A NO-FREE-LUNCH THEOREM FOR MULTITASK LEARNING

A PREPRINT

Steve Hanneke
Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Samory Kpotufe
Columbia University, Statistics
skk2175@columbia.edu

August 5, 2020

ABSTRACT

Multitask learning and related areas such as multi-source domain adaptation address modern settings where datasets from N related distributions $\{P_t\}$ are to be combined towards improving performance on any single such distribution $\mathcal{D}$. A perplexing fact remains in the evolving theory on the subject: while we would hope for performance bounds that account for the contribution from multiple tasks, the vast majority of analyses result in bounds that improve at best in the number n of samples per task, but most often do not improve in N. As such, it might seem at first that the distributional settings or aggregation procedures considered in such analyses might be somehow unfavorable; however, as we show, the picture happens to be more nuanced, with interestingly hard regimes that might appear otherwise favorable.

In particular, we consider a seemingly favorable classification scenario where all tasks P_t share a common optimal classifier h^*, and which can be shown to admit a broad range of regimes with improved *oracle rates* in terms of N and n. Some of our main results are as follows:

- We show that, even though such regimes admit minimax rates accounting for both n and N, no adaptive algorithm exists; that is, without access to distributional information, *no* algorithm can guarantee rates that improve with large N for n fixed.
- With a bit of additional information, namely, a ranking of tasks $\{P_t\}$ according to their *distance* to a target $\mathcal{D}$, a simple rank-based procedure can achieve near optimal aggregations of tasks' datasets, despite a search space exponential in N. Interestingly, the optimal aggregation might exclude certain tasks, even though they all share the same h^*.

1 Introduction

Multitask learning and related areas such as *multi-source domain adaptation* address a statistical setting where multiple datasets $Z_t \sim P_t, t = 1, 2, \ldots,$ are to be aggregated towards improving performance w.r.t. a single (or any of, in the case of multitask) such distributions P_t. This is motivated by applications in the sciences and engineering where data availability is an issue, e.g., medical analytics typically require aggregating data from loosely related subpopulations, while identifying traffic patterns in a given city might benefit from pulling data from somewhat similar other cities.

While these problems have received much recent theoretical attention, especially in classification, a perplexing reality emerges: the bulk of results appear to show little improvement from such aggregation over using a single data source. Namely, given N datasets, each of size n, one would hope for convergence rates in terms of the aggregate data size $N \cdot n$, somehow adjusted w.r.t. *discrepancies* between distributions P_t's, but which clearly improve on rates in terms of just n as would be obtained with a single dataset. However, such clear improvements on rates appear elusive, as typical bounds on excess risk w.r.t. a target $\mathcal{D}$ (i.e., one of the P_t's) are of the form (see e.g., [CKW08, BDB08, BDBC$^+$10])

$$\mathcal{E}_{\mathcal{D}}(\hat{h}) \lesssim (nN)^{-\alpha} + n^{-\alpha} + \text{disc}(\{P_t\}\,;\mathcal{D})\,, \text{ for some } \alpha \in [1/2, 1], \tag{1}$$

where in some results, one of the last two terms is dropped. In other words, typical upper-bounds are either dominated by the rate $n^{-\alpha}$, or altogether might not go to 0 with sample size due to the *discrepancy* terms $\text{disc}(\{P_t\}\,;\mathcal{D}) > 0$, even while the excess risk of a naive classifier $\tilde{h}$ trained on the target dataset would be $\mathcal{E}_{\mathcal{D}}(\tilde{h}) \propto n^{-\alpha} \to 0$. As such,

it might seem at first that there is a gap in either algorithmic approaches, formalism and assumptions, or statistical analyses of the problem. However, as we argue here, no algorithm can guarantee a rate improving in aggregate sample size for n fixed, even under seemingly generous assumptions on how sources P_t's relate to a given target $\mathcal{D}$.

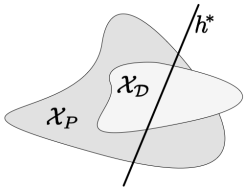
In particular, we consider a seemingly favorable classification setting, where all data distributions P_t's induce the same optimal classifier h^* over a hypothesis class $\mathcal{H}$. This situation is of independent interest, e.g., appearing recently under the name of *invariant risk minimization* (see [ABGLP19] where it is motivated through invariants under causality), but is motivated here by our aim to elucidate basic limits of the multitask problem. As a starting point to understanding the best achievable rates, we first establish minimax upper and lower bounds, up to log terms, for the setting. These oracle rates, as might be expected in such benign settings, do indeed improve with both N and n, as allowed by the level of *discrepancy* between distributions, appropriately formalized (Theorem 1). We then turn to characterizing the extent to which such favorable rates might be achieved by *reasonable* procedures, i.e., adaptive procedures with little to no prior distributional information. Many interesting messages arise, some of which we highlight below:

- *No adaptive procedure exists* outside a threshold $\beta < 1$, where $\beta \in [0, 1]$ (a so-called Bernstein class condition) parametrizes the level of *noise*[1] in label distribution. Namely, while oracle rates might decrease fast in N and n, no procedure based on the aggregate samples alone can guarantee a rate better than $n^{-1/(2-\beta)}$ without further favorable restrictions on distributions (Theorem 5).
- At low noise $\beta = 1$ (e.g., so called Massart's noise) even the naive, yet common approach of *pooling* all datasets, i.e., treating them as if identically distributed, is nearly minimax optimal, achieving rates improving in both N and n. This of course would not hold if optimal classifiers h^*'s differ considerably across P_t's (Theorem 3).
- At any noise level, a ranking of sources P_t's according to discrepancy to a target $\mathcal{D}$ is sufficient information for (partial) adaptivity. While a precise ranking is probably unlikely in practice, an approximate ranking might be available, as domain knowledge on how sources rank in fidelity w.r.t. to an ideal target data: e.g., in settings such as learning with slowly drifting distributions. Here we show that a simple rank-based procedure, using such ranking, efficiently achieves a near optimal aggregation rate, despite the exponential search space of over 2^N possible aggregations (Theorem 4).

Interestingly, even assuming all h^*'s are the same, the optimal aggregation of datasets can change with the choice of target $\mathcal{D}$ in $\{P_t\}$ (see Theorem 2), due to the inherent asymmetry in the information tasks might have on each other, e.g., some P_t might yield data in P_s-dense regions but not the other way around. Hence, to capture a proper picture of the problem, we cannot employ a symmetric notion of discrepancy as is common in the literature on the subject. Instead, we proceed with a notion of *transfer exponent*, which we recently introduced in [HK19] for the setting of domain adaptation with $N = 1$ source distribution, and which we show here to also successfully capture performance limits in the present setting with $N \gg 1$ sources (see definition in Section 3.1).

We note that in the case $N = 1$, there is *always* a minimax adaptive procedure (as shown in [HK19]), while this is not the case here for $N \gg 1$. In hindsight, the reason is simple: considering $N = O(1)$, i.e., as a constant, there is no consequential difference between a rate in terms of $N \cdot n$ and one in terms of n. In other words, the case $N = O(1)$ does not yield adequate insight into multitask with $N \gg 1$, and therefore does not properly inform practice as to the best approaches to aggregating and benefiting from multiple datasets.

Background and Related Work



The bulk of theoretical work on multitask and related areas build on early work on domain adaptation (i.e., $N = 1$) such as [BDBCP07, CMRR08, BDBC$^+$10], which introduce notions of discrepancy such as the $d_\mathcal{A}$-divergence, $\mathcal{Y}$-discrepancy, that specialize the *total-variation* metric to the setting of domain adaptation. These notions often result in bounds of the form (1), starting with [CKW08, BDBC$^+$10]. Such bounds can in fact be shown to be tight w.r.t. the discrepancy term $\mathrm{disc}(\{P_t\}; \mathcal{D})$, for given distributional settings and sample sizes, owing for instance to early work on the limits of learning under distribution drift (see e.g. [Bar92]). However, the rates of (1) appear pessimistic when we consider settings of interest here where h^*'s remain the same (or nearly so) across tasks, as they suggest no general improvement on risk with larger sample size. Consider for instance simple situations as depicted on the right, where a source P and target $\mathcal{D}$ (with respective supports $\mathcal{X}_P, \mathcal{X}_\mathcal{D}$) might differ considerably in the mass they assign to regions of data space, thus inducing large discrepancies, but where both assign sufficient mass to decision boundaries to help identify h^* with enough samples from *either* distribution. Proper parametrization resolves this issue, as we show through new multitask rates $\mathcal{E}_\mathcal{D}(\hat{h}) \to 0$ in natural situations with $\mathrm{disc}(\{P_t\}; \mathcal{D}) > 0$, and even with no target sample.

[1] The term *noise* is used here to indicate nondeterminism in the label Y of a sample point X, and so is rather non-adversarial.

We remark that other notions of discrepancy, e.g., *Maximum Mean Discrepancy* [GSH$^+$09], *Wasserstein distance* [RHS17, SQZY18] are employed in domain adaptation; however they appear relatively less often in the theoretical literature on multitask and related areas. For multitask, the work of [BDB08] proposes a more-structured notion of task relatedness, whereby a source P_t is induced from a target $\mathcal{D}$ through a transformation of the domain $\mathcal{X}_{\mathcal{D}}$; that work also incurs an $n^{-1/2}$ term in the risk bound, but no discrepancy term. The work of [MMR09] considers Rényi divergences in the context of optimal aggregation in multitask under population risk, but does not study sample complexity.

The use of a non-metric discrepancy such as Rényi divergence brings back an important point: two distributions might have asymmetric information on each other w.r.t. domain adaptation. Such insight was raised recently, independently in [KM18, HK19, APMS19], with various natural examples therein (see also Section 3.1). In particular, it motivates a more unified view of multitask and *multisource domain* adaptation, which are often treated separately. Namely, if the goal in multitask is to perform as well as possible on each task P_t in our set, then as we show, such asymmetry in information between tasks calls for different aggregation of datasets for each target P_t: in other words, treating multitask as separate multisource problems, even if the optimal h^* is the same across tasks.

In contrast, a frequent aim in multitask, dating back to [Car97], has been to ideally arrive at a single aggregation of task datasets that simultaneously benefits all tasks P_t's. Following this spirit, many theoretical works on the subject are concerned with bounds on *average risk* across tasks [Bax97, AZ05, MPRP13, PM13, PL14, YHC13, MPRP16], rather than bounding the supremum risk accross tasks – i.e., treating multitask as separate multisources, as of interest here. Some of these average bounds, e.g., [MPRP13, MPRP16, PM13], remove the dependence on discrepancy inherent in bounds of the form (1), but maintain a term of the form $n^{-\alpha}$; in other words, any bound on supremum risk derived from such results would be in terms of a single dataset size n. The work of [BHPQ17] directly addresses the problem of bounding the supremum risk, but also incurs a term of the form $n^{-\alpha}$ in the risk bound.

In the context of multisource domain adaptation, it has been recognized in practice that some datasets that are too *far* from the ideal target might hurt target performance and should be downweighted accordingly, a situation that has been termed *negative transfer*. These situations further motivate the need for *adaptive* procedures that can automatically identify good datasets. As far as theoretical insights, it is clear that negative transfer might happen for instance, under ERM, if optimal classifiers h^*'s are considerably different across tasks. Interestingly, even when h^*'s are allowed arbitrarily close (but not equal), [BDLLP10] shows that, for $N = 1$, the source dataset can be useless without labeled target data. We later derived minimax lower-bounds for the case $N = 1$ with or without labeled target data in [HK19] for a range of situations including those considered in [BDLLP10]. Such results however do not quite confirm negative transfer, as they allow the possibility that useless datasets might remain safe to include. For the multisource case $N \gg 1$, to the best of our knowledge, situations of negative transfer have only been described in adversarial settings with corrupted labels. For instance, the recent papers of [Qia18, MMM19, KFAL20] show limits of multitask under various adversarial corruption of labels in datasets, while [SZ19] derives a positive result, i.e., rates (for Lipschitz loss) decreasing in both N and n, up to excluded or downweighted datasets. The procedure of [SZ19] is however nonadaptive as it requires known noise proportions.

[KFAL20] is of particular interest as they are concerned with *adaptivity* under label corruption. They show that, even if at most $N/2$ of the datasets are corrupted, no procedure can get a rate better than $1/n$. In contrast, in the stochastic setting considered here, there is always an adaptive procedure with significantly faster rates than $1/n$ if at most $N/2$ (in fact any fixed fraction) of distributions P_t's are *far* from the target $\mathcal{D}$ (Theorem 9 of Appendix B). This dichotomy is due to the strength of adversarial corruptions they consider which in effect can flip optimal h^*'s on corrupted sources. What we show here is that, even when h^*'s are fixed across tasks, and datasets are sampled i.i.d. from each P_t, i.e., non-adversarially, no algorithm can achieve a rate better than $1/\sqrt{n}$ while a non-adaptive oracle procedure can (see Theorem 5). In other words, some datasets are indeed unsafe for *any* multisource procedure even in nonadversarial settings, as they can force suboptimal choices w.r.t. a target, absent additional restrictions on the problem setup.

As discussed earlier, such favorable restrictions concern for instance situations where information is available on how sources rank in distance to a target. In particular, the case $n = 1$ intersects with the so-called *distribution drift* setting where each distribution $P_t, t \in [N]$ has bounded discrepancy $\sup_{t \geq 1} \text{dist}(P_t, P_{t+1})$ w.r.t. the next distribution P_{t+1} as *time* t varies. While the notion of discrepancy is typically taken to be total-variation [BDBM89, Bar92, BL97] or related notions [MM12, ZL19, HY19], both our upper and lower-bounds, specialized to the case $n = 1$, provide a new perspective for distribution drift under our distinct parametrization of how P_t's relate to each other. For instance, our results on multisource under ranking (Theorem 4) imply new rates for distribution drift, when all h^*'s are the same across time, with excess error at time N going to 0 with N, even in situations where $\sup_{t \geq 1} \text{dist}(P_t, P_{t+1}) > 0$ in total variation. Such consistency in N is unavailable in prior work on distribution drift.

Finally we note that there has been much recent theoretical efforts towards other formalisms of relations between distributions in a multitask or multisource scenario, with particular emphasis on distributions sharing com-

mon latent substructures, both as applied to classification [MBS13, MB17, AKK+19], or to regression settings [JSRR10, LPVDG+11, NW11, DHK+20, TJJ20]. The present work does not address such settings.

Paper Outline We start with setup and definitions in Sections 2 and 3. This is followed by a technical overview of results, along with discussions of the analysis and novel proof techniques, in Section 4. Minimax lower and upper bounds are derived in Sections 5 and 6. Constrained regimes allowing partial adaptivity are discussed in Section 7. This is followed by impossibility theorems for adaptivity in Section 8.

2 Basic Classification Concepts

We consider a classification setting $X \mapsto Y$, where X, Y are drawn from some space $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} \doteq \{-1, 1\}$. We focus on *proper learning* where, given data, a learner is to return a classifier $h : \mathcal{X} \mapsto \mathcal{Y}$ from some fixed class $\mathcal{H} \doteq \{h\}$.

Assumption 1 (Bounded VC)**.** Throughout we will let $\mathcal{H} \in 2^{\mathcal{X}}$ denote a hypothesis class of finite VC dimension $d_{\mathcal{H}}$, which we consider fixed in all subsequent discussions. To focus on nontrivial cases, we assume $|\mathcal{H}| \geq 3$ throughout.

We note that our algorithmic techniques and analysis extend to more general $\mathcal{H}$ through Rademacher complexity or empirical covering numbers. We focus on VC classes for simplicity, and to allow simple expressions of minimax rates.

The performance of any classifier h will be captured through the 0-1 risk and excess risk as defined below.

Definition 1. Let $R_P(h) \doteq P_{X,Y}\left(h(X) \neq Y\right)$ denote the **risk** of any $h : \mathcal{X} \mapsto \mathcal{Y}$ under a distribution $P = P_{X,Y}$. The **excess risk** of h over any $h' \in \mathcal{H}$ is then defined as $\mathcal{E}_P(h; h') \doteq R_P(h) - R_P(h')$, while for the excess risk over the best in class we simply write $\mathcal{E}_P(h) \doteq R_P(h) - \inf_{h' \in \mathcal{H}} R_P(h')$. We let h^*_P denote any element of $\operatorname{argmin}_{h \in \mathcal{H}} R_P(h)$ (which we will assume exists); if P is clear from context we might just write h^* for a minimizer (a.k.a. *best in class*). Also define the pseudo-distance $P(h \neq h') \doteq P_X(h(X) \neq h'(X))$.

Given a finite dataset S of (x, y) pairs in $\mathcal{X} \times \mathcal{Y}$, we let $\hat{R}_S(h) \doteq \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}\{h(x) \neq y\}$ denote the **empirical risk** of h under S; if $S = \emptyset$, define $\hat{R}_S(h) \doteq 0$. The **excess empirical risk** over any h' is $\hat{\mathcal{E}}_S(h; h') \doteq \hat{R}_S(h) - \hat{R}_S(h')$. Also define the empirical pseudo-distance $\hat{\mathbb{P}}_S(h \neq h') \doteq \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}\{h(x) \neq h'(x)\}$.

The following condition is a classical way to capture a continuum from easy to hard classification. In particular, in vanilla classification, the best rate excess risk achievable by a classifier trained on data of size m, can be shown to be of order $m^{-1/(2-\beta)}$, i.e, interpolates between $1/n$ and $1/\sqrt{n}$, as controlled by $\beta \in [0, 1]$ defined below.

Definition 2. Let $\beta \in [0, 1], C_\beta \geq 2$. A distribution P is said to satisfy a **Bernstein class condition** with parameters (C_β, β) if the following holds:

$$\forall h \in \mathcal{H}, \quad P(h \neq h^*_P) \leq C_\beta \cdot \mathcal{E}^\beta_P(h). \tag{2}$$

Notice that the above always holds with at least $\beta = 0$.

The condition can be viewed as quantifying the *amount of noise* in Y since always have $P_X(h \neq h^*_P) \geq \mathcal{E}_P(h)$ with equality when $Y = h^*(X)$. In particular, it captures the so-called *Tsybakov noise margin condition* when the Bayes classifier is in $\mathcal{H}$, that is, let $\eta(x) \doteq \mathbb{E}_P[Y \mid X = x]$, then the margin condition

$$P_X\left(x : |\eta(x) - 1/2| \leq \tau\right) \leq C\tau^\kappa, \forall \tau > 0, \text{ and some } \kappa \geq 0,$$

implies that P satisfies (2) with $\beta = \kappa/(1 + \kappa)$ for some C_β [Tsy04].

The importance of such margin considerations become evident as we consider which multitask learner is able to automatically adapt optimally to unknown relations between distributions; interestingly, hardness of adaptation has to do not only with relations (or discrepancies) between distributions, but also with β.

3 Multitask Setting

We consider a setting where multiple datasets are drawn independently from (related) distributions $P_t, t \in [N + 1]$, with the aim to return hypotheses h's from $\mathcal{H}$ with low excess error $\mathcal{E}_{P_t}(h)$ w.r.t. any of these distributions. W.l.o.g. we fix attention to a single *target* distribution $\mathcal{D} \doteq P_{N+1}$, and therefore reduce the setting to that of *multisource*[2].

[2]The term *multisource* is often used for situations where the learner has no access to target data, which is not required to be the case here, although is handled simply by setting the target sample size to 0 in our bounds.

Definition 3. A **multisource learner** $\hat{h}$ is any function $\prod_{t\in[N+1]}(\mathcal{X}\times\mathcal{Y})^{n_t}\mapsto\mathcal{H}$ (for some sample sizes $n_t\in\mathbb{N}$), i.e., given a multisample $Z=\{Z_t\}_{t\in[N+1]}, |Z_t|=n_t$, returns a hypothesis in $\mathcal{H}$. In an abuse of notation, we will often conflate the learner $\hat{h}$ with the hypothesis it returns.

A multitask setting can then be viewed as one where $N+1$ multisource learners $\hat{h}_t$ (targeting each P_t) are trained on the same multisample Z. Much of the rest of the paper will therefore discuss learners for a given target $\mathcal{D}\doteq P_{N+1}$.

3.1 Relating Sources to Target

Clearly, how well one can do in multitask depend on how distributions relate to each other. The following parametrization will serve to capture the relation between sources and target distributions.

