Proceedings of Machine Learning Research vol 195:1–32, 2023 36th Annual Conference on Learning Theory

Limits of Model Selection under Transfer Learning

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Samory Kpotufe SAMORY@COLUMBIA.EDU
Columbia University

Yasaman Mahdaviyeh YM2898@COLUMBIA.EDU
Columbia University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Theoretical studies on *transfer learning* (or *domain adaptation*) have so far focused on situations with a known hypothesis class or *model*; however in practice, some amount of model selection is usually involved, often appearing under the umbrella term of *hyperparameter-tuning*: for example, one may think of the problem of *tuning* for the right neural network architecture towards a target task, while leveraging data from a related *source* task.

In addition to the usual tradeoffs on approximation vs. estimation errors involved in model selection, this problem brings in a new complexity term, namely, the *transfer distance* between source and target distributions, which is known to vary with the choice of hypothesis class.

We present a first study of this problem, focused on classification. Remarkably, the analysis reveals that *adaptive rates*, i.e., those achievable with no distributional information, can be arbitrarily slower than *oracle rates*, i.e., when given knowledge on *distances*.

Keywords: Transfer Learning, Domain adaptation, Model Selection, Lepski's Method.

1. Introduction

Domain adaptation or Transfer learning concern settings where data from a *source* distribution P is to be leveraged to improve learning on a target distribution Q where perhaps less data is available. While this problem has received much renewed attention of late, theoretical studies have focused on settings where a suitable hypothesis (or model) class $\mathcal{H}$ is already known. However, this is rarely the case in practice where some amount of model selection is required, as often referred to as *hyperparameter tuning*: one wishes, e.g., to tune for the right architecture with neural networks, a suitable polynomial degree in regression, or an appropriate kernel for kernel machines, all while leveraging both source and target data. Importantly, as target data is often limited in these settings, it ideally should not be used alone to drive model selection, even though it is a priori unclear how to leverage the source data.

We present a first study of this problem, in the context of classification, under a simple formalism where we assume a hierarchy of models $\{\mathcal{H}_i\}, \mathcal{H}_i \subset \mathcal{H}_{i+1}$, each with known complexity d_i (here VC-dimension); the problem is then to try and understand the achievable target Q-risk in modern transfer settings with access to both source and target data, as opposed to just target data. We note however that our analysis allows for no target data, as in fact we have no restriction on data sizes from either source nor target.

To establish a baseline performance, assume the hierarchy $\{\mathcal{H}_i\}$ admits a global Q-risk minimizer h_Q^* from an unknown model $\mathcal{H}_{i_Q^*} \in \{\mathcal{H}_i\}$. Then it is known that, using n_Q data from Q, an excess risk $\mathcal{E}_Q(\hat{h}) \doteq \mathcal{E}_Q(\hat{h}; h^*_Q) \lesssim \sqrt{d_{i^*_Q}/n_Q}$ is achievable without prior knowledge of $\mathcal{H}_{i^*_Q}$, e.g., via structural risk minimization (SRM), a.k.a., complexity regularization, which essentially tradeoff *estimation error* $\sqrt{d_i/n_Q}$, and *approximation error* $\min_{h\in\mathcal{H}_i} \mathcal{E}(h, h^*_Q)$ over models $\{\mathcal{H}_i\}$.

Now, model selection in a transfer scenario, i.e., given related source data from P, involves an additional tradeoff parameter: the *distance* or information that P yields on Q, which is now well understood to be tied to the choice of hypothesis class $\mathcal{H}_i$. Early notions of distance $P \to Q$, e.g., from seminal works of Mansour et al. (2009a); Ben-David et al. (2010) already formalize the idea that the differences between P and Q are only relevant in regions of space in line with $\mathcal{H}$, e.g., disagreement regions between given hypotheses in $\mathcal{H}$. In other words, while a model choice $\mathcal{H}$ out of the hierarchy $\{\mathcal{H}_i\}$ may balance estimation and approximation errors, it may fail to maximally leverage the data from P if it induces a large *distance* $P \to Q$.

As the distances $P \to Q$ induced over models in $\{\mathcal{H}_i\}$ are a priori unknown (however formalized), our analysis especially distinguishes betwen *adaptive* model selection rates—i.e., rates achievable from P and Q samples alone without distributional information—and usual minimax *oracle* rates. Remarkably, unlike in usual model selection, these can be significantly different.

Main Results. For a fixed class $\mathcal{H}$, we adopt a recent notion of *distance* $P \to Q$ from (Hanneke and Kpotufe, 2019) comprised of two components: (1) the excess risk $\mathcal{E}_Q(h^*_P)$ of a risk minimizer h^*_P under P, and (2) a *transfer-exponent* ρ which essentially measures the effective sample size contributed by P to the target problem Q. Thus suppose access to n_P samples from P and n_Q samples from Q, the following upper-bound was shown to be achievable adaptively:

$$\mathcal{E}_Q(\hat{h}) \lesssim \min\left\{(d/n_P)^{1/2\rho} + \mathcal{E}_Q(h^*_P)\ ;\ (d/n_Q)^{1/2}\right\},\ \text{where } d \text{ is the VC dimension of } \mathcal{H}. \quad (1)$$

For sanity check, note that (1) is of order $(d/(n_P + n_Q))^{1/2}$ when $P = Q$, i.e., $\rho = 1$, $\mathcal{E}_Q(h^*_P) = 0$. Also notice that the rate is faster with smaller ρ and $\mathcal{E}_Q(h^*_P)$.

Now, if we knew the above rate to be tight in general, we then get a first sense of the best rates we might expect for any fixed model choice $\mathcal{H}_i$ out of the hierarchy.

• *Tightness of* (1). As a first basic result, we show that the above adaptive rate on a fixed choice $\mathcal{H}$, admits matching lower-bounds over any parameter value (Theorem 1). This complements a lower-bound of (Hanneke and Kpotufe, 2019) which only holds for $\mathcal{E}_Q(h^*_P) = 0$. This is especially important in our setting in order to cover a rich variety of situations.

• *Adaptive Upper-Bounds and Speedups.* Now suppose that each $\mathcal{H}_i$ in the hierarchy admits *transfer distance* $(\rho_i, \mathcal{E}_Q(h^*_{P,i}))$, a priori unknown. Together with known class complexity d_i, and sample sizes n_p, n_Q, these distances induce subtle tradeoffs on model choices $\mathcal{H}_i$ for the Q task.

— First, we verify through some technical examples, namely basic neural-networks, that indeed some rich set of tradeoffs are captured through the above parametrization. That is, rich combinations of $(\rho_i, \mathcal{E}_Q(h^*_{P,i}))$ emerge from the interaction between (P, Q) and nested network architectures.

— Having established the tightness of the above equation 1, and given the baseline of model selection under target, we can show (see Lemma 1) that selecting any fixed $\mathcal{H}_i$ would yield an adaptive upper-bound of

$$\mathcal{E}_Q(\hat{h}_i) \lesssim \phi(i),\ \text{where } \phi(i) \approx \min\left\{(d_i/n_P)^{1/2\rho_i} + \mathcal{E}_Q(h^*_{P,i})\ ;\ \left(d_{i^*_Q}/n_Q\right)^{1/2}\right\}.$$

Unfortunately, as we discussed in the next bullet point, no algorithm exist that can minimize $\phi(i)$ in general and achieve optimal tradeoff on *distance*. Instead, we establish the following adaptive guarantee (see Theorem 2). Suppose that $\{\mathcal{H}_i\}$ admits a global P-risk minimizer h^*_P at unknown level i^*_P; then there exists a procedure $\hat{h}$ achieving

$$\mathcal{E}_Q(\hat{h}) \lesssim \phi(i^*_P) \text{ from samples alone.}$$

In other words, the procedure automatically favors model selection under source P—at least commensurate with the unknown model $\mathcal{H}_{i^*_P}$—if P is thus informative on Q, and falls back on leveraging target data otherwise, all without prior knowledge of distributional parameters.

We emphasize that in contrast, popular SRM *approaches yield no clear such guarantee:* suppose $n_Q = 0$, SRM can only guarantee low P-risk, but no specific choice of model class.

• *Oracle Rates are Unachievable.* With knowledge of distance parameters $\left\{(\rho_i, \mathcal{E}_Q(h^*_{P,i}))\right\}$ (or at least of the ranking they induce on $\{\phi(i)\}$), an *oracle* procedure can achieve the rate $\min_i \phi(i)$, which can be arbitrarily faster than $\phi(i^*_P)$.

Interestingly, as we show in Theorem 4, no *adaptive* procedure, i.e., without such domain knowledge, can achieve a bound better than $\phi(i^*_P)$, without further structural conditions on the hierarchy $\{\mathcal{H}_i\}$, even in situations where $\min_i \phi(i) \ll \phi(i^*_P)$. This result holds even when the learner $\hat{h}$ is *improper*, i.e., when $\hat{h}$ is allowed to return a hypothesis outside of $\cup_i \mathcal{H}_i$.

Related Work. Transfer Learning has received much attention over the years, with studies, both in the context of classification and regression, considering various notions of relations between P and Q. Early works include (Ben-David et al., 2007; Crammer et al., 2008; Cortes et al., 2008; Ben-David et al., 2010; Gretton et al., 2009; Mansour et al., 2009b) which already recognize the importance of the choice of hypothesis class in quantifying the information the source P has on the target Q. These ealrier works have been refined over time, e.g., considering multiple source distributions rather than just one (Maurer et al., 2013; Pentina and Lampert, 2014; Yang et al., 2013; Maurer et al., 2016).

More recently, *assymetric* notions of discrepancy have been proposed, noting that P may have information on Q but not the other way around (Kpotufe and Martinet, 2018; Hanneke and Kpotufe, 2019; Achille et al., 2019; Mousavi Kalan et al., 2020). We adopt such a notion in this work.

Despite much of the attention on this problem, a single hypothesis class $\mathcal{H}$ has been commonly assumed. However, a separate line of work on *meta-learning* can be seen as somewhat related, as they often assume relationship between optimal predictors, often in the form of a shared *low-dimensional substructure*; these settings may be recast as learning a target hypothesis class of lower complexity (Ando and Zhang, 2005; Muandet et al., 2013; McNamara and Balcan, 2017; Arora et al., 2019; Jalali et al., 2010; Lounici et al., 2011; Negahban and Wainwright, 2011; Du et al., 2020; Tripuraneni et al., 2020). This however does not embody the full richness of model selection.

Paper Organization. We start in Section 2 with basic definitions and setup. This is followed by an overview of results in Section 3. Much of the proofs are discussed in Section 4 with more technical results relegated to the Appendix.

2. Preliminaries

2.1. Setup

Basic Definitions. Let X, Y be jointly distributed according to some measure μ (later P or Q), where X is in some domain $\mathcal{X}$ and $Y \in \mathcal{Y} \doteq \{\pm 1\}$. A *hypothesis class* or *model* is a set $\mathcal{H}$ of functions $\mathcal{X} \mapsto \mathcal{Y}$. All these objects are assumed to be measurable, so that we may consider classification risks of the form $R(h) \doteq \mathbb{E}[h(X) \neq Y]$, as measured under μ.

Definition 1 *The* **excess risk** *of a classifier, w.r.t. $\mathcal{H}$, is defined as $\mathcal{E}(h) = R(h) - \inf_{h' \in \mathcal{H}} R(h')$. Furthermore we use the notation $\mathcal{E}(h, h') \doteq R(h) - R(h') = \mathcal{E}(h) - \mathcal{E}(h')$.*

We adopt the following classical noise conditions (see e.g. (Massart and Nédélec, 2006; Koltchinskii, 2006; Bartlett et al., 2006)).

Definition 2 *Assume $R(h)$ is minimized at $h^* \in \mathcal{H}$. We say that $\mathcal{H}$ satisfies a* **Bernstein Class Condition** *(BCC), as measured under μ, with parameters (C_β, β),$C_\beta > 0$ and $\beta \in [0, 1]$, if $\forall h \in \mathcal{H}$*

$$\mathbb{P}(h \neq h^*) \leq C_\beta \cdot \mathcal{E}(h)^\beta.$$

Note that the condition trivially holds for $\beta = 0$, $C_\beta = 1$. The condition captures the hardness of the learning problem: when $\beta = 1$, which formalizes *low noise* regimes, we expect fast rates of the form n^{-1}, in terms of sample size n, while for $\beta = 0$, rates are of the more common form $n^{-1/2}$.

When h^* is not unique, BCC remains well defined (i.e., the definition is invariant to the choice of h^*), as it imposes (when $\beta > 0$) that all h^*'s differ on a set of measure 0 under the data distribution.

Transfer Setting. We consider a *source* and *target* distributions P and Q on (X, Y), where we let $\mathcal{E}_P, \mathcal{E}_Q$ denote excess-risks under P and Q. We are interested in excess risk $\mathcal{E}_Q(\hat{h})$ of classifiers trained jointly on n_P i.i.d samples from P, and n_Q i.i.d. samples from Q. Achievable such excess risks necessarily depend on the *distance* $P \to Q$ appropriately formalized.

We adopt some recent notion of *distance* from (Hanneke and Kpotufe, 2019); for ease of exposition, we make the following simplifying assumptions.

Assumption 1 *We assume for any $\mathcal{H}$ considered henceforth that $\mathcal{E}_P$ and $\mathcal{E}_Q$ are minimized in $\mathcal{H}$. We let h^*_P, h^*_Q denote any such respective risk minimizers. Furthermore, if multiple minimizers $\{h^*_P\}$ exist under P, we assume that one of them achieves $\sup_{\{h^*_P\}} \mathcal{E}_Q(h^*_P)$, and denote it h^*_P.*

The distance $P \to Q$ is then given by $\mathcal{E}_Q(h^*_P)$, and the following quantity ρ:

Definition 3 *We call $0 < \rho \leq \infty$* **transfer exponent** *from P to Q with respect to $\mathcal{H}$ if there exists $C_\rho > 0$ such that for all $h \in \mathcal{H}$,*

$$C_\rho \cdot \mathcal{E}_P(h, h^*_P)^{1/\rho} \geq \mathcal{E}_Q(h, h^*_P).$$

We say that ρ is **minimal** *when no $0 < \rho' < \rho$ is a transfer exponent from P to Q w.r.t. $\mathcal{H}$.*

Notice that the above parametrization holds trivially for $\rho = \infty$, $C_\rho = 1$. Larger values of the pair $(\rho, \mathcal{E}_Q(h^*_P))$ denote higher discrepancy $P \to Q$. For intuition on ρ, consider the case $h^*_P = h^*_Q = h^*$; then ρ simply describes how well P reveals the *decision boundary* defined by h^*, i.e., whether hypotheses h with small P-excess risk also have small Q-excess risk. Various examples of the continuum $\rho \to \infty$ are given in (Hanneke and Kpotufe, 2019, 2022). We build on the intuition therein to derive Examples 1 and 2 of Section 2.2 below for our specific setting with a hierarchy of hypothesis classes.

Model Selection Setting. We consider a situation where the learner has access to a hierarchy $\{\mathcal{H}_i\}, \mathcal{H}_i \subset \mathcal{H}_{i+1}$ of hypothesis classes, where each $\mathcal{H}_i$ has VC dimension d_i, $d_i \le d_{i+1}$. We let $h^*_{P,i}$, $h^*_{Q,i}$ denote the P and Q risk minimizers over model $\mathcal{H}_i$ (according to Assumption 1).

