(c) one of at most $n_P + 1$ possible split points.

Hence, $|\{\ell|_{\mathcal{S}} : \ell \in \mathbb{L}_L\}| \leq \{(L - 1)d(n_P + 1)\}^{L-1}$. Moreover, there are at most $(n_P + 1)^L$ possible choices for $(\tau_1, \ldots, \tau_L) \in \{0, 1/n_P, 2/n_P, \ldots, 1\}^L$. Since each $h|_{\mathcal{S}}$ is of the form $x \mapsto \tau_{\ell|_{\mathcal{S}}(x)}$ for some $\ell|_{\mathcal{S}}$ with $\ell \in \mathbb{L}_L$ and $(\tau_1, \ldots, \tau_L) \in \{0, 1/n_P, 2/n_P, \ldots, 1\}^L$, the result follows. □

COROLLARY 14. *Fix* $\mathcal{D}_P = ((X_1^P, Y_1^P), \dots, (X_{n_P}^P, Y_{n_P}^P)) \in (\mathbb{R}^d \times \{0,1\})^{n_P}$. *For every* $L \in \mathbb{N}$ *and* $\sigma > 0$, *we have* $|\{\hat{f}_{\sigma,h}^P : h \in \mathcal{H}_L\}| \leq \{Ld(n_P+1)\}^{2L}$, *where* $\hat{f}_{\sigma,h}^P$ *is defined in* (16).

PROOF. Let $\mathcal{S} = \{X_i^P\}_{i=1}^{n_P}$. Observe from the definition (10) that $\hat{m}_{k,h_0}^P = \hat{m}_{k,h_1}^P$ whenever $h_0|_{\mathcal{S}} = h_1|_{\mathcal{S}}$. Hence, by (11) and (16) the same is true of $\hat{k} \equiv \hat{k}_{\sigma,h}^P(\cdot)$ and $\hat{f}_{\sigma,h}^P$. Thus, by Lemma 13, we have

$$\left|\{\hat{f}_{\sigma,h}^P : h \in \mathcal{H}_L\}\right| = \left|\{\hat{f}_{\sigma,h|_{\mathcal{S}}}^P : h \in \mathcal{H}_L\}\right| \leq \left|\{h|_{\mathcal{S}} : h \in \mathcal{H}_L\}\right| \leq \{Ld(n_P+1)\}^{2L},$$

as required. □

Recall from (16) that $\hat{f}_{\sigma,L}^P(\cdot) = \mathbb{1}_{\{\hat{m}_{\hat{k},\hat{h}}^P(\cdot) \geq 0\}}$, where $\hat{k} \equiv \hat{k}_{\sigma,\hat{h}}^P(\cdot)$ is defined in (11), and where $\hat{h} \in \mathcal{H}_L$ is selected by empirical risk minimisation over $\mathcal{D}_Q^0$ as in (12). We are now in position to apply Proposition 12 to obtain the main conclusion of this subsection.

PROPOSITION 15. *Fix* $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$, *where* $\theta = (d_Q, \gamma_Q, d_P, \gamma_P, C_{P,Q}, \alpha, C_{\mathrm{M}}, \beta, C_{\mathrm{S}})$, *with* $\beta/(2\beta + d_P) < \gamma_P$, *and* $(P, Q) \in \mathcal{P}_{\theta^\sharp}$. *There exists* $C'_\theta > 0$, *depending only on* θ, *such that for every* $\delta \in (0,1)$, *if we set* $\sigma^* = \min\{\lceil 3\log_+^{1/2}(n_P/\delta)\rceil, n_P\}$, *then with probability at least* $1 - 2\delta$, *we have*

$$\mathcal{E}(\hat{f}_{\sigma^*,L^*}^P) \leq C'_\theta \left\{ \left(\frac{\log_+(n_P/\delta)}{\phi^2 \cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta + \gamma_P(2\beta+d_P)}} + \left(\frac{\Delta}{\phi}\right)^{1+\alpha} + \left(\frac{L^*\log_+(L^* d n_P/\delta)}{n_Q}\right)^{\frac{1+\alpha}{2+\alpha}} \right\}.$$

PROOF. Recalling $\tilde{C}_\theta$ from Proposition 3, by Proposition 12 combined with Corollary 14, we can find $C'_\theta \geq 2\tilde{C}_\theta$, depending only on θ, such that

$$\mathbb{P}\left\{ \mathcal{E}(\hat{f}_{\sigma^*,L^*}^P) > 2\mathcal{E}(\hat{f}_{\sigma^*,h^*}^P) + C'_\theta \left(\frac{L^*\log_+(L^* d n_P/\delta)}{n_Q}\right)^{\frac{1+\alpha}{2+\alpha}} \middle| \mathcal{D}_P \right\} \leq \delta.$$

Moreover, by Proposition 3,

$$\mathbb{P}\left[\mathcal{E}(\hat{f}_{\sigma^*,h^*}^P) > \tilde{C}_\theta \left\{ \left(\frac{\log_+(n_P/\delta)}{\phi^2 \cdot n_P}\right)^{\frac{\beta\gamma_P(1+\alpha)}{\alpha\beta + \gamma_P(2\beta+d_P)}} + \left(\frac{\Delta}{\phi}\right)^{1+\alpha} \right\} \right] \leq \delta.$$

The result follows. □

5.3. *Completion of the proofs of Theorem 2 and upper bound in Theorem 1.* PROOF OF THEOREM 2. Since $\hat{\mathcal{F}}^P$ and $\hat{\mathcal{F}}^Q$ were constructed using only $\mathcal{D}_P \cup \mathcal{D}_Q^0$ (and not $\mathcal{D}_Q^1$), we may apply Proposition 12 conditionally on $\mathcal{D}_P \cup \mathcal{D}_Q^0$ and take expectations to obtain that with probability at least $1 - \delta/4$, we have

$$\mathcal{E}(\hat{f}_{\mathrm{ATL}}) \leq 2\min\{\mathcal{E}(\hat{f}_{\sigma^*,0}^P), \mathcal{E}(\hat{f}_{\sigma^*,L^*\wedge n_Q}^P), \mathcal{E}(\hat{f}_{\tilde{\sigma}}^Q)\} + 64 C_{\mathrm{M}}^{\frac{1}{2+\alpha}} \left(\frac{\log(8|\hat{\mathcal{F}}^P \cup \hat{\mathcal{F}}^Q|/\delta)}{\lceil n_Q/2\rceil}\right)^{\frac{1+\alpha}{2+\alpha}},$$

where $\sigma^* = \min\{\lceil 3\log_+^{1/2}(n_P/\delta)\rceil, n_P\}$ and $\tilde{\sigma} = \min\{\lceil 3\log_+^{1/2}(n_Q/\delta)\rceil, n_Q\}$. Now $|\hat{\mathcal{F}}^P \cup \hat{\mathcal{F}}^Q| \leq n_P^2(n_Q+1) + n_Q^2$, so the result follows from Corollary 10, Proposition 15 and Corollary 11, which give the required high-probability bounds for $\mathcal{E}(\hat{f}_{\sigma^*,0}^P)$, $\mathcal{E}(\hat{f}_{\sigma^*,L^*\wedge n_Q}^P)$ and $\mathcal{E}(\hat{f}_{\tilde{\sigma}}^Q)$, respectively. □

PROOF OF UPPER BOUND IN THEOREM 1. We consider four cases. First, if $A^{\mathrm{U}}_{n_P,n_Q} = \min\{A^{\mathrm{U}}_{n_P,n_Q}, B^{\mathrm{U}}_{n_Q}, 1\}$ and $(L^* a^{\mathrm{U}}_1 / n_Q)^{\frac{1+\alpha}{2+\alpha}} \le (1-\phi)^{1+\alpha}$, then the result follows by taking $\delta = n_Q^{-\frac{1+\alpha}{2+\alpha}}$ in Proposition 15. Second, if $A^{\mathrm{U}}_{n_P,n_Q} = \min\{A^{\mathrm{U}}_{n_P,n_Q}, B^{\mathrm{U}}_{n_Q}, 1\}$ and $(L^* a^{\mathrm{U}}_1 / n_Q)^{\frac{1+\alpha}{2+\alpha}} > (1-\phi)^{1+\alpha}$, then the result follows by taking $\delta = n_P^{-\frac{1+\alpha}{2+\alpha}}$ in Corollary 10. Third, if $B^{\mathrm{U}}_{n_Q} = \min\{A^{\mathrm{U}}_{n_P,n_Q}, B^{\mathrm{U}}_{n_Q}, 1\}$, then the result follows by taking $\delta = n_Q^{-\frac{1+\alpha}{2+\alpha}}$ in Corollary 11. Finally, $\min\{A^{\mathrm{U}}_{n_P,n_Q}, B^{\mathrm{U}}_{n_Q}\} > 1$, then the result follows from the fact that the excess risk of any data-dependent classifier is at most 1. $\square$

6. Proof of the lower bound in Theorem 1. The proof of the lower bound in Theorem 1 begins with a version of Assouad's lemma for transfer learning (Section 6.1) that translates the problem into one of constructing an appropriate family of distributions indexed by a hypercube. To apply this lemma, we first construct the respective marginal distributions (Section 6.2) and then the corresponding regression functions (Section 6.3). The lower bound is finally obtained via two applications of these results, reflecting the different challenges of estimating the decision tree function (Section 6.4) and the source regression function (Section 6.5).