Definition 4. Let $\rho>0$ (up to $\rho=\infty$), and $C_\rho\geq 2$. We say that a distribution P has **transfer exponent** (C_ρ,ρ) w.r.t. a distribution $\mathcal{D}$ (under $\mathcal{H}$), if

$$\forall h\in\mathcal{H},\quad \mathcal{E}_{\mathcal{D}}(h)\leq C_\rho\cdot\mathcal{E}_P^{1/\rho}(h).$$

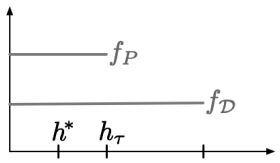
Notice that the above always holds with at least $\rho=\infty$.

We have shown in earlier work [HK19] that the *transfer exponent* manages to tightly capture the minimax rates of transfer in various situations with a single source P and target $\mathcal{D}$ ($N=1$), including ones where the best hypotheseses $h^*_P, h^*_{\mathcal{D}}$ are different for source and target; in the case of main interest here where P and $\mathcal{D}$ share a same best in class $h^*\doteq h^*_P\doteq h^*_{\mathcal{D}}$, for respective data sizes $n_P, n_{\mathcal{D}}$, the best possible excess risk $\mathcal{E}_{\mathcal{D}}$ is of order $(n_P^{1/\rho}+n_{\mathcal{D}})^{-1/(2-\beta)}$ which is tight for any values of ρ and β (a Bernstein class parameter on P and $\mathcal{D}$). In other words, ρ captures an effective data size $n_P^{1/\rho}$ contributed by the source to the target: this decreases as $\rho\to\infty$, delineating a continuum from easy to hard transfer. Interestingly, $\rho<1$ reveals the fact that source data could be more useful than target data, for instance if the classification problem is easier under the source (e.g., P has more mass at the decision boundary).

Altogether, the transfer exponent ρ appears as a more optimistic measure of discrepancy between source and target, as it reveals the possibility of transfer – even at fast rates – in many situations where traditional measures are pessimistically large. To illustrate this, we recall some of the examples from [HK19]. In all examples below, we assume for simplicity that $Y=h^*(X)$ for some $h^*\doteq h^*_P\doteq h^*_{\mathcal{D}}$.

Example 1. (Discrepancies $d_{\mathcal{A}}, d_{\mathcal{Y}}$ can be too large) Let $\mathcal{H}$ consist of one-sided thresholds on the line, and let $P_X\doteq\mathcal{U}[0,2]$ and $Q_X\doteq\mathcal{U}[0,1]$. Let h^* be thresholded at $1/2$. We then see that for all h_τ thresholded at $\tau\in[0,1]$, $2P_X(h_\tau\neq h^*)=\mathcal{D}_X(h_\tau\neq h^*)$, where for $\tau>1$, $P_X(h_\tau\neq h^*)=\frac{1}{2}(\tau-1/2)\geq\frac{1}{2}\mathcal{D}_X(h_\tau\neq h^*)=\frac{1}{4}$. Thus, the transfer exponent $\rho=1$ with $C_\rho=2$, so we have fast transfer at the same rate $1/n_P$ as if we were sampling from $\mathcal{D}$.

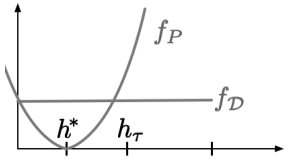
On the other hand, recall that the $d_{\mathcal{A}}$-divergence takes the form[3] $d_{\mathcal{A}}(P,\mathcal{D})\doteq\sup_{h\in\mathcal{H}}|P_X(h\neq h^*)-\mathcal{D}_X(h\neq h^*)|$, while the $\mathcal{Y}$-discrepancy takes the form $d_{\mathcal{Y}}(P,\mathcal{D})\doteq\sup_{h\in\mathcal{H}}|\mathcal{E}_P(h)-\mathcal{E}_{\mathcal{D}}(h)|$. The two coincide when $Y=h^*(X)$.



Now, take h_τ as the threshold at $\tau=1/2$, and $d_{\mathcal{A}}=d_{\mathcal{Y}}=\frac{1}{4}$ which would wrongly imply that transfer is not feasible at a rate faster than $\frac{1}{4}$; we can in fact make this situation worse, i.e., let $d_{\mathcal{A}}=d_{\mathcal{Y}}\to\frac{1}{2}$ by letting h^* correspond to a threshold close to 0. A first issue is that these divergences get large in large disagreement regions; this is somewhat mitigated by *localization* – i.e., defining these discrepancies w.r.t. h's in a vicinity of h^*, but does not quite resolve the issue, as discussed in earlier work [HK19].

Example 2. (Minimum ρ, and the inherent asymmetry of transfer) Suppose $\mathcal{H}$ is the class of one-sided thresholds on the line, h^* is a threshold at 0. The marginal $\mathcal{D}_X$ has uniform density $f_{\mathcal{D}}$ (on an interval containing 0), while, for some $\rho\geq 1$, P_X has density $f_P(\tau)\propto\tau^{\rho-1}$ on $\tau>0$ (and uniform on the rest of the support of $\mathcal{D}$, not shown). Consider any h_τ at threshold $\tau>0$, we have $P_X(h_\tau\neq h^*)=\int_0^\tau f_P\propto\tau^\rho$, while $\mathcal{D}_X(h_\tau\neq h^*)\propto\tau$. Notice that for any fixed $\epsilon>0$, $\lim_{\tau>0,\,\tau\to 0}\frac{\mathcal{D}_X(h_\tau\neq h^*)^{\rho-\epsilon}}{P_X(h_\tau\neq h^*)}=\lim_{\tau>0,\,\tau\to 0}C\frac{\tau^{\rho-\epsilon}}{\tau^\rho}=\infty$.

We therefore see that ρ is the smallest possible transfer-exponent. Interestingly, now consider transferring instead from $\mathcal{D}$ to P: we would have $\rho(\mathcal{D}\to P)=1\leq\rho\doteq\rho(P\to\mathcal{D})$; in other words, there are natural situations where it is easier to transfer from $\mathcal{D}$ to P than from P to $\mathcal{D}$, as in the case here where P gives relatively little mass



[3] Note that these divergences are often defined w.r.t. every pair $h,h'\in\mathcal{H}$ rather than w.r.t. h^* which makes them smaller.

to the decision boundary. This is *not captured by symmetric notions of distance*, e.g., metrics or semi-metrics such as $d_{\mathcal{A}}$, $d_{\mathcal{Y}}$, MMD, TV, or Wasserstein.

Finally note that the above examples can be extended to more general hypothesis classes $\mathcal{H}$ as the examples merely play on how fast f_P decreases w.r.t. $f_{\mathcal{D}}$ in regions of space.

3.2 Multisource Class

We are now ready to formalize the main class of distributional settings considered in this work.

Definition 5 (Multisource class)**.** We consider classes $\mathcal{M} = \mathcal{M}\left(C_\rho, \{\rho_t\}_{t\in[N]}, \{n_t\}_{t\in[N+1]}, C_\beta, \beta\right)$ of product distributions of the form $\Pi = \prod_{t\in[N]} P_t^{n_t} \times \mathcal{D}^{n_{\mathcal{D}}},\ n_t \geq 1,\ n_{\mathcal{D}} \doteq n_{N+1} \geq 0$, satisfying:

(A1). There exists $h^* \in \mathcal{H}, \quad \forall t \in [N+1],\ h^* \in \operatorname{argmin}_{h\in\mathcal{H}} \mathcal{E}_{P_t}(h)$,

(A2). Sources P_t's have transfer exponent (C_ρ, ρ_t) w.r.t. the target $\mathcal{D}$,

(A3). All sources P_t and target $\mathcal{D}$ admit a Bernstein class condition with parameter (C_β, β).

For notational simplicity, we will often let $P_{N+1} \doteq \mathcal{D}$. Also, although it is not a parameter of the class, we will also refer to $\rho_{N+1} = 1$, as $\mathcal{D}$ always has transfer exponent $(C_\rho, 1)$ w.r.t. itself.

Remark 1 (h^* is almost unique)**.** Note that, for $\beta > 0$, the Bernstein class condition implies that h^* above satisfies $P_t(h^* \neq h_t^*) = 0$ for any other $h_t^* \in \operatorname{argmin}_{h\in\mathcal{H}} \mathcal{E}_{P_t}(h)$. Furthermore for any $t \in [N]$ such that $\rho_t < \infty$, we also have $\mathcal{E}_{\mathcal{D}}(h_t^*) = 0$, implying $\mathcal{D}(h^* \neq h_t^*) = 0$ for $\beta > 0$.

3.3 Additional Notation

Implicit target $\mathcal{D} = \mathcal{D}(\Pi)$. Every time we write Π, we will implicitly mean $\Pi = \prod_{t\in[N]} P_t^{n_t} \times \mathcal{D}^{n_{N+1}}$, so that in any context where a Π is introduced, we may write, for instance, $\mathcal{E}_{\mathcal{D}}(h)$, which then refers to the $\mathcal{D}$ distribution in Π.

Indices and Order Statistics. For any $Z = \{Z_t\}_{t\in[N+1]} \sim \Pi$, and indices $I \subset [N+1]$, we let $Z^I \doteq \bigcup_{s\in I} Z_s$.

We will often be interested in order statistics $\rho_{(1)} \leq \rho_{(2)} \ldots \leq \rho_{(t)}$ of ordered $\rho_t, t \in [N+1]$ values, in which case $P_{(t)}, Z_{(t)}, n_{(t)}$ will denote distribution, sample and sample size at index (t). We will then let $Z^{(t)} \doteq Z^{\{(1),\ldots,(t)\}}$.

Average transfer exponent. For any $t \in [N+1]$, define $\bar{\rho}_t \doteq \sum_{s\in[t]} \alpha_{(s)} \cdot \rho_{(s)}$, where $\alpha_{(s)} \doteq \frac{n_{(s)}}{\sum_{r\in[t]} n_{(r)}}$.

Aggregate ERM. For any $Z^I, I \subset [N+1]$, we let $\hat{h}_{Z^I} \doteq \operatorname{argmin}_{h\in\mathcal{H}} \hat{R}_{Z^I}(h)$, and correspondingly we also define $\hat{h}_{Z^{(t)}}, t \in [N+1]$ as the ERM over $Z^{(t)}$. When $t = N+1$ we simply write $\hat{h}_Z$ for $\hat{h}_{Z^{(N+1)}}$.

Min and Max. We often use the short notations $a \wedge b \doteq \min\{a, b\}$, $a \vee b \doteq \max\{a, b\}$.

Positive Logarithm. For any $x \geq 0$, define $\log(x) \doteq \max\{\ln(x), 1\}$.

$1/0$ Convention. We adopt the convention that $1/0 = \infty$.

Asymptotic Order. We often write $a \lesssim b$ or $a \asymp b$ in the statements of key results, to indicate inequality, respectively, equality, up to constants and logarithmic factors. The precise constants and logarithmic factors are always presented in supporting results.

4 Results Overview

We start by investigating what the best possible transfer rates are for multisource classes $\mathcal{M}$, and then investigate the extent to which these rates are attainable by adaptive procedures, i.e., procedures with little access to class information such as transfer exponents ρ_t from sources to target.

From this point on, we let $\mathcal{M} = \mathcal{M}\left(C_\rho, \{\rho_t\}_{t\in[N]}, \{n_t\}_{t\in[N+1]}, C_\beta, \beta\right)$ denote any multisource class, with *any admissible value* of relevant parameters, unless these parameters are specifically constrained a result's statement.

4.1 Minimax Rates

Theorem 1 (Minimax Rates)**.** *Let $\mathcal{M}$ denote any multisource class where $\rho_t \geq 1, \forall t \in [N]$. Let $\hat{h}$ denote any multisource learner with knowledge of $\mathcal{M}$. We have:*

$$\inf_{\hat{h}} \sup_{\Pi \in \mathcal{M}} \mathbb{E}_{\Pi} \left[\mathcal{E}_{\mathcal{D}} \left(\hat{h} \right) \right] \asymp \min_{t \in [N+1]} \left(\sum_{s=1}^{t} n_{(s)} \right)^{-1/(2-\beta)\bar{\rho}_t}.$$

Remark 2. We remark that the proof of the above result (see Theorems 6 and 7 of Sections 5 and 6 for matching lower and upper bounds) imply that, in fact, we could replace $\bar{\rho}_t$ with simply $\rho_{(t)}$ and the result would still be true. In other words, although intuitively $\bar{\rho}_t$ might be much smaller than $\rho_{(t)}$ for any fixed t, the minimum values over $t \in [N+1]$ can only differ up to logarithmic terms.

We also note that the constraint that $\rho_t \geq 1$ is only needed for the lower bound (Theorem 6), whereas all of our upper bounds (Theorems 7, 3, and 4) hold for any values $\rho_t > 0$. Moreover, there exist classes $\mathcal{H}$ where the lower bound also holds for all $\rho_t > 0$, so that the form of the bound is generally not improvable. The case $\rho_t \in (0, 1)$ represents a kind of *super transfer*, where the source samples are actually *more* informative than target samples.

It follows from Theorem 1 that, despite there being 2^{N+1} possible ways of aggregating datasets (or more if we consider general weightings of datasets), it is sufficient to search over $N+1$ possible aggregations – defined by the ranking $\rho_{(1)} \leq \rho_{(2)} \ldots \leq \rho_{(N+1)}$ – to nearly achieve the minimax rate.

The lower-bound (Theorem 6) relies on constructing a subset (of $\mathcal{M}$) of product distributions $\Pi_h, h \in \mathcal{H}$, which are mutually *close* in KL-divergence, but far under the pseudo-metric $\mathcal{D}_X(h \neq h')$. For illustration, considering the case $\beta = 1$, for any h, h' that are sufficiently far under $\mathcal{D}_X$ so that $\mathcal{E}_{\mathcal{D}}(h'; h) \gtrsim \epsilon$, the lower-bound construction is such that

$$\mathrm{KL}(\Pi_h \| \Pi_{h'}) \lesssim \sum_{t=1}^{N+1} n_t \epsilon^{\rho_t} \lesssim 1, \tag{3}$$

ensuring that $\Pi_h, \Pi_{h'}$ are hard to distinguish from finite sample. Thus, the largest ϵ satisfying the second inequality above, say ϵ_{N+1} is a minimax lower-bound. On the other-hand, the upper-bound (Theorem 7) relies on a uniform Bernstein's inequality that holds for non-identically distributed r.v.s (Lemma 1); in particular, by accounting for *variance* in the risk, such Bernstein-type inequality allows us to extend (to the multisource setting) usual fixed-point arguments that capture the effect of the noise parameter β. Now, again for illustration, let $\beta = 1$, and consider the ERM $\hat{h} \doteq \hat{h}_Z$ combining all $N+1$ datasets. Let $\mathcal{E}_\alpha \doteq \sum_{t=1}^{N+1} \alpha_t \mathcal{E}_t(\hat{h})$, $\alpha_t \doteq n_t / \left(\sum_s n_s\right)$, then the concentration arguments described above ensure that $\mathcal{E}_\alpha(\hat{h}) \lesssim 1/\left(\sum_t n_t\right)$. Now notice that, by definition of ρ_t, $\mathcal{E}_\alpha(\hat{h}) \gtrsim \sum_t \alpha_t \mathcal{E}_{\mathcal{D}}^{\rho_t}(\hat{h})$, in other words, $\mathcal{E}_{\mathcal{D}}^{\rho_t}(\hat{h})$ satisfies the second inequality in (3), and must therefore be at most of order ϵ_{N+1}. This establishes the tightness of ϵ_{N+1} as a minimax rate, all that is left being to elucidate its exact form in terms of sample sizes. Similar, but somewhat more involved arguments apply for general β, though in that case we find that pooling all of the data does not suffice to achieve the minimax rate.

Notice that the rates of Theorem 1 immediately imply minimax rates for multitask under the assumption (A1) of sharing a same h^* (with appropriate ρ_t's w.r.t. any target $\mathcal{D} \doteq P_s, s \in [N+1]$). It is then natural to ask whether the minimax rate for various targets P_s might be achieved by the same algorithm, i.e., the same aggregation of tasks, in light of a common approach in the literature (and practice) of optimizing for a single classifier that does well simultaneously on all tasks. We show that even when all h^*'s are the same, the optimal aggregation might differ across targets P_s, simply due to the inherent assymmetry of transfer. We have the following theorem, proved in Appendix C.

Theorem 2 (Target affects aggregation)**.** *Set $N = 1$. There exists $P, \mathcal{D}$ satisfying a Bernstein class condition with parameters (C_β, β) for some $0 \leq \beta < 1$, and sharing the same $h^* = h_P^* = h_{\mathcal{D}}^*$ such that the following holds. Consider a multisample $Z = \{Z_P, Z_{\mathcal{D}}\}$ consisting of independent datasets $Z_P \sim P^{n_P}, Z_{\mathcal{D}} \sim \mathcal{D}^{n_{\mathcal{D}}}$.*

Let $\hat{h}_Z, \hat{h}_{Z_P}, \hat{h}_{Z_{\mathcal{D}}}$ denote the ERMs over Z, Z_P, $Z_{\mathcal{D}}$ respectively. Suppose $1 \leq n_{\mathcal{D}}^2 \leq \frac{1}{8} n_P^{(2-2\beta)/(2-\beta)}$. Then

$$\mathbb{E}\left[\mathcal{E}_{\mathcal{D}}(\hat{h}_{Z_P})\right] \wedge \mathbb{E}\left[\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)\right] \geq \frac{1}{4}, \quad \textit{while } \mathbb{E}\left[\mathcal{E}_{\mathcal{D}}(\hat{h}_{Z_{\mathcal{D}}})\right] \lesssim n_{\mathcal{D}}^{-1/(2-\beta)};$$
$$\textit{however, } \mathbb{E}\left[\mathcal{E}_P(\hat{h}_{Z_P})\right] \vee \mathbb{E}\left[\mathcal{E}_P(\hat{h}_Z)\right] \lesssim n_P^{-1/(2-\beta)}.$$

Remark 3 (Suboptimality of *pooling*)**.** A common practice is to *pool* all datasets together and return an ERM as in $\hat{h}_Z$. We see from the above result that this might be optimal for some targets while suboptimal for other targets. However, pooling is near optimal (simultaneously for all targets P_s) whenever $\beta = 1$, as discussed in Section 4.2 below.

4.2 Some Regimes of (Semi) Adaptivity

It is natural to ask whether the above minimax rates for $\mathcal{M}$ are attainable by *adaptive procedures*, i.e., a reasonable procedure with no access to prior information on (the parameters) of $\mathcal{M}$, but only access to a multisample $Z \sim \Pi$ for some unknown $\Pi \in \mathcal{M}$. As we will see in Section 4.3, this is not possible in general, i.e., outside of the regimes considered here. Our work however leaves open the existence of more refined regimes of adaptivity.

• **Low Noise** $\beta = 1$**.** To start, when the Bernstein class parameter $\beta = 1$ (which would often be a priori unknown to the learner), *pooling* of all datasets is near minimax optimal as stated in the next result. This corresponds to low noise situations, e.g., so-called *Massart's noise* (where $\mathbb{P}(Y \neq h^*(X)|X) \leq (1/2) - \tau$, for some $\tau > 0$), including the *realizable case* (where $Y = h^*(X)$ deterministically). Note that ρ_t's are nonetheless nontrivial (see examples of Section 3.1), however the distributions P_t's are then sufficiently related that their datasets are mutually valuable.

Theorem 3 (Pooling under low noise)**.** *Suppose $\beta = 1$. Consider any $\Pi \in \mathcal{M}$ and let $\hat{h}_Z$ denote the ERM over $Z \sim \Pi$. Let $\delta \in (0,1)$. There exists a universal constant $c > 0$, such that, with probability at least $1 - \delta$,*

$$\mathcal{E}_{\mathcal{D}}(\hat{h}_Z) \leq \min_{t \in [N+1]} C_\rho \left(c\, C_\beta \frac{d_{\mathcal{H}} \log\left(\frac{1}{d_{\mathcal{H}}} \sum_{s=1}^{N+1} n_s \right) + \log(1/\delta)}{\sum_{s=1}^{t} n_{(s)}} \right)^{1/\bar{\rho}_t}.$$

The theorem is proven in Section 7.1. We also state a general bound on $\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)$, holding for any β, in Corollary 2 of Appendix B; the implied rates are not always optimal, though interestingly they are near-optimal in the case that $\sum_{t=1}^{t^*} n_{(t)} \propto \sum_{t=1}^{N} n_t$, where t^* is the minimizer of the r.h.s. in Theorem 1. We note that, unlike in the oracle upper-bounds of Theorem 7, the logarithmic term in the above result is in terms of the entire sample size, rather than the sample size at which the minimizer in $t \in [N+1]$ is attained.