Assumption 2 *We assume $\{\mathcal{H}_i\}$ admits global risk minimizers h^*_P and h^*_Q w.r.t P and Q; let i^*_P, i^*_Q, unknown to the learner, denote the indices of the smallest classes containing an h^*_P, resp. h^*_Q.*

Definition 4 (Noise and Transfer Parameters) *We let $(C_{\beta_{P,i}}, \beta_{P,i})$ and $(C_{\beta_{Q,i}}, \beta_{Q,i})$ denote BCC parameters for $\mathcal{H}_i$ w.r.t. P and Q. For simplicity, we let $\beta_P \doteq \beta_{P,i^*_P}$ and $\beta_Q \doteq \beta_{Q,i^*_Q}$.*

Finally, we let (C_{ρ_i}, ρ_i) denotes transfer-exponents from P to Q under class $\mathcal{H}_i$.

Assumption 3 *We assume for simplicity that all $C_{\beta_{P,i}}, C_{\beta_{Q,i}}$ are upper-bounded by some C_β.*

2.2. Examples and Intuition on Tradeoffs.

We start with the following remark.

Remark 1 (Implicit Structure on $\left\{(\rho_i, \mathcal{E}_Q(h^*_{P,i}))\right\}$) *To get some intuition, let's consider a simpler situation where $h^*_{P,i}$ is unique for each class $\mathcal{H}_i$. It then follows by definition, and the fact that the classes are nested, that for $i > i^*_P$, we have that ρ_i is also a transfer-exponent for i^*_P. Also, by Assumption 1, for $i > i^*_P$, $h^*_{P,i} = h^*_P$ so we have $\mathcal{E}_Q(h^*_{P,i}) = \mathcal{E}_Q(h^*_{P,i^*_P})$.*

In other words, model selection would not favor $\mathcal{H}_i$ over $\mathcal{H}_{i^*_P}$ *if $i > i^*_P$. However, for $i < i^*_P$, the distance parameters $(\rho_i, \mathcal{E}_Q(h^*_{P,i}))$ are unrestricted—i.e., either term may increase or decrease as i increases to i^*_P—if we impose no further condition on the hierarchy $\{\mathcal{H}_i\}$, thus inducing subtle tradeoffs. Such unrestricted increase or dicrease in distance below i^*_P is illustrated by the examples below and further by the lower-bound construction for Theorem 5.*

Note that, similarly, for μ denoting either P or Q, the BCC parameters $\beta_{\mu,i}$'s are nondecreasing for $i \ge i^*$. Thus, following from the remark, suppose for instance that the *distances* $(\rho_i, \mathcal{E}_Q(h^*_{P,i}))$ were decreasing with $i = 1, 2, \ldots i^*_P$, either in the first or second terms. Then, while usual model selection (as in a non-transfer setting) would favor the smallest class with small error, now it could be that a larger class transfers better. On the flip side, we could have situations were all ρ_i's increase, while $\mathcal{E}_Q(h^*_{P,i})$'s decrease, leading to similarly complicated tradeoffs.

The examples below illustrate such richness of situations in the case of simple two-layer neural networks for $X \in \mathbb{R}$, where the nested classes $\mathcal{H}_i \subset \mathcal{H}_{i+1}$ correspond to increasing width. We emphasize that the main point of these examples is to illustrate the basic thesis that *distance* between source P and target Q may change with given classes in the hierarchy, in particular for model classes that speak to contemporary interest. We will revisit some such examples in Section 3.2 when discussing achievable bounds.

Example 1 (Two Layer Neural Nets with Threshold Activation) *Define $\mathcal{H}_i = \{h_\theta : \mathbb{R} \mapsto \pm 1\}$, indexed over $\theta \doteq (i, a, r, w, b)$, for $a, w, b \in \mathbb{R}^i$ and $r \in \mathbb{R}$, and where h_θ is of the form*

$$h_\theta(x) \doteq sign\left(\sum_{j=1}^{i} a_j \; sign\left(w_j x - b_j\right) + r\right). \tag{2}$$

Proposition 1 *For every finite sequence $1 \leq \rho_1 \leq \cdots \leq \rho_L$, there exists source and target distributions P, Q over $[0,1] \times \{-1,+1\}$ such that $\forall 1 \leq i \leq L$, ρ_i is the minimal transfer exponent from P to Q w.r.t. $\mathcal{H}_i$, $i_P^* = i_Q^* = L$ and $\beta_{P,i} = \beta_{Q,i} = 1$. Furthermore, the sequence of values $\mathcal{E}_Q(h_{P,i}^*)$, $i = 1, 2, \ldots, L$, is strictly decreasing, depends only on L, but not on $\{\rho_i\}$; finally we have that C_{ρ_i}s are upper and lower bounded by functions that depend on i and L only, but not on the choice of $\{\rho_i\}$.*

In particular, as we may have ρ_i's increasing while $\mathcal{E}(h_{P,i}^*$ decrease, we see that nontrivial tradeoffs may indeed occur in practice.

Example 2 (Two Layer (Residual) Neural Net with Relu Activation) *Let $\tilde{\mathcal{H}}_i \doteq \{h_\theta : \mathbb{R} \to \pm 1\}$ indexed over $\theta \doteq (i, a, r, w, b, \alpha)$, for $a, w, b \in \mathbb{R}^i$ and $r, \alpha \in \mathbb{R}$, and where h_θ is of the form*

$$h_\theta(x) = sign\left(\left(\sum_{j=1}^{i} a_j[w_j x + b_j]_+\right) + \alpha x + r\right), \text{ using the notation } [\cdot]_+ \doteq \max(0, \cdot).$$

Next proposition uses results of Aliprantis et al. (2006) to connect ReLu residual neural nets to threshold neural nets in one dimension.

Proposition 2 *Let $\tilde{\mathcal{H}}_i$ be the class of Relu neural nets of Example 2, and let $\mathcal{H}_i$ be the class of neural nets from Example 1. We have $\tilde{\mathcal{H}}_i = \mathcal{H}_{i+1}$, and consequently, Proposition 1 still holds.*

The proofs of the propositions above are given in Appendix A. In particular, the proof of Proposition 1 illustrates how the behavior of P and Q around decision boundaries (defined by optimal classifiers at each level $\mathcal{H}_i$) affects model-transferability; as such, even though for simplicity we focus on $X \in \mathbb{R}$ for these examples, the same insights extend to $\mathbb{R}^d$.

3. Overview of results

For intuition behind the analysis, we start with trying to understand *adaptive* transfer rates at a single level $\mathcal{H}_i$ of the hierarchy. A result of (Hanneke and Kpotufe, 2019) (see Proposition 2 therein) offers a first glimpse. It states roughly that, for a fixed class $\mathcal{H}$, there exists an adaptive $\hat{h}$ with access to n_P samples from P and n_Q samples from Q, such that, w.h.p.

$$\mathcal{E}_Q(\hat{h}) \lesssim \min\left\{\left(\frac{d}{n_P}\right)^{\frac{1}{(2-\beta_P)\rho}} + \mathcal{E}_Q(h_P^*), \left(\frac{d}{n_Q}\right)^{\frac{1}{2-\beta_Q}}\right\}, \tag{3}$$

where β_P, β_Q denote BCC parameters for P and Q. While they show that this is tight (for all ρ, β_P, β_Q), their construction assumes $\mathcal{E}_Q(h_P^*) = 0$, which is too restrictive in our setting.

We start our analysis by first showing that (3) is indeed tight in all parameters.

3.1. Lower Bound for a Fixed $\mathcal{H}$

We consider the following class of pairs of distributions P, Q w.r.t. a fixed $\mathcal{H}$.

Definition 5 (Ξ class) *Let $\mathcal{H}$ denote a hypothesis class, and let $\beta_P, \beta_Q \in [0,1), \rho > 0, \alpha < 1$. We then define $\Xi = \Xi(\mathcal{H}, \beta_P, \beta_Q, \rho, \alpha)$ as the set of pairs of distributions (P, Q) satisfying the following conditions. (i) Assumption 1 holds, (ii) both P, Q satisfy a BCC with respective parameters $(1, \beta_P)$, $(1, \beta_Q)$ (iii) ρ is a transfer exponent P to Q w.r.t. $\mathcal{H}$, with $C_\rho \le 1$, and (iv) $\mathcal{E}_Q(h_P^*) \le \alpha$.*

Theorem 1 *Fix some hypothesis class $\mathcal{H}$ with VC dimension $d \ge 9$. Pick any $\rho \ge 1$, and $\beta_P, \beta_Q \in [0,1)$ and let Ξ denote the corresponding class. For every n_P, n_Q where $\max\{n_P, n_Q\} > d$, let $\hat{h}$ be any classifier that has access to n_P and n_Q source and target samples. Then, there exists a universal constant $c > 0$ s.t.*

$$\sup_{(P,Q)\in\Xi} \mathop{\mathbb{P}}_{P^{n_P}\times Q^{n_Q}} \left[\mathcal{E}_Q(\hat{h}) \ge c \cdot \min\left\{\left(\frac{d}{n_P}\right)^{\frac{1}{(2-\beta_P)\rho}} + \alpha, \left(\frac{d}{n_Q}\right)^{\frac{1}{2-\beta_Q}}\right\}\right] \ge \frac{3-2\sqrt{2}}{8}.$$

The result extends a lower-bound construction of (Hanneke and Kpotufe, 2019) by *randomizing* the relation between a fixed h_P^* and candidates h_Q^*'s. The proof is given in Appendix D.

3.2. Upper Bound

Having established the tightness of (3) over the range of parameters (except for $0 < \rho < 1$), we now have a sense of the rates achievable if we fixed a level $\mathcal{H}_i$. However, as we already know that, ignoring samples from source P, a baseline rate of $(d_{i_Q^*}/n_Q)^{(1/2-\beta_Q)}$ is attainable (up to log factors) by standard model selection techniques (Koltchinskii, 2006, Theorem 7), we will aim for a transfer rate $\phi^\sharp(i)$, defined below, that incorporates this term at level $\mathcal{H}_i$.

We fix some $\delta > 0$, and sequence of $\delta_i > 0$ satisfying $\sum_i \delta_i \le \delta$. For instance, $\delta_i = \frac{1}{i(i+1)}\delta$.

Definition 6 *Define the following quantity, for some C_0 independent of all model parameters:*

$$\phi^\sharp(i) \doteq \min\left\{\mathcal{E}_Q(h_{P,i}^*) + C_0 \cdot C_{\rho_i}\left(\frac{d_i\ \log(n_P/\delta_i)}{n_P}\right)^{\frac{1}{(2-\beta_{P,i})\rho_i}}, C_0\left(\frac{d_{i_Q^*}\ \log(n_Q/\delta_{i_Q^*})}{n_Q}\right)^{\frac{1}{(2-\beta_Q)}}\right\}.$$

Since $C_{\rho_i}, \rho_i, \beta_{P,i}$ are not uniquely defined, without loss of generality we may take them to be the valid values which minimize $\phi^\sharp(i)$. We have the following adaptive upper-bound.

Theorem 2 (Adaptive Upper-bound) *There exists a proper learner $\hat{h}$, with no prior distributional knowledge beyond $\{d_i\}$, which, with probability at least $1 - 3\delta$, for a suitable value of C_0 achieves:*

$$\mathcal{E}_Q(\hat{h}) \le \phi^\sharp(i_P^*).$$

For sanity check, notice that if P were equal to Q, then $\rho_{i_P^*} = 1$ is admissible and we recover the usual model selection bound in terms of $\max\{n_P, n_Q\} \propto (n_P + n_Q)$. The bound is never worse than model selection under Q alone, and can improve significantly for P's *close* to Q, i.e., with small $\rho_{i_P^*}, \mathcal{E}_Q(h_P^*)$.

As stated in the introduction, while SRM, a.k.a. *complexity regularization* approaches are prevalent in the literature and in practice, it is unclear whether such approaches can adaptively achieve the above rate of $\phi^\sharp(i_P^*)$. Instead we employ an approach, similar to so-called *Lepski's method*, based on intersections of empirical confidence balls (see Algorithm 1).

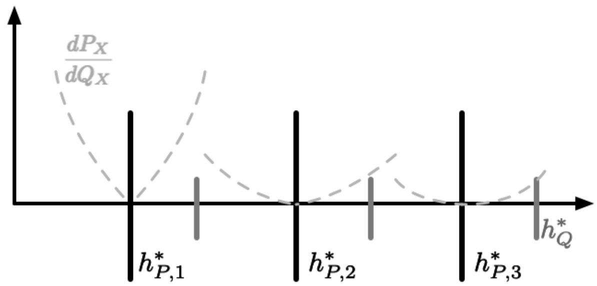

Figure 1: A simple example, following up on NN Examples 1 and 2, where $i^\sharp \doteq \arg\min_i \phi^\sharp(i)$ is different from i^*_P. Here, decision boundaries under P are depicted in black, whereby each $h^*_{P,i}$, $i = 1, 2, 3$, corresponds to the i boundaries on the left of it, including those of $h^*_{P,i-1}$ (level $\mathcal{H}_i$ allows up to i boundaries). Now decision boundaries under Q (as depicted in gray) are shifted to the right of boundaries under P: as a consequence all $h^*_{P_i}$'s have similar excess Q-error $\mathcal{E}_Q$, so that $i^\sharp$ is determined by ρ_i's. Now for this hierarchy, ρ_i may decrease (better transferability) for smaller levels i simply by virtue of P assigning more mass to corresponding decision boundaries as i decreases, as suggested by the density dP_X/dQ_X which is depicted in dashed lines.

We now turn to whether the rate $\phi^\sharp(i^*_P)$ is the best achievable. First, recalling the simple neural-nets Examples 1 and 2, we remark that *there exists situations, i.e., pairs of distributions* (P, Q) *for which* $i^\sharp \doteq \arg\min_i \phi^\sharp(i)$ *is smaller than* $i^\star_P$. The simplest way to see this is to notice in these examples that we may have all $\mathcal{E}_Q(h^*_{P,i})$ equal (or nearly equal) across levels, while at the same time ρ_i's are non-decreasing in these examples, forcing a choice of $i^\sharp$ anywhere below i^*_P. This is illustrated in Section 3.2, and formalized in Proposition 4 of Appendix B.

The next result, relying on a second Algorithm 2, states that the better rate $\min_i \phi^\sharp(i)$ is indeed achievable *given some distributional knowledge*.

Theorem 3 (Oracle Upper-bound) *There exists a proper learner* $\hat{h}$ *which, given knowledge of* $\arg\min_i \phi^\sharp(i)$*, guarantees with probability of at least* $1 - 3\delta$*,*

$$\mathcal{E}_Q(\hat{h}) \leq \min_i \phi^\sharp(i).$$

Unfortunately, as we discuss in the next section, this oracle bound is not achievable adaptively.

3.3. Adaptivity Gap

The following quantity $\phi_\flat(i)$ is of similar order as $\phi^\sharp(i)$ up to log terms, provided $\log(1/\delta_i) \propto d_i$.

Definition 7 *Define the following quantity:*

$$\phi_\flat(i) = \min\left\{ C_{\rho_i}\left(\frac{d_i}{n_P}\right)^{1/(2-\beta_{P,i})\rho_i} + \mathcal{E}_Q(h^*_{P,i}), \left(\frac{d_{i^*_Q}}{n_Q}\right)^{1/(2-\beta_Q)} \right\}.$$

Our aim is to not only establish the un-achievability of the above oracle rate $\min_i \phi(i)$ by adaptive procedures, but also to try and pinpoint the sources of such hardness, i.e., decouple the effect of ρ_i's and $E_Q(h^*_{P,i})$'s. To this end, since these terms only pertain to transfer from P, we need only consider situations where the terms in $\phi(i)$ involving i, achieves the min in the definition of $\phi(i)$.