6.1. *Assouad's lemma for transfer learning.* The following result is a variant of Assouad's lemma (e.g., Yu (1997), Lemma 2, Kim (2020), Section 3.12), adapted to our setting.

LEMMA 16. *Let $\mathcal{P}$ be a set of pairs of distributions (P, Q), each on $\mathbb{R}^d \times \{0,1\}$. Let n_P, $n_Q \in \mathbb{N}_0$, $m \in \mathbb{N}$, $\Sigma = \{-1,1\}^m$, $(x_t)_{t\in[m]} \in (\mathbb{R}^d)^m$, $\epsilon_P, \epsilon_Q \in [0,1/4]$, $u_P, u_Q \in [0,1/m]$, $v_P, v_Q \in [0,1]$ and $\{(P^\sigma, Q^\sigma) : \sigma \in \Sigma\} \subseteq \mathcal{P}$ with respective regression functions $\eta^\sigma_P : \mathbb{R}^d \to [0,1]$, $\eta^\sigma_Q : \mathbb{R}^d \to [0,1]$ and marginals μ_P, μ_Q on $\mathbb{R}^d$ satisfy*:

(i) $2^5(n_P u_P \epsilon_P^2 + n_Q u_Q \epsilon_Q^2) \le 1$;
(ii) $\epsilon_P(2v_P - 1) = \epsilon_Q(2v_Q - 1) = 0$;
(iii) *for $t \in [m]$, we have $\mu_P(\{x_t\}) = u_P$ and $\mu_Q(\{x_t\}) = u_Q$*;
(iv) *for $\sigma = (\sigma_1, \dots, \sigma_m) \in \Sigma$ and $t \in [m]$, we have $\eta^\sigma_P(x_t) = v_P + \sigma_t \cdot \epsilon_P$ and $\eta^\sigma_Q(x_t) = v_Q + \sigma_t \cdot \epsilon_Q$*;
(v) *for $\sigma, \sigma' \in \Sigma$, $x \in \mathrm{supp}(\mu_P) \setminus \{x_t\}_{t\in[m]}$, we have $\eta^\sigma_P(x) = \eta^{\sigma'}_P(x)$; moreover, for $x \in \mathrm{supp}(\mu_Q) \setminus \{x_t\}_{t\in[m]}$, we have $\eta^\sigma_Q(x) = \eta^{\sigma'}_Q(x)$.*

Then

$$\inf_{\hat{f}\in\hat{\mathcal{F}}_{n_P,n_Q}} \sup_{(P,Q)\in\mathcal{P}} \mathbb{E}\{\mathcal{E}(\hat{f})\} \ge \frac{m u_Q \epsilon_Q}{2}.$$

To prove Lemma 16, we introduce some additional notation and provide a preliminary lemma. For $\sigma \in \Sigma$, let ν^σ denote the product measure $(P^\sigma)^{n_P} \times (Q^\sigma)^{n_Q}$. In addition, given $\sigma = (\sigma_1, \dots, \sigma_m) \in \Sigma$ and $t \in [m]$, we define $\sigma^t = (\sigma^t_1, \dots, \sigma^t_m) \in \Sigma$ by $\sigma^t_t := -\sigma_t$ and $\sigma^t_{t'} := \sigma_{t'}$ for $t' \in [m] \setminus \{t\}$.

LEMMA 17. *In the setting of Lemma 16, we have $\mathrm{TV}(\nu^\sigma, \nu^{\sigma^t}) \le 1/2$ for every $\sigma \in \Sigma$ and $t \in [m]$.*

PROOF. We first show that $\mathrm{KL}(P^\sigma, P^{\sigma^t}) \le 16 u_P \epsilon_P^2$. Without loss of generality, we assume that $\epsilon_P > 0$, since otherwise $P^{\sigma^t} = P^\sigma$. Thus, $\eta^{\sigma^t}_P(x_t) = 1 - \eta^\sigma_P(x_t) = 1/2 - \epsilon_P \cdot \sigma_t$

and $\eta_P^{\sigma^t}(x) = \eta_P^{\sigma}(x)$ for all $x \in \mathrm{supp}(\mu_P) \setminus \{x_t\}$. Hence,

$$\begin{aligned}\mathrm{KL}(P^\sigma, P^{\sigma^t}) &= \int_{\mathbb{R}^d \times \{0,1\}} \log\biggl(\frac{dP^\sigma}{dP^{\sigma^t}}\biggr)\, dP^\sigma \\ &= \int_{\mathbb{R}^d} \biggl\{\eta_P^\sigma(x) \log\biggl(\frac{\eta_P^\sigma(x)}{\eta_P^{\sigma^t}(x)}\biggr) + \bigl(1-\eta_P^\sigma(x)\bigr)\log\biggl(\frac{1-\eta_P^\sigma(x)}{1-\eta_P^{\sigma^t}(x)}\biggr)\biggr\}\, d\mu_P(x) \\ &= \bigl(2\eta_P^\sigma(x_t)-1\bigr)\log\biggl(\frac{\eta_P^\sigma(x_t)}{1-\eta_P^\sigma(x_t)}\biggr)\cdot \mu_P\bigl(\{x_t\}\bigr) \\ &= 2\epsilon_P\sigma_t \cdot \log\biggl(\frac{1+2\epsilon_P\sigma_t}{1-2\epsilon_P\sigma_t}\biggr)\cdot u_P = 2u_P\epsilon_P \cdot \log\biggl(\frac{1+2\epsilon_P}{1-2\epsilon_P}\biggr) \\ &\le \frac{8u_P\epsilon_P^2}{1-2\epsilon_P} \le 16 u_P \epsilon_P^2,\end{aligned}$$

where the penultimate inequality uses the inequality $\log a \le a-1$ for $a \ge 1$. By the same argument, we also have $\mathrm{KL}(Q^\sigma, Q^{\sigma^t}) \le 16u_Q\epsilon_Q^2$. By the additive property of Kullback–Leibler divergence for product measures, we conclude that

$$\mathrm{KL}\bigl(\nu^\sigma, \nu^{\sigma^t}\bigr) = n_P \mathrm{KL}\bigl(P^\sigma, P^{\sigma^t}\bigr) + n_Q \mathrm{KL}\bigl(Q^\sigma, Q^{\sigma^t}\bigr) \le 16\bigl(n_P u_P \epsilon_P^2 + n_Q u_Q \epsilon_Q^2\bigr) \le \frac{1}{2}.$$

Thus, by Pinsker's inequality (e.g., Tsybakov (2009), Lemma 2.5),

$$\mathrm{TV}\bigl(\nu^\sigma, \nu^{\sigma^t}\bigr) \le \sqrt{\mathrm{KL}\bigl(\nu^\sigma, \nu^{\sigma^t}\bigr)/2} \le 1/2,$$

as required. □

We now return to the proof of Lemma 16.

PROOF OF LEMMA 16. Without loss of generality, we assume that $\epsilon_Q > 0$, so $v_Q = 1/2$. Fix $\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}$. Given $z \in \mathcal{Z} := (\mathbb{R}^d \times \{0,1\})^{n_P} \times (\mathbb{R}^d \times \{0,1\})^{n_Q}$, let $\hat{f}_z : \mathbb{R}^d \to \{0,1\}$ denote the mapping obtained by taking z as the first argument in $\hat{f}$. Then

$$\begin{aligned}\sup_{(P,Q)\in\mathcal{P}} \mathbb{E}\{\mathcal{E}(\hat{f})\} &\ge \max_{\sigma\in\Sigma} \int_{\mathcal{Z}} \int_{\{x\in\mathbb{R}^d : \hat{f}_z(x) \ne f^*_{Q^\sigma}(x)\}} \bigl|2\eta_Q^\sigma(x) - 1\bigr|\, d\mu_Q(x)\, d\nu^\sigma(z) \\ &\ge \frac{1}{2^m} \sum_{\sigma\in\Sigma} \int_{\mathcal{Z}} \sum_{t=1}^m \bigl|2\eta_Q^\sigma(x_t) - 1\bigr| \cdot \mathbb{1}_{\{\hat{f}_z(x_t) \ne f^*_{Q^\sigma}(x_t)\}} \mu_Q\bigl(\{x_t\}\bigr)\, d\nu^\sigma(z) \\ &= \frac{u_Q\epsilon_Q}{2^{m-1}} \sum_{t=1}^m \sum_{\sigma\in\Sigma} \nu^\sigma\bigl(\bigl\{\hat{f}_z(x_t) \ne f^*_{Q^\sigma}(x_t)\bigr\}\bigr) \\ &= \frac{u_Q\epsilon_Q}{2^m} \sum_{t=1}^m \sum_{\sigma\in\Sigma} \bigl\{\nu^\sigma\bigl(\bigl\{\hat{f}_z(x_t) \ne f^*_{Q^\sigma}(x_t)\bigr\}\bigr) + \nu^{\sigma^t}\bigl(\bigl\{\hat{f}_z(x_t) \ne f^*_{Q^{\sigma^t}}(x_t)\bigr\}\bigr)\bigr\} \\ &\ge \frac{u_Q\epsilon_Q}{2^m} \sum_{t=1}^m \sum_{\sigma\in\Sigma} \bigl\{1 - \mathrm{TV}\bigl(\nu^\sigma, \nu^{\sigma^t}\bigr)\bigr\} \ge \frac{m u_Q \epsilon_Q}{2},\end{aligned}$$

where the penultimate inequality uses the fact that $f^*_{Q^{\sigma^t}}(x_t) = 1 - f^*_{Q^\sigma}(x_t)$ and the final inequality follows from Lemma 17. □