• **Available ranking information.** Now, assume that on top of $Z \sim \Pi \in \mathcal{M}$, we have access to ranking information $\rho_{(1)} \leq \rho_{(2)} \leq \ldots \leq \rho_{(N+1)}$, but no additional information on $\mathcal{M}$. Namely, C_β, β, and the actual values of $\rho_t, t \in [N]$ are unknown to the learner. We show that, in this case, a simple rank-based procedure achieves the minimax rates of Theorem 1, without knowledge of the additional distributional parameters.

Define $\varepsilon(m, \delta) \doteq \frac{d_{\mathcal{H}}}{m} \log\left(\frac{m}{d_{\mathcal{H}}}\right) + \frac{1}{m} \log\left(\frac{1}{\delta}\right)$, for $\delta \in (0,1)$ and $m \in \mathbb{N}$. Let $\mathbf{n}_t \doteq \sum_{s=1}^{t} n_{(s)}$ for each $t \in [N+1]$, ansd recall that $\hat{h}_{Z^{(t)}}$ denotes the ERM over the aggregate sample $Z^{(t)}$.

Rank-based Procedure $\hat{h}$: let $\delta_t \doteq \delta/(6t^2)$, and C_0 as in Lemma 1. For any $t \in [N+1]$, define:

$$\mathcal{H}_{(t)} \doteq \left\{ h \in \mathcal{H} : \hat{\mathcal{E}}_{Z^{(t)}}(h; \hat{h}_{Z^{(t)}}) \leq C_0 \sqrt{\hat{\mathbb{P}}_{Z^{(t)}}(h \neq \hat{h}_{Z^{(t)}}) \varepsilon(\mathbf{n}_t, \delta_t)} + C_0 \varepsilon(\mathbf{n}_t, \delta_t) \right\}, \qquad (4)$$

Return any h in $\bigcap_{s=1}^{N+1} \mathcal{H}_{(s)}$, if not empty, otherwise return any h.

We have the following result for this learning algorithm.

Theorem 4 (Semi-adaptive method under ranking information)**.** *Let $\delta \in (0,1)$. Let $\hat{h}$ denote the above procedure, trained on $Z \sim \Pi$, for any $\Pi \in \mathcal{M}$. There exists a universal constant $c > 0$, such that, with probability at least $1 - \delta$,*

$$\mathcal{E}_{\mathcal{D}}(\hat{h}) \leq \min_{t \in [N+1]} C_\rho \left(c\, C_\beta \frac{d_{\mathcal{H}} \log\left(\frac{1}{d_{\mathcal{H}}} \sum_{s=1}^{t} n_{(s)} \right) + \log(1/\delta)}{\sum_{s=1}^{t} n_{(s)}} \right)^{1/(2-\beta)\bar{\rho}_t}.$$

The result is shown in Section 7.2 by arguing that each h in $\mathcal{H}_{(t)}$ has $\mathcal{E}_{\mathcal{D}}(h)$ of order $\left(\sum_{s \in [t]} n_{(s)} \right)^{-1/(2-\beta)\bar{\rho}_t}$, so the minimizer is attained whenever $\hat{t} = N+1$ (which is probable, as h^* is likely to be in each $\mathcal{H}_{(t)}$).

Recalling our earlier discussion of Section 1, the above result applies to the case of *drifting distributions* by letting $n_t = 1$, and $(t) = N + 2 - t$, i.e., the t^{th} previous example is considered the t^{th} most-relevant to the target $\mathcal{D}$. In this case, in contrast to the lower bounds proven by [Bar92], the above result of Theorem 4 reveals situations where the risk at time N approaches 0 as $N \to \infty$, even though the total-variation distance between adjacent distributions may be bounded away from zero (see examples of Section 3.1). In other words, by constraining the sequence of P_t distributions by the sequence of ρ_t values, we can describe scenarios where the traditional upper bound analysis of drifting distributions from [Bar92, BDBM89, BL97] can be improved.

Remark 4 (approximate ranking)**.** Theorem 4, and its proof, also have implications for scenarios where we may only have access to an *approximate* ranking: that is, where the ranking indices (t) don't strictly order the tasks by their respective minimum valid ρ_t values. For instance, one immediate observation is that, since Definition 4 does not require ρ_t to be minimal, any larger value of ρ_t would also be valid; therefore, for any permutation $\sigma : [N+1] \to [N+1]$, there always exists *some* valid choices of ρ_t values so that the sequence $\rho_{\sigma(t)}$ is non-decreasing, and hence we can define $(t) = \sigma(t)$, so that Theorem 4 holds (with these ρ_t values) for the rank-based procedure applied with this $\sigma(t)$ ordering of the tasks. For instance, this means that if we use an ordering $\sigma(t)$ that only swaps the order of some $\rho_{(t)}$ values that are within some ϵ of each other, then the result in the theorem remains valid aside from replacing $\bar{\rho}_t$ with $\bar{\rho}_t + \epsilon$. A second observation is that, upon inspecting the proof, it is clear that the result is only truly sensitive to the ranking relative to the index t^* achieving the minimum in the bound: that is, if some indices t, t' with $(t) < (t^*)$ and $(t') < (t^*)$ are incorrectly ordered, while still both ranked before t^* (or likewise for indices with $(t) > (t^*)$ and $(t') > (t^*)$), the result remains valid as stated nonetheless.

4.3 Impossibility of Adaptivity in General Regimes

Adaptivity in general is not possible even though the rates of Theorem 1 appear to reduce the exponential search space over aggregations to a more manageable one based on rankings of data. As seen in the previous section, easier settings such as ones with ranking information, or low noise, allow adaptivity to the remaining unknown distributional parameters. The result below states that, outside of these regimes, no algorithm can guarantee a rate better than a dependence on the number of samples $n_{\mathcal{D}}$ from the target task, even when a semi-adaptive procedure using ranking information can achieve any desired rate ϵ.

Theorem 5 (Impossibility of adaptivity)**.** *Pick any $0 \leq \beta < 1$, $C_\beta \geq 2$, and let $1 \leq n < 2/\beta - 1$, $n_{\mathcal{D}} \geq 0$, and $C_\rho = 3$. Pick any $\epsilon > 0$. The following holds for N sufficiently large.*

• Let $\hat{h}$ denote any multisource learner with no knowledge of $\mathcal{M}$. There exists a multisource class $\mathcal{M}$ with parameters $n_{N+1} = n_{\mathcal{D}}$, $n_t = n, \forall t \in [N]$, and all other parameters satisfying the above, such that

$$\sup_{\Pi \in \mathcal{M}} \mathbb{E}_\Pi \left[\mathcal{E}_{\mathcal{D}}(\hat{h}) \right] \geq c \cdot \left(1 \wedge n_{\mathcal{D}}^{-1/(2-\beta)} \right) \text{ for a universal constant } c > 0.$$

• On the other hand, there exists a semi-adaptive classifier $\tilde{h}$, which, given data $Z \sim \Pi$, along with a ranking $\rho_{(1)} \leq \rho_{(2)} \ldots \leq \rho_{(N+1)}$, but no other information on the above $\mathcal{M}$, achieves the rate

$$\sup_{\Pi \in \mathcal{M}} \mathbb{E}_\Pi \left[\mathcal{E}_{\mathcal{D}}(\tilde{h}) \right] \leq \epsilon.$$

The first part of the result follows from Theorem 8 of Section 8, while the second part follows from both Theorem 8 and Theorem 4 of Section 8. The main idea of the proof is to inject enough randomness into the choice of ranking, while at the same time allowing *bad* datasets from distributions with large ρ_t – which would force a wrong choice of h^* – to appear as benign as *good* samples from distributions with small $\rho_t = 1$. Hence, we let N large enough in our constructions so that the bulk of bad datasets would significantly *overwhelm* the information from *good* datasets.

As a technical point, *no knowledge of* $\mathcal{M}$ simply means that a minimax analysis is performed over a larger family containing such $\mathcal{M}$'s, indexed over choices of ranking each corresponding to a fixed $\mathcal{M}$.

Finally, we note that the result leaves open the possibility of adaptivity under further distributional restrictions on $\mathcal{M}$, for instance requiring that the number of samples n per source task be large w.r.t. other parameters such as β and N; although this remains unclear, large values of n could perhaps yield sufficient information to compare and rank tasks.

Another possible restriction towards adaptivity is to require that a large proportion of the samples are from *good* datasets w.r.t. $N+1$. In particular, we show in Theorem 9 of Appendix B that the ERM $\hat{h}_Z$ which pools all datasets achieves a target excess risk $\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)$ depending on the (weighted) *median* of the ρ_t values (or more generally, any *quantile*); in other words as long as a constant fraction of all datapoints (pooled from all datasets) are from tasks with relatively small ρ_t, the bound will be small. However, this is not a safe algorithm in general, as per Theorem 2.

5 Lower Bound Analysis

Theorem 6. *Suppose $|\mathcal{H}| \geq 3$. If every $\rho_t \geq 1$, then for any learning rule $\hat{h}$, there exists $\Pi \in \mathcal{M}$ such that, with probability at least $1/50$,*

$$\mathcal{E}_{\mathcal{D}}(\hat{h}) > c_1 \min_{t\in[N+1]} \left(\frac{c_2 d_{\mathcal{H}}}{\left(\sum_{s=1}^{t} n_{(s)}\right) \log^2\left(\sum_{s=1}^{t} n_{(s)}\right)} \right)^{1/(2-\beta)\bar{\rho}_t}$$

for numerical constants $c_1, c_2 > 0$.

Proof. We will prove this result with $C_\beta = C_\rho = 2$; the general cases $C_\beta \geq 2$ and $C_\rho \geq 2$ follow immediately (since distributions satisfying $C_\beta = C_\rho = 2$ also satisfy the respective conditions for any $C_\beta \geq 2$ and $C_\rho \geq 2$). As in the proof of Theorem 7, for each $t \in [N+1]$, define $\mathbf{n}_t \doteq \sum_{s=1}^{t} n_{(s)}$. We prove the theorem using a standard approach to lower bounds by the probabilistic method. Let $x_0, x_1, \ldots, x_d$ be a sequence in $\mathcal{X}$ such that, $\exists y_0 \in \mathcal{Y}$ for which every $y_1, \ldots, y_d \in \mathcal{Y}$ can be realized by $h(x_1), \ldots, h(x_d)$ for some $h \in \mathcal{H}$ with $h(x_0) = y_0$. If $d_{\mathcal{H}} = 1$, we let $d = 1$, and we can find such a sequence since $|\mathcal{H}| \geq 3$; if $d_{\mathcal{H}} \geq 2$, then we let $d = d_{\mathcal{H}} - 1$ and any shattered sequence of $d_{\mathcal{H}}$ points will suffice. We now specify a family of 2^d distributions, as follows.

Fix $\epsilon \in (0,1)$. For each $\sigma \in \{-1,1\}^d$, for each $t \in [N+1]$, let $P_t^\sigma(\{x_0, y_0\}) = 1 - \epsilon^{\rho_t \beta}$, $P_t^\sigma(\{x_0, -y_0\}) = 0$, and for each $i \in [d]$ and $y \in \{-1,1\}$ let $P_t^\sigma(\{x_i, y\}) = \frac{1}{d}\epsilon^{\rho_t\beta}\left(\frac{1}{2} + \frac{y\sigma_i}{2}\epsilon^{\rho_t(1-\beta)}\right)$. In other words, the marginal probability of each x_i is $\frac{1}{d}\epsilon^{\rho_t\beta}$ and the conditional probability of label 1 given x_i is $\frac{1}{2} + \frac{\sigma_i}{2}\epsilon^{\rho_t(1-\beta)}$.

We first verify that $\Pi_\sigma = \prod_{t\in[N+1]} P_t^\sigma$ is in $\mathcal{M}$. For every σ and every t, the Bayes optimal label at x_0 is always y_0, and the Bayes optimal label at x_i ($i \in [d]$) is always σ_i; since this is the same for every t, the classifier $h^*_\sigma \doteq h^*_{P^\sigma_{N+1}}$ is optimal under every P_t^σ. Next, we check the Bernstein class condition. Define $\ell(h,\sigma) \doteq |\{i \in [d] : h(x_i) \neq \sigma_i\}|$. For any $t \in [N+1]$ and $h \in \mathcal{H}$, we have

$$P_t^\sigma(h \neq h^*_\sigma) = (1 - \epsilon^{\rho_t\beta})\mathbb{1}\{h(x_0) \neq y_0\} + \frac{\ell(h,\sigma)}{d}\epsilon^{\rho_t\beta}$$

while

$$\mathcal{E}_{P_t^\sigma}(h) = (1 - \epsilon^{\rho_t\beta})\mathbb{1}\{h(x_0) \neq y_0\} + \frac{\ell(h,\sigma)}{d}\epsilon^{\rho_t}.$$

Thus,

$$2\mathcal{E}^\beta_{P_t^\sigma}(h) \geq 2\max\left\{ (1-\epsilon^{\rho_t\beta})^\beta \mathbb{1}\{h(x_0) \neq y_0\}, \epsilon^{\rho_t\beta}\left(\frac{\ell(h,\sigma)}{d}\right)^\beta \right\} \geq P_t^\sigma(h \neq h^*_\sigma).$$

Finally, we verify the ρ_t values are satisfied. First note that any t with $\rho_t = \infty$ trivially satisfies the condition. Denote by $\mathcal{D}^\sigma$ the distribution P^σ_{N+1}. Then for each $t \in [N]$ with $1 \leq \rho_t < \infty$, and each $h \in \mathcal{H}$,

$$\begin{aligned}
\mathcal{E}_{\mathcal{D}^\sigma}(h) &= \left(\left((1-\epsilon^\beta)\mathbb{1}\{h(x_0) \neq y_0\} + \epsilon \frac{\ell(h,\sigma)}{d} \right)^{\rho_t} \right)^{1/\rho_t} \\
&\leq 2\left((1-\epsilon^\beta)^{\rho_t}\mathbb{1}\{h(x_0) \neq y_0\} + \left(\epsilon\frac{\ell(h,\sigma)}{d}\right)^{\rho_t} \right)^{1/\rho_t} \leq 2\mathcal{E}^{1/\rho_t}_{P_t^\sigma}(h),
\end{aligned}$$

where the final inequality follows from $\rho_t \geq 1$ and a bit of calculus to verify that $(1-x)^{\rho_t} \leq 1 - x^{\rho_t}$ for all $x \in [0,1]$.

Now the Varshamov-Gilbert bound (see Proposition 10 of Appendix E) implies there exists a subset $\{\sigma^0, \sigma^1, \ldots, \sigma^M\} \subset \{-1,1\}^d$ with $M \geq 2^{d/8}$, $\sigma^0 = (1,1,\ldots,1)$, and every distinct i, j have $\sum_t \mathbb{1}\left\{\sigma_t^i \neq \sigma_t^j\right\} \geq \frac{d}{8}$. Furthermore, for any $i \neq 0$,

$$\begin{aligned}
\mathrm{KL}(\Pi_{\sigma^i} \| \Pi_{\sigma^0}) &= \sum_{t\in[N+1]} n_t \mathrm{KL}(P_t^{\sigma^i} \| P_t^{\sigma^0}) = \sum_{t\in[N+1]} n_t \epsilon^{\rho_t\beta} \frac{1}{d} \sum_{j\in[d]} \mathbb{1}\left\{\sigma_j^i \neq 1\right\} \mathrm{KL}(\frac{1}{2} - \frac{1}{2}\epsilon^{\rho_t(1-\beta)} \| \frac{1}{2} + \frac{1}{2}\epsilon^{\rho_t(1-\beta)}) \\
&\leq c_3 \sum_{t\in[N+1]} n_t \epsilon^{\rho_t\beta} \frac{1}{d} \sum_{j\in[d]} \mathbb{1}\left\{\sigma_j^i \neq 1\right\} \epsilon^{2\rho_t(1-\beta)} \leq c_3 \sum_{t\in[N+1]} n_t \epsilon^{\rho_t(2-\beta)}
\end{aligned}$$

for a numerical constant c_3, where we have used a quadratic approximation of KL divergence between Bernoullis (Lemma 8 of Appendix E). Consider any choice of $\epsilon > 0$ making this last expression less than $\frac{d}{64}$.

Since $\frac{d}{64} \leq (1/8)\log(M)$, Theorem 2.5 of [Tsy09] (see Proposition 9 of Appendix E) implies that for any (possibly randomized) estimator $\hat{\sigma} : (\mathcal{X} \times \mathcal{Y})^{\sum_t n_t} \to \{-1, 1\}^d$, there exists a choice of σ such that a sample $Z \sim \Pi_\sigma$ will have $|\{i : \hat{\sigma}(Z)_i \neq \sigma_i\}| \geq \frac{d}{16}$ with probability at least $\frac{1}{50}$.

In particular, for any learning algorithm $\hat{h}$, we can take $\hat{\sigma} = (\hat{h}(x_1), \ldots, \hat{h}(x_d))$, and this result implies that there is a choice of σ such that, for $\hat{h}$ trained on $Z \sim \Pi_\sigma$, with probability at least $\frac{1}{50}$, there are at least $\frac{d}{16}$ points x_i with $\hat{h}(x_i) \neq \sigma_i = h^*_\sigma(x_i)$, which implies $\mathcal{E}_{\mathcal{D}^\sigma}(\hat{h}) \geq \frac{\epsilon}{16}$.

It remains only to identify an explicit value of ϵ for which $\sum_{t \in [N+1]} n_t \epsilon^{\rho_t(2-\beta)} \leq \frac{d}{64}$.

In particular, fix any positive function $\phi(m)$ such that $\sum_{m=1}^\infty \phi(m)^{-1} \leq cd$ for a finite numerical constant c: for instance, $\phi(m) = \frac{1}{d} m \log^2 m$ will suffice to prove the theorem. Then consider

$$\epsilon = \min_{t \in [N+1]} \left(c_2^{-1}\phi(\mathbf{n}_t)\right)^{-1/(2-\beta)\rho_{(t)}}$$

for a numerical constant $c_2 \in (0, 1)$. Denote by t_* the value of t obtaining the minimum in this expression. Then note that every other $t \neq t_*$ has

$$\left(c_2^{-1}\phi(\mathbf{n}_t)\right)^{-1/\rho_{(t)}} \geq \left(c_2^{-1}\phi(\mathbf{n}_{t_*})\right)^{-1/\rho_{(t_*)}},$$

which implies

$$\frac{\rho_{(t)}}{\rho_{(t_*)}} \geq \frac{\ln\left(c_2^{-1}\phi(\mathbf{n}_t)\right)}{\ln\left(c_2^{-1}\phi(\mathbf{n}_{t_*})\right)},$$

so that

$$\left(c_2^{-1}\phi(\mathbf{n}_{t_*})\right)^{-\rho_{(t)}/\rho_{(t_*)}} \leq \left(c_2^{-1}\phi(\mathbf{n}_{t_*})\right)^{-\frac{\ln\left(c_2^{-1}\phi(\mathbf{n}_t)\right)}{\ln\left(c_2^{-1}\phi(\mathbf{n}_{t_*})\right)}} = \left(c_2^{-1}\phi(\mathbf{n}_t)\right)^{-1}.$$

Thus, we have

$$\sum_{t \in [N+1]} n_t \epsilon^{\rho_t(2-\beta)} \leq \sum_{t \in [N+1]} n_{(t)} \left(c_2^{-1}\phi(\mathbf{n}_t)\right)^{-1} \leq c_2 \sum_{m=1}^\infty \frac{1}{\phi(m)} \leq c_2 cd$$

for a finite numerical constant c. Thus, choosing any $c_2 < 1/(64c)$, we have $\sum_{t \in [N+1]} n_t \epsilon^{\rho_t(2-\beta)} < \frac{d}{64}$, as desired.

Altogether, we have that for any learning rule $\hat{h}$, there exists $\Pi \in \mathcal{M}$ such that if $\hat{h}$ is trained on $Z \sim \Pi$, then with probability at least $1/50$,

$$\mathcal{E}_{\mathcal{D}}(\hat{h}) \geq \frac{\epsilon}{16} = \frac{1}{16} \min_{t \in [N+1]} \left(\frac{c_2}{\phi(\mathbf{n}_t)}\right)^{1/(2-\beta)\rho_{(t)}}.$$

In particular, since we always have $\bar{\rho}_t \leq \rho_{(t)}$, the theorem immediately follows. □

Remark 5. We note that it is clear from the proof that the $\left(\sum_{s=1}^t n_{(s)}\right)\log^2\left(\sum_{s=1}^t n_{(s)}\right)$ denominator in the lower bound is not strictly optimal. It can be replaced by any function $\phi'\left(\sum_{s=1}^t n_{(s)}\right)$ satisfying $\sum_{m=1}^\infty \phi'(m)^{-1} < \infty$: for instance, $\phi'(m) = m\log(m)(\log\log(m))^2$.