Our first result below holds every parameter other than ρ_i's fixed, and show that even then $\min_i \phi(i)$ cannot be achieved adaptively. In particular the construction sets $\mathcal{E}_Q(h^*_{P,i}) = 0$ for all i in the hierarchy, but confuses the learner by randomizing which level below i^*_P admits larger ρ_i's. Proofs of the next theorem is in Section 4.2.

For simplicity, the construction in the next theorem sets all β's to 1. We give a similar result to the next theorem for a richer model class in Appendix E.

Theorem 4 (Oracle Rate is Not Achievable) *There exists a hierarchy $\mathcal{H}_1 \subset \mathcal{H}_2$, with $d_1, d_2 = 1$ satisfying the following. Pick any $\rho_a > \rho_b \geq 1$, and any n_P and n_Q, where $\left(\frac{1}{32 n_P}\right)^{1/\rho_a} \leq \frac{1}{32 n_Q}$. Then there is a family of distributions $\{(P_\sigma, Q_\sigma)\}$, indexed by some σ, such that the following hold.*

(i) $\forall \sigma$, transfer exponents from P_σ to Q_σ are the set $\{\rho_1, \rho_2\} = \{\rho_a, \rho_b\}$, with $C_{\rho_1} = C_{\rho_2} = 1$.

(ii) $\forall \sigma$, we have $\min_i \phi_\flat(i) = \left(\frac{1}{n_P}\right)^{1/\rho_b}$, strictly less than $\max_i \phi_\flat(i) = \left(\frac{1}{n_P}\right)^{1/\rho_a}$.

$$\text{We have that, } \forall \hat{h}, \quad \sup_\sigma \mathbb{P}_{P_\sigma^{n_P} \times Q_\sigma^{n_Q}} \left[\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{1}{256} \cdot \max_i \phi_\flat(i) \right] \geq 1/8.$$

The construction fixes $i^*_P = 2$, and randomizes which of $\rho_1 \neq \rho_2$ takes the largest value in $\{\rho_a, \rho_b\}$. We note that our adaptive upper-bound $\max_\sigma \phi^\sharp(i^*_P)$ matches the lower bound $\max_i \phi_\flat(i)$ up to $\log$ terms. Also notice that, as ρ_a, ρ_b are arbitrary, the lower bound can be arbitrarily worse than the Oracle upper-bound, i.e., we can construct any gap in $[0, 1]$.

The next class of distributions instead fixes ρ_i's and allows $\mathcal{E}_Q(h^*_{P,i})$ to vary. It builds on a similar intuition as for the proof of Theorem 1, and is included for completeness.

Theorem 5 *Let $\mathcal{H}_1 \subset \mathcal{H}_2$ be a model class hierarchy such that there exists a set of two points that $\mathcal{H}_2$ shatters but $\mathcal{H}_1$ does not, and assume that $\mathcal{H}_1$ is non empty. Then for any $1 \geq \alpha \geq 0$, n_P, and n_Q such that $\frac{1}{2n_Q} \geq \alpha$, there exists a class of distribution $\{(P_\sigma, Q_\sigma)\}$ parameterized by $\sigma \in \{1, 2\}$, with $\beta_{P,\sigma} = \beta_{Q,\sigma} = 1$, where, for every σ, $\alpha = \max\left\{\mathcal{E}_{Q_\sigma}(h^*_{P,1}), \mathcal{E}_{Q_\sigma}(h^*_{P,2})\right\}$, satisfying the following. For any classifier $\hat{h}$ that has access to n_P source and n_Q target samples,*

$$\sup_{\sigma \in \{1,2\}} \mathbb{P}_{P_\sigma^{n_P} \times Q_\sigma^{n_Q}} \left[\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \alpha \right] \geq \frac{1}{4}.$$

The proof of this theorem is given in Appendix D.

4. Analysis

4.1. Proofs for Upper-bounds

Definition 8 (Empirical Minimal Sets) *Let $A(n_\mu, \delta, \mathcal{C}(\mathcal{H}_i)) \doteq \frac{d_i \log(n_\mu/d_i) + \log 1/\delta}{n_\mu}$. Given n_μ samples from distribution μ, define the empirical minimal set for hypothesis class $\mathcal{H}_i$ to be*

$$\hat{\mathcal{H}}_i^\mu \doteq \left\{ h \in \mathcal{H}_i \mid \hat{R}_\mu(h) - \hat{R}_\mu(\hat{h}_{\mu,i}) \leq C \left(\mathbb{P}_{n_\mu}[h \neq \hat{h}_{\mu,i}] \cdot A(n_\mu, \delta_i, \mathcal{C}(\mathcal{H}_i)) \right)^{1/2} + cA(n_\mu, \delta_i, \mathcal{C}(\mathcal{H}_i)) \right\},$$

where $\hat{h}_{\mu,i}$ denotes an ERM over $\mathcal{H}_i$ computed using samples from distribution μ.

We assume that in addition to the source and target training sets, we are also given a hold-out target sample set S'_Q of size n_Q. Let $\hat{R}'_Q(\cdot)$ denote the empirical risk and $\mathbb{P}'_{n_Q}$ denote the empirical distribution on these held out samples.

The main algorithm is presented next, and relies on Algorithm 2.

Algorithm 1 Adaptive Trade-off

Input: S_P, S_Q, S'_Q
Compute $\hat{i}_P = \min i \quad \text{s.t.} \bigcap_{j\geq i}^{\infty} \hat{\mathcal{H}}_j^P \neq \emptyset$.
Compute $\tilde{\mathcal{H}}^P = \bigcap_{j\geq \hat{i}_P}^{\infty} \hat{\mathcal{H}}_j^P$.
Return output of Algorithm 2 with S_P, S_Q, S'_Q and the set $\tilde{\mathcal{H}}^P$

Algorithm 2 Tradeoff on Q, at level $\mathcal{H}_i$

Require: Any subset $\tilde{\mathcal{H}}^P \subseteq \hat{\mathcal{H}}_i^P$
Input: $S_Q, S'_Q, \tilde{\mathcal{H}}^P$
Compute $\hat{i}_Q = \min i \quad \text{s.t.} \bigcap_{j\geq i}^{\infty} \hat{\mathcal{H}}_j^Q \neq \emptyset$.
Pick $\hat{h}_Q \in \bigcap_{j\geq \hat{i}_Q}^{\infty} \hat{\mathcal{H}}_j^Q$ and pick $\hat{h}_{P,i} \in \tilde{\mathcal{H}}^P$.
If $\hat{R}'_Q(\hat{h}_{P,i}) - \hat{R}'_Q(\hat{h}_Q) \leq \left(\mathbb{P}'_{n_Q}[\hat{h}_{P,i} \neq \hat{h}_Q] \cdot A(n_Q, \delta, 1)\right)^{1/2} + cA(n_Q, \delta, 1)$:
then $\hat{h}_i \leftarrow \hat{h}_{P,i}$ **else:** $\hat{h}_i \leftarrow \hat{h}_Q$

The next lemma gives guarantees for Algorithm 2, and is essential to our main upper-bounds.
.

Lemma 1 *Let $\hat{h}_i$ be the output of Algorithm 2. With probability of at least $1 - 3\delta$ over the samples S_Q, S'_Q and S_P (which is used to construct $\tilde{\mathcal{H}}^P$)*

$$\mathcal{E}_Q(\hat{h}) \leq \phi^\sharp(i).$$

The proof is given in Appendix C. The proofs of main upper-bound results are given next.
Proof of Theorem 2 Let $\hat{h}$ be the output of Algorithm 1. Note that under the same events where the bound in Lemma 1 holds, using the same arguments as in Claim 8 we can conclude that with probability of at least $1-\delta$, $\hat{i}_P \leq i_P^*$. Consequently $\tilde{\mathcal{H}}^P \subseteq \hat{\mathcal{H}}_{i_P^*}^P$. Since Algorithm 1 returns the output of Algorithm 2 on a subset of $\tilde{\mathcal{H}}^P$, it enjoys the guarantees as Algorithm 2 for level i_P^*. Therefore, the bound in Lemma 1 applies to the output of Algorithm 1 at $i = i_P^*$. ■

The proof of the oracle upper is also a simple application of Lemma 1.
Proof of Theorem 3 Let $i^\sharp \doteq \operatorname{argmin}_i \phi^\sharp(i)$. Given $i^\sharp$, oracle would then run Algorithm 2 with given samples and $\hat{\mathcal{H}}_{i^\sharp}^P$ as input. Applying Lemma 1 to the output would prove the statement of the theorem. ■

4.2. Proofs for Lower-bounds

4.2.1. Proof of Theorem 4

We start with a construction, defining a suitable hierarchy and distributions.

Construction. Let $\mathcal{X} = [0,1]$. We let $\mathcal{H}_1$ contain only two one sided threshold classifiers, and $\mathcal{H}_2$ contains $\mathcal{H}_1$ plus two one sided interval classifiers. Let $r = 1/9$. The one sided threshold classifiers in $\mathcal{H}_1$ are $h_1(x) = \text{sign}(x - 2/3)$ and $h_1'(x) = \text{sign}(x - 2/3 + r)$. The one sided intervals h_2 and h_2' only positively label the set of points in $[1/9, 1/3]$ and $[1/9, 1/3 + r]$ respectively.

We construct a family of four distributions $\{(P_\sigma, Q_\sigma)\}_{\sigma\in\{\pm 1\}^2}$, where each P_σ and Q_σ is supported over $[1/9, 1] \times \{\pm 1\}$. Throughout this section we drop the subscript σ when a quantity is the same for all distributions in the family. We refer to the intervals $[1/9, 1/3], [1/3, 1/3 + r], [2/3 - r, 2/3]$ and $[2/3, 1]$ as L_{out}, L_{in}, R_{in} and R_{out} respectively.

For the marginals, we assume that within each interval the mass is uniformly distributed. Let P_X and $Q_{X,\sigma}$, be the marginal distributions under source and target respectively. All the distributions in the family have the same source marginal distribution P_X, which has $P_X(L_{out}) = 1/3$, $P_X(L_{in}) = P_X(R_{in}) = \frac{1}{c_1 n_P}$, $P_X([1/3 + r, 2/3 - r]) = \frac{5}{12} - \frac{2}{c_1 n_P}$, and $P_X(R_{out}) = \frac{1}{4}$. The constant c_1 is set to 32, the reason for which becomes clear in Claim 3 . The labels for the source are $Y_P(L_{out}) = Y_P(R_{out}) = +1$, and for the rest of intervals the labels are the same as Q for σ.

The target marginal distribution $Q_{X,\sigma}(L_{in})$ and $Q_{X,\sigma}(R_{in})$ depends on σ_2. If $\sigma_2 = +1$, set $Q_{X,(\sigma_1,+1)}(L_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$ and $Q_{X,(\sigma_1,+1)}(R_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$, while if $\sigma_2 = -1$, $Q_{X,(\sigma_1,-1)}(L_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$ and $Q_{X,(\sigma_1,-1)}(R_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$.

Let $\Delta \doteq \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a} - \left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$, for the rest of the intervals, $Q_X(L_{out}) = Q_X(R_{out}) = \frac{1}{2}\Delta$, and finally $Q_X([1/3 + r, 2/3 - r]) = 1 - 2\left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$.

For all σ, labels are noiseless. Let $Y_\sigma(A)$ denote the label of the set A under σ. We set $Y_{Q,\sigma}(L_{out}) = -\sigma_1\sigma_2$, $Y_{Q,\sigma}(R_{out}) = \sigma_2$, $Y_\sigma(L_{in}) = Y_\sigma(R_{in}) = \sigma_1$ and $Y([1/3 + r, 2/3 - r]) = -1$.

We make the following two claims, which imply statements (i) and (ii) of the theorem. Additionally, $\beta_{P,\sigma} = \beta_{Q,\sigma} = 1$, since the labels are noiseless.

Claim 1 *For every σ and $i \in \{1, 2\}$, $\mathcal{E}_{Q_\sigma}(h^*_{P_\sigma,i}) = 0$.*

Proof For every σ under Q_σ there are two risk minimizers, one in $\mathcal{H}_1$ and another in $\mathcal{H}_2 \setminus \mathcal{H}_1$. Specifically, when $\sigma_1 = +1$, both h_2' and h_1' are risk minimizers, with risk $\left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$, since the one that mislabels the inner interval with mass $\left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$ will also mislabel $L_{out} \cup R_{out}$.

On the other hand, when $\sigma_1 = -1$, since the regions L_{out} and R_{out} have the same sign, each of h_1 and h_2 will mislabel exactly one of them, which results in the minimum risk of $\frac{\Delta}{2}$. It is easy to see that both of h_1' and h_2' have a strictly larger risk. ■

Claim 2 *The following holds for every value of σ_1. If $\sigma_2 = 1$, we have $\rho_1 = \rho_b$ and $\rho_2 = \rho_a$. Otherwise, for $\sigma = -1$, we have $\rho_1 = \rho_a$ and $\rho_2 = \rho_b$. Furthermore, for all σ, $C_{\rho_1} = C_{\rho_2} = 1$.*

Proof First consider $\mathcal{H}_1$. Suppose that $\sigma_2 = 1$, since $Q_{X,\sigma}(R_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$, whichever of h_1 or h_1' that is not a risk minimizer under source and target, will have excess risk of $\frac{1}{c_1 n_P}$ under source and $\left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$ under target, which means that ρ_b is a transfer exponent with coefficient one. When $\sigma_2 = -1$, since the region where h_1 and h_1' differ has mass $\frac{1}{c_1 n_P}$ under source and $\left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$ under target, ρ_a is a transfer exponent with respect to $\mathcal{H}_1$ with coefficient one.

For $\mathcal{H}_2$, note that every $h \in \mathcal{H}_1$ has an excess risk of at least $1/3 - 1/4 = 1/12 > \frac{1}{c_1 n_P} = \frac{1}{32 n_P}$ under source and excess risk of at most $\left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$ under target, so the transfer exponent condition with ρ_b or ρ_a and coefficient one holds trivially. For hypotheses that are in $\mathcal{H}_2 \setminus \mathcal{H}_1$, since one of them is a risk minimizer, and the region they differ has mass $\left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$ or $\left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$ under target and $\frac{1}{c_1 n_P}$ under source, then depending on σ_2, either ρ_b or ρ_a would be a transfer exponent with coefficient one. ■

The next proposition shows that for every possibly improper learner, there is a distribution in the family under which the learner has high excess risk. The proof is given in the appendix.

Proposition 3 *Let $c_1 = 32$ in the construction. For any classifier $\tilde{h}$, possibly improper, there exists $\sigma \in \{\pm 1\}^2$ such that $\mathcal{E}_{Q_\sigma}(\tilde{h}) \geq \frac{1}{256} \cdot \left(\frac{1}{n_P}\right)^{1/\rho_a}$.*

Let $\Pi_\sigma \doteq P^{n_P} \times Q^{n_Q}$ and $S_\sigma \sim \Pi_\sigma$ be the source and target samples. The next claim defines the event B and lower bounds its probability.