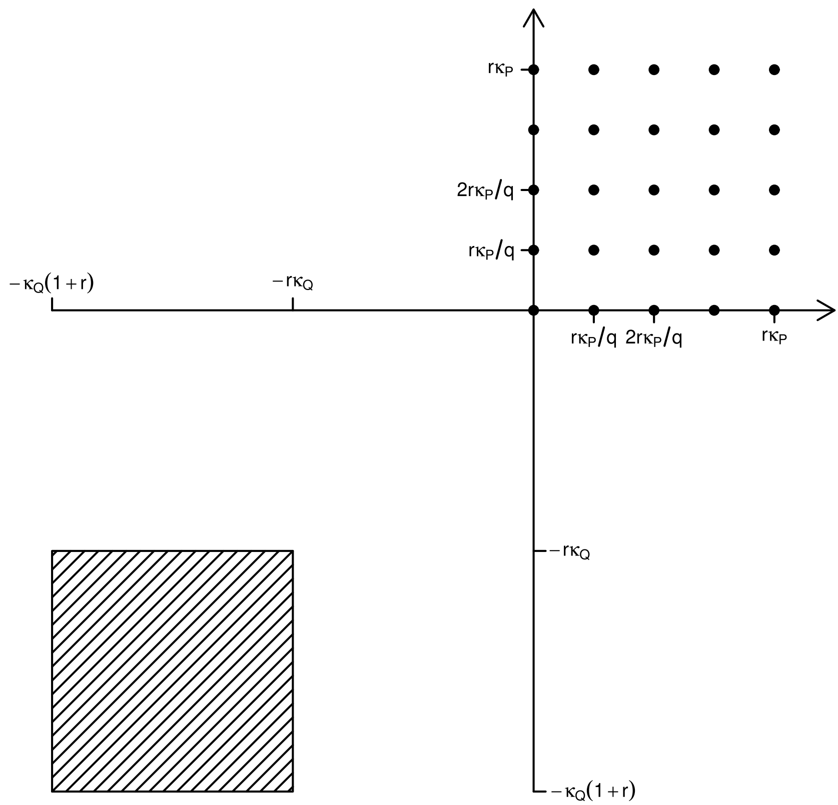

FIG. 2. *Illustration of the support of the measure* μ_{q,r,w,d_0} *in* (23).

6.2. *Marginal construction*. The marginal distributions μ_P and μ_Q in our lower bound construction will not vary with the vertices of our hypercube Σ in Lemma 16. An interesting consequence of this fact is that our lower bound in Theorem 1 will continue to hold, even if these marginal distributions were known (or equivalently, if we were also provided with an infinite sample of unlabelled training data from either distribution). These measures will consist of a mixture of a discrete uniform distribution on a lattice of points in the nonnegative orthant in $\mathbb{R}^d$ and a component consisting of a uniform distribution on a d_Q-dimensional hyper-rectangle in the opposite orthant, as illustrated in Figure 2. Moreover, the lattice component of the support of μ_Q will be a d_Q-dimensional slice within the d_P-dimensional lattice component of the support of μ_P. The structure of these marginals is designed to put as much probability mass as possible on the lattice points, as these will be the points that are difficult to classify, and will maximise the lower bound in Lemma 16. On the other hand, Condition (i) of Lemma 16 constrains us to have a sufficiently large lattice that no individual point provides too much information to the learner. The component supported on the hyper-rectangle is used to ensure that the margin condition (Assumption 3) is satisfied.

To describe the construction more formally, define $\kappa_P := 1/(2d_P^{1/2})$ and $\kappa_Q := 1/(2d_Q^{1/2})$. For $q \in \mathbb{N}$ and $d_0 \in \{d_P, d_Q\}$, let $\widetilde{\mathcal{T}}_{q,d_0} := \{0, 1, \dots, q-1\}^{d_0} \times \{0\}^{d-d_0} \subseteq \mathbb{R}^d$. Now let $(\tilde{x}_t^q)_{t=1}^{q^{d_Q}}$ be an enumeration of the set $\widetilde{\mathcal{T}}_{q,d_Q}$ and let $(\tilde{x}_t^q)_{t=q^{d_Q}+1}^{q^{d_P}}$ be an enumeration of $\widetilde{\mathcal{T}}_{q,d_P} \setminus \widetilde{\mathcal{T}}_{q,d_Q}$. For each $q \in \mathbb{N}$, $r > 0$, $t \in [q^{d_P}]$ and $d_0 \in \{d_P, d_Q\}$, we let $x_t^{q,r} := (r/q) \cdot \kappa_P \cdot \tilde{x}_t^q$ and $\mathcal{T}_{q,r,d_0} := \{x_t^{q,r} : t \in [q^{d_0}]\}$. For a Borel subset A of $\mathbb{R}^d$, let $A_{d_Q} := \{(x_1, \dots, x_{d_Q}) : (x_1, \dots, x_{d_Q}, 0, \dots, 0) \in A\}$ and let $A_{d_Q,d_P} := A \cap (\mathbb{R}^{d_Q} \times [0,1)^{d_P-d_Q} \times \{0\}^{d-d_P})$. Given $q \in \mathbb{N}$, $w \in [0, 1/2]$, $r > 0$, $d_0 \in \{d_P, d_Q\}$, we define a probability measure μ_{q,r,w,d_0} on $\mathbb{R}^d$

by

$$\mu_{q,r,w,d_0}(A) := \frac{(1-w)}{\kappa_Q^{d_Q}} \mathcal{L}_{d_Q}\big(A_{d_Q} \cap [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q}\big) + \frac{w}{N_{q,r,d_0}} \sum_{t=1}^{q^{d_0}} \mathbb{1}_{\{x_t^{q,r} \in A_{d_Q,d_P}\}}, \tag{23}$$

for Borel subsets A of $\mathbb{R}^d$, where $N_{q,r,d_0} := q^{d_Q} \min\{\lceil q/(r\kappa_P)\rceil, q\}^{d_0-d_Q} = |(\mathcal{T}_{q,r,d_0})_{d_Q,d_P}|$. Our marginal measures μ_P and μ_Q will be chosen as instances of μ_{q,r,w,d_0} for particular choices of q, r, w and d_0; see Corollary 21.

We begin with a couple of preliminary lemmas, before presenting the main properties of these marginal distributions in Lemma 20 and Corollary 21 below. This latter result provides sufficient conditions for the marginals to satisfy Assumption 2.

LEMMA 18. *For $\kappa \in (0, d_Q^{-1/2})$ and $x \in [0,\kappa]^{d_Q}$, we have $\mathcal{L}_{d_Q}(\{y \in \mathbb{R}^{d_Q} : \|y-x\| < s\} \cap [0,\kappa]^{d_Q}) \geq (\kappa s)^{d_Q}$ for all $s \in [0,1]$, and $\mathcal{L}_{d_Q}(\{y \in \mathbb{R}^{d_Q} : \|y-x\| < s\} \cap [0,\kappa]^{d_Q}) \leq V_{d_Q} \cdot s^{d_Q}$ for all $s > 0$.*

PROOF. To prove the lower bound, we take $x \in [0,\kappa]^{d_Q}$, $s \in [0,1]$, and consider the map $\psi_{x,s} : z \mapsto s \cdot (z-x) + x$ on $\mathbb{R}^{d_Q}$. Observe that $\psi_{x,s}([0,\kappa]^{d_Q}) \subseteq [0,\kappa]^{d_Q}$. On the other hand, since $x \in \psi_{x,s}([0,\kappa]^{d_Q})$ and $\operatorname{diam}(\psi_{x,s}([0,\kappa]^{d_Q})) \leq s\kappa d_Q^{1/2} < s$, we also have $\psi_{x,s}([0,\kappa]^{d_Q}) \subseteq \{y \in \mathbb{R}^{d_Q} : \|y-x\| < s\}$. Hence,

$$\mathcal{L}_{d_Q}\big(\{y \in \mathbb{R}^{d_Q} : \|y-x\| < s\} \cap [0,\kappa]^{d_Q}\big) \geq \mathcal{L}_{d_Q}\big(\psi_{x,s}([0,\kappa]^{d_Q})\big) \geq (\kappa s)^{d_Q}.$$

The upper bound follows from the fact that $\{y \in \mathbb{R}^{d_Q} : \|y-x\| < s\} \cap [0,1]^{d_Q} \subseteq \{y \in \mathbb{R}^{d_Q} : \|y-x\| < s\}$. $\square$