Remark 6. We also note that there exist classes $\mathcal{H}$ for which the lower bound can be extended to the full range of $\{\rho_t\}$ (i.e., any values $\rho_t > 0$). Specifically, in this general case, if there exist two points $x_0, x_1 \in \mathcal{X}$ such that *all* $h \in \mathcal{H}$ agree on $h(x_0)$ while $\exists h, h' \in \mathcal{H}$ with $h(x_1) \neq h'(x_1)$, then by the same construction (with $d = 1$) used in the proof we would have, with probability at least $1/50$,

$$\mathcal{E}_{\mathcal{D}}(\hat{h}) > c_1 \min_{t \in [N+1]} \left(\frac{c_2}{\left(\sum_{s=1}^t n_{(s)}\right)\log^2\left(\sum_{s=1}^t n_{(s)}\right)}\right)^{1/(2-\beta)\bar{\rho}_t}.$$

6 Upper Bound Analysis

We will in fact establish the upper bound as a bound holding with high probability $1-\delta$, for any $\delta \in (0,1)$. Throughout this subsection, let $\mathcal{M} = \mathcal{M}\left(C_\rho, \{\rho_t\}_{t\in[N]}, \{n_t\}_{t\in[N+1]}, C_\beta, \beta\right)$, for any admissible values of the parameters. Let

$$t^* \doteq \operatorname*{argmin}_{t\in[N+1]} C_\rho \left(2^{10} C_0^4 C_\beta \frac{d_{\mathcal{H}} \log\left(\frac{1}{d_{\mathcal{H}}}\sum_{s=1}^t n_{(s)}\right) + \log(1/\delta)}{\sum_{s=1}^t n_{(s)}} \right)^{1/(2-\beta)\bar{\rho}_t},$$

for C_0 as in Lemma 1 below, and for $\delta \in (0,1)$. The oracle procedure just returns $\hat{h}_{Z^{(t^*)}}$, the ERM over $Z^{(t^*)}$.

We have the following theorem.

Theorem 7. *For any $\Pi \in \mathcal{M}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\mathcal{E}_{Q_t}(\hat{h}_{Z^{(t^*)}}) \leq \min_{t\in[N+1]} C_\rho \left(C \frac{d_{\mathcal{H}} \log\left(\frac{1}{d_{\mathcal{H}}}\sum_{s=1}^t n_{(s)}\right) + \log(1/\delta)}{\sum_{s=1}^t n_{(s)}} \right)^{1/(2-\beta)\bar{\rho}_t}$$

for a constant $C = 2^{10} C_0^4 C_\beta$, where C_0 is a numerical constant from Lemma 1 below.

The proof will rely on the following lemma: a uniform Bernstein inequality for independent but non-identically distributed data. Results of this type are well known for i.i.d. data (see e.g., [Kol06]). For completeness, we include a proof of the extension to non-identically distributed data in Appendix A. The main technical modification of the proof, compared to the i.i.d. case, is employing a generalization of Bousquet's inequality for non-identical data, due to [KR05].

Lemma 1. *(Uniform Bernstein inequality for non-identical distributions) For any $m \in \mathbb{N}$ and $\delta \in (0,1)$, define $\varepsilon(m,\delta) \doteq \frac{d_{\mathcal{H}}}{m}\log\left(\frac{m}{d_{\mathcal{H}}}\right) + \frac{1}{m}\log\left(\frac{1}{\delta}\right)$ and let $S = \{(X_1,Y_1),\ldots,(X_m,Y_m)\}$ be independent samples. With probability at least $1-\delta$, $\forall h, h' \in \mathcal{H}$,*

$$\mathbb{E}\left[\hat{\mathcal{E}}_S(h;h')\right] \leq \hat{\mathcal{E}}_S(h;h') + C_0\sqrt{\min\left\{\mathbb{E}\left[\hat{\mathbb{P}}_S(h\neq h')\right], \hat{\mathbb{P}}_S(h\neq h')\right\}\varepsilon(m,\delta)} + C_0\varepsilon(m,\delta), \tag{5}$$

and

$$\frac{1}{2}\mathbb{E}\left[\hat{\mathbb{P}}_S(h\neq h')\right] - C_0\varepsilon(m,\delta) \leq \hat{\mathbb{P}}_S(h\neq h') \leq 2\mathbb{E}\left[\hat{\mathbb{P}}_S(h\neq h')\right] + C_0\varepsilon(m,\delta), \tag{6}$$

for a universal numerical constant $C_0 \in (0,\infty)$.

In particular, we will use this lemma via the following implication.

Lemma 2. *For any Π as in Theorem 7, for any $I \subseteq [N+1]$, letting $\mathbf{n}_I \doteq \sum_{t\in I} n_t$ and $\bar{P}_I \doteq \mathbf{n}_I^{-1}\sum_{t\in I} n_t P_t$, for any $\delta \in (0,1)$, on the event (of probability at least $1-\delta$) from Lemma 1 (for $S = Z^I$ there), for any $h \in \mathcal{H}$ satisfying*

$$\hat{\mathcal{E}}_{Z^I}(h;\hat{h}_{Z^I}) \leq C_0\sqrt{\hat{\mathbb{P}}_{Z^I}(h\neq\hat{h}_{Z^I})\varepsilon(\mathbf{n}_I,\delta)} + C_0\varepsilon(\mathbf{n}_I,\delta), \tag{7}$$

it holds that

$$\mathcal{E}_{\bar{P}_I}(h) \leq 32C_0^2\left(C_\beta\varepsilon(\mathbf{n}_I,\delta)\right)^{1/(2-\beta)}.$$

Proof. On the event from Lemma 1 for $S = Z^I$, it holds that $\forall h, h' \in \mathcal{H}$, (since $\mathcal{E}_{\bar{P}_I}(h;h') = \mathbb{E}[\hat{\mathcal{E}}_{Z^I}(h;h')]$)

$$\mathcal{E}_{\bar{P}_I}(h;h') \leq \hat{\mathcal{E}}_{Z^I}(h;h') + C_0\sqrt{\min\left\{\bar{P}_I(h\neq h'), \hat{\mathbb{P}}_{Z^I}(h\neq h')\right\}\varepsilon(\mathbf{n}_I,\delta)} + C_0\varepsilon(\mathbf{n}_I,\delta) \tag{8}$$

and

$$\hat{\mathbb{P}}_{Z^I}(h\neq h') \leq 2\bar{P}_I(h\neq h') + C_0\varepsilon(\mathbf{n}_I,\delta). \tag{9}$$

Suppose this event occurs.

By the Bernstein class condition and Jensen's inequality, for any $h \in \mathcal{H}$ we have

$$\bar{P}_I(h\neq h^*) \leq C_\beta \mathbf{n}_I^{-1}\sum_{t\in I} n_t \mathcal{E}_{P_t}^\beta(h) \leq C_\beta\left(\mathcal{E}_{\bar{P}_I}(h)\right)^\beta. \tag{10}$$

Furthermore, applying (8) with $h = \hat{h}_{Z^I}$ and $h' = h^*$ reveals that

$$\mathcal{E}_{\bar{P}_I}(\hat{h}_{Z^I}) \le C_0\sqrt{\bar{P}_I(\hat{h}_{Z^I} \ne h^*)\varepsilon(\mathbf{n}_I,\delta)} + C_0\varepsilon(\mathbf{n}_I,\delta).$$

Combining this with (10) (for $h = \hat{h}_{Z^I}$) implies that

$$\mathcal{E}_{\bar{P}_I}(\hat{h}_{Z^I}) \le C_0\sqrt{C_\beta\left(\mathcal{E}_{\bar{P}_I}(\hat{h}_{Z^I})\right)^\beta \varepsilon(\mathbf{n}_I,\delta)} + C_0\varepsilon(\mathbf{n}_I,\delta), \tag{11}$$

which implies

$$\mathcal{E}_{\bar{P}_I}(\hat{h}_{Z^I}) \le 2C_0\left(C_\beta\varepsilon(\mathbf{n}_I,\delta)\right)^{1/(2-\beta)}. \tag{12}$$

Next, applying (8) with any $h \in \mathcal{H}$ satisfying (7) and with $h' = \hat{h}_{Z^I}$ implies

$$\mathcal{E}_{\bar{P}_I}(h;\hat{h}_{Z^I}) \le 2C_0\sqrt{\hat{\mathbb{P}}_{Z^I}(h \ne \hat{h}_{Z^I})\varepsilon(\mathbf{n}_I,\delta)} + 2C_0\varepsilon(\mathbf{n}_I,\delta), \tag{13}$$

and (9) implies the right hand side is at most (after some simplifications of the resulting expression)

$$2C_0\sqrt{2\bar{P}_I(h \ne \hat{h}_{Z^I})\varepsilon(\mathbf{n}_I,\delta)} + 2C_0(1+\sqrt{C_0})\varepsilon(\mathbf{n}_I,\delta). \tag{14}$$

By the triangle inequality and (10), together with (12), we have

$$\begin{aligned}\bar{P}_I(h \ne \hat{h}_{Z^I}) \le \bar{P}_I(h \ne h^*) + \bar{P}_I(\hat{h}_{Z^I} \ne h^*) &\le C_\beta\left(\mathcal{E}_{\bar{P}_I}(h)\right)^\beta + C_\beta\left(\mathcal{E}_{\bar{P}_I}(\hat{h}_{Z^I})\right)^\beta \\ &\le C_\beta\left(\mathcal{E}_{\bar{P}_I}(h)\right)^\beta + C_\beta(2C_0)^\beta\left(C_\beta\varepsilon(\mathbf{n}_I,\delta)\right)^{\beta/(2-\beta)}.\end{aligned}$$

Combining this with (14) and (13) implies (with some simplification of the resulting expression)

$$\mathcal{E}_{\bar{P}_I}(h;\hat{h}_{Z^I}) \le 2C_0\sqrt{2C_\beta\left(\mathcal{E}_{\bar{P}_I}(h)\right)^\beta\varepsilon(\mathbf{n}_I,\delta)} + 8C_0^2\left(C_\beta\varepsilon(\mathbf{n}_I,\delta)\right)^{1/(2-\beta)}.$$

Since $\mathcal{E}_{\bar{P}_I}(h) = \mathcal{E}_{\bar{P}_I}(h;\hat{h}_{Z^I}) + \mathcal{E}_{\bar{P}_I}(\hat{h}_{Z^I})$, together with (12) this implies

$$\mathcal{E}_{\bar{P}_I}(h) \le 2C_0\sqrt{2C_\beta\left(\mathcal{E}_{\bar{P}_I}(h)\right)^\beta\varepsilon(\mathbf{n}_I,\delta)} + 10C_0^2\left(C_\beta\varepsilon(\mathbf{n}_I,\delta)\right)^{1/(2-\beta)}.$$

In particular, this inequality immediately implies the claimed inequality in the lemma. □

We now present the proof of Theorem 7.

Proof of Theorem 7. For each $t \in [N+1]$, define $\mathbf{n}_t \doteq \sum_{s=1}^{t} n_{(s)}$. For brevity, let $\bar{\rho} = \bar{\rho}_{t^*}$ and $\hat{h} = \hat{h}_{Z^{(t^*)}}$. Let $\bar{P} = \mathbf{n}_{t^*}^{-1}\sum_{t=1}^{t^*} n_{(t)}P_{(t)}$. First note that

$$\mathcal{E}_{\bar{P}}(\hat{h}) \ge \frac{1}{\mathbf{n}_{t^*}}\sum_{t=1}^{t^*} n_{(t)}\left(C_\rho^{-1}\mathcal{E}_{\mathcal{D}}(\hat{h})\right)^{\rho_{(t)}} \ge \left(C_\rho^{-1}\mathcal{E}_{\mathcal{D}}(\hat{h})\right)^{\bar{\rho}},$$

where the final inequality is due to Jensen's inequality. In particular, this implies

$$\mathcal{E}_{\mathcal{D}}(\hat{h}) \le C_\rho\mathcal{E}_{\bar{P}}^{1/\bar{\rho}}(\hat{h}), \tag{15}$$

so that it suffices to upper bound the expression on the right hand side. Toward this end, note that $\hat{h}$ trivially satisfies (7) for $I = \{(1),\ldots,(t^*)\}$. Therefore, Lemma 2 implies that with probability at least $1-\delta$,

$$\mathcal{E}_{\bar{P}}(\hat{h}) \le 32C_0^2(C_\beta\varepsilon(\mathbf{n}_{t^*},\delta))^{1/(2-\beta)}.$$

Combining this with (15) immediately implies the theorem. □

Remark 7. In particular, the upper bound for Theorem 1 follows from Theorem 7 by plugging in $\delta = 1/\sum_{s=1}^{t^*} n_{(s)}$.

7 Partially Adaptive Procedures

Fix any N, C_ρ, $\{\rho_t\}_{t\in[N]}$, $\{n_t\}_{t\in[N+1]}$, C_β, and $\beta \in [0,1]$, and let $\mathcal{M} = \mathcal{M}\left(C_\rho, \{\rho_t\}_{t\in[N]}, \{n_t\}_{t\in[N+1]}, C_\beta, \beta\right)$.

7.1 Pooling Under Low Noise $\beta = 1$

We now present the proof of Theorem 3 which states the near-optimality of *pooling*, independent of the choice of target $\mathcal{D}$, whenever $\beta = 1$.

Proof of Theorem 3. For any $t \in [N+1]$, let $\mathbf{n}_t \doteq \sum_{s=1}^{t} n_{(s)}$ and $\bar{P}_t \doteq (\mathbf{n}_t)^{-1} \sum_{s=1}^{t} n_{(s)} P_{(s)}$. Suppose the event from Lemma 1 holds for $Z = Z^{(N+1)}$, which occurs with probability at least $1 - \delta$. By Lemma 2, we know that on this event,

$$\mathcal{E}_{\bar{P}_{N+1}}(\hat{h}_Z) \leq 32 C_0^2 C_\beta \cdot \varepsilon(\mathbf{n}_{N+1}, \delta).$$

Combining this with the definition of ρ_t, we have

$$\mathbf{n}_{N+1}^{-1} \sum_{t \in [N+1]} n_t \left(C_\rho^{-1} \mathcal{E}_{\mathcal{D}}(\hat{h}_Z) \right)^{\rho_t} \leq 32 C_0^2 C_\beta \cdot \varepsilon(\mathbf{n}_{N+1}, \delta).$$

Since the left hand side is monotonic in $\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)$, if we wish to bound $\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)$ by some value ϵ, it suffices to take any value of ϵ such that

$$\mathbf{n}_{N+1}^{-1} \sum_{t \in [N+1]} n_t \left(C_\rho^{-1} \epsilon \right)^{\rho_t} \geq 32 C_0^2 C_\beta \cdot \varepsilon(\mathbf{n}_{N+1}, \delta),$$

or more simply

$$\sum_{t \in [N+1]} n_t \left(C_\rho^{-1} \epsilon \right)^{\rho_t} \geq 32 C_0^2 C_\beta \left(d_{\mathcal{H}} \log\left(\frac{\mathbf{n}_{N+1}}{d_{\mathcal{H}}} \right) + \log\left(\frac{1}{\delta} \right) \right).$$

In particular, let us take

$$\epsilon \doteq C_\rho \left(32 C_0^2 C_\beta \frac{d_{\mathcal{H}} \log(\mathbf{n}_{N+1}/d_{\mathcal{H}}) + \log(1/\delta)}{\mathbf{n}_{t^*}} \right)^{1/\bar{\rho}_{t^*}}$$

where t^* is the value of t that minimizes the right hand side of this definition of ϵ. Then

$$\begin{aligned} \sum_{t \in [N+1]} n_t \left(C_\rho^{-1} \epsilon \right)^{\rho_t} &\geq \mathbf{n}_{t^*} \sum_{t=1}^{t^*} \frac{n_{(t)}}{\mathbf{n}_{t^*}} \left(C_\rho^{-1} \epsilon \right)^{\rho_{(t)}} \geq \mathbf{n}_{t^*} (C_\rho^{-1} \epsilon)^{\bar{\rho}_{t^*}} \\ &= 32 C_0^2 C_\beta \left(d_{\mathcal{H}} \log(\mathbf{n}_{N+1}/d_{\mathcal{H}}) + \log(1/\delta) \right). \end{aligned}$$

□

7.2 Aggregation Under Ranking Information

Proof of Theorem 4. For any $t \in [N+1]$ let $\bar{P}_t = \mathbf{n}_t^{-1} \sum_{s=1}^{t} n_{(s)} P_{(s)}$. Let t^* be as in Theorem 7, and by the same argument from the proof of Theorem 7, (15) holds for $\hat{h}$ in the present context as well (where $\bar{P}$ there is $\bar{P}_{t^*}$ in the present notation). Thus, the theorem will be proven if we can bound $\mathcal{E}_{\bar{P}_{t^*}}(\hat{h})$ by $(c C_\beta \varepsilon(\mathbf{n}_{t^*}, \delta_{t^*}))^{1/(2-\beta)}$ for some numerical constant c.

By a union bound, with probability at least $1 - \sum_{t=1}^{N+1} \delta_t \geq 1 - \delta$, for every $t \in [N+1]$, the event from Lemma 1 holds for $S = Z^{(t)}$ (and with δ_t in place of δ there). In particular, this implies that for every $t \in [N+1]$,

$$\hat{\mathcal{E}}_{Z^{(t)}}(h^*; \hat{h}_{Z^{(t)}}) \leq C_0 \sqrt{\hat{\mathbb{P}}_{Z^{(t)}}(h^* \neq \hat{h}_{Z^{(t)}}) \varepsilon(\mathbf{n}_t, \delta_t)} + C_0 \varepsilon(\mathbf{n}_t, \delta_t),$$

so that there do exist classifiers h in $\mathcal{H}$ satisfying (4), and hence $\hat{h}$ satisfies (4). In particular, this implies $\hat{h}$ satisfies the inequality in (4) for $t = t^*$. By Lemma 2, this implies that on this same event from above, it holds that

$$\mathcal{E}_{\bar{P}_{t^*}}(\hat{h}) \leq 32 C_0^2 (C_\beta \varepsilon(\mathbf{n}_{t^*}, \delta_{t^*}))^{1/(2-\beta)},$$

which completes the proof (for instance, taking $c = 2^{10} C_0^4$). □

8 Impossibility of Adaptivity over Multisource Classes $\mathcal{M}$

Theorem 5 follows as corollary to Theorem 8, the main result of this section. In particular, the second part of Theorem 5 follows from the condition on $\mathcal{M}$ that it contains enough tasks with $\rho_t = 1$, and calling on Theorem 4.

Theorem 8 (Impossibility of Adaptivity)**.** *Pick $C_\rho = 3$, and any $0 \le \beta < 1$, $C_\beta \ge 2$, a number of samples per source task $1 \le n < 2/\beta - 1$, and a number of target samples $n_{\mathcal{D}} \ge 0$. Let the number of source tasks $N = N_P + N_Q$, for N_P, N_Q as specified below. There exist universal constants $C_0, C_1, c > 0$ such that the following holds.*

Choose any $N_Q \ge C_0$, and suppose N_P is sufficiently large so that $N_P \ge 3N_Q$, and furthermore

$$N_P^{(2-(n+1)\beta)/(2-\beta)} \ge C_1 \cdot N_Q^2 \cdot 2^{15n}.$$

Let $\boldsymbol{\mathcal{M}}$ denote the family of multisource classes $\mathcal{M}$ satisfying the above conditions, with parameters $n_t = n, \forall t \in [N]$, $n_{N+1} = n_{\mathcal{D}}$, and, in addition, such that at least $\frac{1}{2}N_Q$ of the exponents $\{\rho_t\}_{t\in[N]}$ are at most 1.