Claim 3 *Let B be the event that of all n_P source and n_Q target samples fall in the intervals $L_{out} \cup [1/3 + r, 2/3 - r] \cup R_{out}$ under source and $[1/3 + r, 2/3 - r]$ under target. Then we may choose c_1 (from the definition of marginal distributions) such that for all $\sigma \in \{\pm 1\}^2$, $\Pi_\sigma[B] \geq 7/8$.*

Proof For any σ,

$$\Pi_\sigma[B] = \left(1 - \frac{2}{c_1 n_P}\right)^{n_P} \left(1 - 2\left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}\right)^{n_Q} \geq \left(1 - \frac{2 n_P}{c_1 n_P}\right)\left(1 - 2 n_Q (\frac{1}{c_1 n_P})^{1/\rho_a}\right),$$

where the inequality follows by Bernoulli's inequality. By the assumption that $(\frac{1}{c_1 n_P})^{1/\rho_a} \leq \frac{1}{32 n_Q}$ and picking $c_1 = 32$, we can ensure $\Pi_\sigma[B] \geq 7/8$. ■

Proof of Theorem 4

Let $\hat{h}$ be a classifier that is output by a learning algorithm that has access to samples S_σ. The lower bound follows by randomizing the choice of σ. Suppose that $\hat{\sigma}$ is sampled uniformly at random from $\{\pm 1\}^2$, then

$$
\begin{aligned}
\sup_{\sigma} \mathop{\mathbb{P}}_{\Pi_\sigma}\left[\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{1}{256}\cdot\left(\frac{1}{n_P}\right)^{1/\rho_a}\right] \geq & \mathbb{E}_{\hat{\sigma}}\,\mathbb{E}_{S_{\hat{\sigma}}}\left[\mathbb{1}\left(\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{1}{256}\cdot\left(\frac{1}{n_P}\right)^{1/\rho_a}\right)\right] \\
= & \mathbb{E}_{S_{\hat{\sigma}}}\,\mathbb{E}_{\hat{\sigma}|S_{\hat{\sigma}}}\left[\mathbb{1}\left(\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{1}{256}\cdot\left(\frac{1}{n_P}\right)^{1/\rho_a}\right)\right] \\
\geq & \mathbb{E}_{S_{\hat{\sigma}}}\,\mathbb{E}_{\hat{\sigma}|S_{\hat{\sigma}}}\left[\mathbb{1}\left(\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{1}{256}\cdot\left(\frac{1}{n_P}\right)^{1/\rho_a}\right)\cdot\mathbb{1}\left(B\right)\right].
\end{aligned}
$$

By construction, $\mathbb{P}_{\hat{\sigma}|S_{\hat{\sigma}},B}(\sigma) = \mathbb{P}_{\hat{\sigma}}(\sigma) = 1/4$. Let $\tilde{\sigma}$ index the distribution that results in high $\hat{h}$ excess risk as in Proposition 3. We have

$$
\begin{aligned}
\mathbb{E}_{S_{\hat{\sigma}}}\,\mathbb{E}_{\hat{\sigma}|S_{\hat{\sigma}}}\left[\mathbb{1}\left(\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{1}{256}\cdot\left(\frac{1}{n_P}\right)^{1/\rho_a}\right)\cdot\mathbb{1}\left(B\right)\right] & \geq \mathbb{E}_{S_{\hat{\sigma}}}\,\mathbb{E}_{\hat{\sigma}|S_{\hat{\sigma}}}\left[\mathbb{1}\left(\hat{\sigma}=\tilde{\sigma}\right)\cdot\mathbb{1}\left(B\right)\right] \\
& = \mathbb{E}_{S_{\hat{\sigma}}}\,\mathbb{E}_{\hat{\sigma}|S_{\hat{\sigma}}}\left[\mathbb{1}\left(\hat{\sigma}=\tilde{\sigma}\right)\cdot\mathbb{1}\left(B\right)\right] \\
& = \frac{1}{4}\cdot\mathop{\mathbb{P}}_{\Pi_{\hat{\sigma}}}\left[B\right] \geq \frac{7}{32}.
\end{aligned}
$$

■

Conclusion

We have shown that source data can help significantly improve target risk under model selection; however, adaptive rates do not always match oracle rates in the model selection setting, as we exhibit situations where no procedure can attain oracle rates without distributional knowledge. Even more striking is that the gap between optimal adaptive rates and oracle minimax rates can be arbitrary, which is not often the case in minimax theory. However this leaves open the possibility of smaller or more controlled gaps under, e.g., further structural assumptions on the model hierarchy.

Acknowledgments

We thank COLT reviewers and AC for useful comments that help improve the manuscript.

References

Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. The information complexity of learning tasks, their structure and their distance. *arXiv:1904.03292*, 2019.

Charalambos D. Aliprantis, David Harris, and Rabee Tourky. Continuous piecewise linear functions. *Macroeconomic Dynamics*, 10(1):77–99, 02 2006.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv:1902.09229*, 2019.

P. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, 2008.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv:2002.09434*, 2020.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning*, pages 131–160, 2009.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, 2019.

Steve Hanneke and Samory Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, 2010.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. *arXiv:1803.01833*, 2018.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv:0902.3430*, 2009a.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009b.

P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5): 2326–2366, 2006.

Andreas Maurer, Massi Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *International Conference on Machine Learning*, 2013.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

Daniel McNamara and Maria-Florina Balcan. Risk bounds for transferring representations with and without fine-tuning. In *International Conference on Machine Learning*, 2017.

Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems*, 33:1959–1969, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 2013.

S. N. Negahban and M. J. Wainwright. Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Transactions on Information Theory*, 57(6):3841–3863, 2011.

Anastasia Pentina and Christoph Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

Nilesh Tripuraneni, Michael I Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *arXiv:2006.11650*, 2020.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their expectation. *Theory of Probability and its Applications*, 16:264–280, 1971.

Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90(2):161–189, 2013.

Appendix A. Proofs of Propositions for Examples

A.1. Proof of Proposition 1

Proof of Proposition 1

We construct P and Q such that marginal distributions Q_X and P_X are supported on $[0,1]$. Let Q_X be the uniform distribution over $[0,1]$. To define the source marginal distribution, we pick L points $V=\{v_k\}_{k=1}^{L}$ on the unit interval so that each $v_k=\frac{k}{L+1}$. Then define

$$\mathrm{f}_P(x) \propto \rho_m \cdot 2^{-2m\cdot\rho_m} \left|x - v_m\right|^{\rho_m - 1}, \tag{4}$$

where $v_m \in V$ is the closest point to x, and ties are broken by picking the smaller one, except when x is in the first interval, in which case we set $m = 1$. This leads to L partitions $R_1, \ldots R_L$ of the interval $[0, 1]$ such that for every $x \in R_i$, $\mathrm{f}_P(x) = \frac{\rho_i \cdot 2^{-(2i \cdot \rho_i + i)}}{Z} |x - v_i|^{\rho_i - 1}$, where Z is a normalizing constant. See Figure 2 for an example with $L = 3$.

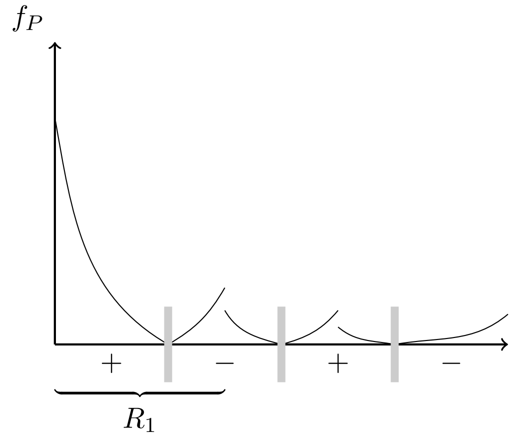

Figure 2: Construction of the marginal density of P for the threshold neural neural net example, with $L = 3$.

The labels for both source and target are given by f_{θ^*}, where $\theta^* = (L, a^*, r^*, w^*, b^*)$ is a set of parameters for a risk minimizer. We pick these parameters such that $i_P^* = i_Q^* = L$ as follows. First, set w^* to the all ones vector and $b_i^* = v_i$ for all $i \leq L$. The boundaries b_i^* divide $[0, 1]$ into $L+1$ intervals, distinct from the L regions $R_1, \cdots R_L$. Next claim shows that we can pick a^*, r^* such that the label for these intervals are alternating, so that every point $b_i^* = v_i$ is indeed a decision boundary. Note that this leads to $\beta_{P,i} = \beta_{Q,i} = 1$. Proofs for all the claims in this proof appear at the end of this section.

Claim 4 *Let $0 < b'_1 < \cdots < b'_i < 1$ be an increasing sequence of points in $[0, 1]$ that partition the unit interval into $i + 1$ intervals $I_1 = [0, b'_1], I_2 = (b'_1, b'_2], \ldots,$ and $I_{i+1} = (b'_i, 1]$. For any sign pattern $\sigma \in \{\pm 1\}^{i+1}$, there exists a set of parameters θ, such that f_θ maps any $x \in I_j$ to σ_j for all $j \in [i + 1]$. Furthermore, any two layer threshold neural net of the form eq. (2), that is, any θ, with i hidden units can lead to at most i decision boundaries.*

Next claim shows that there is risk minimizer in $\mathcal{H}_i$ that has the same decision boundaries as the smallest i decision boundaries in θ^* and correctly labels the first $i + 1$ intervals, by matching their signs with the signs of the first $i + 1$ intervals generated by θ^*.

Claim 5 *For any $i \leq L$, let $\theta_i^* = (i, a^{*,i}, r^{*,i}, w^{*,i}, b^{*,i})$, where for every $j \in [i]$, $b_j^{*,i} = b_j^*$, and $w^{*,i}$ is the all ones vector. The first i intervals generated by $b^{*,i}$ are the same as those of θ^*, and by Claim 4 we can pick $a^{*,i}$ and $r^{*,i}$ such that θ_i^* makes no error in the first $i + 1$ intervals generated by θ^*. Then $f_{\theta_i^*}$ is a unique risk minimizer over the class $\mathcal{H}_i$ under P.*

In the next proposition we show that for any $i \leq L$, ρ_i is a transfer exponent from P to Q with respect to $\mathcal{H}_i$. Intuitively, we show that whenever there is error, it is dominated by the error in the regions determined by the first i thresholds.

Claim 6 *For every $1 \leq i \leq L$, there exists a constant $0 < C_{\rho_i} < \infty$ such that ρ_i is a transfer exponent from P to Q with respect to $\mathcal{H}_i$ with coefficient $C_{\rho_i} \leq (L+1)2^{3i+1}$.*

Next, we show that for every $1 \leq i \leq L$ a transfer exponent from P to Q with respect to $\mathcal{H}_i$ is lower bounded by ρ_i. Fix a level $1 \leq i \leq L$, and consider a sequence of classifiers $f_{\theta(t)}$ constructed so that $\theta(t)$ matches θ_i^* everywhere except for the last decision boundary $b_i^{*,i}$. That is, $\theta(t) = (i, a_i^*, r^{*,i}, w^{*,i}, b^t)$, where $b_j^t = b_j^{*,i}$ for all $j < i$, and for $t < \frac{1}{2(L+1)}$,

$$b_i^t = b_i^{*,i} + t.$$

It is easy to see that the interval $[b_i^{*,i}, b_i^t]$ is the only disagreement region between θ_i^* and $\theta(t)$ and has length t. By the construction of P_X given in eq. (4), and integrating over this region, the excess risk is

$$\mathcal{E}_P(f_{\theta(t)}, f_{\theta_i^*}) = Ct^{\rho_i},$$

for some constant C. Since Q_X is the uniform distribution,

$$\mathcal{E}_Q(f_{\theta(t)}, f_{\theta_i^*}) = t.$$

Now we argue that the minimal transfer exponent for this level is at least ρ_i. Suppose for contradiction, that there exists a transfer exponent $\tilde{\rho}_i < \rho_i$. That would imply that there exists a constant $C_{\tilde{\rho}_i}$ such that for every $f_\theta \in \mathcal{H}_i$,

$$\frac{\mathcal{E}_Q(f_\theta, f_{\theta_i^*})}{\mathcal{E}_P(f_\theta, f_{\theta_i^*})^{1/\tilde{\rho}_i}} \leq C_{\tilde{\rho}_i}.$$

However, for the sequence of f_{θ_t} constructed above

$$\lim_{t \to 0} \frac{\mathcal{E}_Q(f_{\theta_t}, f_{\theta_i^*})}{\mathcal{E}_P(f_{\theta_t}, f_{\theta_i^*})^{1/\tilde{\rho}_i}} = \frac{t}{(Ct^{\rho_i})^{1/\tilde{\rho}_i}} = C' t^{1 - \frac{\rho_i}{\tilde{\rho}_i}} = \infty.$$

Now that we have shown that ρ_i are indeed minimal transfer exponents, we will show that for these minimal transfer exponents, C_{ρ_i} is lower bounded by a function that depends only on $L+1$. Fix some $t_0 < \frac{1}{2(L+1)}$ and consider $f_{\theta_{t_0}}$. By the same calculations as in eq. (5) we have $\mathcal{E}_P(f_{\theta_t}, f_{\theta_i^*}) = \frac{2^{-(2i \cdot \rho_i + i)}}{Z} t^{\rho_i}$ and

$$\frac{\mathcal{E}_Q(f_\theta, f_{\theta_i^*})}{\mathcal{E}_P(f_\theta, f_{\theta_i^*})^{1/\rho_i}} = \frac{tZ^{1/\rho_i}}{2^{-(2i+i/\rho_i)}t} = \frac{Z^{1/\rho_i}}{2^{-(2i+i/\rho_i)}}.$$

Since $2^{-(2i+i/\rho_i)} \leq 2^{-2i}$, it suffices to lower bound Z^{1/ρ_i}, note that $Z \geq 2^{-(2i \cdot \rho_i + i)} \left(\frac{1}{L+1}\right)^{\rho_i}$ and plugging this into the expression above we get

$$\frac{\mathcal{E}_Q(f_\theta, f_{\theta_i^*})}{\mathcal{E}_P(f_\theta, f_{\theta_i^*})^{1/\rho_i}} \geq \frac{1}{L+1}.$$

Finally, to see that $i_P^* = i_Q^* = L$, note that for every $1 \leq i < L$, $f_{\theta_i^*}$ labels the interval I_{i+2}, which starts at v_{i+1}, incorrectly, so it cannot achieve zero excess risk. ■

Proof of claim 4
We will argue that there exists $a_\sigma \in \mathbb{R}^i$ and $r_\sigma \in \mathbb{R}$ such functions of the

$$h_{a_\sigma, r_\sigma}(x) = \text{sign}\left(\sum_{j=1}^{i} \alpha_{\sigma,j} \text{ sign}\left(x - b'_j\right) + r_\sigma\right),$$

can produce the sign pattern σ. Since these functions are a restricted form of the two layer neural nets introduced in eq. (2), this would prove the first part of the claim.

Define the function $g : [0,1] \mapsto \{\pm 1\}^{i+1}$ where

$$g(x)_j = \text{sign}\left(x - b'_j\right).$$

The function g maps the $i+1$ intervals to $i+1$ points on the unit cube. Let $x_1, \ldots, x_{i+1}$ be a set of arbitrary points from each interval $I_1, \ldots, I_{i+1}$, then we have

$$\begin{aligned} g(x_1) &= [-1, -1, -1, \ldots, -1], \\ g(x_2) &= [+1, -1, -1, \ldots, -1], \\ &\ldots \\ g(x_{i+1}) &= [+1, +1, +1, \ldots, +1]. \end{aligned}$$

Note that $g(x_1) = -g(x_2)$, but the set of vectors $g(x_1), \ldots, g(x_i)$ are linearly independent. The functions

$$h_{a,r}(x) = \text{sign}\left(a^\top g(x) + r\right)$$

are affine halfspaces parameterized by a and r, so they can shatter any set of $i+1$ points where there is at most two colinear points. We take a_σ and r_σ to be coefficients of the affine halfspace that produces the labels σ.