LEMMA 19. *For $q \in \mathbb{N}$, $d_0 \in \{d_P, d_Q\}$, $x \in \widetilde{\mathcal{T}}_{q,d_0}$ and $s \leq 4qd_P^{1/2}$, we have $|\widetilde{\mathcal{T}}_{q,d_0} \cap B_s(x)| \geq \{s/(2^4 d_P^{1/2})\}^{d_0}$.*

PROOF. First, observe that if $q = 1$, then $|\widetilde{\mathcal{T}}_{q,d_0} \cap B_s(x)| = |\{0\}| = 1 \geq \{s/(2^4 d_P^{1/2})\}^{d_0}$. For $q \geq 2$, we have $s/(2^4 d_P^{1/2}) \leq (q-1)/2$. Hence, for each $x \in \widetilde{\mathcal{T}}_{q,d_0}$, we can find a d_P-dimensional, axis-aligned cube A with vertex x and side length $s/(2^4 d_P^{1/2})$ containing at least $\lceil s/(2^4 d_P^{1/2})\rceil^{d_0}$ elements of $\widetilde{\mathcal{T}}_{q,d_0}$. Thus, $|\widetilde{\mathcal{T}}_{q,d_0} \cap B_s(x)| \geq |\widetilde{\mathcal{T}}_{q,d_0} \cap A| \geq \{s/(2^4 d_P^{1/2})\}^{d_0}$. $\square$

LEMMA 20. *Let $q \in \mathbb{N}$, $r > 0$, $w \in [0,1/2]$ and $d_0 \in \{d_P, d_Q\}$. We have:*

(i) $\omega_{\mu_{q,r,w,d_0},d_0}(x) \geq 1 - w$ *for all* $x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}$;

(ii) $\omega_{\mu_{q,r,w,d_0},d_0}(x) \geq 2^{-3d_0} \cdot \min\{1, w \cdot q^{d_0} \cdot N_{q,r,d_0}^{-1} \cdot r^{-d_0}\}$ *for all* $x \in \mathcal{T}_{q,r,d_Q}$.

PROOF. To prove (i), we take $x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}$ and $s \in (0,1)$. As shorthand, we write $B := B_s(x)$, so that $B_{d_Q} = \{(x_1,\ldots,x_{d_Q}) : (x_1,\ldots,x_{d_Q},0,\ldots,0) \in B\}$. By Lemma 18 combined with the translation invariance of Lebesgue measure, we have

$$\mu_{q,r,w,d_0}(B) \geq (1-w) \cdot \kappa_Q^{-d_Q} \cdot \mathcal{L}_{d_Q}\big(B_{d_Q} \cap [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q}\big) \geq (1-w) \cdot s^{d_0}.$$

The claim (i) follows.

To prove (ii), we take $x = x_t^{q,r} \in \mathcal{T}_{q,r,d_Q}$. If $s \in (0, (2r) \wedge 1]$, then $\tilde{s} := \{q/(r\kappa_P)\} \cdot s \le \min\{4qd_P^{1/2}, q/(r\kappa_P)\}$, so by Lemma 19 we have

$$\begin{aligned}\mu_{q,r,w,d_0}(B_s(x)) &\ge \frac{w}{N_{q,r,d_0}} \cdot |\mathcal{T}_{q,r,d_0} \cap B_s(x_t^{q,r})| = \frac{w}{N_{q,r,d_0}} \cdot |\widetilde{\mathcal{T}}_{q,d_0} \cap B_{\tilde{s}}(\tilde{x}_t^q)| \\ &\ge \frac{w}{N_{q,r,d_0}} \cdot \{\tilde{s}/(2^4 d_P^{1/2})\}^{d_0} = \frac{w}{N_{q,r,d_0}} \cdot \{qs/(2^4 r\kappa_P d_P^{1/2})\}^{d_0} \\ &= 2^{-3d_0} \cdot w \cdot r^{-d_0} \cdot \frac{q^{d_0}}{N_{q,r,d_0}} \cdot s^{d_0}.\end{aligned}$$

On the other hand, if $s \in (2r, 1]$ then with $z_r := (\overbrace{-r\kappa_Q, \dots, -r\kappa_Q}^{d_Q}, \overbrace{0, \dots, 0}^{d-d_Q}) \in \mathbb{R}^d$, we have $\|x - z_r\| \le \|x\| + \|z_r\| \le r/2 + r/2 < s/2$. Hence, by (i), we have $\mu_{q,r,w,d_0}(B_s(x)) \ge \mu_{q,r,w,d_0}(B_{s/2}(z_r)) \ge (1 - w) \cdot (s/2)^{d_0} \ge 2^{-(d_0+1)} s^{d_0} \ge 2^{-3d_0} s^{d_0}$, and the conclusion follows. □

COROLLARY 21. *Take $C_{P,Q} > 1$, $d_Q \in [1, d]$, $d_P \in [d_Q, d]$ and $\gamma_P, \gamma_Q > 0$. Suppose that $q \in \mathbb{N}$, $r > 0$ and $w_P, w_Q \in [0, 2^{-3d_P(\gamma_P \vee \gamma_Q)} \wedge (1 - C_{P,Q}^{-1/(\gamma_P \wedge \gamma_Q)})]$ satisfy $w_Q(w_P \cdot q^{d_P} \cdot N_{q,r,d_P}^{-1})^{-\gamma_P} r^{d_P\gamma_P} \le 2^{-3d_P\gamma_P}$ and $w_Q^{1-\gamma_Q} r^{d_Q\gamma_Q} \le 2^{-3d_Q\gamma_Q}$. Then Assumption 2 is satisfied for $\mu_P = \mu_{q,r,w_P,d_P}$ and $\mu_Q = \mu_{q,r,w_Q,d_Q}$.*

PROOF. For the first condition of Assumption 2 consider initially $\xi \in (0, 2^{-3d_Q} \cdot \min\{1, w_Q \cdot r^{-d_Q}\}]$. Then by Lemma 20,

$$\mu_Q(\{x \in \mathbb{R}^d : \omega_{\mu_Q,d_Q}(x) < \xi\}) = 0 \le C_{P,Q} \cdot \xi^{\gamma_Q}.$$

If $\xi \in (2^{-3d_Q} \cdot \min\{1, w_Q \cdot r^{-d_Q}\}, 1 - w_Q]$, then by Lemma 20 again,

$$\mu_Q(\{x \in \mathbb{R}^d : \omega_{\mu_Q,d_Q}(x) < \xi\}) = w_Q \le C_{P,Q} \cdot 2^{-3d_Q\gamma_Q}(1 \wedge w_Q^{\gamma_Q} r^{-d_Q\gamma_Q}) \le C_{P,Q} \cdot \xi^{\gamma_Q}.$$

Finally, if $\xi \in (1 - w_Q, \infty)$, then

$$\mu_Q(\{x \in \mathbb{R}^d : \omega_{\mu_Q,d_Q}(x) < \xi\}) = 1 \le C_{P,Q}(1 - w_Q)^{\gamma_Q} \le C_{P,Q} \cdot \xi^{\gamma_Q},$$

as required.

For the second condition of Assumption 2 let $\xi \in (0, 2^{-3d_P} \cdot \min\{1, w_P \cdot q^{d_P} \cdot N_{q,r,d_P}^{-1} \cdot r^{-d_P}\}]$. Then by Lemma 20,

$$\mu_Q(\{x \in \mathbb{R}^d : \omega_{\mu_P,d_P}(x) < \xi\}) = 0 \le C_{P,Q} \cdot \xi^{\gamma_P}.$$

If $\xi \in (2^{-3d_P} \cdot \min\{1, w_P \cdot q^{d_P} \cdot N_{q,r,d_P}^{-1} \cdot r^{-d_P}\}, 1 - w_P]$, then by Lemma 20 again,

$$\begin{aligned}\mu_Q(\{x \in \mathbb{R}^d : \omega_{\mu_P,d_P}(x) < \xi\}) = w_Q &\le C_{P,Q} \cdot 2^{-3d_P\gamma_P}\{1 \wedge (w_P \cdot q^{d_P} \cdot N_{q,r,d_P}^{-1} \cdot r^{-d_P})^{\gamma_P}\} \\ &\le C_{P,Q} \cdot \xi^{\gamma_P}.\end{aligned}$$

Finally, if $\xi \in (1 - w_P, \infty)$, then

$$\mu_Q(\{x \in \mathbb{R}^d : \omega_{\mu_P,d_P}(x) < \xi\}) = 1 \le C_{P,Q}(1 - w_P)^{\gamma_P} \le C_{P,Q} \cdot \xi^{\gamma_P},$$

as required. □

6.3. *Target regression function construction.* We now describe a construction of a family of target regression functions that are indexed by the vertices of a hypercube as in Lemma 16. We begin by defining the restrictions of the elements of this family to the support of μ_P; on this set, these restrictions will be perturbations of the uninformative regression function that takes the constant value $1/2$. The perturbations should be as large as possible, to maximise the quantity ϵ_Q in Lemma 16 and to ensure that the margin condition (Assumption 3) holds, but need to be small enough that the restrictions can be extended to functions on $\mathbb{R}^d$ that satisfy the Hölder continuity condition (Assumption 4).