• Let $\hat{h}$ denote any classification procedure having access to $Z \sim \Pi$, $\Pi \in \mathcal{M}$, but without knowledge of $\mathcal{M}$. We have:

$$\inf_{\hat{h}} \sup_{\mathcal{M}\in\boldsymbol{\mathcal{M}}} \sup_{\Pi\in\mathcal{M}} \mathbb{P}_\Pi \left(\mathcal{E}_{\mathcal{D}}(\hat{h}) \ge \frac{1}{4} \cdot \left(1 \wedge n_{\mathcal{D}}^{-1/(2-\beta)} \right) \right) \ge c.$$

• On the other hand, there exists a semi-adaptive classifier $\tilde{h}$, which, given data $Z \sim \Pi$, along with a ranking $\{\rho_{(t)}\}_{t\in[N+1]}$ of increasing exponents values, achieves the rate

$$\sup_{\mathcal{M}\in\boldsymbol{\mathcal{M}}} \sup_{\Pi\in\mathcal{M}} \mathbb{E}_\Pi \left[\mathcal{E}_{\mathcal{D}}(\tilde{h}) \right] \lesssim (n \cdot N_Q)^{-1/(2-\beta)}.$$

As it turns out, the above theorem in fact holds for even smaller classes $\mathcal{M}$ admitting just two choices of distributions, namely P_σ and Q_σ (for fixed $\sigma \in \{\pm 1\}$) below, to which sources P_t's might be assigned to. In fact the result holds even if $\hat{h}$ has knowledge of the construction below, but of course not of σ.

Construction. We build on the following distributions supported on 2 points x_0, x_1: here we simply assume that $\mathcal{H}$ contains at least 2 classifiers that disagree on x_1 but agree on x_0 (this will be the case, for some choice of x_0, x_1, if $\mathcal{H}$ contains at least 3 classifiers). Therefore, w.l.o.g., assume that x_0 has label 1. Let $n \ge 1, n_{\mathcal{D}} \ge 0, N_P, N_Q \ge 1$, $0 \le \beta < 1$, and define $\epsilon \doteq (n \cdot N_P)^{-1/(2-\beta)}$ and $\epsilon_0 \doteq 1 \wedge n_{\mathcal{D}}^{-1/(2-\beta)}$. Let $\sigma \in \{\pm 1\}$ – which we will often abbreviate as $\pm$. In all that follows, we let $\eta_\mu(X)$ denote the regression function $\mathbb{P}_\mu[Y = 1 \mid X]$ under distribution μ.

- **Target** $\mathcal{D}_\sigma = \mathcal{D}_X \times \mathcal{D}^\sigma_{Y|X}$**:** Let $\mathcal{D}_X(x_1) = \frac{1}{2} \cdot \epsilon_0^\beta$, $\mathcal{D}_X(x_0) = 1 - \frac{1}{2} \cdot \epsilon_0^\beta$; finally $\mathcal{D}^\sigma_{Y|X}$ is determined by $\eta_{\mathcal{D},\sigma}(x_1) = 1/2 + \sigma \cdot c_0 \epsilon_0^{1-\beta}$, and $\eta_{\mathcal{D},\sigma}(x_0) = 1$, for some c_0 to be specified later.
- **Noisy** $P_\sigma = P_X \times P^\sigma_{Y|X}$**:** Let $P_X(x_1) = c_1\epsilon^\beta$, $P_X(x_0) = 1 - c_1\epsilon^\beta$; finally $P^\sigma_{Y|X}$ is determined by $\eta_{P,\sigma}(x_1) = 1/2 + \sigma \cdot \epsilon^{1-\beta}$, and $\eta_{P,\sigma}(x_0) = 1$, for an appropriate constant c_1 specified later.
- **Benign** $Q_\sigma = Q_X \times Q^\sigma_{Y|X}$**:** Let $Q_X(x_1) = 1$; finally $Q^\sigma_{Y|X}$ is determined by $\eta_{Q,\sigma}(x_1) = 1/2 + \sigma/2$.

The above construction is such that P_σ can be pushed far from $\mathcal{D}_\sigma$, while Q_σ remains close to $\mathcal{D}_\sigma$. This is formalized in the following proposition, which can be verified by checking the required inequalities (from Definitions 2 and 4) are satisfied in all 4 cases of classifications of x_0, x_1.

Proposition 1 (Exponents of P and Q w.r.t. $\mathcal{D}$)**.** *P_σ and Q_σ have transfer-exponents $(3, \rho_P)$ and $(3, \rho_Q)$, respectively w.r.t. $\mathcal{D}_\sigma$, with $\rho_P \ge \frac{\log(c_1^{-(2-\beta)} n \cdot N_P)}{\log(c_0^{-(2-\beta)}(1 \vee n_{\mathcal{D}}))}$ and $\rho_Q \le 1$. Furthermore, the 3 distributions $P_\sigma, Q_\sigma, \mathcal{D}_\sigma$ satisfy the Bernstein class condition with parameters $(C_\beta, \beta), C_\beta = \max\{(1/2)c_0^{-\beta}, 2\}$.*

Approximate Multitask. Let $N \doteq N_P + N_Q$, $\alpha_P \doteq N_P/N$ and $\alpha_Q = N_Q/N$. Now consider the distribution

$$\Gamma_\sigma = (\alpha_P P_\sigma^n + \alpha_Q Q_\sigma^n)^N \times \mathcal{D}_\sigma^{n_{\mathcal{D}}}.$$

Proof Strategy. Henceforth, let $Z = \{Z_t\}_{t=1}^{N+1}$ denote a random draw from Γ_σ, where $Z_t = \{(X_{t,i}, Y_{t,i})\}_{i=1}^n \sim \Gamma_\sigma$ for $t \in [N]$, and $Z_{N+1} \sim \mathcal{D}_\sigma$. Much of the effort will be in showing that any learner has nontrivial error in guessing σ from a random Z; we then reduce this fact to a lower-bound on the risk of an adaptive classifier that takes as input a multitask sample of the form $Z' \sim \prod_{t=1}^N P_t^n \times \mathcal{D}_\sigma^n$, where P_t's denote P_σ or Q_σ.

For intuition, the true label σ is hard to guess from samples $Z_t \sim P_\sigma^n$ alone since $\eta_{P,\sigma} \approx 1/2$, but easy to guess from samples $Z_t \sim Q^n$, each being a homogeneous vector of n points x_1 with identical labels $Y = \sigma$. However, such homogeneous vectors can be produced by P_σ^n also, with identical but flipped labels $Y = -\sigma$, and the product of mixtures Γ_σ adds in enough randomness to make it hard to guess the source of a given vector Z_t. In fact, this intuition becomes evident by considering the *likelihood-ratio* between Γ_+ and Γ_- which decomposes into the contribution of homogeneous vs non-homogenous vectors. This is formalized in the following proposition.

Proposition 2 (Likelihood ratio and sufficient statistics)**.** *Let $Z \sim \Gamma_-$, and let and $\hat{N}_+$ and $\hat{N}_-$ denote the number of homogeneous vectors Z_t in Z with marginals at x_1, that is:*

$$\hat{N}_\sigma(Z) \doteq \sum_{t\in[N]} \mathbb{1}\{\forall i \in [n], X_{t,i} = x_1 \wedge Y_{t,i} = \sigma\}.$$

Next, let $\hat{n}_+$ and $\hat{n}_-$ denote the total number of ± 1 labels in Z, over occurrences of x_1. That is:

$$\hat{n}_\sigma(Z) \doteq \sum_{t\in[N], i\in[n]} \mathbb{1}\{X_{t,i} = x_1 \wedge Y_{t,i} = \sigma\}.$$

We then have that (where $\mathbb{P}$ is under the randomness of $Z \sim \Gamma_-$, and we let $\mathcal{D}_\sigma(Z_{N+1}) = 1$ when $n_\mathcal{D} = 0$):

$$\mathbb{P}\left(\frac{\Gamma_+(Z)}{\Gamma_-(Z)} > 1\right) \geq \mathbb{P}\left(\hat{N}_+(Z) > \hat{N}_-(Z) \wedge \hat{n}_+(Z) \geq \hat{n}_-(Z)\right) \cdot \mathbb{P}\left(\frac{\mathcal{D}_+(Z_{N+1})}{\mathcal{D}_-(Z_{N+1})} \geq 1\right). \tag{16}$$

Remark 8 (Likelihoood Ratio and Optimal Discriminants)**.** Consider sampling Z from the mixture $\frac{1}{2}\Gamma_+ + \frac{1}{2}\Gamma_-$. Then the Bayes classifier (for identifying $\sigma = \pm$) returns $+1$ if $\Gamma_+(Z) > \Gamma_-(Z)$, and therefore, given the symmetry between $\Gamma_\pm$, has probability of misclassification at least $\mathbb{P}_{\Gamma_-}(\Gamma_+(Z) > \Gamma_-(Z))$. We emphasize here that enforcing that Γ_σ be defined in terms of a mixture – rather than as product of P_σ and Q_σ terms – allows us to bound $\mathbb{P}_{\Gamma_-}(\Gamma_+(Z) > \Gamma_-(Z))$ below as in Proposition 2 above; otherwise this probability is always 0 for a product distribution containing Q_σ since $Q_\sigma(Z_t) = 0$ whenever some $Y_{t,i} = -\sigma$. In fact a product distribution inherently encodes the idea that the learner *knows* the positions of P_σ and Q_σ vectors in Z, and can therefore simply focus on Q_σ vectors to easily discover σ.

We now further reduce the r.h.s. of (16) to events that are simpler to bound: the main strategy is to properly condition on intermediate events to reveal independences that induce simple i.i.d. Bernouilli's we can exploit. Towards this end, we first consider the high-probability events of the following proposition.

Proposition 3. *Let $Z \sim \Gamma_-$. Define $\hat{N}_P(Z), \hat{N}_Q(Z)$ as the number of homogeneous vectors $\{Z_t : \forall i, j \in [n], X_{t,i} = X_{t,j} = x_1 \wedge Y_{t,i} = Y_{t,j}\}$ generated respectively by P_-, and Q_-.*

Define the events (on Z):

$$\mathbf{E}_P \doteq \left\{\mathbb{E}\left[\hat{N}_P\right]/2 \leq \hat{N}_P \leq 2\mathbb{E}\left[\hat{N}_P\right]\right\} \text{ and } \mathbf{E}_Q \doteq \left\{\hat{N}_Q \leq 2\mathbb{E}\left[\hat{N}_Q\right]\right\}.$$

We have that $\mathbb{P}(\mathbf{E}_P^\mathsf{c}) \leq 2\exp\left(-\mathbb{E}\left[\hat{N}_P\right]/8\right)$ while $\mathbb{P}(\mathbf{E}_Q^\mathsf{c}) \leq \exp\left(-\mathbb{E}\left[\hat{N}_Q\right]/3\right)$.

Proof. The proposition follows from multiplicative Chernoff bounds. □

Notice that, in the above proposition, for $Z \sim \Gamma_-$, the expectations in the events are given by

$$\mathbb{E}\left[\hat{N}_P(Z)\right] = N_P P_X^n(x_1)\left(\eta_{P,-}^n(x_1) + \eta_{P,+}^n(x_1)\right) \text{ and } \mathbb{E}\left[\hat{N}_Q(Z)\right] = N_Q. \tag{17}$$

By the proposition, we just need these quantities large enough for the events to hold with sufficient probability. The next proposition conditions on these events to reduce the likelihood to simpler events.

Proposition 4 (Further Reducing (16))**.** *Let $Z \sim \Gamma_-$. For $\sigma = \pm$, let $\hat{N}_\sigma(Z)$ and $\hat{n}_\sigma(Z)$ as defined in Proposition 2. Furthermore, let $\tilde{n}_\sigma(Z) \doteq \hat{n}_\sigma(Z) - (n \cdot \hat{N}_\sigma(Z))$ denote the total number of $\pm$ labels at x_1 excluding homogeneous vectors. In all that follows we let $\mathbb{P}$ denote $\mathbb{P}_{\Gamma_-}$, and we drop the dependence on Z for ease of notation.*

Let $\hat{N}_P, \hat{N}_Q$ and the events $\mathbf{E}_P, \mathbf{E}_Q$ as defined in Proposition 3. Suppose that for some $\delta_1, \delta_2 > 0$, we have

(i) $\mathbb{P}\left(\hat{N}_+ > \hat{N}_- \mid \hat{N}_P, \hat{N}_Q\right) \mathbb{1}\{\mathbf{E}_P \cap \mathbf{E}_Q\} \geq \delta_1 \cdot \mathbb{1}\{\mathbf{E}_P \cap \mathbf{E}_Q\}$.

(ii) $\mathbb{P}\left(\tilde{n}_+ > \tilde{n}_-\right) \geq \delta_2$*, further assuming that* $n > 1$ *so that the event is well defined (i.e.,* $\tilde{n}_\pm$ *are not both* 0*).*

We then have that:

$$\mathbb{P}\left(\hat{N}_+ \geq \hat{N}_- \wedge \hat{n}_+ \geq \hat{n}_-\right) \geq \delta_1 \cdot \left(\delta_2 - \mathbb{P}\left(\mathbf{E}_P^{\mathsf{c}}\right) - \mathbb{P}\left(\mathbf{E}_Q^{\mathsf{c}}\right)\right). \tag{18}$$

Proof. Let $A \doteq \{\hat{n}_+ \geq \hat{n}_-\}$, $B \doteq \left\{\hat{N}_+ > \hat{N}_-\right\}$, and $\tilde{A} \doteq \{\tilde{n}_+ > \tilde{n}_-\}$. First notice that, by definition, $\tilde{A} \cap B \implies A \cap B$, so we just need to bound $\mathbb{P}(\tilde{A} \cap B)$ below. We have:

$$\begin{aligned}
\mathbb{P}\left(\tilde{A} \cap B\right) &= \mathbb{E}\left[\mathbb{P}\left(\tilde{A} \cap B \mid \hat{N}_P, \hat{N}_Q\right)\right] = \mathbb{E}\left[\mathbb{P}\left(\tilde{A} \mid \hat{N}_P, \hat{N}_Q\right) \cdot \mathbb{P}\left(B \mid \hat{N}_P, \hat{N}_Q\right)\right] && (19)\\
&\geq \delta_1 \cdot \mathbb{E}\left[\mathbb{P}\left(\tilde{A} \mid \hat{N}_P, \hat{N}_Q\right) \cdot \mathbb{1}\{\mathbf{E}_P \cap \mathbf{E}_Q\}\right]\\
&\geq \delta_1 \cdot \mathbb{E}\left[\mathbb{P}\left(\tilde{A} \mid \hat{N}_P, \hat{N}_Q\right) - \mathbb{1}\left\{\mathbf{E}_P^{\mathsf{c}} \cup \mathbf{E}_Q^{\mathsf{c}}\right\}\right] = \delta_1 \cdot \left(\mathbb{P}\left(\tilde{A}\right) - \mathbb{P}\left(\mathbf{E}_P^{\mathsf{c}} \cup \mathbf{E}_Q^{\mathsf{c}}\right)\right) && (20)\\
&\geq \delta_1 \cdot \left(\delta_2 - \mathbb{P}\left(\mathbf{E}_P^{\mathsf{c}}\right) - \mathbb{P}\left(\mathbf{E}_Q^{\mathsf{c}}\right)\right).
\end{aligned}$$

In (19), we used the fact that, all vectors in Z being independent, $\tilde{A}$ is independent of B given any value of $\left\{\hat{N}_P, \hat{N}_Q\right\}$; in (20) we used the fact that for any $p \in [0,1]$, we have $p \cdot \mathbb{1}\{\mathbf{E}\} = p - p \cdot \mathbb{1}\{\mathbf{E}^{\mathsf{c}}\} \geq p - \mathbb{1}\{\mathbf{E}^{\mathsf{c}}\}$. □

The following is a known counterpart of Chernoff bounds, following from Slud's inequalities [Slu77] for Binomials.

Lemma 3 (Anticoncentration (Slud's Inequality))**.** *Let* $\{X_i\}_{i\in[m]}$ *denote i.i.d. Bernouilli's with parameter* $0 < p \leq 1/2$*. Then for any* $0 \leq m_0 \leq m(1-2p)$*, we have*

$$\mathbb{P}\left(\sum_{i\in[m]} X_i > mp + m_0\right) \geq \frac{1}{4}\exp\left(\frac{-m_0^2}{mp(1-p)}\right).$$

Proof. This form of the lower bound follows from Theorem 2.1 of [Slu77] – which here implies a lower bound $1 - \Phi(m_0/\sqrt{mp(1-p)})$, for Φ the standard normal CDF – together with the well-known bound: $1 - \Phi(x) \geq \frac{1}{2}\left(1 - \sqrt{1 - e^{-x^2}}\right) \geq \frac{1}{4}e^{-x^2}$ [Tat53]. See [Mou10] for a detailed derivation of a similar expression. □

Proposition 5 (δ_1 from Proposition 4)**.** *Pick any* $0 \leq \beta < 1$*,* $1 \leq n < 2/\beta - 1$*, and* $N_Q \geq 1$*. Let* $0 < c_1 \leq 1/64$*, in the construction of* $\eta_{P,\sigma}$*,* $\sigma = \pm$*. Suppose* N_P *is sufficiently large so that*

$$(n \cdot N_P)^{(2-(n+1)\beta)/(2-\beta)} \geq 4 \cdot N_Q^2 \cdot n \cdot 2^{4n} c_1^{-n}.$$

Let $\mathbf{E} = \mathbf{E}_P \cap \mathbf{E}_Q$ *as defined in Proposition 3 over* $Z \sim \Gamma_-$*. Adopting the notation of Proposition 4, we have*

$$\mathbb{P}\left(\hat{N}_+ > \hat{N}_- \mid \hat{N}_P, \hat{N}_Q\right) \mathbb{1}\{\mathbf{E}\} \geq \frac{1}{12} \cdot \mathbb{1}\{\mathbf{E}\}.$$

Proof. Consider $N_H = \left\{\hat{N}_P, \hat{N}_Q\right\}$ such that $\mathbf{E} \doteq \mathbf{E}_P \cap \mathbf{E}_Q$ holds. Let $\tilde{B} \doteq \left\{\hat{N}_+ > \frac{1}{2}\hat{N}_P + \mathbb{E}\left[\hat{N}_Q\right]\right\}$, and notice that $\tilde{B} \subset \left\{\hat{N}_+ > \hat{N}_-\right\}$ under $\mathbf{E}_Q \doteq \left\{\mathbb{E}\left[\hat{N}_Q\right] \geq \hat{N}_Q/2\right\}$. Therefore we only need to bound $\mathbb{P}\left(\tilde{B} \mid N_H\right) = \mathbb{P}\left(\tilde{B} \mid \hat{N}_P\right)$. Now, for $\sigma = \pm$, let $\mathcal{Z}_{P,\sigma}$ denote the set of homogeneous vectors in $\{Z_t : \forall i \in [n], X_{t,i}x_1 \wedge Y_{t,i} = \sigma\}$ generated by P_-, and notice that, conditioned on $\hat{N}_P$, $\hat{N}_+$ is distributed as Binomial$(\hat{N}_P, p)$, where

$$p = \mathbb{P}\left(Z_t \in \mathcal{Z}_{P,+} \mid Z_t \in \mathcal{Z}_{P,+} \cup \mathcal{Z}_{P,-}, \hat{N}_P\right) = \mathbb{P}\left(Z_t \in \mathcal{Z}_{P,+} \mid Z_t \in \mathcal{Z}_{P,+} \cup \mathcal{Z}_{P,-}\right) = \frac{\eta_{P,-}^n(x_1)}{\eta_{P,-}^n(x_1) + \eta_{P,+}^n(x_1)}.$$

Therefore, applying Lemma 3,

$$\mathbb{P}\left(\hat{N}_+ > \hat{N}_P \cdot p + \sqrt{\hat{N}_P \cdot p(1-p)} \mid \hat{N}_P\right) \geq \frac{1}{12}.$$

We now just need to show that the event under the probability implies $\tilde{B}$, in other words, that

$$\sqrt{\hat{N}_P \cdot p(1-p)} \geq \frac{1}{2}\hat{N}_P - \hat{N}_P \cdot p + \mathbb{E}\left[\hat{N}_Q\right] = \hat{N}_P \cdot \left(\frac{1}{2} - p\right) + N_Q. \tag{21}$$

Next we upper bound the r.h.s of (21) and lower-bound its l.h.s. Under $\mathbf{E}_P$ and using (17), we have:

$$\begin{aligned}\hat{N}_P \cdot \left(\frac{1}{2} - p\right) &\leq \mathbb{E}\left[\hat{N}_P\right] \cdot ((1-p) - p) = N_P P_X^n(x_1) \cdot \left(\eta_{P,+}^n(x_1) - \eta_{P,-}^n(x_1)\right) \\ &\leq N_P P_X^n(x_1) \cdot n \cdot (\eta_{P,+}(x_1) - \eta_{P,-}(x_1)), \end{aligned} \tag{22}$$

where for the last inequality we used the fact that, for $a \geq b > 0$, we have $a^n - b^n = (a-b)\sum_{k=0}^{n-1} a^{n-1} \cdot \left(\frac{b}{a}\right)^k$. On the other hand, the conditions on N_P let us lower-bound $\eta_{P,-}(x_1)$ by $1/4$, and we have:

$$\begin{aligned}\hat{N}_P \cdot p(1-p) &\geq \frac{1}{2}\mathbb{E}\left[\hat{N}_P\right] \cdot p(1-p) = \frac{1}{2} N_P P_X^n(x_1) \cdot \frac{\eta_{P,-}^n(x_1) \cdot \eta_{P,+}^n(x_1)}{\eta_{P,-}^n(x_1) + \eta_{P,+}^n(x_1)} \\ &\geq \frac{1}{2} N_P P_X^n(x_1) \cdot 2^{-3n}. \end{aligned} \tag{23}$$

Combining (22) and (23), we see that $\tilde{B}$ is implied by

$$2^{-(2n)}\sqrt{N_P P_X^n(x_1)} \geq N_P P_X^n(x_1) \cdot n \cdot (\eta_{P,+}(x_1) - \eta_{P,-}(x_1)) + N_Q, \text{ which is in turn implied by}$$

$$n^{-1} 2^{-(4n)} c_1^n \left(n \cdot N_P\right)^{\frac{2-(n+1)\beta}{2-\beta}} \geq 2c_1^{2n}\left(n \cdot N_P\right)^{\frac{2-2n\beta}{2-\beta}} + 2N_Q^2.$$

The conditions of the proposition ensure that the above inequality holds. □

We obtain a bound on $\mathbb{P}(N_+ > \hat{N}_-)$ as an immediate corollary to the above proposition.