To see that two layer neural nets of the form eq. (2) parameterized by $\theta = (i, a, r, w, b)$ can have at most i decision boundaries, note that adding a hidden unit can add at most one decision boundary, and when $i = 0$, there are no decision boundaries. ■

Proof of claim 5 Recall the regions $R_1, \ldots, R_i$. We first argue that if a classifier does not place a decision boundary in some region $i' \in [i]$, then its' error is larger than θ_i^*, and then argue that among all the classifiers that place exactly one decision boundary in each of those regions, only the ones that have exactly the same decision boundaries as $\theta^{*,i}$ can be risk minimizers. Suppose θ_i is some classifier that doesn't place any decision boundaries in $R_{i'}$, then it must mislabel one of the intervals to the left or right of $b_{i'}^*$, that is either the interval $I_{i'+1} \cap R_{i'}$ or $I_{i'} \cap R_{i'}$. Then the risk can be lower bounded by

$$\begin{aligned} R_P(h_{\theta_i}) &\geq P_X(I_{i'+1} \cap R_{i'}) \qquad (5) \\ &= \int_{b_{i'}^*}^{\frac{v_{i'}+v_{i'+1}}{2}} \frac{\rho_{i'} \cdot 2^{-(2i' \cdot \rho_{i'} + i')}}{Z} \cdot |x - b_{i'}^*|^{\rho_{i'}-1} \, dx = \frac{\rho_{i'} \cdot 2^{-(2i' \cdot \rho_{i'} + i')}}{Z \rho_{i'}} \left(\frac{|I_{i'+1}|}{2}\right)^{\rho_{i'}} \\ &= \frac{2^{-(2i' \cdot \rho_{i'} + i')}}{Z} \left(\frac{1}{2(L+1)}\right)^{\rho_{i'}} \\ &\geq \frac{2^{-(2i \cdot \rho_i + i)}}{Z} \left(\frac{1}{2(L+1)}\right)^{\rho_i}. \end{aligned}$$

On the other hand, θ_i^* labels the first $i+1$ intervals correctly, so

$$R_P(h_{\theta_i^*}) \leq \sum_{j=i+2}^{L+1} P_X(I_j). \tag{6}$$

Note that when $i = L$, by construction $R_P(h_{\theta_i^*}) = 0$.

For every $i+2 \leq j \leq L$, we can write $P_X(I_j) = P_X(I_j \cap R_j) + P_X(I_j \cap R_{j-1}) = \frac{2^{-(2j\cdot\rho_j+j)}}{Z}\left(\frac{1}{2(L+1)}\right)^{\rho_j} + \frac{2^{-(2(j-1)\cdot\rho_{j-1}+j-1)}}{Z}\left(\frac{1}{2(L+1)}\right)^{\rho_{j-1}}$. Since for all j, $\rho_j \geq 1$, and $\rho_{j-1} \leq \rho_j$, we get that

$$P_X(I_j) \leq \frac{2^{-2((j-1)\cdot\rho_{j-1}+j-1)}}{Z}\left(\frac{1}{L+1}\right)^{\rho_{j-1}}. \tag{7}$$

Going back to eq. (6), for $i < L$ we get

$$\begin{aligned}
R_P(h_{\theta_i^*}) &\leq \sum_{j=i+2}^{L+1} \frac{2^{-(2(j-1)\cdot\rho_{j-1}+j-1)}}{Z}\left(\frac{1}{L+1}\right)^{\rho_{j-1}} \leq \frac{2^{-2(i+1)\cdot\rho_{i+1}}}{Z}\sum_{j=i+2}^{L+1} 2^{-j+1}\cdot\left(\frac{1}{L+1}\right)^{\rho_{i+1}} \\
&\leq \frac{2^{-(2(i+1)\cdot\rho_{i+1}+i)}}{Z}\left(\frac{1}{L+1}\right)^{\rho_{i+1}} \\
&< \frac{2^{-(2i\cdot\rho_i+i)}}{Z}\left(\frac{1}{2(L+1)}\right)^{\rho_i} \leq R_P(h_{\theta_i}),
\end{aligned}$$

so the excess risk can be lower bounded by

$$R_P(h_{\theta_i}) - R_P(h_{\theta_i^*}) \geq \frac{2^{-((2i+1)\cdot\rho_i+i)}}{Z}\left(\frac{1}{L+1}\right)^{\rho_i}\left(1-2^{-\rho_{i+1}}\right) \geq \frac{2^{-((2i+1)\cdot\rho_i+i)}}{2Z}\left(\frac{1}{L+1}\right)^{\rho_i}$$

and consequently for some $C_1 = \frac{2^{-((2i+1)\cdot\rho_i+i)}}{2Z}$,

$$\mathcal{E}_P(h_{\theta_i}, h^*_{\theta_i}) \geq C_1\left(\frac{1}{L+1}\right)^{\rho_i}. \tag{8}$$

Therefore, we have shown that any classifier f_{θ_i} that doesn't place a decision boundary in any of the first i regions has a strictly larger excess risk than excess risk of $f_{\theta_i^*}$. Then it would suffice to show that among the classifiers that place exactly one decision boundary in each of $R_1, \ldots, R_i$, $f_{\theta_i^*}$ is a risk minimizer.

Let θ'_i be a function that places one decision boundary in each of the regions $R_1, \ldots, R_i$. Let $v' \in [0,1]^i$ denote the location of its' decision boundaries, and let R_k be some region where $v'_k \neq b^*_k$, that is a region where the decision boundary is different from that of θ^*. Now we will argue that θ'_i cannot be a risk minimizer, since it can be modified to another classifier that has a strictly smaller risk. Without loss of generality assume that $v'_k < b^*_k$, the other direction follows by the same argument. It must be that either $f_{\theta'_i}$ labels all the points in the interval (v'_k, b^*_k) incorrectly and is correct on the rest of R_k, or it mislabels at least all the points in $R_k \cap I_{k+1}$. In the latter case, as we have seen before in eq. (5), since $k \in i$, excess risk of $f_{\theta'_i}$ would be strictly larger than excess risk of $f_{\theta_i^*}$. If the only set of points that are mislabelled in R_k are in the interval (v'_k, b^*_k), then moving v'_k right by increasing it to b^*_k would eliminate error in this interval without affecting other intervals, thus strictly decreasing the excess risk. Therefore, any classifier that places decision boundaries not on $b_1^*, \ldots, b_i^*$ cannot be a risk minimizer.

Note that in terms of parameters, the risk minimizer is not unique, since there can be a family of parameters that give the same decision boundaries and signs. ■

Proof of Claim 6 Recall that he boundaries $v_1, \ldots, v_L$ partitioned the unit interval into $L+1$ intervals $I_1, \ldots, I_{L+1}$, where $I_1 = [0, v_1], I_2 = (v_2, v_3]$ and so on. Recall that θ_i^* is the parameters of a risk minimizer in $\mathcal{H}_i$, as described in claim 5. Based on the points in V, we also defined the regions $R_1, \ldots, R_L$, where all the points in each region shared the same density function under P. Let f_{θ_i} be an arbitrary element of $\mathcal{H}_i$, by Claim 4, f_{θ_i} can have at most i decision boundaries. We can break down the excess risk into the contributions from each region,

$$\mathcal{E}_Q(f_{\theta_i}, f_{\theta_i^*}) = \sum_{j=1}^{L} Q_X(f_{\theta_i}(X) \neq f_{\theta_i^*}(X) \wedge X \in R_j) = \sum_{j=1}^{L} Q_X(E_j),$$

where E_j is the set of points in R_j that θ_i labels differently from θ_i^*. Consider the first i regions, since θ_i has at most i decision boundaries and the regions are disjoint, it must be the case that either

1. θ_i does not place a decision boundary in at least one of the regions $R_1, \ldots, R_i$, or
2. θ_i places exactly one decision boundary in every region $R_1, \ldots, R_i$.

We break down the proof into the two cases above, and start with the simpler case 1. In this case, there exists at least one region $R_{i'}$, for some $i' \leq i$, such that θ_i has placed no decision boundary there and consequently the whole interval $R_{i'}$ has the same label, while under θ_i^*, there would be a boundary at $v_{i'} \in R_{i'}$, which implies that θ_i must have mislabelled an interval on at least one side of $v_{i'}$. Then under source $P_X(E_{i'}) \geq P_X(I_{i'+1} \cap R_{i'})$ and by eq. (5) and eq. (8) the excess risk

$$\mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*}) = \sum_{j=1}^{i} P_X(E_j) \geq P_X(I_{i'+1} \cap R_{i'}) \geq C_1 \left(\frac{1}{L+1}\right)^{\rho_i}, \tag{9}$$

where C_1 is some positive constant that could depend on i and ρ_i. On the other hand,

$$\mathcal{E}_Q(f_{\theta_i}, f_{\theta_i^*}) \leq 1.$$

Now by eq. (9)

$$\mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*})^{1/\rho_i} \geq \frac{C_1^{1/\rho_i}}{L+1} = \frac{2^{-(2i+1)-i/\rho_i}}{(2Z)^{1/\rho_i} \cdot (L+1)}.$$

Using eq. (7), we have that $Z \leq (L+1)2^{-4}\left(\frac{1}{L+1}\right)^{\rho_L} \leq 2^{-4}$, and consequently for any $\rho_i \geq 1$ and fixed L, and $\mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*})^{1/\rho_i} \geq \frac{2^{-3i-1}}{L+1}$ setting $C_{\rho_i}^{(1)} = \frac{L+1}{2^{-3i-1}}$, we can conclude that in case 1

$$\mathcal{E}_Q(f_{\theta_i}, f_{\theta_i^*}) \leq C_{\rho_i}^{(1)} \mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*})^{1/\rho_i}. \tag{10}$$

In case 2, when $i < L$, θ_i has no decision boundaries in regions $R_{i+1}, \ldots, R_L$, so they will all have the same label, since they also have the same label under θ_i^*. it must be that either all their labels agree with those of θ_i^*, or their label disagrees with the label θ_i^* assigns to those regions. If they are all labelled incorrectly, we will argue that,

$$P_X(E_i) \geq C_1 \left(\frac{1}{L+1}\right)^{\rho_i}.$$

To see this, note that θ_i places its' decision boundary in region R_i either to left of b_i^* or to the right. In the former case, then the interval $I_{i+1} \cap R_i$ is also labelled incorrectly, while in the latter case the interval $I_i \cap R_i$ would have incorrect labels, so in either case by eq. (5) eq. (8) holds. Consequently, we can make the same arguments as in case 1 to get eq. (10).

Going back to case 2, suppose that the regions $R_{i+1}, \ldots, R_L$ are labelled according to θ_i^*, or $i = L$ (they don't exist), so that they don't contribute to the excess risk. Let $m \in \operatorname{argmax}_{j \in [i]} Q_X(E_j)$, so that R_m is a region that has large contribution to the excess risk. Then

$$\mathcal{E}_Q(f_{\theta_i}, f_{\theta_i^*}) = \sum_{j=1}^{i} Q_X(E_j) \leq i\, Q_X(E_m), \tag{11}$$

while for the source, we can lower bound

$$\mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*}) \geq P_X(E_m).$$

Let b_m be the point that is a decision boundary in R_m under θ_i. Then E_m is either the interval that has b_m^*, b_m as its' end points or it is a union of two intervals, one of which has size at least $|I_{m+1}|/2$. If it is a union of two intervals, since source excess risk will be bounded away from zero by a constant, then we can use the same arguments as in item 1. If E_m is an interval that has b_m^* and b_m as its' end points, then $P_X(E_m) = C_3 Q_X(E_m)^{\rho_m}$, where by similar calculations as those in eq. (5), $C_3 = \frac{2^{-(2m \cdot \rho_m + m)}}{Z}$. Since $Z \leq 2^{-4}$ for any value of ρ_is and L, we have $C_3^{1/\rho_m} \geq 2^{-2m - m/\rho_m + 4/\rho_m} \geq 2^{-3m}$. Then

$$\mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*})^{1/\rho_m} \geq P_X(E_m)^{1/\rho_m} = C_3^{1/\rho_m} Q_X(E_m),$$

and by eq. (11)

$$i2^{3i} \cdot \mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*})^{1/\rho_i} \geq \frac{i}{C_3^{1/\rho_m}} \mathcal{E}_P(f_{\theta_i}, f_{\theta_i^*})^{1/\rho_m} \geq \mathcal{E}_Q(f_{\theta_i}, f_{\theta_i^*}).$$

Finally, setting $C_{\rho_i} = \max\left\{C_\rho^{(1)}, i2^{3i}\right\} = \max\left\{(L+1) \cdot 2^{3i+1}, i2^{3i}\right\}$, we have shown that $C_{\rho_i} \leq (L+1)2^{3i+1}$ is a transfer exponent with respect to $\mathcal{H}_i$. ■

A.2. Proof of Proposition 2

Proof of Proposition 2

First we define a convenient parameterization of the set of all classifiers over the real line.

Definition 9 (Class of k decision boundaries) *Let $\mathcal{H}_k^B \doteq \left\{h^B_{b_1,\ldots,b_k,\sigma_1}\right\}$ be the set of classifiers over $\mathbb{R}$ that have at most k decision boundaries, given by points $b_1 < b_2 < \cdots < b_k \in \mathbb{R}$, and $\sigma_1 \in \{\pm 1\}$, which is the sign of the first interval $(-\infty, b_1]$.*

Any two layer threshold neural net with i activation units can have at most i decision boundaries, so it belongs to $\mathcal{H}_i^B$. Claim 4, shows that that the class of threshold neural nets with i hidden units can generate i boundaries, and all possible labellings of the corresponding intervals, so we can conclude that $\mathcal{H}_i = \mathcal{H}_i^B$.

Let $\mathcal{F}_k^{CPWL}$ be the class of continuous piecewise linear (CPWL) functions with at most k linear pieces and consequently at most $k-1$ knots.

Claim 7 *We have that* $\text{sign} \circ \mathcal{F}_i^{CPWL} = \mathcal{H}_i^B$.

Proof It is easy to see that once we fix $\text{sign}\,(0)$, thresholding each linear piece can result in at most one decision boundary, so a CPWL function with i pieces can generate at most i decision boundaries.

We argue that any set of i decision boundaries $b_1 < \cdots < b_i$ and label assignment on the line can be generated by taking the sign of some CPWL function with at most i pieces. To see this, consider the i intervals $I_1, \dots I_{i-1}$ whose end points are the boundaries. Let $m_1, \dots, m_{i-1}$ be the mid points of these intervals, and consider points $(m_1, \text{sign}\,(I_1)), \dots, (m_{i-1}, \text{sign}\,(I_{i-1}))$ on the xy-plane. The CPWL function can be constructed by passing the first line through the pair of points $((b_1, 0), (m_1, \text{sign}\,(I_1)))$, the second line through $((b_2, 0), (m_2, \text{sign}\,(I_2)))$ and so on. The last line interpolates the points $(m_{i-1}, \text{sign}\,(I_{i-1}))$ and $(0, b_i)$. ■

Following Lemma, which is adapted from Aliprantis et al. (2006) (Corollary 3.5), states that any CPWL function with at most $k+1$ linear pieces can be written as a two layer ReLu Residual neural network with at most k hidden units.