Given $\epsilon \in (0, 1/8]$, $q \in \mathbb{N}$, $r > 0$, $\sigma = (\sigma_t)_{t=1}^{q^{d_Q}} \in \{-1, 1\}^{q^{d_Q}}$, we first define $\eta^\circ_{\epsilon,q,r,\sigma} : [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q} \cup \mathcal{T}_{q,r,d_P} \to \mathbb{R}$ by

$$\eta^\circ_{\epsilon,q,r,\sigma}(x) := \begin{cases} \frac{1}{2} - 2\epsilon - \frac{1}{4}\|x - z_r\|^\beta & \text{if } x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}, \\ \frac{1}{2} + \sigma_t \cdot \epsilon & \text{if } x = x_t^{q,r} \text{ with } t \le q^{d_Q}, \\ \frac{1}{2} - 2\epsilon & \text{if } x = x_t^{q,r} \text{ with } q^{d_Q} < t \le q^{d_P}, \end{cases}$$

where $z_r := (\overbrace{-r\kappa_Q, \dots, -r\kappa_Q}^{d_Q}, \overbrace{0, \dots, 0}^{d-d_Q})$. The main results of this subsection (Corollary 23 and Lemma 24) provide sufficient conditions for an extension $\eta_{\epsilon,q,r,\sigma}$ of $\eta^\circ_{\epsilon,q,r,\sigma}$ to the whole of $\mathbb{R}^d$ to satisfy Assumptions 4 and 3, respectively. Recalling that $\kappa_P = 1/(2d_P^{1/2})$ and $\kappa_Q = 1/(2d_Q^{1/2})$, we first present a basic property of $\eta^\circ_{\epsilon,q,r,\sigma}$.

LEMMA 22. *Let $q \in \mathbb{N}$, $r > 0$, $\beta \in (0, 1]$, $\sigma = (\sigma_t)_{t=1}^{q^{d_Q}} \in \{-1, 1\}^{q^{d_Q}}$ and $\epsilon \in (0, 1/8 \wedge (1/6) \cdot (r \cdot \kappa_P/q)^\beta]$. Then $|\eta^\circ_{\epsilon,q,r,\sigma}(x) - \eta^\circ_{\epsilon,q,r,\sigma}(x')| \le \|x - x'\|^\beta$ for all $x, x' \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q} \cup \mathcal{T}_{q,r,d_P}$. Moreover, $\eta^\circ_{\epsilon,q,r,\sigma}(x) \in [0, 1]$ for all $x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q} \cup \mathcal{T}_{q,r,d_P}$.*

PROOF. To prove the first part of the lemma, we consider three cases. First, if $x, x' \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}$, then by Minkowski's inequality,

$$|\eta^\circ_{\epsilon,q,r,\sigma}(x) - \eta^\circ_{\epsilon,q,r,\sigma}(x')| = \frac{1}{4}\big|\|x - z_r\|^\beta - \|x' - z_r\|^\beta\big| \le \frac{1}{4}\|x - x'\|^\beta.$$

Second, if $x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}$ and $x' \in \mathcal{T}_{q,r,d_P}$, then

$$\begin{aligned} |\eta^\circ_{\epsilon,q,r,\sigma}(x) - \eta^\circ_{\epsilon,q,r,\sigma}(x')| &\le \frac{1}{4}\|x - z_r\|^\beta + 3\epsilon \\ &\le \frac{1}{2}\big\{\|x - z_r\|^\beta + (r \cdot \kappa_P)^\beta\big\} \le \big\{\|x - z_r\| + (r \cdot \kappa_Q)\big\}^\beta. \end{aligned} \tag{24}$$

Now let $x_r \in \mathbb{R}^d$ denote the point where the line segment joining x and 0 meets the boundary of the convex set $\mathcal{C}_r := [-r\kappa_Q, \infty)^{d_Q} \times \{0\}^{d-d_Q} \subseteq \mathbb{R}^d$, and note that $\|x_r\| \ge r \cdot \kappa_Q$. Observe that z_r is the Euclidean projection of x onto $\mathcal{C}_r$. Hence

$$\|x - z_r\| + r \cdot \kappa_Q \le \|x - x_r\| + \|x_r\| = \|x\| \le \|x - x'\|. \tag{25}$$

The combination of (24) and (25) establishes the desired property in the second case.

Finally, if $x, x' \in \mathcal{T}_{q,r,d_P}$ with $x \neq x'$, then

$$|\eta^\circ_{\epsilon,q,r,\sigma}(x) - \eta^\circ_{\epsilon,q,r,\sigma}(x')| \leq 3\epsilon \leq \left(\frac{r \cdot \kappa_P}{q}\right)^\beta \leq \|x - x'\|^\beta.$$

To prove the second part of the lemma, suppose first that $x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}$. Then, since $\|x - z_r\| \leq 1$ and $\epsilon \in (0, 1/8]$, we must have $\eta^\circ_{\epsilon,q,r,\sigma}(x) \in [0,1]$. On the other hand, if $x \in \mathcal{T}_{q,r,d_P}$, then $\eta^\circ_{\epsilon,q,r,\sigma}(x) \in \{1/2 - 2\epsilon, 1/2 - \epsilon, 1/2 + \epsilon\} \subseteq [0,1]$. □

COROLLARY 23. *Let $q \in \mathbb{N}$, $r > 0$, $\beta \in (0,1]$, $\sigma = (\sigma_t)_{t=1}^{q^{d_Q}} \in \{-1,1\}^{q^{d_Q}}$ and $\epsilon \in (0, 1/8 \wedge (1/6) \cdot (r \cdot \kappa_P/q)^\beta]$. Then there exists a function $\eta_{\epsilon,q,r,\sigma} : \mathbb{R}^d \to [0,1]$ such that*

$$\eta_{\epsilon,q,r,\sigma}(x) := \begin{cases} \frac{1}{2} - 2\epsilon - \frac{1}{4}\|x - z_r\|^\beta & \text{if } x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}, \\ \frac{1}{2} + \sigma_t \cdot \epsilon & \text{if } x = x_t^{q,r} \text{ with } t \leq q^{d_Q}, \\ \frac{1}{2} - 2\epsilon & \text{if } x = x_t^{q,r} \text{ with } q^{d_Q} < t \leq q^{d_P}, \end{cases} \tag{26}$$

and $|\eta_{\epsilon,q,r,\sigma}(x) - \eta_{\epsilon,q,r,\sigma}(x')| \leq \|x - x'\|^\beta$ for all $x, x' \in \mathbb{R}^d$. In particular, Assumption 4 holds for the regression function $\eta_Q = \eta_{\epsilon,q,r,\sigma}$ with $C_S = 1$.

PROOF. By Lemma 22, the function $\eta^\circ_{\epsilon,q,r,\sigma} : [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q} \cup \mathcal{T}_{q,r,d_P} \to [0,1]$ is Hölder continuous with exponent β and constant 1 on its domain. By McShane's extension theorem McShane (1934), Corollary 1, there exists an extension $\eta'_{\epsilon,q,r,\sigma} : \mathbb{R}^d \to \mathbb{R}$ which is Hölder continuous with exponent β and constant 1, and satisfies $\eta'_{\epsilon,q,r,\sigma}(x) = \eta^\circ_{\epsilon,q,r,\sigma}$ for $x \in [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q} \cup \mathcal{T}_{q,r,d_P}$. The function $\eta_{\epsilon,q,r,\sigma} : \mathbb{R}^d \to [0,1]$ given by $\eta_{\epsilon,q,r,\sigma}(x) := \{\eta'_{\epsilon,q,r,\sigma}(x) \vee 0\} \wedge 1$ has the desired properties. □

LEMMA 24. *Let $q \in \mathbb{N}$, $r > 0$, $\beta \in (0,1]$, $\sigma = (\sigma_t)_{t=1}^{q^{d_Q}} \in \{-1,1\}^{q^{d_Q}}$, $\epsilon \in (0, 1/8 \wedge (1/6) \cdot (r \cdot \kappa_P/q)^\beta]$, $C_M \geq 1 + 2^{2d_Q/\beta} d_Q^{d_Q/2} V_{d_Q}$, $\alpha \in [0, d_Q/\beta]$ and $w_Q \in [0, (1/2) \wedge \epsilon^\alpha]$. Then Assumption 3 holds whenever Q has marginal $\mu_Q = \mu_{q,r,w_Q,d_Q}$ and regression function $\eta_Q = \eta_{\epsilon,q,r,\sigma}$.*

PROOF. Without loss of generality, take $\zeta < 1$. First, suppose $\zeta \geq \epsilon$. By (26), if $x \in \mathrm{supp}(\mu_Q) \setminus \mathcal{T}_{q,r,d_Q}$ and $|\eta_Q(x) - 1/2| < \zeta$, then $\|x - z_r\| \leq (4\zeta)^{1/\beta}$. As shorthand, we write $B := B_{(4\zeta)^{1/\beta}}(z_r)$, so that $B_{d_Q} = \{(x_1, \ldots, x_{d_Q}) : (x_1, \ldots, x_{d_Q}, 0, \ldots, 0) \in B\}$. Hence,

$$\begin{aligned} \mu_Q(\{x \in \mathbb{R}^d : |\eta_Q(x) - 1/2| < \zeta\}) &\leq \mu_Q(\mathcal{T}_{q,r,d_Q}) + \mu_Q(B) \\ &\leq w_Q + \kappa_Q^{-d_Q} \cdot \mathcal{L}_{d_Q}(B_{d_Q} \cap [-\kappa_Q(1+r), -r\kappa_Q]^{d_Q}) \\ &\leq \epsilon^\alpha + (2\kappa_Q)^{-d_Q} V_{d_Q} (4\zeta)^{d_Q/\beta} \leq C_M \cdot \zeta^\alpha. \end{aligned}$$

On the other hand, if $\zeta < \epsilon$, then $\mu_Q(\{x \in \mathbb{R}^d : |\eta_Q(x) - 1/2| < \zeta\}) = 0 \leq C_M \cdot \zeta^\alpha$, as required. □

6.4. *Difficulty of estimating the decision tree function.* Lemma 25 below provides an initial minimax lower bound that arises from the difficulty of estimating the decision tree function. The proof will involve the marginal distributions μ_P and μ_Q constructed in Section 6.2, the family of target regression functions constructed in Section 6.3, and will contain a description of the construction of the corresponding family of source regression functions that is appropriate for this lower bound. Recall the definition of $B^{\mathrm{L}}_{n_Q}$ from Theorem 1.