Corollary 1. *Under the conditions of Proposition 5, we have:*

$$\mathbb{P}\left(\hat{N}_+ > \hat{N}_-\right) \geq \mathbb{E}\left[\mathbb{P}\left(\hat{N}_+ > \hat{N}_- \mid \hat{N}_P, \hat{N}_Q\right) \mathbb{1}\{\mathbf{E}\}\right] \geq \frac{1}{12}\mathbb{P}(\mathbf{E}).$$

We now turn to the second condition of Proposition 4.

Proposition 6 (δ_2 from Proposition 4)**.** *Let $0 \leq \beta < 1$, $n > 1$, and $N_Q \geq 16$. Let $0 < c_1 \leq 2^{-10}$ in the construction of $\eta_{P,\sigma}$, $\sigma = \pm$. Suppose N_P is sufficiently large so that*

$$\textit{(i) } N_P \geq 3N_Q \quad \textit{and} \quad \textit{(ii) } \left(n \cdot N_P\right)^{\frac{2-2\beta}{2-\beta}} \geq 4096 \cdot n^2 \cdot c_1^{-1}.$$

Let $Z \sim \Gamma_-$, and $\tilde{n}_\sigma = \tilde{n}_\sigma(Z)$, $\sigma = \pm$ as defined in Proposition 4. We then have that $\mathbb{P}\left(\tilde{n}_+ > \tilde{n}_-\right) \geq \frac{1}{48}$.

Proof. Under the notation of Proposition 4, let $\hat{N}_{P,-} \doteq \hat{N}_P - \hat{N}_+$, and for homogeneity of notation herein, let $\hat{N}_{P,+} \doteq \hat{N}_+$. Fix $\delta > 0$ to be defined, and notice that if $\delta > n \cdot \left(\hat{N}_{P,+} - \hat{N}_{P,-}\right)$, then the event

$$\tilde{A}_\delta \doteq \left\{\tilde{n}_+ + n \cdot \hat{N}_{P,+} \geq \tilde{n}_- + n \cdot \hat{N}_{P,-} + \delta\right\} \text{ implies } \left\{\tilde{n}_+ > \tilde{n}_-\right\}.$$

As a first step, we want to upper-bound $\left(\hat{N}_{P,+} - \hat{N}_{P,-}\right)$. Let V_δ denote an upper-bound on the variance of this quantity: we have by Bernstein's inequality that, for any $t \leq \sqrt{V_\delta}$, with probability at least $1 - e^{-t^2/4}$,

$$\left(\hat{N}_{P,+} - \hat{N}_{P,-}\right) \leq \mathbb{E}\left[\hat{N}_{P,+} - \hat{N}_{P,-}\right] + t\sqrt{V_\delta} \leq t\sqrt{V_\delta}. \tag{24}$$

We therefore set $\delta = 4n \cdot \sqrt{V_\delta}$, whereby, for $\sqrt{V_\delta} \geq 4$, the event of (24) (with $t = 4$) happens with probability at least $1 - 1/48$. Hence, we set $V_\delta \doteq 16 \vee \mathbb{E}\left[\hat{N}_{P,+} + \hat{N}_{P,-}\right] \geq 16 \vee \text{Var}\left(\hat{N}_{P,+} - \hat{N}_{P,-}\right)$, where $\mathbb{E}\left[\hat{N}_{P,+} + \hat{N}_{P,-}\right] \doteq \mathbb{E}\left[\hat{N}_P\right]$ is given in equation (17). We now proceed with a lower-bound on $\mathbb{P}(\tilde{A}_\delta)$.

Let $\hat{n}_{x_1}$ denote the number of points sampled from P_- that fall on x_1. Notice that, conditioned on these samples' indices, $n_+ \doteq \tilde{n}_+ + n \cdot \hat{N}_{P,+}$ is distributed as Binomial $(\hat{n}_{x_1}, p)$, where $p = \eta_{P,-}(x_1)$, the probability of $+$ given that x_1 is sampled from P_-. Applying Lemma 3, and integrating over $\hat{N}_Q$, we have

$$\mathbb{P}\left(n_+ > \hat{n}_{x_1} \cdot p + \sqrt{\hat{n}_{x_1} \cdot p(1-p)}\right) \geq \frac{1}{12}. \tag{25}$$

Now notice that $\tilde{A}_\delta$ holds whenever $\hat{n}_+ \geq \frac{1}{2}(\hat{n}_{x_1} + \delta)$, since $\hat{n}_{x_1} = \left(\tilde{n}_+ + n \cdot \hat{N}_{P,+}\right) + \left(\tilde{n}_- + n \cdot \hat{N}_{P,-}\right)$. Under the event of (25), we have $\hat{n}_+ > \frac{1}{2}(\hat{n}_{x_1} + \delta)$, whenever it holds that

$$\sqrt{\hat{n}_{x_1} \cdot p(1-p)} \geq \frac{1}{2}(\hat{n}_{x_1} + \delta) - \hat{n}_{x_1} \cdot p = \hat{n}_{x_1}\left(\frac{1}{2} - p\right) + \frac{1}{2}\delta. \tag{26}$$

Next we bound $\hat{n}_{x_1}$ with high probability. Consider any value of $\hat{N}_Q$ such that $\mathbf{E}_Q$ (from Proposition 3) holds, i.e., $\hat{N}_Q \leq 2N_Q$. Conditioned on such $\hat{N}_Q$, $\hat{n}_{x_1}$ is itself a Binomial with

$$\mathbb{E}\left[\hat{n}_{x_1} \mid \hat{N}_Q\right] = n(N - \hat{N}_Q) \cdot P_X(x_1) \geq \frac{1}{2} nN \cdot P_X(x_1) \geq \frac{1}{2} n \cdot N_P \cdot P_X(x_1) = \frac{1}{2} c_1 (n \cdot N_P)^{\frac{2-2\beta}{2-\beta}},$$

where for the first inequality we used the fact that $N_P \geq 3N_Q$. Hence, by a multiplicative Chernoff bound,

$$\mathbb{P}\left(\frac{1}{2}\mathbb{E}\left[\hat{n}_{x_1}|\hat{N}_Q\right] \leq \hat{n}_{x_1} \leq 2\mathbb{E}\left[\hat{n}_{x_1}|\hat{N}_Q\right]\middle|\hat{N}_Q\right) \cdot \mathbb{1}\{\mathbf{E}_Q\} \geq \left(1 - \frac{1}{48}\right)\mathbb{1}\{\mathbf{E}_Q\}, \tag{27}$$

whenever $c_1 (n \cdot N_P)^{\frac{2-2\beta}{2-\beta}} \geq 40$. Now, by Proposition 3, $\mathbf{E}_Q$ holds with probability at least $1 - 1/48$ whenever $\mathbb{E}\left[\hat{N}_Q\right] = N_Q > 16$. Thus, integrating (27) over $\hat{N}_Q$, we get that, with probability at least $1 - 1/24$,

$$\frac{1}{4} n \cdot N_P \cdot P_X(x_1) \leq \hat{n}_{x_1} \leq 2n \cdot N \cdot P_X(x_1) \leq \frac{8}{3} n \cdot N_P \cdot P_X(x_1). \tag{28}$$

Thus, bounding both sides of (26), $\tilde{A}_\delta$ holds whenever a) the events of (25) and (28) hold, and b) the following inequality is satisfied:

$$\sqrt{\frac{1}{4} n \cdot N_P \cdot P_X(x_1) \cdot p(1-p)} \geq \frac{8}{3} n \cdot N_P \cdot P_X(x_1)\left(\frac{1}{2} - p\right) + 2n \cdot \sqrt{V_\delta}, \text{ which holds whenever}$$

$$\frac{1}{4\sqrt{2}} c_1^{1/2} (n \cdot N_P)^{\frac{1-\beta}{2-\beta}} \geq \frac{8}{3} c_1 (n \cdot N_P)^{\frac{1-\beta}{2-\beta}} + 2n \cdot \left(4 \vee n^{-1/2} c_1^{n/2} (n \cdot N_P)^{\frac{1-(n+1)\beta/2}{2-\beta}}\right), \tag{29}$$

where the r.h.s. and l.h.s. of (29) upper and lower bound, respectively, the r.h.s. and l.h.s. of the previous inequality (using the setting of V_δ and (17), and lower-bounding $p(1-p)$ by $1/8$ for N_P as large as assumed). By the conditions of the Proposition, (29) is satisfied. Thus, $\tilde{A}_\delta$ holds with probability at least $1/24$ since the events of (25) and (28) hold together with that probability (using the fact that $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) - \mathbb{P}(B^{\mathsf{c}})$). Finally, we can conclude that $\{\tilde{n}_+ > \tilde{n}_-\}$ with probability at least $1/48$ since $\tilde{A}_\delta$ and the event of (24) hold together with that probability. □

Finally we bound the $\mathcal{D}_\pm$ term in the likelihood equation (16).

Proposition 7 ($\mathcal{D}_+/\mathcal{D}_-$)**.** *Let $n_{\mathcal{D}} > 0$. Again let $Z \sim \Gamma_-$, and let $0 < c_0 \leq 1/4$ in the construction of $\eta_{\mathcal{D},\sigma}, \sigma = \pm$.*

$$\mathbb{P}\left(\frac{\mathcal{D}_+(Z_{N+1})}{\mathcal{D}_-(Z_{N+1})} \geq 1\right) \geq \frac{1}{84}.$$

The proof is given in Appendix D, and follows similar lines as above, namely, isolate sufficient statistics (number of $\pm$ in Z_{N+1}) and concluding by anticoncentration upon proper conditioning.

We can now combine all the above analysis into the following main proposition.

Proposition 8. *Pick any $0 \leq \beta < 1$, $1 \leq n < 2/\beta - 1$, and $N_Q \geq 16$. Let $0 < c_1 \leq 2^{-10}$, and $0 < c_0 \leq 1/4$ in the constructions of $\eta_{P,\sigma}$, $\eta_{\mathcal{D},\sigma}$, $\sigma = \pm$. Suppose N_P is sufficiently large so that $N_P \geq 3N_Q$, and also*

(i) $(n \cdot N_P)^{\frac{2-2\beta}{2-\beta}} \geq 4096 \cdot n^2 \cdot c_1^{-1}$.

(ii) $(n \cdot N_P)^{(2-(n+1)\beta)/(2-\beta)} \geq 4 \cdot N_Q^2 \cdot n \cdot 2^{4n} c_1^{-n}$.

(iii) $N_P^{(2-(n+1)\beta)/(2-\beta)} \geq 22 \cdot n \cdot 2^n \cdot c_1^{-n}$.

Let $\hat{h}$ denote any classification procedure having access to $Z \sim \Gamma_\sigma, \sigma = \pm$. We then have that

$$\inf_{\hat{h}} \sup_{\sigma \in \{\pm\}} \mathbb{P}_{\Gamma_\sigma}\left(\mathcal{E}_{\mathcal{D}}(\hat{h}) \geq c_0 \cdot \left(1 \wedge n_{\mathcal{D}}^{-1/(2-\beta)}\right)\right) \geq \frac{1}{12} \cdot \frac{1}{96} \cdot \frac{1}{84}.$$

Proof. Following the above propositions, again assume w.l.o.g. that $Z \sim \Gamma_-$. Let $\mathbf{E}_P, \mathbf{E}_Q$ as defined in Proposition 3 over $Z \sim \Gamma_-$, and notice that, under our assumptions on N_Q and *(iii) on N_P, each of these events occurs with probability at least* $1 - 1/(2 \cdot 96)$.

Thus, for $n > 1$, by Propositions 2, 4, and 7, we have that $\mathbb{P}_{\Gamma_-}(\Gamma_+(Z) > \Gamma_-(Z))$ is at least $\delta_1(\delta_2 - 1/96)\frac{1}{84}$. Now plug in $\delta_1 = 1/12$ and $\delta_2 = 1/48$ from Propositions 5 and 6. For $n = 1$, using Proposition 2 and 7, and noticing that $\{\hat{N}_+(Z) > \hat{N}_-(Z)\} \implies \{\hat{n}_+(Z) \geq \hat{n}_-(Z)\}$, we can conclude by Corollary 1 that $\mathbb{P}_{\Gamma_-}(\Gamma_+(Z) > \Gamma_-(Z))$ is at least $\frac{1}{12}\mathbb{P}(\mathbf{E}_P \cap \mathbf{E}_Q) \cdot \frac{1}{84}$, again matching the lower-bound in the statement.

Now, if $\hat{h}$ wrongly picks $\sigma = +$ (i.e. picks $h \in \mathcal{H}$, $h(x_1) = +$), then $\mathcal{E}_{\mathcal{D}}(\hat{h}) \geq \mathcal{D}_X(x_1) \cdot (\eta_{\mathcal{D},+} - \eta_{\mathcal{D},-}) = c_0\epsilon_0$. By Remark 8, for any $\hat{h}$, the probability that $\hat{h}$ picks $\sigma = +$ is bounded below by $\mathbb{P}_{\Gamma_-}(\Gamma_+(Z) > \Gamma_-(Z))$. □

We can now conclude with the proof of the main result of this section.

Proof of Theorem 8. The first part of the statement builds on Proposition 8 as follows. Set $c_0 = 1/4$ and $c_1 = 2^{-10}$. First, let $Z \sim \Gamma_\sigma$, and let $\hat{N}_Q$ denote the number of vectors $Z_t \in Z$ that were generated by Q (as in Proposition 3). Let $\mathbf{E}_{Q,\sharp}$ denote the event that $\hat{N}_Q \geq \mathbb{E}[\hat{N}_Q]/2 = N_Q/2$, Let $\mathbf{E}_{\mathcal{E}} \doteq \left\{\mathcal{E}_{\mathcal{D}}(\hat{h}) \geq c_0 \cdot \epsilon_0\right\}$. By Proposition 8, for some $\sigma \in \{\pm\}$, we have that $\mathbb{P}(\mathbf{E}_{\mathcal{E}})$ is bounded below.

Now decouple the randomness in Z as follows. Let $\zeta \doteq \{\zeta_t\}_{t \in [N]}$ denote N i.i.d. choices of P_σ or Q_σ with respective probabilities $\alpha_P = N_P/N$ and $\alpha_Q = N_Q/N$; choose $Z_t \in Z$ according to ζ_t^n. We then have that

$$\begin{aligned}\mathbb{E}\left[\mathbb{P}\left(\mathbf{E}_{\mathcal{E}} \mid \zeta\right) \mid \mathbf{E}_{Q,\sharp}\right] &\geq \mathbb{E}\left[\mathbb{P}\left(\mathbf{E}_{\mathcal{E}} \mid \zeta\right) \cdot \mathbb{1}\{\mathbf{E}_{Q,\sharp}\}\right] \\ &\geq \mathbb{E}\left[\mathbb{P}\left(\mathbf{E}_{\mathcal{E}} \mid \zeta\right)\right] - \mathbb{P}\left(\mathbf{E}_{Q,\sharp}^{\mathsf{c}}\right) = \mathbb{P}\left(\mathbf{E}_{\mathcal{E}}\right) - \mathbb{P}\left(\mathbf{E}_{Q,\sharp}^{\mathsf{c}}\right) \geq c,\end{aligned}$$

where we can bound $\mathbb{P}\left(\mathbf{E}_{Q,\sharp}^{\mathsf{c}}\right)$ how ever small for N_Q sufficiently large (by a multiplicative Chernoff). Now conclude by noticing that the above conditional expectation is a projection of the measure α^N onto $\mathcal{M}$ (via the injection $\zeta \mapsto \Pi \in \mathcal{M}$) and is bounded below, implying $\sup_{\zeta | \mathbf{E}_{Q,\sharp}} \mathbb{P}\left(\mathbf{E}_{\mathcal{E}} \mid \zeta\right)$ must be bounded below.

The second part of the theorem is a direct consequence of the results of Section 7. □

References

[ABGLP19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

[AKK+19] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv:1902.09229*, 2019.

[APMS19] Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. The information complexity of learning tasks, their structure and their distance. *arXiv:1904.03292*, 2019.

[AZ05] 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.

[Bar92] Peter L Bartlett. Learning with a slowly changing distribution. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, 1992.

[Bax97] Jonathan Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39, 1997.

[BDB08] Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine Learning*, 73(3):273–287, 2008.

[BDBC+10] 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.

[BDBCP07] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 2007.

[BDBM89] Shai Ben-David, Gyora M Benedek, and Yishay Mansour. A parametrization scheme for classifying models of learnability. In *Proceedings of the 2nd Annual Workshop on Computational Learning Theory*, 1989.

[BDLLP10] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

[BHPQ17] Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative PAC learning. In *Advances in Neural Information Processing Systems*, 2017.

[BL97] Rakesh D Barve and Philip M Long. On the complexity of learning from drifting distributions. *Information and Computation*, 138(2):170–193, 1997.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

[Car97] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[CKW08] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

[CMRR08] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, 2008.

[DHK+20] 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.

[GSH+09] 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.

[HK19] Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, 2019.

[HY19] Steve Hanneke and Liu Yang. Statistical learning under nonstationary mixing processes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[JSRR10] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, 2010.

[KFAL20] Nikola Konstantinov, Elias Frantar, Dan Alistarh, and Christoph H Lampert. On the sample complexity of adversarial multi-source PAC learning. *arXiv:2002.10384*, 2020.

[KM18] Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. *arXiv:1803.01833*, 2018.

[Kol06] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[Kol11] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. 2011.

[KR05] Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[LPVDG+11] 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.

[MB17] Daniel McNamara and Maria-Florina Balcan. Risk bounds for transferring representations with and without fine-tuning. In *International Conference on Machine Learning*, 2017.

[MBS13] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 2013.

[MM12] Mehryar Mohri and Andres Munoz Medina. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, 2012.

[MMM19] Saeed Mahloujifar, Mohammad Mahmoody, and Ameer Mohammed. Universal multi-party poisoning attacks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[MMR09] 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*, 2009.

[Mou10] Nima Mousavi. How tight is Chernoff bound? `https://ece.uwaterloo.ca/~nmousavi/Papers/Chernoff-Tightness.pdf`, 2010.

[MPRP13] Andreas Maurer, Massi Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *International Conference on Machine Learning*, 2013.

[MPRP16] 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.

[NW11] 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.

[PL14] Anastasia Pentina and Christoph Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

[PM13] Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, 2013.

[Qia18] Mingda Qiao. Do outliers ruin collaboration? *arXiv:1805.04720*, 2018.

[RHS17] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

[Sau72] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[Slu77] Eric V Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, pages 404–412, 1977.

[SQZY18] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *32nd AAAI Conference on Artificial Intelligence*, 2018.

[SZ19] Clayton Scott and Jianxin Zhang. Learning from multiple corrupted sources, with application to learning from label proportions. *arXiv:1910.04665*, 2019.