This implies that $\text{sign} \circ \mathcal{F}_{i+1}^{CPWL} = \tilde{\mathcal{H}}_i$, since it is easy to see that any function of the form $f_\theta(x) = \left(\sum_{i=1}^{i} a_i[w_i x + b_i]_+\right) + \alpha x + r$ can have at most i knots and consequently $i+1$ linear pieces.

Lemma 2 (Aliprantis et al. (2006), Corollary 3.5) *Any CPWL function of the form*

$$f(x) = \begin{cases} m_0 x + c_0 & \text{if } x \leq b_0 \\ m_i x + c_i & \text{if } b_{i-1} \leq x \leq b_i \text{ for } 1 \leq i \leq k \\ m_{k+1} x + c_k & \text{if } x \geq b_k, \end{cases}$$

where $-\infty < b_0 < b_1 < \cdots < b_k < \infty$ *and* $(m_i, c_i), 0 \leq i \leq k+1$ *are real numbers, can also be written in the form*

$$f(x) = c_0 + m_0 x + \sum_{i=0}^{k} (m_{i-1} - m_i)[t - b_i]_+.$$

■

Appendix B. Example for $i^\sharp \doteq \arg\min_i \phi^\sharp(i)$ Below i_P^*

Proposition 4 *Following up on Examples 1 and 2 with $L = 3$, for every $1 \le \rho_1 \le \rho_2 \le \rho_3$, there exists P and Q over $[0, 1] \times \{\pm 1\}$ such that the following holds.* i) $i_P^* = 3$, ii) ρ_1, ρ_2, ρ_3 are minimal transfer exponents from P to Q, where C_{ρ_i}'s are uniformly upper and lower bounded independently of ρ_i's, and iii) $\mathcal{E}_Q(h^*_{P,1}) = \mathcal{E}_Q(h^*_{P,2}) = \mathcal{E}_Q(h^*_{P,3})$. Consequently, while $i_P^* = 3$, we may choose ρ_i's so that $i^\sharp \doteq \text{argmin}_i\, \phi^\sharp(i)$ could be below any of the levels $1, 2, 3$, for n_P sufficiently large.

Proof We use the same construction as in Proposition 1, with the exception that target does not share the same decision boundaries. Specifically, set $L = 3$ and let v_1, v_2, v_3 be the decision boundaries under source. Under target, set $v_1' = \frac{v_1+v_2}{2}$, $v_2' = \frac{v_2+v_3}{2}$, and $v_3' = \frac{v_3+1}{2}$ to be the decision boundaries. Let $I_1, \dots, I_4$, as defined in the proof of Proposition 1, be the intervals defined by decision boundaries under source, and $I_1', \dots, I_4'$ be the intervals $[0, v_1'], [v_1', v_2'], [v_2', v_3'], [v_3', 1]$. For any sequence of labels assigned to $I_1, \dots, I_4$, assign the same sequence of labels to $I_1', \dots, I_4'$.

Since the marginal densities of P, Q and $h^*_{P,i}$ have not changed, the conditions on transfer exponent and coefficient are satisfied. Since $I_1 \subset I_1'$ and the sign patterns under source and target match, $h^*_{P,i}$ don't make any errors under target in the interval I_1. Exactly half of each of the intervals I_2, I_3 and I_4 is labelled $+1$ under target, and $h^*_{P,i}$ give a single label to each of these intervals, so each $h^*_{P,i}$ labels half of each of the intervals I_2, I_3 and I_4 incorrectly under target.

For the last part of the proposition, pick any value $i \in \{2, 3\}$, and to ensure that $\text{argmin}_i\, \phi^\sharp(i) < i$, for all $j \le i$ set $\rho_j = \rho$ and for all $j > i$, set $\rho_j > \rho$.

■

Appendix C. Remaining Upper-bound Proofs

Our analysis relies on the following lemma.

Lemma 3 (Vapnik and Chervonenkis (1971)) *Recall $A(n_\mu, \delta, \mathcal{C}(\mathcal{H}_i)) \doteq \frac{d_i \log(n_\mu/d_i) + \log 1/\delta}{n_\mu}$. For any $\delta > 0$, with probability of at least $1 - \delta$, for all $h, h' \in \mathcal{H}_i$*

$$R_\mu(h) - R_\mu(h') \le \hat{R}_\mu(h) - \hat{R}_\mu(h') + \sqrt{\min\{\mathbb{P}_\mu[h \ne h'], \mathbb{P}_{n_\mu}[h \ne h']\} \cdot A(n_\mu, \delta, \mathcal{C}(\mathcal{H}_i))} + cA(n_\mu, \delta, \mathcal{C}(\mathcal{H}_i)), \text{ and}$$

$$\frac{1}{2}\mathbb{P}_\mu[h \ne h'] - cA(n_\mu, \delta, \mathcal{C}(\mathcal{H}_i)) \le \mathbb{P}_{n_\mu}[h \ne h'] \le 2\mathbb{P}_\mu[h \ne h'] + cA(n_\mu, \delta, \mathcal{C}(\mathcal{H}_i)). \quad (12)$$

Proof of Lemma 1 Let $\hat{h}$ be the output of Algorithm 2, with $\tilde{\mathcal{H}}^P \subseteq \hat{\mathcal{H}}_i^P$. First we state a few useful claims. Proofs of these claims appear in Appendix C

Claim 8 *Let $\hat{i}_Q = \min i$ s.t. $\bigcap_{j \ge i}^{\infty} \hat{\mathcal{H}}_j^Q \ne \emptyset$, then with probability of at least $1 - \delta$, $\hat{i}_Q \le i_Q^*$.*

Next claim can be used to bound the excess risk of $\hat{h}_{P,i}$ with respect to $h^*_{P,i}$. Since a similar statement would also hold for Q under the same high probability event as in claim 8, and $\hat{i}_Q \le i_Q^*$, we could conclude that with probability of at least $1 - \delta$,

$$\mathcal{E}_Q(\hat{h}_Q) \le cA(n_Q, \delta_{i_Q^*}, \mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}. \quad (13)$$

Claim 9 *For any level i, and any $\hat{h}_i \in \hat{\mathcal{H}}_i^P$, with probability of at least $1-\delta_i$,*

$$\mathcal{E}_P(\hat{h}_i, h^*_{P,i}) \leq cA(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))^{\frac{1}{2-\beta_{P,i}}}.$$

Let E_Q and E_P be the events where the bounds in Claims 8 and 9 hold. Let E_H be the event that the bounds given in Lemma 3 hold over the hypothesis class $\{\hat{h}_{P,i}, \hat{h}_Q\}$ and held out samples from Q. Note that complexity of the class $\{\hat{h}_{P,i}, \hat{h}_Q\}$ is one. We first claim that under E_H and E_Q, if $\hat{h} = \hat{h}_{P,i}$, then $\mathcal{E}_Q(\hat{h}_{P,i}) \leq 25cA(n_Q, \delta, \mathcal{C}(\mathcal{H}_{i^*_Q}))^{\frac{1}{2-\beta_Q}}$. Suppose that $\hat{h} = \hat{h}_{P,i}$, which means that the if-statement condition in Algorithm 2 must have been satisfied. Under E_H,

$$\begin{aligned}
\mathcal{E}_Q(\hat{h}_{P,i}) - \mathcal{E}_Q(\hat{h}_Q) =& R_Q(\hat{h}_{P,i}) - R_Q(\hat{h}_Q) \leq \hat{R}'_Q(\hat{h}_{P,i}) - \hat{R}'_Q(\hat{h}_Q) \\
&+ \left(\mathbb{P}'_{n_Q}[\hat{h}_{P,i} \neq \hat{h}_Q] \cdot A(n_Q, \delta, 1)\right)^{1/2} + cA(n_Q, \delta, 1) \\
\leq& 2\left(\mathbb{P}'_{n_Q}[\hat{h}_{P,i} \neq \hat{h}_Q] \cdot A(n_Q, \delta, 1)\right)^{1/2} + 2cA(n_Q, \delta, 1) \\
\leq& 2\left(\mathbb{P}'_{n_Q}[\hat{h}_{P,i} \neq h^*_Q] \cdot A(n_Q, \delta, 1)\right)^{1/2} \\
&+ 2\left(\mathbb{P}'_{n_Q}[\hat{h}_Q \neq h^*_Q] \cdot A(n_Q, \delta, 1)\right)^{1/2} + 2cA(n_Q, \delta, 1).
\end{aligned}$$

By the second part of Lemma 3 and BCC,

$$\begin{aligned}
\mathcal{E}_Q(\hat{h}_{P,i}) - \mathcal{E}_Q(\hat{h}_Q) \leq& 2\left(2\mathbb{P}_Q[\hat{h}_{P,i} \neq h^*_Q] \cdot A(n_Q, \delta, 1)\right)^{1/2} \\
&+ 2\left(2\mathbb{P}_Q[\hat{h}_Q \neq h^*_Q] \cdot A(n_Q, \delta, 1)\right)^{1/2} + 4c \cdot A(n_Q, \delta, 1) \\
\leq& 2\left(2C_{\beta_Q}\mathcal{E}_Q(\hat{h}_{P,i})^{\beta_Q} \cdot A(n_Q, \delta, 1)\right)^{1/2} \\
&+ 2\left(2C_{\beta_Q}\mathcal{E}_Q(\hat{h}_Q)^{\beta_Q} \cdot A(n_Q, \delta, 1)\right)^{1/2} + 4c \cdot A(n_Q, \delta, 1).
\end{aligned} \tag{14}$$

Assuming that $1 \leq \mathcal{C}(\mathcal{H}_{i^*_Q})$, so that $A(n_Q, \delta, 1) \leq A(n_Q, \delta, \mathcal{C}(\mathcal{H}_{i^*_Q}))$ and plugging in the bound in eq. (13), we can upper bound

$$2\left(2C_{\beta_Q}\mathcal{E}_Q(\hat{h}_Q)^{\beta_Q} \cdot A(n_Q, \delta, 1)\right)^{1/2} \leq 8cA(n_Q, \delta_{i^*_Q}, \mathcal{C}(\mathcal{H}_{i^*_Q}))^{\frac{1}{2-\beta_Q}}.$$

Now if $2\left(2C_{\beta_Q}\mathcal{E}_Q(\hat{h}_{P,i})^{\beta_Q} \cdot A(n_Q, \delta, 1)\right)^{1/2} > 8cA(n_Q, \delta_{i^*_Q}, \mathcal{C}(\mathcal{H}_{i^*_Q}))^{\frac{1}{2-\beta_Q}} \geq \mathcal{E}_Q(\hat{h}_Q)$, then going back to eq. (14), we can upper bound

$$\mathcal{E}_Q(\hat{h}_{P,i}) \leq 8\left(2C_{\beta_Q}\mathcal{E}_Q(\hat{h}_{P,i})^{\beta_Q} \cdot A(n_Q, \delta, 1)\right)^{1/2}.$$

Solving for $\mathcal{E}_Q(\hat{h}_{P,i})$ gives the bound

$$\mathcal{E}_Q(\hat{h}_{P,i}) \leq CA(n_Q, \delta, 1)^{\frac{1}{2-\beta_Q}}.$$

On the other hand, if $2\left(2C_{\beta_Q}\mathcal{E}_Q(\hat{h}_{P,i})^{\beta_Q}\cdot A(n_Q,\delta,1)\right)^{1/2} \leq 8cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}$, then the term $8cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}$ dominates the r.h.s. of in eq. (14) and we get

$$\mathcal{E}_Q(\hat{h}_{P,i}) \leq 24cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}} + \mathcal{E}_Q(\hat{h}_Q) \leq 25cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}.$$

In either case, if $\hat{h} = \hat{h}_{P,i}$ then $\mathcal{E}_Q(\hat{h}_{P,i}) \leq 25cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}$, or equivalently, if $\mathcal{E}_Q(\hat{h}_{P,i}) > 25cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}$, then $\hat{h} = \hat{h}_Q$. We can also argue that if $\mathcal{E}_Q(\hat{h}_{P,i}) \leq \mathcal{E}_Q(\hat{h}_Q)$, then $\hat{h} = \hat{h}_{P,i}$. Suppose that $\mathcal{E}_Q(\hat{h}_{P,i}) \leq \mathcal{E}_Q(\hat{h}_Q)$, then under the event E_H,

$$\begin{aligned}\hat{R}'_Q(\hat{h}_{P,i}) - \hat{R}'_Q(\hat{h}_Q) \leq& R_Q(\hat{h}_{P,i}) - R_Q(\hat{h}_Q) + \left(\mathbb{P}'_{n_Q}[\hat{h}_{P,i} \neq \hat{h}_Q]\cdot A(n_Q,\delta,1)\right)^{1/2} \\ &+ cA(n_Q,\delta,1) \\ \leq& \left(\mathbb{P}'_{n_Q}[\hat{h}_{P,i} \neq \hat{h}_Q]\cdot A(n_Q,\delta,1)\right)^{1/2} + cA(n_Q,\delta,1),\end{aligned}$$

which means that the if-statement condition in Algorithm 1 will be satisfied and $\hat{h} = \hat{h}_{P,i}$.