LEMMA 25. *Fix $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$ with $\alpha\beta \leq d_Q$, $\gamma_P(1-\gamma_Q) \leq \gamma_Q$ and $C_{\mathrm{M}} \geq 1 + 2^{2d_Q/\beta} d_Q^{d_Q/2} V_{d_Q}$. Then there exists $c_{\theta,0} > 0$, depending only on θ, such that*

$$\inf_{\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}} \sup_{(P,Q) \in \mathcal{P}_{\theta^\sharp}} \mathbb{E}\{\mathcal{E}(\hat{f})\} \geq c_{\theta,0} \biggl\{ \biggl(\frac{L^*}{n_Q}\biggr)^{\frac{1+\alpha}{2+\alpha}} \wedge B^{\mathrm{L}}_{n_Q} \wedge (1-\phi)^{1+\alpha} \biggr\}. \tag{27}$$

PROOF. Our goal is to define a particular instantiation of the construction in Lemma 16, which requires us to specify $m \in \mathbb{N}$, $(x_t)_{t \in [m]} \in (\mathbb{R}^d)^m$, $\epsilon_P, \epsilon_Q \in [0, 1/4]$, $u_P, u_Q \in [0, 1/m]$, v_P, $v_Q \in [0,1]$, regression functions $\eta_P^\sigma : \mathbb{R}^d \to [0,1]$, $\eta_Q^\sigma : \mathbb{R}^d \to [0,1]$ for $\sigma \in \Sigma = \{-1,1\}^m$, and marginals μ_P, μ_Q on $\mathbb{R}^d$.

To this end, we first define some intermediate quantities that depend only on θ. Let

$$\rho \equiv \rho_\theta := \frac{\gamma_Q(d_Q - \alpha\beta) + \alpha\beta}{\gamma_Q(2\beta + d_Q) + \alpha\beta}; \qquad a_1 \equiv a_{1,\theta} := 2^{-3d_P(\gamma_P \vee \gamma_Q)} \wedge \bigl(1 - C_{P,Q}^{-1/(\gamma_P \wedge \gamma_Q)}\bigr);$$

$$\rho_1 \equiv \rho_{1,\theta} := \frac{d_Q}{\beta(2+\alpha)} + \frac{\alpha}{\gamma_P(2+\alpha)} + 1; \qquad b_1 \equiv b_{1,\theta} := \frac{2^{5\rho_1} \kappa_P^{d_P}}{8^{d_P} \cdot 6^{d_Q/\beta} \cdot 2^{5+d_P-d_Q}};$$

$$\lambda \equiv \lambda_\theta := \frac{\alpha + 2\gamma_Q + d_Q\gamma_Q/\beta}{2+\alpha}; \qquad a_2 \equiv a_{2,\theta} := 2^{5(\lambda - \gamma_Q)} \cdot 2^{-3d_Q\gamma_Q} \cdot \kappa_P^{d_Q\gamma_Q} \cdot 6^{-d_Q\gamma_Q/\beta}.$$

Now let $a \equiv a_\theta := \min\{(a_1 b_1)^{1/\rho_1}, 2^5 a_1^{(2+\alpha)/\alpha}, a_2^{1/\lambda}, 2^{-(1+3\alpha)}, 2^{4-2/\alpha}\} > 0$. This allows us to define

$$q = \lfloor \min(a n_Q^\rho, L^*)^{1/d_Q} \rfloor.$$

Observe that $q \geq 1$ whenever $n_Q \geq a^{-1/\rho}$, and we will therefore first prove the desired lower bound in this case. Now let $m = q^{d_Q}$, let $\epsilon_P = 0$, let

$$\epsilon \equiv \epsilon_Q = \min\biggl\{ \biggl(\frac{m}{2^5 n_Q}\biggr)^{1/(2+\alpha)}, \frac{1-\phi}{4} \biggr\}, \tag{28}$$

let $w_Q = \epsilon^\alpha$, let $u_Q = w_Q/m$, let $r = (6\epsilon)^{1/\beta} q/\kappa_P$, let $w_P = (8r)^{d_P} N_{q,r,d_P} q^{-d_P} w_Q^{1/\gamma_P}$ and let $u_P = w_P/N_{q,r,d_P}$. Set $x_t = x_t^{q,r}$ for $t \in [m]$, where $x_t^{q,r}$ is defined at the beginning of Section 6.2. Further, let $v_P = 1/2 + \epsilon$ and $v_Q = 1/2$. Recalling (23), we will take the marginal distributions to be $\mu_P = \mu_{q,r,w_P,d_P}$ and $\mu_Q = \mu_{q,r,w_Q,d_Q}$, noting that by our choice of the first three terms in the minimum defining a, the conditions of Corollary 21 hold (this uses the hypothesis that $\gamma_P(1-\gamma_Q) \leq \gamma_Q$), and this corollary then tells us that Assumption 2 is satisfied. For $\sigma \in \Sigma$, let $\eta_Q^\sigma = \eta_{\epsilon,q,r,\sigma}$ as defined in Corollary 23, noting that the fourth term in the minimum defining a ensures that the conditions of this corollary hold and, therefore, that each η_Q^σ satisfies Assumption 4. Moreover, the final term in the minimum defining a, together with the hypotheses of the current lemma, guarantees that the conditions of Lemma 24 hold, so the distribution Q^σ on $\mathbb{R}^d \times \{0,1\}$ with marginal distribution μ_Q and regression function η_Q^σ satisfies Assumption 3.

It remains to define η_P^σ for $\sigma \in \Sigma$, and to do this, we first define a decision tree partition and a family of transfer functions. Recalling the definition of $\mathcal{T}_{q,r,d_Q}$ from the beginning of Section 6.2, let $\{\mathcal{X}_1^*, \dots, \mathcal{X}_{L^*}^*\} \in \mathbb{T}_{L^*}$ be such that $\mathcal{X}_\ell^* \cap \mathcal{T}_{q,r,d_Q} = \{x_\ell^{q,r}\}$ for each $\ell \in [m] \subseteq [L^*]$ (the fact that $m \le L^*$ follows from our definition of q). Define $h : [0,1] \to [0,1]$ by

$$h(z) := \begin{cases} z & \text{if } z \in [0, 1/2 - 2\epsilon], \\ 3z + 4\epsilon - 1 & \text{if } z \in [1/2 - 2\epsilon, 1/2 - \epsilon], \\ \dfrac{(1-2\epsilon)z + 4\epsilon}{1+2\epsilon} & \text{if } z \in [1/2 - \epsilon, 1]. \end{cases}$$

Observe that

$$\frac{h(z) - 1/2}{z - 1/2} \ge \frac{1-2\epsilon}{1+2\epsilon} \ge 1 - 4\epsilon \ge \phi \tag{29}$$

for $z \in [0,1] \setminus 1/2$, where the final bound follows from the second term in the minimum defining ϵ. For $\sigma = (\sigma_1, \dots, \sigma_m) \in \Sigma$ and $\ell \in [m]$, define $g_\ell^\sigma : [0,1] \to [0,1]$ by

$$g_\ell^\sigma(z) := \begin{cases} z & \text{if } \sigma_\ell = 1, \\ h(z) & \text{if } \sigma_\ell = -1, \end{cases}$$

and for $\ell \in \{m+1, \dots, L^*\}$, let $g_\ell^\sigma(z) := z$. We can now set $\eta_P^\sigma = g_\ell^\sigma \circ \eta_Q^\sigma$ on $\mathcal{X}_\ell^*$, and note that by (29), Assumption 1 holds for each transfer function g_ℓ^σ.