[Tat53] Robert F Tate. On a double inequality of the normal distribution. *The Annals of Mathematical Statistics*, 24(1):132–134, 1953.

[TJJ20] Nilesh Tripuraneni, Michael I Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *arXiv:2006.11650*, 2020.

[Tsy04] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[Tsy09] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[VC71] 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.

[vW96] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag New York, 1996.

[YHC13] Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90(2):161–189, 2013.

[ZL19] Alexander Zimin and Christoph H Lampert. Tasks without borders: A new approach to online multitask learning. In *Workshop on Adaptive & Multitask Learning*, 2019.

Appendix

A Proof of Lemma 1

Let $W_{1:m} = (W_1, \dots, W_m)$ be a vector of independent $\mathcal{W}$-valued random variables (for some space $\mathcal{W}$), not necessarily identically distributed. Let $\mathcal{F}$ be a set of measurable functions $\mathcal{W} \to [-1, 1]$, and let $|\mathcal{F}(m)|$ be the number of distinct vectors possible on m points. Let $\alpha = \{\alpha_1, \dots, \alpha_m\} \in [0, 1]^m$. Define $\hat{\mu}_\alpha(f) \doteq \sum_{i=1}^m \alpha_i f(W_i)$, and $\mu_\alpha(f) \doteq \mathbb{E}[\hat{\mu}_\alpha(f)]$, and also $\hat{\sigma}^2_\alpha(f) \doteq \sum_{i=1}^m \alpha_i^2 f^2(W_i)$, and $\sigma^2_\alpha(f) \doteq \mathbb{E}\hat{\sigma}^2_\alpha(f)$. Define $\mathcal{F}_\sigma \doteq \{f \in \mathcal{F} : \sigma_\alpha(f) \le \sigma\}$.

The following lemma is immediate from a result of [KR05] (see also [BLM13] Section 12.5).

Lemma 4. *For any $\sigma > 0$ with $\mathcal{F}_\sigma \neq \emptyset$, letting $L_\sigma \doteq \sup_{f \in \mathcal{F}_\sigma} (\hat{\mu}_\alpha(f) - \mu_\alpha(f))$, $\forall \epsilon > 0$,*

$$\mathbb{P}(L_\sigma \ge \mathbb{E}[L_\sigma] + 2\epsilon) \le \exp\left\{-\frac{\epsilon}{4} \ln\left(1 + 2\ln\left(1 + \frac{\epsilon}{2\mathbb{E}[L_\sigma] + \sigma^2}\right)\right)\right\}.$$

In particular, based on the inequality $\ln(1 + x) \ge \frac{x}{1+x}$, this implies

$$\mathbb{P}(L_\sigma \ge \mathbb{E}[L_\sigma] + 2\epsilon) \le \exp\left\{-\frac{\epsilon^2}{6\epsilon + 4\mathbb{E}[L_\sigma] + 2\sigma^2}\right\}.$$

In particular, for any $\delta > 0$, setting

$$\epsilon = 12 \max\left\{\sqrt{(\mathbb{E}[L_\sigma] + \sigma^2) \ln\left(\frac{1}{\delta}\right)}, \ln\left(\frac{1}{\delta}\right)\right\}$$

reveals that, with probability at least $1 - \delta$,

$$L_\sigma \le \mathbb{E}[L_\sigma] + 24 \max\left\{\sqrt{(\mathbb{E}[L_\sigma] + \sigma^2) \ln\left(\frac{1}{\delta}\right)}, \ln\left(\frac{1}{\delta}\right)\right\}. \tag{30}$$

Next, the following lemma bounds $\mathbb{E}[L_\sigma]$ using a standard route.

Lemma 5. *There is a numerical constant $C \ge 1$ such that, for any $\sigma > 0$ with $\mathcal{F}_\sigma \neq \emptyset$, for L_σ as in Lemma 4,*

$$\mathbb{E}[L_\sigma] \le C\sqrt{\sigma^2 \ln(|\mathcal{F}(m)|)} + C \ln(|\mathcal{F}(m)|).$$

Proof. From Lemma 11.4 of [BLM13], for $\epsilon_1, \dots, \epsilon_m$ independent $\mathrm{Uniform}(\{-1, 1\})$ (and independent of $W_{1:m}$),

$$\mathbb{E}[L_\sigma] \le 2\mathbb{E}\left[\sup_{f \in \mathcal{F}_\sigma} \sum_{i=1}^m \epsilon_i \alpha_i \left(f(W_i) - \mathbb{E}[f(W_i)]\right)\right].$$

Furthermore, from [BLM13] (Exercise 13.2, based on a result proved in [vW96]; see also [Kol11]), there is a numerical constant $C > 0$ such that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_\sigma} \sum_{i=1}^m \epsilon_i \alpha_i \left(f(W_i) - \mathbb{E}[f(W_i)]\right)\right] \le C\sigma\sqrt{\log(|\mathcal{F}(m)|)} + C \log(|\mathcal{F}(m)|).$$

□

In particular, the above results imply the following lemma.

Lemma 6. *There exists a numerical constant $C \ge 1$ such that, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, every $f \in \mathcal{F}$ satisfies*

$$\hat{\mu}_\alpha(f) \le \mu_\alpha(f) + C\sqrt{\sigma^2_\alpha(f) \ln\left(\frac{|\mathcal{F}(m)| \log_2(2m)}{\delta}\right)} + C \ln\left(\frac{|\mathcal{F}(m)| \log_2(2m)}{\delta}\right).$$

Proof. Let $\delta_m \doteq \frac{\delta}{\log_2(2m)}$. Combining Lemma 5 with (30) implies that, letting $\sigma_k^2 = 2^k$ ($k \in \{0, \dots, \log_2(m)\}$), with probability at least $1-\delta$ (by a union bound), every $k \in \{0, \dots, \log_2(m)\}$ has (for a numerical constant $C \geq 1$)

$$\begin{aligned}
L_{\sigma_k} &\leq 24 \max\left\{ \sqrt{(\mathbb{E}[L_{\sigma_k}] + \sigma_k^2) \ln\left(\frac{1}{\delta_m}\right)}, \ln\left(\frac{1}{\delta_m}\right) \right\} + C\sqrt{\sigma_k^2 \ln(|\mathcal{F}(m)|)} + C\ln(|\mathcal{F}(m)|) \\
&\leq 24\sqrt{(\mathbb{E}[L_{\sigma_k}] + \sigma_k^2) \ln\left(\frac{1}{\delta_m}\right)} + C\sqrt{\sigma_k^2 \ln(|\mathcal{F}(m)|)} + (24+C)\ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right) \\
&\leq 24\sqrt{2\sigma_k^2 \ln\left(\frac{1}{\delta_m}\right)} + 48\sqrt{C\sqrt{\sigma_k^2 \ln(|\mathcal{F}(m)|)} \ln\left(\frac{1}{\delta_m}\right)} + 48\sqrt{C\ln(|\mathcal{F}(m)|) \ln\left(\frac{1}{\delta_m}\right)} \\
&\quad + C\sqrt{\sigma_k^2 \ln(|\mathcal{F}(m)|)} + (24+C)\ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right) \\
&\leq (72 + 2C + 48\sqrt{C}) \left(\sqrt{\sigma_k^2 \ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right)} + \ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right) \right).
\end{aligned}$$

Suppose this event occurs. In particular, for each $f \in \mathcal{F}$, let $k(f) \doteq \min\{k : f \in \mathcal{F}_{\sigma_k}\}$. Then every $f \in \mathcal{F}$ has (for a universal numerical constant $C' \geq 1$)

$$\begin{aligned}
\hat{\mu}_\alpha(f) &\leq \mu_\alpha(f) + C'\sqrt{\sigma_{k(f)}^2 \ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right)} + C'\ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right) \\
&\leq \mu_\alpha(f) + C'\sqrt{\max\{1, 2\sigma_\alpha^2(f)\} \ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right)} + C'\ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right) \\
&\leq \mu_\alpha(f) + C'\sqrt{2\sigma_\alpha^2(f) \ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right)} + 2C'\ln\left(\frac{|\mathcal{F}(m)|}{\delta_m}\right).
\end{aligned}$$

□

We can also state a concentration result specific to the $\hat{\sigma}_\alpha^2(f)$ values, as follows.

Lemma 7. *There exists a numerical constant $C \geq 1$ such that, for any $\delta \in (0,1)$, with probability at least $1-\delta$, every $f \in \mathcal{F}$ satisfies*

$$\frac{1}{2}\sigma_\alpha^2(f) - C\ln\left(\frac{|\mathcal{F}(m)|\log_2(2m)}{\delta}\right) \leq \hat{\sigma}_\alpha^2(f) \leq 2\sigma_\alpha^2(f) + C\ln\left(\frac{|\mathcal{F}(m)|\log_2(2m)}{\delta}\right).$$

Proof. Define $\mathcal{F}' \doteq \{f^2 : f \in \mathcal{F}\}$ and note that $|\mathcal{F}'(m)| \leq |\mathcal{F}(m)|$. Also define $\alpha_i' \doteq \alpha_i^2$. Applying Lemma 6 with this $\mathcal{F}'$ and α', we have that, with probability at least $1-\delta/2$, every $f \in \mathcal{F}$ satisfies (for some numerical constant $C \geq 1$)

$$\begin{aligned}
\hat{\sigma}_\alpha^2(f) &\leq \sigma_\alpha^2(f) + C\sqrt{\mathbb{E}\left[\sum_{i=1}^m \alpha_i^4 f^4(W_i)\right] \ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)} + C\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right) \\
&\leq \sigma_\alpha^2(f) + C\sqrt{\sigma_\alpha^2(f) \ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)} + C\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right). \qquad (31)
\end{aligned}$$

If $\sigma_\alpha^2(f) \geq C^2 \ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)$ then the expression in (31) is at most

$$2\sigma_\alpha^2(f) + C\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right),$$

while if $\sigma_\alpha^2(f) \leq C^2 \ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)$ then the expression in (31) is at most

$$\sigma_\alpha^2(f) + (C^2+C)\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right).$$

Thus, either way we have

$$\hat{\sigma}_\alpha^2(f) \le 2\sigma_\alpha^2(f) + (C^2+C)\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right).$$

On the other hand, consider the set $\mathcal{F}'' \doteq \{-f^2 : f \in \mathcal{F}\}$ and note that again we have $|\mathcal{F}''(m)| \le |\mathcal{F}(m)|$. Applying Lemma 6 with this $\mathcal{F}''$ and α', we have that, with probability at least $1-\delta/2$, every $f \in \mathcal{F}$ satisfies (for some numerical constant $C \ge 1$)

$$\begin{aligned}-\hat{\sigma}_\alpha^2(f) &\le -\sigma_\alpha^2(f) + C\sqrt{\mathbb{E}\left[\sum_{i=1}^m \alpha_i^4 f^4(W_i)\right]\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)} + C\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)\\ &\le -\sigma_\alpha^2(f) + C\sqrt{\sigma_\alpha^2(f)\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)} + C\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right). \end{aligned} \tag{32}$$

If $\sigma_\alpha^2(f) \ge 4C^2\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)$ then the expression in (32) is at most

$$-\frac{1}{2}\sigma_\alpha^2(f) + C\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right),$$

while if $\sigma_\alpha^2(f) \le 4C^2\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right)$ then the expression in (32) is at most

$$-\sigma_\alpha^2(f) + (2C^2+C)\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right).$$

Thus, either way we have

$$-\hat{\sigma}_\alpha^2(f) \le -\frac{1}{2}\sigma_\alpha^2(f) + (2C^2+C)\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right),$$

which implies

$$\hat{\sigma}_\alpha^2(f) \ge \frac{1}{2}\sigma_\alpha^2(f) - (2C^2+C)\ln\left(\frac{|\mathcal{F}(m)|2\log_2(2m)}{\delta}\right).$$

The lemma now follows by a union bound, so that these two events (each of probability at least $1-\delta/2$) occur simultaneously, with probability at least $1-\delta$. □

We will now show that Lemma 1 follows directly from Lemmas 6 and 7.

Proof of Lemma 1. Set $\mathcal{W} = \mathcal{X}\times\mathcal{Y}$, $W_i = (X_i, Y_i)$, and $\alpha_i = 1$. For each $h, h' \in \mathcal{H}$, define $f_{h,h'}(x,y) \doteq \mathbb{1}\{h'(x)\ne y\} - \mathbb{1}\{h(x)\ne y\}$. Note that $\hat{\sigma}_\alpha^2(f_{h,h'}) = m\hat{\mathbb{P}}_S(h \ne h')$. $\mathcal{F} \doteq \{f_{h,h'} : h, h' \in \mathcal{H}\}$ and note that $|\mathcal{F}(m)| \le |\mathcal{H}(m)|^2$. Applying Lemma 6 with this $\mathcal{F}$, we have that with probability at least $1-\delta/2$, every $h, h' \in \mathcal{H}$ satisfy (for some universal numerical constant $C \ge 1$)

$$\hat{\mathcal{E}}_S(h';h) \le \mathbb{E}\left[\hat{\mathcal{E}}_S(h';h)\right] + C\sqrt{\mathbb{E}\left[\hat{\mathbb{P}}_S(h\ne h')\right]\frac{1}{m}\ln\left(\frac{2|\mathcal{H}(m)|^2\log_2(2m)}{\delta}\right)} + \frac{C}{m}\ln\left(\frac{2|\mathcal{H}(m)|^2\log_2(2m)}{\delta}\right). \tag{33}$$

Furthermore, Lemma 7 implies that, with probability at least $1-\delta/2$, every $h, h' \in \mathcal{H}$ satisfy (for some universal numerical constant $C' \ge 1$)

$$\frac{1}{2}\mathbb{E}\left[\hat{\mathbb{P}}_S(h\ne h')\right] - \frac{C'}{m}\ln\left(\frac{2|\mathcal{H}(m)|^2\log_2(2m)}{\delta}\right) \le \hat{\mathbb{P}}_S(h\ne h') \le 2\mathbb{E}\left[\hat{\mathbb{P}}_S(h\ne h')\right] + \frac{C'}{m}\ln\left(\frac{2|\mathcal{H}(m)|^2\log_2(2m)}{\delta}\right). \tag{34}$$

By a union bound, with probability at least $1-\delta$, every $h, h' \in \mathcal{H}$ satisfy both of (33) and (34). In particular, combining (33) with the left inequality in (34), this also implies every $h, h' \in \mathcal{H}$ satisfy

$$\begin{aligned}\hat{\mathcal{E}}_S(h';h) \le \mathbb{E}\left[\hat{\mathcal{E}}_S(h';h)\right] &+ 2C\sqrt{\hat{\mathbb{P}}_S(h\ne h')\frac{1}{m}\ln\left(\frac{2|\mathcal{H}(m)|^2\log_2(2m)}{\delta}\right)}\\ &+ \left(2\sqrt{C'}+1\right)C\frac{1}{m}\ln\left(\frac{2|\mathcal{H}(m)|^2\log_2(2m)}{\delta}\right).\end{aligned} \tag{35}$$

Finally, recall from Sauer's lemma [VC71, Sau72] that $|\mathcal{H}(m)| \leq \left(\frac{e\max\{m,d_{\mathcal{H}}\}}{d_{\mathcal{H}}}\right)^{d_{\mathcal{H}}}$, and therefore

$$\begin{aligned}\frac{1}{m}\ln\left(\frac{2|\mathcal{H}(m)|^2\log_2(2m)}{\delta}\right) &\leq \frac{1}{m}\ln\left(\left(\frac{e\max\{m,d_{\mathcal{H}}\}}{d_{\mathcal{H}}}\right)^{2d_{\mathcal{H}}}\frac{2\log_2(2m)}{\delta}\right)\\ &\leq \frac{1}{m}\ln\left(\left(\frac{e\max\{m,d_{\mathcal{H}}\}}{d_{\mathcal{H}}}\right)^{3d_{\mathcal{H}}}\frac{4}{\delta}\right) \leq C''\varepsilon(m,\delta)\end{aligned}$$

for some numerical constant $C'' \geq 1$.

Altogether (and subtracting $\mathbb{E}\left[\hat{\mathcal{E}}_S(h';h)\right]$ and $\hat{\mathcal{E}}_S(h';h)$ from both sides of (33) and (35)), we have that with probability at least $1-\delta$, every $h, h' \in \mathcal{H}$ satisfy

$$\mathbb{E}\left[\hat{\mathcal{E}}_S(h;h')\right] \leq \hat{\mathcal{E}}_S(h;h') + 2C\sqrt{C''}\sqrt{\min\left\{\mathbb{E}\left[\hat{\mathbb{P}}_S(h\neq h')\right], \hat{\mathbb{P}}_S(h\neq h')\right\}\varepsilon(m,\delta)} + \left(2\sqrt{C'}+1\right)CC''\varepsilon(m,\delta)$$

and

$$\frac{1}{2}\mathbb{E}\left[\hat{\mathbb{P}}_S(h\neq h')\right] - C'C''\varepsilon(m,\delta) \leq \hat{\mathbb{P}}_S(h\neq h') \leq 2\mathbb{E}\left[\hat{\mathbb{P}}_S(h\neq h')\right] + C'C''\varepsilon(m,\delta),$$

which completes the proof. □

B Pooling is Optimal if Enough Tasks are Good

While our results in Sections 4.3 and 8 imply that, in general, one cannot achieve optimal rates by simply pooling all of the data and using the global ERM $\hat{h}_Z$, in this section we find that in some special cases this naive approach can actually be successful: namely, cases where *most* of the tasks have ρ_t below the cut-off value $\rho_{(t^*)}$ chosen by the optimization in the optimal proceedure from Section 6. We in fact show a general result for pooling, arguing that it always achieves a rate depending on the (weighted) *median* value of $\bar{\rho}_t$, or more generally any *quantile* of $\bar{\rho}_t$ values.

Theorem 9 (Pooling Beyond $\beta = 1$)**.** *For any $\alpha \in (0,1]$, let $t(\alpha)$ be the smallest value in $[N+1]$ such that $\sum_{t\in[t(\alpha)]} n_{(t)} \geq \alpha\sum_{t=1}^{N+1} n_t$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\mathcal{E}_{\mathcal{D}}\left(\hat{h}_Z\right) \leq C_\rho\left(C\frac{d_{\mathcal{H}}\log\left(\frac{1}{d_{\mathcal{H}}}\sum_{t=1}^{N+1}n_t\right)+\log(1/\delta)}{\sum_{t=1}^{N+1}n_t}\right)^{1/(2-\beta)\bar{\rho}_{t(\alpha)}}$$

for a constant $C = (32C_0^2/\alpha)^{2-\beta}C_\beta$.

Proof. Letting $\mathbf{n}_{[N+1]} \doteq \sum_{t=1}^{N+1} n_t$, $\bar{P} \doteq \mathbf{n}_{[N+1]}^{-1}\sum_{t=1}^{N+1} n_t P_t$, and $\mathbf{n}_\alpha \doteq \sum_{t\in[t(\alpha)]} n_{(t)}$, we have

$$\begin{aligned}\mathcal{E}_{\bar{P}}(\hat{h}_Z) &\geq \mathbf{n}_{[N+1]}^{-1}\sum_{t=1}^{N+1} n_t\left(C_\rho^{-1}\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)\right)^{\rho_t}\\ &\geq \frac{\mathbf{n}_\alpha}{\mathbf{n}_{[N+1]}}\sum_{t\in[t(\alpha)]}\frac{n_{(t)}}{\mathbf{n}_\alpha}\left(C_\rho^{-1}\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)\right)^{\rho_{(t)}} \geq \alpha\left(C_\rho^{-1}\mathcal{E}_{\mathcal{D}}(\hat{h}_Z)\right)^{\bar{\rho}_{t(\alpha)}},\end{aligned}$$

where the third inequality is due to Jensen's inequality. This implies

$$\mathcal{E}_{\mathcal{D}}(\hat{h}_Z) \leq C_\rho\left((1/\alpha)\mathcal{E}_{\bar{P}}(\hat{h}_Z)\right)^{1/\bar{\rho}_{t(\alpha)}}. \tag{36}$$

Lemma 2 implies that with probability at least $1-\delta$,

$$\mathcal{E}_{\bar{P}}(\hat{h}_Z) \leq 32C_0^2(C_\beta\varepsilon(\mathbf{n}_{[N+1]},\delta))^{1/(2-\beta)}.$$

Combining this with (36) completes the proof. □

For instance, if all n_t are equal some common value n, and we take $\alpha = 1/2$, then we can take $t(\alpha) = \lceil (N+1)/2\rceil$, so that the optimal rate will be achieved as long as at least half of the tasks have $\bar{\rho}_t$ below the value $\bar{\rho}_{t^*}$ for t^* the minimizer of the bound in Theorem 1.