We can then conclude that under the events E_H and E_Q,

$$\mathcal{E}_Q(\hat{h}) \leq \min\left\{\mathcal{E}_Q(\hat{h}_{P,i}), 25cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}\right\}.$$

Using the transfer exponent condition described in Definition 3 we would get

$$\mathcal{E}_Q(\hat{h}_{P,i}) \leq \mathcal{E}_Q(h^*_{P,i}) + \mathcal{E}_Q(\hat{h}_{P,i}, h^*_{P,i}) \leq \mathcal{E}_Q(h^*_{P,i}) + C_{\rho_{i_P^*}}\mathcal{E}_P(\hat{h}_{P,i}, h^*_{P,i})^{\frac{1}{\rho_{i_P^*}}}. \tag{15}$$

Applying Claim 9, under the event E_P

$$\mathcal{E}_P(\hat{h}_{P,i}) \leq CA(n_P,\delta_i,\mathcal{C}(\mathcal{H}_i))^{\frac{1}{2-\beta_{P,i}}}.$$

Plugging this back into eq. (15), we get

$$\mathcal{E}_Q(\hat{h}_{P,i}) \leq \mathcal{E}_Q(h^*_{P,i}) + C\cdot C_{\rho_i}A(n_P,\delta_i,\mathcal{C}(\mathcal{H}_i))^{\frac{1}{(2-\beta_P)\rho_i}}$$

Finally, we can conclude that under events E_H, E_Q, and E_P, which hold simultaneously with probability of at least $1-3\delta$,

$$\mathcal{E}_Q(\hat{h}) \leq \min\left\{\mathcal{E}_Q(h^*_{P,i}) + C\cdot C_{\rho_i}A(n_P,\delta_i,\mathcal{C}(\mathcal{H}_i))^{\frac{1}{(2-\beta_P)\rho_i}}, cA(n_Q,\delta_{i_Q^*},\mathcal{C}(\mathcal{H}_{i_Q^*}))^{\frac{1}{2-\beta_Q}}\right\}.$$

■

Proof of Claim 8 Since the hypothesis classes are nested, $h^*_Q \in \mathcal{H}_j$ for every $j \geq i^*_Q$. First, we argue that with probability of at least $1-\delta$, for every $j \geq i^*_Q$,

$$h^*_Q \in \hat{\mathcal{H}}^Q_j,$$

which would then imply that $\bigcap_{j \geq i_Q^*}^{\infty} \hat{\mathcal{H}}_j^Q \neq \emptyset$, and consequently $\hat{i}_Q \leq i_Q^*$. Let $\hat{h}_{Q,j}$ be an empirical risk minimizer for level j. Let E_Q be the event where the bounds in Lemma 3 hold for every level of the hierarchy with $\delta_i = \delta w_i$ for each level, so that E_Q occurs with probability of at least $1 - \sum_{j=1}^{\infty} \delta_j = 1 - \delta$. Under E_Q, for every $j \geq i_Q^*$

$$R_Q(\hat{h}_{Q,j}) - R_Q(h_Q^*) \leq \hat{R}_Q(\hat{h}_{Q,j}) - \hat{R}_Q(h_Q^*) + \sqrt{\mathbb{P}_{n_Q}[\hat{h}_{Q,j} \neq h_Q^*] \cdot A(n_Q, \delta_j, \mathcal{C}(\mathcal{H}_j))} + A(n_Q, \delta_j, \mathcal{C}(\mathcal{H}_j)),$$

since h_Q^* is a risk minimizer, moving the risk difference to the right hand side of the inequality and the empirical risk difference to the left hand side gives

$$\hat{R}_Q(h_Q^*) - \hat{R}_Q(\hat{h}_{Q,j}) \leq C\sqrt{\mathbb{P}_{n_Q}[\hat{h}_{Q,j} \neq h_Q^*] \cdot A(n_Q, \delta_j, \mathcal{C}(\mathcal{H}_j))} + cA(n_Q, \delta_j, \mathcal{C}(\mathcal{H}_j)).$$

Therefore, by Definition 8, under E_Q, $h_Q^* \in \hat{\mathcal{H}}_j^Q$ for every $j \geq i_Q^*$, implying that $\hat{i}_Q \leq i_Q^*$. ■

Proof of Claim 9 Let E_P be the event that the bound in Lemma 3 holds over P samples.

$$\mathcal{E}_P(\hat{h}_i, h_{P,i}^*) = R_P(\hat{h}_i) - R_P(h_{P,i}^*) \leq \hat{R}_P(\hat{h}_i) - \hat{R}_P(\hat{h}_{P,i}) \tag{16}$$

$$+ \left(\mathbb{P}_{n_P}[\hat{h}_i \neq h_{P,i}^*] \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} + cA(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i)). \tag{17}$$

Since $\hat{h}_i \in \hat{\mathcal{H}}_i^P$, the expression in 16 can be upper bounded by

$$\left(\mathbb{P}_{n_P}[\hat{h}_i \neq \hat{h}_{P,i}] \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} + cA(n_Q, \delta_i, \mathcal{C}(\mathcal{H}_i)).$$

Furthermore, under E_P we can upper bound

$$\begin{aligned} \mathbb{P}_{n_P}[\hat{h}_i \neq \hat{h}_{P,i}] &\leq \mathbb{P}_{n_P}[\hat{h}_i \neq h_{P,i}^*] + \mathbb{P}_{n_P}[h_{P,i}^* \neq \hat{h}_{P,i}] \\ &\leq 2\mathbb{P}_P[\hat{h}_i \neq h_{P,i}^*] + 2\mathbb{P}_P[h_{P,i}^* \neq \hat{h}_{P,i}] + cA(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i)), \end{aligned}$$

where the second inequality followed by applying the second part of Lemma 3, which is stated in equation 12.

By Bernstein class noise condition (Definition 2), the first two terms above can be upper bounded by

$$C\mathcal{E}_P(\hat{h}_i, h_{P,i}^*)^{\beta_{P,i}} + C\mathcal{E}_P(\hat{h}_{P,i}, h_{P,i}^*)^{\beta_{P,i}},$$

then going back to equation 16, we get

$$\hat{R}_P(\hat{h}_i) - \hat{R}_P(\hat{h}_{P,i}) \leq C\left(\mathcal{E}_P(\hat{h}_i, h_{P,i}^*)^{\beta_{P,i}} \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} \tag{18}$$

$$+ C\left(\mathcal{E}_P(\hat{h}_{P,i}, h_{P,i}^*)^{\beta_{P,i}} \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} + cA(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i)).$$

Again , using Definition 2 and under E_P, we can also upper bound the first term in equation 17,

$$\begin{aligned}\left(\mathbb{P}_{n_P}[\hat{h}_i \neq h^*_{P,i}] \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} &\leq \left(\mathbb{P}_P[\hat{h}_i \neq h^*_{P,i}] \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} \\ &\quad + cA(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i)) \\ &\leq C\left(\mathcal{E}_P(\hat{h}_i, h^*_{P,i})^{\beta_{P,i}} \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} \\ &\quad + CA(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i)).\end{aligned}$$

Note that this upper bound can be absorbed into the bound given in 18 by adjusting the constants. In total, we get

$$\begin{aligned}\mathcal{E}_P(\hat{h}_i, h^*_{P,i}) \leq& C\left(\mathcal{E}_P(\hat{h}_i, h^*_{P,i})^{\beta_{P,i}} \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} \\ &+ C\left(\mathcal{E}_P(\hat{h}_{P,i}, h^*_{P,i})^{\beta_{P,i}} \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} + cA(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i)).\end{aligned}$$

Now since $\hat{h}_{P,i}$ is an ERM over the class $\mathcal{H}_i$ under P, under Definition 2 and event E_P, using lemma 3 we can upper bound its' excess risk by

$$C\left(A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{\frac{1}{2-\beta_{P,i}}},$$

which leads to the upper bound

$$\begin{aligned}\mathcal{E}_P(\hat{h}_i, h^*_{P,i}) \leq& C\left(\mathcal{E}_P(\hat{h}_i, h^*_{P,i})^{\beta_{P,i}} \cdot A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{1/2} \\ &+ C\left(A(n_P, \delta_i, \mathcal{C}(\mathcal{H}_i))\right)^{\frac{1}{2-\beta_{P,i}}}.\end{aligned}$$

Consider the inequality above without the term on the second line, we can then solve for $\mathcal{E}_Q(\hat{h}_Q)$ and get the bound in the statement of this claim, since the solution will in the same order as the second line. ■

Appendix D. Remaining Lower Bound Proofs

D.1. Proof of Theorem 1

The proof builds on Theorem 1 in Hanneke and Kpotufe (2019), simply by enriching the family of distributions used therein. Let $\varepsilon_Q \doteq \left(\frac{d}{n_Q}\right)^{\frac{1}{2-\beta_Q}}$ and $\varepsilon_P \doteq \left(\frac{d}{n_P}\right)^{\frac{1}{(2-\beta_P)\rho}}$, note that

$$\min\left\{\varepsilon_P + \alpha, \varepsilon_Q\right\} \geq \min\left\{\max\left\{\varepsilon_P, \alpha\right\}, \varepsilon_Q\right\} = \max\left\{\min\left\{\varepsilon_P, \varepsilon_Q\right\}, \min\left\{\alpha, \varepsilon_Q\right\}\right\},$$

where the equality follows by distributing the min. Let $c_1, c_2 \leq 2$ be constants that will be determined later; define

$$\epsilon_1 \doteq c_1 \min\left\{\varepsilon_P, \varepsilon_Q\right\},$$

and

$$\epsilon_2 \doteq c_2 \min\left\{\alpha, \varepsilon_Q\right\}.$$

Theorem 1 of Hanneke and Kpotufe (2019) gives a lower bound of order $c\epsilon_1$ which holds with probability of at least $\frac{3-2\sqrt{2}}{8}$, for some universal constant c. Here, we will construct another family of distributions that would lead to a lower bound of order $\tilde{c}\epsilon_2$ for a universal constant $\tilde{c}$. In fact, the only difference is that the source distribution is the same for all the members of hard family of distributions.

Source and target marginal distributions are supported on a set of points $x_0, x_1, \ldots, x_{d-1}$ in the domain $\mathcal{X}$ that is shattered by $\mathcal{H}$. Only the target distribution in the family of hard distributions $\{P^{n_P} \times Q_\sigma^{n_Q}\}$ depends on $\sigma \in \{\pm 1\}^{d-1}$. Source marginal distribution P_X is the uniform distribution on $x_0, x_1, \ldots x_{d-1}$, and $P_{Y|X=x_i}(Y=1) = 1$. For the target, let $Q_{X,\sigma}(x_0) = 1 - \epsilon_2^{\beta_Q}$ and $Q_{X,\sigma}(x_i) = \frac{\epsilon_2^{\beta_Q}}{d-1}$for $i \geq 1$. The target labels for $i \geq 1$ are given by $Q_{\sigma,Y|X=x_i}(Y=1) = \frac{1}{2} + \frac{\sigma_i}{4} \cdot \epsilon_2^{1-\beta_Q}$, and $Q_{\sigma,Y|X=x_0}(Y=1) = 1$.

Now we can verify the Bernstein class noise condition. Let $h_\sigma \in \mathcal{H}$ be the Bayes classifier under Q_σ and let $\delta(.,.)$ denote number of coordinates σ, σ' differ, or equivalently Hamming distance of $\frac{\sigma+1^{d-1}}{2}$ and $\frac{\sigma'+1^{d-1}}{2}$ Note that for any distinct pair $\sigma, \sigma' \in \{\pm 1\}^{d-1}$,

$$\mathcal{E}_{Q_{\sigma'}}(h_\sigma) = \delta(\sigma', \sigma) \cdot \frac{\epsilon_2^{\beta_Q}}{2(d-1)} \cdot \epsilon_2^{1-\beta_Q} = \frac{\delta(\sigma', \sigma)}{2(d-1)} \cdot \epsilon_2, \tag{19}$$

while

$$\mathbb{P}_{Q_{\sigma'}}(h_\sigma \neq h_{\sigma'}) = \delta(\sigma', \sigma) \cdot \frac{\epsilon_2^{\beta_Q}}{d-1}.$$

Additionally, for every σ', $\mathcal{E}_{Q_{\sigma'}}(h_P^*) \leq \epsilon_2/2 \leq c_2/2 \cdot \alpha$.

By Proposition 5 of Hanneke and Kpotufe (2019), there exists a $(d-1)/8$ packing $\mathcal{N}$ of the $d-1$ dimensional hypercube such that the all ones vector $1^{d-1} \in \mathcal{N}$ and $|\mathcal{N}| \geq 2^{\frac{d-1}{8}}$, where the metric used for the packing is Hamming distance. Suppose that $\mathcal{N}'$ is such a packing over the $\{\pm 1\}^{d-1}$ hypercube. Now consider the restriction of the family of distributions to $\sigma \in \mathcal{N}'$, by eq. (19), for every distinct $\sigma, \sigma' \in \mathcal{N}'$, $\mathcal{E}_{Q_\sigma}(h_{\sigma'}) \geq \frac{\epsilon_2}{16}$.

Next, we show that the KL divergence between distributions parameterized by any two distinct σ, σ' is small. First, write

$$\mathcal{D}_{\text{kl}}\left(P^{n_P} \times Q_\sigma^{n_Q} | P^{n_P} \times Q_{\sigma'}^{n_Q}\right) = n_Q \mathcal{D}_{\text{kl}}(Q_\sigma | Q_{\sigma'}) = n_Q \sum_{i=1}^{d-1} \frac{\epsilon_2^{\beta_Q}}{d-1} \cdot \mathcal{D}_{\text{kl}}\left(Q_{\sigma,Y|X=x_i} | Q_{\sigma',Y|X=x_i}\right). \tag{20}$$

Now we use Lemma 2 in Hanneke and Kpotufe (2019), which gives an upper bound on KL divergence of two Bernoulli distributions with small bias, to get that $\mathcal{D}_{\text{kl}}\left(Q_{\sigma,Y|X=x_i} | Q_{\sigma',Y|X=x_i}\right) \leq c_0 \epsilon_2^{2-2\beta_Q}$, as long as $\epsilon_2^{1-\beta_Q} < 1/2$. Going back to eq. (20), we get

$$\mathcal{D}_{\text{kl}}\left(P^{n_P} \times Q_\sigma^{n_Q} | P^{n_P} \times Q_{\sigma'}^{n_Q}\right) \leq n_Q \cdot \epsilon_2^{\beta_Q} \cdot c_0/4 \cdot \epsilon_2^{2-2\beta_Q} \leq c_0/4 \cdot n_Q \cdot \epsilon_2^{2-\beta_Q} \leq c_0/4 \cdot c_2^{2-\beta_Q} \cdot d.$$

Now pick c_2 such that $c_0/4 \cdot c_2^{2-\beta_Q} \cdot d < 1/8 \log(\frac{d-1}{8})$, so that we can apply Proposition 4 of Hanneke and Kpotufe (2019) (which is Theorem 2.5 of Tsybakov (2009)) to get that

$$\sup_{\sigma \in \mathcal{N}} \mathbb{P}_{P^{n_P} \times Q_\sigma^{n_Q}} \left[\mathcal{E}_{Q_\sigma}(\hat{h}) \geq 1/32 \cdot \epsilon_2\right] \geq \frac{3-2\sqrt{2}}{8}.$$

D.2. Proof of Theorem 5

Let x_0, x_1 be the set of points that are exclusively shattered by $\mathcal{H}_2$, then it is possible to pick $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2 \setminus \mathcal{H}_1$ such that h_1 and h_2 disagree on exactly one of x_0, x_1. Without loss of generality assume that $h_1(x_0) = h_2(x_0) = y_0$ and $h_2(x_1) \neq h_1(x_1)$. Since there is no label noise, $\beta_{P,i} = \beta_{Q,i} = 1$, for $i \in \{1, 2\}$.

Source distribution does not depend on b, so we have $P_{X,\sigma}(x_0) = P_{X,\sigma}(x_1) = 1/2$, and the labels are given by h_2, Target marginal distribution also does not depend on b, and is given by $Q_{X,\sigma}(x_0) = 1 - \alpha$, and $Q_{X,\sigma}(x_1) = \alpha$. The labels for target are set so that when $\sigma = 1$, h_1 is a risk minimizer, and when $\sigma = 2$, h_2 is a risk minimizer. That is, $Q_{Y|X=x,\sigma}(1) = h_\sigma(x)$. Note that for every $\sigma, i \in \{1, 2\}$

$$\mathcal{E}_{Q_\sigma}(h^*_{P,i}) = \begin{cases} 0 & \sigma = i \\ \alpha & b \neq i. \end{cases}$$

For any classifier $\hat{h}$, define

$$\hat{\sigma} \doteq \begin{cases} 1 & \hat{h}(x_1) = h_1(x_1) \\ 2 & \hat{h}(x_1) = h_2(x_1). \end{cases}$$

Let E_σ be the event where under target all the samples are (x_0, y_0), then for any σ

$$\underset{P_\sigma^{n_P} \times Q_\sigma^{n_Q}}{\mathbb{P}} [E_B] = (1-\alpha)^{n_Q} \geq \left(1 - \frac{1}{c_1 n_Q}\right)^{n_Q} \geq 1 - 1/c_1.$$

Under the event E_B, $\hat{h}$ cannot distinguish between Q_1 and Q_2. So under the event E_B, no classifier can output the correct answer more than half the times, so with probability of at least $\frac{1}{2} \cdot (1 - 1/c_1)$

$$\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \mathcal{E}_{Q_\sigma}(h_{\hat{\sigma}}) = \mathcal{E}_{Q_\sigma}(h_{P,\hat{\sigma}}) = \max_i \left\{\mathcal{E}_{Q_\sigma}(h^*_{P,i})\right\}.$$

Setting $c_1 = 2$ proves the statement.