We are now in a position to verify that our constructed marginals and family of source and target regression functions satisfy the conditions of Lemma 16 with $\mathcal{P} = \mathcal{P}_{\theta^\sharp}$. Condition (i) holds because $\epsilon_P = 0$ and $2^5 n_Q u_Q \epsilon_Q^2 = 2^5 n_Q \epsilon_Q^{2+\alpha} \le 1$ by definition of ϵ_Q in (28). The verification of Condition (ii) again uses the fact that $\epsilon_P = 0$, and also that $v_Q = 1/2$. Condition (iii) follows immediately by definition of μ_P, μ_Q, x_t, u_P and u_Q. The second part of Condition (iv) holds by definition of η_Q^σ, together with the definitions of $\eta_{\epsilon,q,r,\sigma}$ in (23), v_Q and ϵ_Q. The first part of this condition uses this second part, together with the facts that $v_P = 1/2 + \epsilon$ and $h(1/2 - \epsilon) = 1/2 + \epsilon$. Finally, Condition (v) holds because the restriction of $\eta_{\epsilon,q,r,\sigma}$ in (26) to $[-\kappa_Q(1+r), -r\kappa_Q]^{d_Q} \times \{0\}^{d-d_Q}$ does not depend on σ, and because g_ℓ^σ is the identity function for $\ell \in \{m+1, \dots, L^*\}$.

Writing $c'_{\theta,0} := a^{\frac{1+\alpha}{2+\alpha}}/2^{(6+d_Q)(1+\alpha)}$, we conclude from Lemma 16 that

$$\inf_{\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}} \sup_{(P,Q) \in \mathcal{P}} \mathbb{E}\{\mathcal{E}(\hat{f})\} \ge \frac{m u_Q \epsilon_Q}{2} \ge c'_{\theta,0} \biggl\{ \biggl(\frac{L^*}{n_Q} \biggr)^{\frac{1+\alpha}{2+\alpha}} \wedge B_{n_Q}^{\mathrm{L}} \wedge (1-\phi)^{1+\alpha} \biggr\} \tag{30}$$

for $n_Q \ge a^{-1/\rho}$. But the left-hand side of (30) is decreasing in n_Q, so the full result holds on setting $c_{\theta,0} := c'_{\theta,0} \cdot 2^{-(1+\alpha)} a^{\frac{1+\alpha}{\rho(2+\alpha)}}$. □

6.5. *Difficulty of estimating the source regression function and completion of the proof of the lower bound in Theorem 1.*

LEMMA 26. *Fix $\theta^\sharp = (\Delta, \phi, L^*, \theta) \in \Theta^\sharp$ with $\alpha\beta \le d_Q$, $\gamma_P(1-\gamma_Q) \le \gamma_Q$ and $C_{\mathrm{M}} \ge 1 + 2^{2d_Q/\beta} d_Q^{d_Q/2} V_{d_Q}$. Then there exists $c_{\theta,1} > 0$, depending only on θ, such that*

$$\inf_{\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}} \sup_{(P,Q) \in \mathcal{P}_{\theta^\sharp}} \mathbb{E}\{\mathcal{E}(\hat{f})\} \ge c_{\theta,1} \min\biggl\{ \biggl(\frac{1}{\phi^2 \cdot n_P} \biggr)^{\frac{\beta\gamma_P(1+\alpha)}{\gamma_P(2\beta+d_P)+\alpha\beta}} + \biggl(\frac{\Delta}{\phi} \biggr)^{1+\alpha}, B_{n_Q}^{\mathrm{L}}, 1 \biggr\}.$$

PROOF. Recalling the definition of $a_1 = 2^{-3d_P(\gamma_P \vee \gamma_Q)} \wedge (1 - C_{P,Q}^{-1/(\gamma_P \wedge \gamma_Q)})$ from the proof of Lemma 25, we define $a_3 := a_1^{1/d_Q} \cdot 6^{-1/\beta} \cdot (\kappa_P/16)^{d_P/d_Q}$ and let

$$a_4 := \min\Big\{a_3^{\frac{\beta d_Q \gamma_Q}{\gamma_Q(d_Q - \alpha\beta)+\alpha\beta}}, \frac{1}{8}, a_1^{1/\alpha}, 2^{-1/\alpha}, \big(\kappa_P^{d_P} \cdot 2^{-(6+3d_P)} \cdot 6^{-d_P/\beta}\big)^{\frac{\beta\gamma_P}{\gamma_P(2\beta+d_P)+\alpha\beta}},$$
$$\big(a_3^{d_Q} \cdot 2^{-(6+d_Q)}\big)^{\frac{\beta\gamma_Q}{\gamma_Q(2\beta+d_Q)+\alpha\beta}}\Big\},$$

$$\epsilon \equiv \epsilon_Q := a_4 \cdot \min\bigg(\max\bigg\{\Big(\frac{1}{\phi^2 \cdot n_P}\Big)^{\frac{\beta\gamma_P}{\gamma_P(2\beta+d_P)+\alpha\beta}}, \frac{\Delta}{\phi}\bigg\}, \Big(\frac{1}{n_Q}\Big)^{\frac{\beta\gamma_Q}{\gamma_Q(2\beta+d_Q)+\alpha\beta}}\bigg).$$

Take $q := \lfloor a_3 \cdot \epsilon^{-\frac{\gamma_Q(d_Q-\alpha\beta)+\alpha\beta}{\beta d_Q \gamma_Q}} \rfloor$. We will initially assume that $n_Q \geq 1$, which means that $\epsilon \leq a_3^{\frac{\beta d_Q \gamma_Q}{\gamma_Q(d_Q-\alpha\beta)+\alpha\beta}}$ so $q \geq 1$. Further define $m = q^{d_Q}$, let $\epsilon_P = (\phi \cdot \epsilon - \Delta) \vee 0$, let $w_Q = \epsilon^\alpha$, let $u_Q = w_Q/m$, let $r = (6\epsilon)^{1/\beta} q/\kappa_P$, let $w_P = (8r)^{d_P} N_{q,r,d_P} q^{-d_P} w_Q^{1/\gamma_P}$, let $u_P = w_P/N_{q,r,d_P}$ and let $v_P = v_Q = 1/2$. Set $x_t = x_t^{q,r}$ for $t \in [m]$, where $x_t^{q,r}$ is defined at the beginning of Section 6.2. We will take the marginal distributions to be $\mu_P = \mu_{q,r,w_P,d_P}$ and $\mu_Q = \mu_{q,r,w_Q,d_Q}$, which as in the proof of Lemma 25, satisfy the conditions of Corollary 21, and hence Assumption 2. For $\sigma \in \Sigma = \{-1, 1\}^m$, let $\eta_Q^\sigma = \eta_{\epsilon,q,r,\sigma}$. Then, as in the proof of Lemma 25, the conditions of Corollary 23 and Lemma 24 hold, so Assumptions 4 and 3 are also satisfied. For $\delta \in (0, \phi/2]$, define $h_{\phi,\delta} : [0, 1] \to [0, 1]$ by

$$h_{\phi,\delta}(z) := \begin{cases} \phi \cdot (z - 1/2) + 1/2 + \delta & \text{if } z \in [0, 1/2 - \delta/\phi], \\ 1/2 & \text{if } z \in [1/2 - \delta/\phi, 1/2 + \delta/\phi], \\ \phi \cdot (z - 1/2) + 1/2 - \delta & \text{if } z \in [1/2 + \delta/\phi, 1], \end{cases}$$

and for $\delta > \phi/2$, let $h_{\phi,\delta}(\cdot) := 1/2$. For $\sigma \in \Sigma$, we take $\eta_P^\sigma = h_{\phi,\Delta} \circ \eta_Q^\sigma$, and $g_\ell := h_{\phi,0}$ for $\ell \in [L^*]$. Note that these definitions ensure that each g_ℓ satisfies (5), and $|\eta_P^\sigma(x) - g_\ell(\eta_Q^\sigma(x))| \leq \|h_{\phi,\Delta} - h_{\phi,0}\|_\infty \leq \Delta$ for $x \in \mathcal{X}_\ell$, so Assumption 1 holds.

Finally, similar (but slightly simpler) arguments to those used in the proof of Lemma 25 verify that the assumptions of Lemma 16 hold with $\mathcal{P} = \mathcal{P}_{\theta\sharp}$, so writing $c_{\theta,1} := a_4^{1+\alpha}/4$, we conclude from Lemma 16 that

$$\inf_{\hat{f} \in \hat{\mathcal{F}}_{n_P,n_Q}} \sup_{(P,Q)\in\mathcal{P}} \mathbb{E}\{\mathcal{E}(\hat{f})\} \geq \frac{m u_Q \epsilon_Q}{2} \tag{31}$$
$$\geq c_{\theta,1} \min\bigg\{\Big(\frac{1}{\phi^2 \cdot n_P}\Big)^{\frac{\beta\gamma_P(1+\alpha)}{\gamma_P(2\beta+d_P)+\alpha\beta}} + \Big(\frac{\Delta}{\phi}\Big)^{1+\alpha}, B_{n_Q}^{\mathrm{L}}\bigg\}$$

whenever $n_Q \geq 1$. But the left-hand side of (31) is decreasing in n_Q, so the full result follows. □

PROOF OF THE LOWER BOUND IN THEOREM 1. This follows immediately from Lemmas 25 and 26. □

Acknowledgements. We are very grateful for the constructive feedback from the anonymous reviewers, which helped to improve the paper.