Optimizing the bound in Theorem 9 over the choice of α yields the following result.

Corollary 2 (General Pooling Bound)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\mathcal{E}_{\mathcal{D}}\left(\hat{h}_Z\right) \leq \min_{t\in[N+1]} C_\rho \left(C \frac{d_{\mathcal{H}} \log\left(\frac{1}{d_{\mathcal{H}}}\sum_{s=1}^{N+1} n_s\right) + \log(1/\delta)}{\left(\sum_{s=1}^{t} n_{(s)}\right)^{2-\beta}\left(\sum_{s=1}^{N+1} n_s\right)^{-(1-\beta)}} \right)^{1/(2-\beta)\bar{\rho}_t}$$

for a constant $C = (32C_0^2)^{2-\beta}C_\beta$.

Remark 9. In particular, note that this result recovers the bound of Theorem 3 in the special case of $\beta = 1$.

C Different Optimal Aggregations of Tasks in Multitask

We present a proof of Theorem 2 in this section. Recall that this result states that different choices of target in the same multitask setting can induce different optimal aggregation of tasks. As a consequence, the naive approach of pooling of all tasks can adversely affect target risk even when all h^*'s are the same.

We employ a similar construction to that of Section 8.

Setup. We again build on distributions supported on 2 datapoints x_0, x_1. W.l.o.g., assume that x_0 has label 1. Let $n_p, n_q \geq 1$, $0 \leq \beta < 1$, and define $\epsilon \doteq n_P^{-1/(2-\beta)}$. Let $\sigma \in \{\pm 1\}$ – which we will often abbreviate as $\pm$. In all that follows, we let $\eta_\mu(X)$ denote the regression function $\mathbb{P}_\mu[Y = 1 \mid X]$ under distribution μ.

- **Target $\mathcal{D}_\sigma = \mathcal{D}_X \times \mathcal{D}^\sigma_{Y|X}$:** Let $\mathcal{D}_X(x_1) = 1/2$, $\mathcal{D}_X(x_0) = 1/2$; finally $\mathcal{D}^\sigma_{Y|X}$ is determined by $\eta_{\mathcal{D},\sigma}(x_1) = 1/2 + \sigma\cdot(1/4)$, and $\eta_{\mathcal{D},\sigma}(x_0) = 1$.
- **Source $P_\sigma = P_X \times P^\sigma_{Y|X}$:** Let $P_X(x_1) = \epsilon^\beta$, $P_X(x_0) = 1-\epsilon^\beta$; finally $P^\sigma_{Y|X}$ is determined by $\eta_{P,\sigma}(x_1) = 1/2 + \sigma\cdot c_2\epsilon^{1-\beta}$, and $\eta_{P,\sigma}(x_0) = 1$, for an appropriate constant c_1 specified in the proof.

Proof of Theorem 2. Let $P \doteq P_-$ and $\mathcal{D} \doteq \mathcal{D}_-$. The above construction then ensures that any h s.t. $h(x_1) = +1$ has excess error $\mathcal{E}_{\mathcal{D}}(h) = 1/4$. Thus we just have to show that number of $+1$ labels at x_1 exceeds the number of -1 labels at x_1 with non-zero probability (so that both $\hat{h}_Z, \hat{h}_{Z_P}$ would pick $+1$ at x_1). Let $\hat{n}_{P,1}$ denote the number of samples from P at x_1, and $\hat{n}^+_{P,1}$ denote those samples from P having label $+1$. Notice that if $\hat{n}^+_{P,1} > \frac{1}{2}\left(\hat{n}_{P,1} + n_{\mathcal{D}}\right)$, then $+1$ necessarily dominates -1 at x_1.

Now, conditioned on $\hat{n}_{P,1}$, $\hat{n}^+_{P,1}$ is distributed as Binomial$(\hat{n}_{P,1}, p)$ with $p = \eta_{P,-}(x_1)$. Applying Lemma 3, we have

$$\mathbb{P}\left(\hat{n}^+_{P,1} > n_{P,1}\cdot p + \sqrt{\hat{n}_{P,1}\cdot p(1-p)}\right) \geq \frac{1}{12}.$$

In other words, under the above binomial event, $+1$ dominates whenever (the second inequality below holds)

$$\sqrt{\hat{n}_{P,1}\cdot p(1-p)} \geq \frac{1}{4}\sqrt{\hat{n}_{P,1}} \geq \hat{n}_{P,1}\left(\frac{1}{2} - p\right) + \frac{1}{2}n_{\mathcal{D}},$$

in other words, if we have both $\hat{n}_{P,1}\cdot c_1^2\cdot\epsilon^{2(1-\beta)} \leq \frac{1}{64}$ and $n_{\mathcal{D}}^2 \leq \frac{1}{4}\hat{n}_{P,1}$. Let $\mathbf{E}_P \doteq \left\{\frac{1}{2}\mathbb{E}\left[\hat{n}_{P,1}\right] \leq \hat{n}_{P,1} \leq 2\mathbb{E}\left[\hat{n}_{P,1}\right]\right\}$, where $\mathbb{E}\left[\hat{n}_{P,1}\right] = n_P\cdot P_X(x_1) = n_P\cdot\epsilon^\beta$. Under this event, we just need that $2c_1^2\cdot n_P\cdot\epsilon^{(2-\beta)} = 2c_1^2 \leq \frac{1}{64}$, and that $n_{\mathcal{D}}^2 \leq \frac{1}{8}n_P\cdot\epsilon^\beta = \frac{1}{8}n_P^{(2-2\beta)/(2-\beta)}$, requiring $\beta < 1$. Hence, integrating over $\hat{n}_{P,1}$, we have that

$$\mathbb{P}\left(\hat{n}^+_{P,1} > \frac{1}{2}\left(\hat{n}_{P,1} + n_{\mathcal{D}}\right)\right) \geq \frac{1}{12}\mathbb{P}\left(\mathbf{E}_P\right) > 0,$$

where $\mathbb{P}\left(\mathbf{E}_P\right)$ is bounded below by a multiplicative Chernoff whenever $\mathbb{E}\left[\hat{n}_{P,1}\right] = n_P^{(2-2\beta)/(2-\beta)} \geq 1$. □

D Supporting Results for Section 8 on Impossibility of Adaptivity

Recall that from our construction, we set

$$\Gamma_\sigma = \left(\alpha_P P_\sigma^n + \alpha_Q Q_\sigma^n\right)^N \times \mathcal{D}_\sigma^{n_{\mathcal{D}}}.$$

Proof of Proposition 2. Let $\Gamma_{\sigma,\alpha} \doteq \alpha_P P_\sigma^n + \alpha_Q Q_\sigma^n$. Clearly

$$\frac{\Gamma_+(Z)}{\Gamma_-(Z)} = \frac{\mathcal{D}_+(Z_{N+1})}{\mathcal{D}_-(Z_{N+1})} \cdot \prod_{t=1}^{N} \frac{\Gamma_{+,\alpha}(Z_t)}{\Gamma_{-,\alpha}(Z_t)}.$$

Define $\hat{n}_{t,0}$, $\hat{n}_{t,+}$, and $\hat{n}_{t,-}$ as the number of points $(X_{t,i}, i \in [n])$ in Z_t which, respectively fall on x_0, or fall on x_1 with $Y_{t,i} = +1$, or fall on x_1 with $Y_{t,i} = -1$. Recall that $\hat{n}_\sigma = \sum_{t\in[N]} \hat{n}_{t,\sigma}$ for $\sigma = \pm$. Next, define $\mathcal{Z}_\sigma \doteq \{Z_t, t \in [N] : \hat{n}_{t,\sigma} = n\}$, i.e., the set of σ-homogeneous vectors. We have:

$$\Gamma_{+,\alpha}(Z_t) = \alpha_P \cdot \left(\eta_{P,-}^{\hat{n}_{t,-}}(x_1) \cdot \eta_{P,+}^{\hat{n}_{t,+}}(x_1) \cdot P_X^{(\hat{n}_{t,+}+\hat{n}_{t,-})}(x_1) \cdot P_X^{\hat{n}_{t,0}}(x_0)\right) + \alpha_Q \cdot \mathbb{1}\{Z_t \in \mathcal{Z}_+\},$$
$$\Gamma_{-,\alpha}(Z_t) = \alpha_P \cdot \left(\eta_{P,-}^{\hat{n}_{t,+}}(x_1) \cdot \eta_{P,+}^{\hat{n}_{t,-}}(x_1) \cdot P_X^{(\hat{n}_{t,+}+\hat{n}_{t,-})}(x_1) \cdot P_X^{\hat{n}_{t,0}}(x_0)\right) + \alpha_Q \cdot \mathbb{1}\{Z_t \in \mathcal{Z}_-\}.$$

We can then consider homogeneous and non-homogeneous vectors separately. Let $\mathcal{Z}_\pm \doteq \mathcal{Z}_+ \cup \mathcal{Z}_-$,

$$\begin{aligned}\prod_{Z_t\in\mathcal{Z}_\pm} \frac{\Gamma_{+,\alpha}(Z_t)}{\Gamma_{-,\alpha}(Z_t)} &= \frac{\alpha_P^{\hat{N}_-}\left(\eta_{P,-}(x_1)\cdot P_X(x_1)\right)^{n\cdot\hat{N}_-}}{\left(\alpha_P\left(\eta_{P,+}(x_1)\cdot P_X(x_1)\right)^n + \alpha_Q\right)^{\hat{N}_-}} \cdot \frac{\left(\alpha_P\left(\eta_{P,+}(x_1)\cdot P_X(x_1)\right)^n + \alpha_Q\right)^{\hat{N}_+}}{\alpha_P^{\hat{N}_+}\left(\eta_{P,-}(x_1)\cdot P_X(x_1)\right)^{n\cdot\hat{N}_+}} \\ &= \alpha_P^{\left(\hat{N}_- - \hat{N}_+\right)} \cdot \left(\eta_{P,-}(x_1)\cdot P_X(x_1)\right)^{n\left(\hat{N}_- - \hat{N}_+\right)} \cdot \left(\alpha_P\left(\eta_{P,+}(x_1)\cdot P_X(x_1)\right)^n + \alpha_Q\right)^{\hat{N}_+ - \hat{N}_-} \end{aligned} \tag{37}$$

where we broke up the product over $\mathcal{Z}_-$ and $\mathcal{Z}_+$. Now, canceling out similar terms in the fraction,

$$\begin{aligned}\prod_{Z_t\notin\mathcal{Z}_\pm} \frac{\Gamma_{+,\alpha}(Z_t)}{\Gamma_{-,\alpha}(Z_t)} &= \left(\frac{\eta_{P,-}(x_1)}{\eta_{P,+}(x_1)}\right)^{\left(\hat{n}_- - n\cdot\hat{N}_-\right)} \cdot \left(\frac{\eta_{P,+}(x_1)}{\eta_{P,-}(x_1)}\right)^{\left(\hat{n}_+ - n\cdot\hat{N}_+\right)} \\ &= \left(\frac{\eta_{P,+}(x_1)}{\eta_{P,-}(x_1)}\right)^{\left(\hat{n}_+ - \hat{n}_-\right)} \cdot \left(\frac{\eta_{P,-}(x_1)}{\eta_{P,+}(x_1)}\right)^{n\left(\hat{N}_+ - \hat{N}_-\right)} \end{aligned} \tag{38}$$

Now, the second factor in (38) can be expanded as follows to cancel out the second factor in (37):

$$\left(\frac{\eta_{P,-}(x_1)}{\eta_{P,+}(x_1)}\right)^{n\left(\hat{N}_+ - \hat{N}_-\right)} = \left(\frac{\eta_{P,-}(x_1)\cdot P_X(x_1)}{\eta_{P,+}(x_1)\cdot P_X(x_1)}\right)^{n\left(\hat{N}_+ - \hat{N}_-\right)}.$$

That is, we have:

$$\left(\frac{\eta_{P,-}(x_1)}{\eta_{P,+}(x_1)}\right)^{n\left(\hat{N}_+ - \hat{N}_-\right)} \cdot \prod_{Z_t\in\mathcal{Z}_\pm} \frac{\Gamma_{+,\alpha}(Z_t)}{\Gamma_{-,\alpha}(Z_t)} = \left(\frac{\alpha_P\left(\eta_{P,+}(x_1)\cdot P_X(x_1)\right)^n + \alpha_Q}{\alpha_P\left(\eta_{P,+}(x_1)\cdot P_X(x_1)\right)^n}\right)^{\left(\hat{N}_+ - \hat{N}_-\right)}.$$

In other words, we have:

$$\prod_{t=1}^{N} \frac{\Gamma_{+,\alpha}(Z_t)}{\Gamma_{-,\alpha}(Z_t)} = \left(\frac{\eta_{P,+}(x_1)}{\eta_{P,-}(x_1)}\right)^{\left(\hat{n}_+ - \hat{n}_-\right)} \cdot \left(\frac{\alpha_P\left(\eta_{P,+}(x_1)\cdot P_X(x_1)\right)^n + \alpha_Q}{\alpha_P\left(\eta_{P,+}(x_1)\cdot P_X(x_1)\right)^n}\right)^{\left(\hat{N}_+ - \hat{N}_-\right)}.$$

We then conclude by noticing that each of the above fractions is greater than 1. □

Proof of Proposition 7. For $t = N+1$, define $\hat{n}_{t,0}$, $\hat{n}_{t,+}$, and $\hat{n}_{t,-}$ as the number of points $(X_{t,i}, i \in [n])$ in Z_t which, respectively fall on x_0, or fall on x_1 with $Y_{t,i} = +1$, or fall on x_1 with $Y_{t,i} = -1$. Thus, for $t = N+1$ fixed, we have

$$\frac{\mathcal{D}_+(Z_{N+1})}{\mathcal{D}_-(Z_{N+1})} = \frac{\mathcal{D}_X^{\hat{n}_{t,0}}(x_0)\cdot\mathcal{D}_X^{(\hat{n}_{t,+}+\hat{n}_{t,-})}(x_1)\cdot\eta_{\mathcal{D},+}^{\hat{n}_{t,+}}(x_1)\cdot\eta_{\mathcal{D},-}^{\hat{n}_{t,-}}(x_1)}{\mathcal{D}_X^{\hat{n}_{t,0}}(x_0)\cdot\mathcal{D}_X^{(\hat{n}_{t,+}+\hat{n}_{t,-})}(x_1)\cdot\eta_{\mathcal{D},+}^{\hat{n}_{t,-}}(x_1)\cdot\eta_{\mathcal{D},-}^{\hat{n}_{t,+}}(x_1)} = \left(\frac{\eta_{\mathcal{D},+}(x_1)}{\eta_{\mathcal{D},-}(x_1)}\right)^{\left(\hat{n}_{t,+} - \hat{n}_{t,-}\right)}. \tag{39}$$

Now (39) ≥ 1 whenever $\hat{n}_{t,+} \geq \hat{n}_{t,-}$, so we proceed to bounding the probability of this event under $\mathcal{D}_-$. In particular, when $n = 1$, the event has probability at least $\mathbb{P}(\hat{n}_{t,0} = 1) = \mathcal{D}_X(x_0) = 1 - \frac{1}{2} = \frac{1}{2}$. Assume henceforth that $n > 1$, and let $\hat{n}_t \doteq \hat{n}_{t,+} + \hat{n}_{t,-}$. Conditioned on $\hat{n}_t$, $\hat{n}_{t,+}$ is distributed as Binomial$(\hat{n}_t, p)$, with $p = \eta_{\mathcal{D},-}(x_1)$. By the anticoncentration Lemma 3, we then have

$$\mathbb{P}\left(\hat{n}_{t,+} > \hat{n}_t \cdot p + \sqrt{\hat{n}_t \cdot p(1-p)}\right) \geq \frac{1}{12}, \tag{40}$$

so the event $\hat{n}_{t,+} \geq \hat{n}_{t,-}$ holds whenever

$$\sqrt{\hat{n}_t \cdot p(1-p)} \geq \frac{1}{2}\hat{n}_t - \hat{n}_t \cdot p = \hat{n}_t \left(\frac{1}{2} - p\right) = \hat{n}_t \cdot c_0 \cdot \epsilon_0^{1-\beta}, \text{in other words, whenever}$$

$$\hat{n}_t \cdot c_0^2 \cdot \epsilon_0^{2(1-\beta)} \leq \frac{1}{8}, \text{ (assuming } c_0 \leq 1/4). \tag{41}$$

Now, consider the event $\mathbf{E}_{\mathcal{D}} \doteq \{\hat{n}_t \leq 2\mathbb{E}[\hat{n}_t]\}$, where $\mathbb{E}[\hat{n}_t] = n_{\mathcal{D}} \cdot \mathcal{D}_X(x_1) = \frac{1}{2} n_{\mathcal{D}} \cdot \epsilon_0^{\beta}$. Under $\mathbf{E}_{\mathcal{D}}$, (41) is satisfied whenever $c_0^2 \leq \frac{1}{8}$, recalling $\epsilon_0 \doteq n_{\mathcal{D}}^{-1/(2-\beta)}$. In other words, under this condition, we have

$$\mathbb{P}\left(\hat{n}_{t,+} \geq \hat{n}_{t,-}\right) \geq \mathbb{E}\left[\mathbb{P}\left(\hat{n}_{t,+} \geq \hat{n}_{t,-} \mid \hat{n}_t\right) \mathbb{1}\{\mathbf{E}_{\mathcal{D}}\}\right] \geq \frac{1}{12}\mathbb{P}(\mathbf{E}_{\mathcal{D}}).$$

Finally, by multiplicative Chernoff, the event $\mathbf{E}_{\mathcal{D}}$ holds with probability at least $1 - \exp(-1/6) > 1/7$. □

E Auxiliary Lemmas

The following propositions are taken verbatim from [HK19].

Proposition 9 (Thm 2.5 of [Tsy09])**.** *Let $\{\Pi_h\}_{h \in \mathcal{H}}$ be a family of distributions indexed over a subset $\mathcal{H}$ of a semi-metric $(\mathcal{F}, \mathrm{dist})$. Suppose $\exists\, h_0, \ldots, h_M \in \mathcal{H}$, where $M \geq 2$, such that:*

(i) $\mathrm{dist}(h_i, h_j) \geq 2s > 0, \quad \forall 0 \leq i < j \leq M,$

(ii) $\Pi_{h_i} \ll \Pi_{h_0} \quad \forall i \in [M],$ *and the average KL-divergence to Π_{h_0} satisfies*

$$\frac{1}{M}\sum_{i=1}^{M} \mathcal{D}_{kl}(\Pi_{h_i} | \Pi_{h_0}) \leq \alpha \log M, \text{ where } 0 < \alpha < 1/8.$$

Let $Z \sim \Pi_h$, and let $\hat{h} : Z \mapsto \mathcal{F}$ denote any improper *learner of $h \in \mathcal{H}$. We have for any $\hat{h}$:*

$$\sup_{h \in \mathcal{H}} \Pi_h \left(\mathrm{dist}\left(\hat{h}(Z), h\right) \geq s\right) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log(M)}}\right) \geq \frac{3 - 2\sqrt{2}}{8}.$$

Proposition 10 (Varshamov-Gilbert bound)**.** *Let $d \geq 8$. Then there exists a subset $\{\sigma_0, \ldots, \sigma_M\}$ of $\{-1, 1\}^d$ such that $\sigma_0 = (1, \ldots, 1)$,*

$$\mathrm{dist}(\sigma_i, \sigma_j) \geq \frac{d}{8}, \quad \forall\, 0 \leq i < j \leq M, \quad \textit{and} \quad M \geq 2^{d/8},$$

where $\mathrm{dist}(\sigma, \sigma') \doteq \mathrm{card}(\{i \in [m] : \sigma(i) \neq \sigma'(i)\})$ is the Hamming distance.

Lemma 8 (A basic KL upper-bound)**.** *For any $0 < p, q < 1$, we let $\mathcal{D}_{kl}(p|q)$ denote $\mathcal{D}_{kl}(\mathrm{Ber}(p)|\mathrm{Ber}(q))$. Now let $0 < \epsilon < 1/2$ and let $z \in \{-1, 1\}$. We have*

$$\mathcal{D}_{kl}(1/2 + (z/2)\cdot\epsilon \,|\, 1/2 - (z/2)\cdot\epsilon) \leq c_0 \cdot \epsilon^2, \textit{ for some } c_0 \textit{ independent of } \epsilon.$$