D.3. Proof of Proposition 3

Proper estimators. Let $\hat{h} \in \mathcal{H}$ be some proper estimator. If $\hat{h} \in \{h_1, h_2\}$, set $\sigma_1 = +1$ and σ_2 arbitrary. If $\hat{h} \in \{h_1', h_2'\}$, then set $\sigma_1 = -1$, and σ_2 such that the region that has mass $(\frac{1}{c_1 n_P})^{1/\rho_a}$ is labelled incorrectly. That is, if $\hat{h} = h_1'$, $\sigma_2 = +1$. It is easy to see that with this choice of σ, $\mathcal{E}_{Q_\sigma}(\hat{h}, h^*_{Q_\sigma}) \geq (\frac{1}{32 n_P})^{1/\rho_a}$.

Improper estimators. Let $\hat{h}$ be an improper estimator. For $-1/2 \leq \epsilon \leq 1/2$, we say that $\hat{h}$ has bias ϵ on an interval I if it classifies $1/2 + \epsilon$ fraction of the interval under uniform measure as positive. That is, $\mathbb{E}_{U(I)}\left[\hat{h}(X)\right] = 2\epsilon$. Note that if a classifier has bias ϵ on an interval I, and $\operatorname{sign}(\epsilon) \neq \operatorname{sign}(I)$, then the error of the classier on that interval is $1/2 + |\epsilon|$. Even if $\operatorname{sign}(\epsilon) = \operatorname{sign}(I)$, as long all of the interval has the same label, the error will be at least $1/2 - |\epsilon|$.

For simplicity let $a \doteq (\frac{1}{c_1 n_P})^{1/\rho_a}$, $b \doteq (\frac{1}{c_1 n_P})^{1/\rho_b}$, and recall that $\Delta = a - b$. Also note that in our construction, the risk minimizer has risk equal to a under $\sigma = (+1, \sigma_2)$, while the risk minimizer under $\sigma = (-1, \sigma_2)$ has risk equal to $\Delta/2$.

Now fix some improper estimator $\hat{h}$, and let $\epsilon_L, \epsilon_{L_{in}}, \epsilon_{R_{in}}, \epsilon_R$ be biases of $\hat{h}$ on L_{out}, L_{in}, R_{in} and R_{out} respectively. We break down the proof to three cases.

Case 1: $\epsilon_{L_{in}}, \epsilon_{R_{in}} \leq -1/4$.

Let $\sigma_1 = +1$ so that the intervals L_{in}, R_{in} have positive labels and $\hat{h}$ has risk of at least $3/4(a+b)$ on the intervals R_{in}, L_{in}. We can pick σ_2 such that the error on the region $L_{out} \cup R_{out}$ is at least

$$\left((\frac{1}{2} + \max(|\epsilon_L|, |\epsilon_R|)) + (\frac{1}{2} - \min(|\epsilon_L|, |\epsilon_R|)) \right) \cdot \frac{\Delta}{2} \geq \frac{\Delta}{2},$$

by making sure that $\hat{h}$ makes more error on the L_{out} or R_{out} interval that has the maximum absolute bias. Then for this σ, $R_\sigma(\hat{h}) \geq \frac{a-b}{2} + \frac{3(a+b)}{4} \geq \frac{5a}{4}$, while the risk minimizer has error of a, so $\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{a}{4}$.

Case 2:$|\epsilon_L + \epsilon_R| \geq 1/4$, and Case 1 condition does not hold.

Set $\sigma_1 = -1$, and pick σ_2 such that $\sigma_2 = -\text{sign}\,(\epsilon_L + \epsilon_R)$. Note that total bias over the region $L_{out} \cup R_{out}$ would be $\frac{\epsilon_R+\epsilon_L}{2}$, since $(1/2 + \epsilon_L) \cdot \frac{\Delta}{2} + (1/2 + \epsilon_R) \cdot \frac{\Delta}{2} = \left(\frac{1}{2} + \frac{\epsilon_R+\epsilon_L}{2}\right) \cdot \Delta$. On the other hand, since we are in case 2, it must be that either $\epsilon_{L_{in}} > -1/4$ or $\epsilon_{R_{in}} > -1/4$, which would mean that the error over the intervals L_{in} and R_{in} is at least $1/4b$

Then we can ensure that

$$\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \left(\frac{1}{2} + \frac{|\epsilon_R + \epsilon_L|}{2}\right) \cdot (a-b) + \frac{b}{4} - \frac{a-b}{2} \geq \frac{a-b}{8} + \frac{b}{4} \geq \frac{a}{8}.$$

Case 3:$|\epsilon_R + \epsilon_L| < 1/4$ and the condition in Case 1 does not hold.

Set $\sigma_1 = -1$, and pick σ_2 such that whichever of R_{in} or L_{in} that has more positive bias is assigned mass a. Since we are not in Case 1, $\max(\epsilon_{R_{in}}, \epsilon_{L_{in}}) > -1/4$, leading to error of at least $\frac{a}{4}$ over $R_{in} \cup L_{in}$. On the other hand, since the bias in the regions R_{out} and L_{out} is $\frac{\epsilon_R+\epsilon_L}{2}$, we have

$$\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \left(\frac{1}{2} - \frac{|\epsilon_R + \epsilon_L|}{2}\right) \cdot (a-b) + \frac{a}{4} - \frac{a-b}{2} \geq \frac{a}{8}.$$

The statement of the proposition follows by lower bounding $\left(\frac{1}{32}\right)^{1/\rho_a} \geq \frac{1}{32}$.

Appendix E. Adaptivity Lower Bounds for a Larger Class

In this section, restricting to proper learners, we show similar adaptivity lower bounds as in Theorem 4 for a larger model class. Let $\bar{\mathcal{H}}_1 = \{h_t\}$ be the class of one sided thresholds, where $h_t(x) = \text{sign}\,(x - t)$. Let $\bar{\mathcal{H}}_2$ additionally include one sided intervals, where only the points inside a closed interval are labelled positive.

Theorem 6 *Let $\bar{\mathcal{H}}_1$ and $\bar{\mathcal{H}}_2$ be the class of one sided thresholds and intervals as described above. Pick any $\rho_a > \rho_b \geq 1$, and any n_P and n_Q, where $\left(\frac{1}{32n_P}\right)^{1/\rho_a} \leq \min\left\{\frac{1}{24}, \frac{1}{32n_Q}\right\}$. There exists a family of distributions $\{(P_\sigma, Q_\sigma)\}$, indexed by some σ, such that the following hold.*

(i) For all σ, minimal transfer exponents from P_σ to Q_σ are the set $\{\rho_1, \rho_2\} = \{\rho_a, \rho_b\}$.

(ii) For all σ, we have $\min_i \phi_\flat(i) = \left(\frac{1}{n_P}\right)^{1/\rho_b}$, strictly less than $\max_i \phi_\flat(i) = \left(\frac{1}{n_P}\right)^{1/\rho_a}$.

$$\textit{We have that, } \forall \hat{h}, \quad \sup_{\sigma} \mathbb{P}_{P_\sigma^{n_P} \times Q_\sigma^{n_Q}} \left[\mathcal{E}_{Q_\sigma}(\hat{h}) \geq \frac{1}{64} \cdot \max_i \phi_\flat(i) \right] \geq 1/8.$$

Proof of Theorem 6

In this proof, since the construction is very similar to the one in Theorem 4, we will use the same notation and refer to the objects defined there.

The family of distributions. We divide the unit interval as in the proof of Theorem 4 and let $\sigma \in \{\pm 1\}$. Recall $\Delta \doteq \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a} - \left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$, where c_1 is constant that will be picked later . Source distributions P_σ are the same as in the construction in Theorem 4. The target marginals are as follows.

- $Q_{X,(1,1)}(L_{out}) = 0$, and $Q_{X,(1,1)}(R_{out}) = \Delta$.
- $Q_{X,(1,-1)}(L_{out}) = \Delta$, and $Q_{X,(1,1)}(R_{out}) = 0$.
- $Q_{X,(-1,\cdot)}(L_{out}) = Q_{X,(-1,\cdot)}(R_{out}) = \Delta/2$.
- $Q_{X,(\cdot,+1)}(L_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$ and $Q_{X,(\cdot,+1)}(R_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$.
- $Q_{X,(\cdot,-1)}(L_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_b}$ and $Q_{X,(\cdot,-1)}(R_{in}) = \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$.
- The remaining mass is in the middle interval, so $Q_X([1/3+r, 2/3-r]) = 1 - 2\left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$.

The masses in all intervals except for R_{in} and L_{in} are distributed uniformly within that interval. For intervals L_{in} and R_{in}, the densities are

- $\mathrm{f}_L(x) \propto |x - (1/3 + r/2)|^{\frac{1}{\rho_a}-1}$ and $\mathrm{f}_R(x) \propto |x - (2/3 - r/2)|^{\frac{1}{\rho_b}-1}$ when $\sigma_2 = +1$, and
- $\mathrm{f}_L(x) \propto |x - (1/3 + r/2)|^{\frac{1}{\rho_b}-1}$ and $\mathrm{f}_R(x) \propto |x - (2/3 - r/2)|^{\frac{1}{\rho_a}-1}$ if $\sigma_2 = -1$.

In this construction, only the labels of the intervals R_{in} and L_{in} depend on σ, and are given by $Y_{Q,\sigma}(L_{in}) = Y_{Q,\sigma}(R_{in}) = \sigma_1$. If the intervals L_{out} and R_{out} have non zero mass under σ, then they are labelled $+1$. The middle interval $[1/3+r, 2/3-r]$ is labelled -1 for every σ.

Claim 10 *Recall $\mathcal{H}_1 \subset \mathcal{H}_2$ from Theorem 4. For every σ and $i \in \{1,2\}$, we have $\mathcal{E}_{Q_\sigma}(h^*_{P_\sigma,i}) = 0$ and the risk minimizers over the classes $\mathcal{H}_1$ and $\mathcal{H}_2$ under both source and target are the same as the risk minimizers over classes $\bar{\mathcal{H}}_1, \bar{\mathcal{H}}_2$.*

Proof Since the middle interval has a large negative mass and $\bar{\mathcal{H}}_1$ is the class of one sided thresholds, any one sided threshold that positively labels the middle interval cannot be a risk minimizer. Since the threshold is in the intervals $L_{in} \cup L_{out}$, we can see that the risk minimizers are either h_1 or h_1' and are shared between source and target, implying that $\mathcal{E}_{Q_\sigma}(h^*_{P_\sigma,1}) = 0$.

Under source a one sided interval that is a risk minimizer would choose to label intervals L_{out}, L_{in} accurately, since there is large negative mass in the middle interval, and the mass in R_{out} is small than the mass in L_{out} by a constant factor. Under target, there are multiple one sided intervals that are risk minimizers, but since the total positive mass in the left side ($L_{out} \cup L_{in}$) is equal to the total positive mass in the right side $R_{out} \cup R_{in}$, and the negative mass in the center interval is very large, one of h_2 or h_2' would also be a risk minimizer under target depending on σ_1, and it would be shared with source, so $\mathcal{E}_{Q_\sigma}(h^*_{P_\sigma,2}) = 0$.

■

Claim 11 *For every $\sigma = (\sigma_1, \sigma_2) \in \{\pm 1\}^2$, if $\sigma_2 = 1$, then ρ_b and ρ_a are transfer exponents from P_σ to Q_σ with respect to $\bar{\mathcal{H}}_1$ and $\bar{\mathcal{H}}_2$ respectively. If $\sigma_2 = -1$, then they are transfer exponents with respect to $\bar{\mathcal{H}}_2$ and $\bar{\mathcal{H}}_1$ instead.*

Proof To see that ρ_a and ρ_b are transfer exponents, note that the labels are always the same under source and target, and the only intervals where ratio of densities of source and target is not a constant are L_{in} and R_{in}. In the case of one sided thresholds, if some $h_t \in \bar{\mathcal{H}}_1$ has source excess risk that is $\epsilon < \frac{1}{c_1 n_P}$, it must be that $t \in R_{in}$. Which then implies that its' target excess risk is going to be of order $c\left(\frac{\epsilon}{c_1 n_P}\right)^{1/\rho_b}$ or $c\left(\frac{\epsilon}{c_1 n_P}\right)^{1/\rho_a}$ depending on σ_2. Similarly, any $h \in \bar{\mathcal{H}}_2$ that has source excess risk $\epsilon < \frac{1}{c_1 n_P}$ must be a once sided interval with both of its' end points in the region $L_{out} \cup L_{in}$. If the region that it makes error on is not in L_{in}, then the ratio of source and target excess risks is bounded by a constant, while if the error region is in L_{in}, h will have excess risk of order $\left(\frac{\epsilon}{c_1 n_P}\right)^{1/\rho_b}$ or $c\left(\frac{\epsilon}{c_1 n_P}\right)^{1/\rho_a}$ depending on σ_2.

To argue that ρ_a and ρ_b are minimal transfer exponents, fix $\sigma_2 = -1$ and consider a sequence of one sided thresholds $h_{2/3-r+t}$ as $t \to 0$. Target excess risk for this sequence decreases at the rate $\left(\frac{t}{c_1 n_P}\right)^{1/\rho_a}$, while under source it would be $\frac{t}{c_1 n_P}$. If $\rho' < \rho_a$ is a transfer exponent, the ratio of the excess risks $\frac{t^{1/\rho_a}}{t^{1/\rho'}}$ would not be bounded by a constant as $t \to 0$. A similar argument works for $\bar{\mathcal{H}}_2$ and $\sigma_2 = +1$, since $h^*_{P,2} \in \mathcal{H}_2$. ■

Next, we show that for every proper learner $\hat{h} \in \bar{\mathcal{H}}_2$, there is a distribution in the family where $\hat{h}$ incurs large excess risk.

Proposition 5 *Let $c_1 = 32$. For any proper learner $\hat{h}$, there exists $\sigma \in \{\pm 1\}^2$ such that $\mathcal{E}_{Q_\sigma}(\hat{h}) \geq 1/2 \cdot \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$*

Proof By construction, for every proper learner $\tilde{h} \in \mathcal{H}_2$, there exists σ such that $\mathcal{E}_{Q_\sigma}(\tilde{h}) \geq \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$. We project every proper learner $\hat{h} \in \bar{\mathcal{H}}_2$ by picking $h \in \mathcal{H}_2$ whose labeling on the regions R_{in} and L_{in} agrees the most with h, under the uniform measure over L_{in} and R_{out}. In the case that $\hat{h}$ has positive labels in both of the regions, its' excess risk will be a large constant. So $\hat{h}$ agrees with its' projection h on at least one of the intervals R_{in} or L_{out} plus at least half of the other interval. Thus, if σ is such that $\mathcal{E}_{Q_\sigma}(h) \geq \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$, then $\mathcal{E}_{Q_\sigma}(\hat{h}) \geq 1/2 \cdot \left(\frac{1}{c_1 n_P}\right)^{1/\rho_a}$. ■

We define the event B and randomize the choice of σ as in the proof of Theorem 4. The constructions are such that the event B has exactly the same probability as in the proof of Theorem 4, and the rest of the proof follows by exactly the same argument. ■