Funding. The second author was supported by Engineering and Physical Sciences Research Council (EPSRC) New Investigator Award EP/V002694/1. The third author was supported by Engineering and Physical Sciences Research Council (EPSRC) Programme grant EP/N031938/1 and EPSRC Fellowship EP/P031447/1.

SUPPLEMENTARY MATERIAL

Supplement to "Adaptive transfer learning" (DOI: 10.1214/21-AOS2102SUPP; .pdf). Supplementary information.

REFERENCES

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

BEN-DAVID, S., BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F. and VAUGHAN, J. W. (2010a). A theory of learning from different domains. *Mach. Learn.* **79** 151–175. MR3108150 https://doi.org/10.1007/s10994-009-5152-4

BEN-DAVID, S., LU, T., LUU, T. and PÁL, D. (2010b). Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics* 129–136.

BERNSTEIN, S. (1924). On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine*, *Sect. Math* **1** 38–49.

BIAU, G. and DEVROYE, L. (2013). Cellular tree classifiers. *Electron. J. Stat.* **7** 1875–1912. MR3084675 https://doi.org/10.1214/13-EJS829

BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method*. *Springer Series in the Data Sciences*. Springer, Cham. MR3445317 https://doi.org/10.1007/978-3-319-25388-6

BLANCHARD, G., FLASKA, M., HANDY, G., POZZI, S. and SCOTT, C. (2016). Classification with asymmetric label noise: Consistency and maximal denoising. *Electron. J. Stat.* **10** 2780–2824. MR3549019 https://doi.org/10.1214/16-EJS1193

BLANCHARD, G., DESHMUKH, A. A., DOGAN, U., LEE, G. and SCOTT, C. (2017). Domain generalization by marginal transfer learning. Preprint. Available at arXiv:1711.07910.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128. MR4206671 https://doi.org/10.1214/20-AOS1949

CANDELA, J. Q., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press, Cambridge.

CANNINGS, T. I., BERRETT, T. B. and SAMWORTH, R. J. (2020). Local nearest neighbour classification with applications to semi-supervised learning. *Ann. Statist.* **48** 1789–1814. MR4124344 https://doi.org/10.1214/19-AOS1868

CANNINGS, T. I., FAN, Y. and SAMWORTH, R. J. (2020). Classification with imperfect training labels. *Biometrika* **107** 311–330. MR4108933 https://doi.org/10.1093/biomet/asaa011

CARUANA, R. (1997). Multitask learning. *Mach. Learn.* **28** 41–75.

CHAUDHURI, K. and DASGUPTA, S. (2014). Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems* 3437–3445.

CHRISTIANSEN, R., PFISTER, N., JAKOBSEN, M. E., GNECCO, N. and PETERS, J. (2020). The difficult task of distribution generalization in nonlinear models. Preprint. Available at arXiv:2006.07433.

CORTES, C., MOHRI, M. and MUÑOZ MEDINA, A. (2019). Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.* **20** 1–30. MR3911408

COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.

EVANS, L. C. and GARIEPY, R. F. (2015). *Measure Theory and Fine Properties of Functions*, Revised ed. *Textbooks in Mathematics*. CRC Press, Boca Raton, FL. MR3409135

FIX, E. and HODGES, J. L. (1951). *Discriminatory Analysis*: *Nonparametric Discrimination*, *Consistency Properties*. USAF School of Aviation Medicine.

FRÉNAY, B. and VERLEYSEN, M. (2014). Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **25** 845–869.

GADAT, S., KLEIN, T. and MARTEAU, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.* **44** 982–1009. MR3485951 https://doi.org/10.1214/15-AOS1395

GERMAIN, P., LACASSE, A., LAVIOLETTE, F., MARCHAND, M. and ROY, J.-F. (2015). Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *J. Mach. Learn. Res.* **16** 787–860. MR3361305

GRETTON, A., SMOLA, A., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. and SCHÖLKOPF, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning* **3** 5.

HALL, P., PARK, B. U. and SAMWORTH, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.* **36** 2135–2152. MR2458182 https://doi.org/10.1214/07-AOS537

HANNEKE, S. and KPOTUFE, S. (2019). On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems* 9867–9877.

HARDY, G. H. and LITTLEWOOD, J. E. (1930). A maximal theorem with function-theoretic applications. *Acta Math*. **54** 81–116. MR1555303 https://doi.org/10.1007/BF02547518

KIM, A. K. H. (2020). Obtaining minimax lower bounds: A review. *J. Korean Statist. Soc*. **49** 673–701. MR4147680 https://doi.org/10.1007/s42952-019-00027-7

KPOTUFE, S. and MARTINET, G. (2018). Marginal singularity, and the benefits of labels in covariate-shift. In *Proceedings of the* 31*st Conference on Learning Theory* **75** 1882–1886.

KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inf. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248

LEDRAPPIER, F. and YOUNG, L.-S. (1985). The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula. *Ann. of Math*. (2) **122** 509–539. MR0819556 https://doi.org/10.2307/1971328

LIPTON, Z. C., WANG, Y.-X. and SMOLA, A. (2018). Detecting and correcting for label shift with black box predictors. In *Proceedings of the* 35*th International Conference on Machine Learning* **80**.

MAITY, S., SUN, Y. and BANERJEE, M. (2020). Minimax optimal approaches to the label shift problem. Preprint. Available at arXiv:2003.10443.

MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist*. **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of the* 22*nd Annual Conference on Learning Theory*.

MAURER, A., PONTIL, M. and ROMERA-PAREDES, B. (2016). The benefit of multitask representation learning. *J. Mach. Learn. Res*. **17** Paper No. 81. MR3517104

MCDIARMID, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*. *Algorithms Combin*. **16** 195–248. Springer, Berlin. MR1678578 https://doi.org/10.1007/978-3-662-12788-9_6

MCSHANE, E. J. (1934). Extension of range of functions. *Bull. Amer. Math. Soc*. **40** 837–842. MR1562984 https://doi.org/10.1090/S0002-9904-1934-05978-0

MENON, A. K., VAN ROOYEN, B. and NATARAJAN, N. (2018). Learning from binary labels with instance-dependent noise. *Mach. Learn*. **107** 1561–1595. MR3835279 https://doi.org/10.1007/s10994-018-5715-3

MOHRI, M. and MUÑOZ MEDINA, A. (2012). New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **7568** 124–138. Springer, Heidelberg. MR3042886 https://doi.org/10.1007/978-3-642-34106-9_13

PAN, S. J. and YANG, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. **22** 1345–1359.

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist*. **23** 855–881. MR1345204 https://doi.org/10.1214/aos/1176324626

REEVE, H. W. J. and BROWN, G. (2017). Minimax rates for cost-sensitive learning on manifolds with approximate nearest neighbours. In *International Conference on Algorithmic Learning Theory*. *Proc. Mach. Learn. Res*. (*PMLR*) **76** 11–56. Proceedings of Machine Learning Research PMLR. MR3776687

REEVE, H. W., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Supplement to "Adaptive transfer learning." https://doi.org/10.1214/21-AOS2102SUPP

REEVE, H. W. J. and KABÁN, A. (2019a). Classification with unknown class-conditional label noise on non-compact feature spaces. In *Proceedings of the* 32*nd Annual Conference on Learning Theory* **99** 1–28.

REEVE, H. W. J. and KABÁN, A. (2019b). Fast rates for a kNN classifier robust to unknown asymmetric label noise. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.) **97** 5401–5409.

SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist*. **40** 2733–2763. MR3097618 https://doi.org/10.1214/12-AOS1049

SCOTT, C. (2019). A generalized Neyman–Pearson criterion for optimal domain adaption. In *Algorithmic Learning Theory* 2019. *Proc. Mach. Learn. Res*. (*PMLR*) **98** 1–24. Proceedings of Machine Learning Research PMLR. MR3932867

SCOTT, C. and NOWAK, R. D. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inf. Theory* **52** 1335–1353. MR2241192 https://doi.org/10.1109/TIT.2006.871056

SCOTT, C. and ZHANG, J. (2019). Learning from multiple corrupted sources, with application to learning from label proportions. Preprint. Available at arXiv:1910.04665.

SINHA, A., NAMKOONG, H. and DUCHI, J. C. (2018). Certifying some distributional robustness with principled adversarial training. In *Sixth International Conference on Learning Representations* (*ICLR* 2018).

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist*. **5** 595–645. MR0443204

STORKEY, A. (2009). When training and test sets are different: Characterizing learning transfer. In *Dataset Shift in Machine Learning* 3–28. MIT Press, Cambridge.

SUGIYAMA, M., SUZUKI, T. and KANAMORI, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge Univ. Press, Cambridge. MR2895762 https://doi.org/10.1017/CBO9781139035613

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. Preprint. Available at arXiv:1503.06388.

WEICHWALD, S. and PETERS, J. (2021). Distributional robustness as a guiding principle for causality in cognitive neuroscience. *J. Cogn. Neurosci.* **33** 226–247.

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *Journal of Big Data* **3** 9.

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963

ZHANG, K., SCHÖLKOPF, B., MUANDET, K. and WANG, Z. (2013). Domain adaptation under target and conditional shift. In *Proceedings of the* 30*th International Conference on Machine Learning* **28** 819–827.