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 $D$. 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:

- 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 $D$, 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 $D$ (i.e., one of the $P_t$’s) are of the form (see e.g., Crammer, Kearns and Wortman, 2008; Ben-David...
(1) \[ \mathcal{E}_D(\hat{h}) \leq (nN)^{-\alpha} + n^{-\alpha} + \text{disc}(\{P_i\}; D), \text{ for some } \alpha \in [1/2, 1], \]

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 \( \text{disc}(\{P_i\}; D) > 0 \), even while the excess risk of a naive classifier \( \hat{h} \) trained on the target dataset would be \( \mathcal{E}_D(\hat{h}) \propto n^{-\alpha} \to 0 \). 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 \( P_i \)'s relate to \( D \).

In particular, we consider a seemingly favorable classification setting, where all data distributions \( P_i \)'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 Arjovsky et al., 2019, 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 are highlight next:

- **No adaptive procedure exists.** 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 \( n^{-1/2} \) without further favorable restrictions, or information, on distributions (Theorem 5).
- In low noise regimes, e.g., so-called Massart’s noise (which we parametrize by a classical Bernstein class condition), even the naive but 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 \( P_i \)'s (Theorem 3).
- At any noise level, a ranking of sources \( P_i \)'s according to discrepancy to a target \( D \) is sufficient information for (partial) adaptivity. While a precise ranking is probably unlikely in practice, an approximate ranking might be available, based on domain knowledge on how sources rank in fidelity w.r.t. to an ideal target sample: e.g., in settings such as learning with slowly drifting distributions. Here we show that a simple 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 \( D \) in \( \{P_i\} \) (see Theorem 2), due to the inherent asymmetry in the information tasks might have on each other, e.g., some \( P_i \) 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 (Hanneke and Kpotufe, 2019) 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 Hanneke and Kpotufe, 2019), 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 (Ben-David et al.,
2007; Cortes et al., 2008; Ben-David et al., 2010a), which
introduce notions of discrepancy such as the $d_A$-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 (Crammer, Kearns and Wortman, 2008; Ben-David
et al., 2010a). Such bounds can in fact be shown to be tight w.r.t. the discrepancy term
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., Bartlett, 1992). 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 $X_P, X_\mathcal{D}$) 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 multitask
rates $\mathbb{E}_\mathcal{D}(\hat{h}) \to 0$ in natural situations with disc$(\{P_t\}; \mathcal{D}) > 0$, even with no target sample.

We remark that other notions of discrepancy, e.g., Maximum Mean Discrepancy (Gretton
et al., 2009), Wasserstein distance (Redko, Habrard and Sebban, 2017; Shen et al., 2018)
are employed in domain adaptation; however they appear relatively less often in the theo-
retical literature on multitask and related areas. For multitask, the work of Ben-David and
Borbely (2008) 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 $X_\mathcal{D}$; that work also incurs
an $n^{-1/2}$ term in the risk bound, but no discrepancy term. The work of Mansour, Mohri and
Rostamizadeh (2009) 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 im-
portant point: two distributions might have asymmetric information on each other w.r.t. domain
adaptation. Such insight was raised recently in (Kpotufe and Martinet, 2018; Hanneke and
Kpotufe, 2019) and independently in (Achille et al., 2019), 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 (Caruana, 1997), 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 (Baxter, 1997; Ando and Zhang, 2005; Maurer, Pontil and Romera-
Paredes, 2013; Pentina and Lampert, 2014; Yang, Hanneke and Carbonell, 2013; Maurer,
Pontil and Romera-Paredes, 2016), rather than bounding the supremum risk across tasks –
i.e., treating multitask as separate multisources, as of interest here. Some of these average bounds, e.g., (Maurer, Pontil and Romera-Paredes, 2013, 2016), remove the dependence on discrepancy inherent in bounds of the form (1), but maintain a term of the form $n^{-\alpha}$; in other words, bounds on supremum risk derived from such results would be in terms of a single dataset size $n$. The work of Blum et al. (2017) 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), (Ben-David et al., 2010b) 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 (Hanneke and Kpotufe, 2019) for a range of situations including those considered in (Ben-David et al., 2010b). 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 Qiao (2018); Mahloujifar, Mahmoody and Mohammed (2019); Konstantinov et al. (2020) show limits of multitask under various adversarial corruption of labels in datasets, while Scott and Zhang (2019) derive a positive result, i.e., rates (for Lipschitz loss) decreasing in both $N$ and $n$, up to excluded or downweighted datasets. The procedure of Scott and Zhang (2019) is however nonadaptive as it requires known noise proportions.

(Konstantinov et al., 2020) 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 $D$ (Theorem 9 of Appendix C). 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 $n = 1$ interacts 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 (Ben-David, Benedek and Mansour, 1989; Bartlett, 1992; Barve and Long, 1997) or related notions (Mohri and Medina, 2012; Hanneke and Yang, 2019), 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.
In this work, we have chosen to focus on the setting of learning guarantees phrased with respect to a given hypothesis class $\mathcal{H}$ of finite VC dimension – a family of classifiers – where the interest is in achieving low excess risk relative to the best $h$ in the class $\mathcal{H}$ – which is often appealing, as guarantees can be stated with minimal assumptions on the data distribution and on the quality of the best-in-class. While this perspective is perhaps the most dominant in the classification literature, it does not provide a theory of the approximation error of the class $\mathcal{H}$, which requires stronger modeling assumptions on the form of the Bayes classifier, as done for instance in the nonparametric literature on classification.

There have been a significant number of works on multitask and transfer learning in the nonparametric setting, which consider the case of smooth regression functions and margin noise conditions (Cai and Wei, 2021; Kpotufe and Martinet, 2018; Reeve, Cannings and Samworth, 2021). In addition to the fact that the nonparametric setting is generally incomparable to the present setting of learning relative to a hypothesis class, an even more important distinction between the present work and all of these works is in our treatment of the number of tasks $N$. For one, the works of Kpotufe and Martinet (2018) and Reeve, Cannings and Samworth (2021) both consider the special case $N = 1$ (a.k.a. domain adaptation or transfer learning), whereas $N$ is implicitly being treated as a constant in the work of Cai and Wei (2021) – which is evident, e.g., by the fact that Theorem 5 therein cannot otherwise hold for general $n$ (e.g., $n = 1$). In contrast, our rates capture the interplay between any size $N$ and dataset sizes $n_t$ within our setting, which better captures the benefits of having a number of data sources $N \gg 1$, and which turns out to be crucial in establishing our main result on the dichotomy between oracle and adaptive rates.

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 common latent substructures, both as applied to classification (Muandet, Balduzzi and Schölkopf, 2013; McNamara and Balcan, 2017; Arora et al., 2019), or to regression settings (Jalali et al., 2010; Lounici et al., 2011; Negahban and Wainwright, 2011; Du et al., 2020; Tripuraneni, Jordan and Jin, 2020). 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 $\mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} = \{-1, 1\}$. We focus on statistical learning where, given data, a learner is to return a classifier $h : \mathcal{X} \mapsto \mathcal{Y}$, and we are interested in performing nearly as well as the best function in a given hypothesis class $\mathcal{H}$ of such functions.

**ASSUMPTION 1 (Bounded VC).** Throughout we will let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ 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 $|\mathcal{H}| \geq 3$.

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 $R_P(h) \doteq P_{X,Y}(h(X) \neq Y)$ 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 $\arg\min_{h \in \mathcal{H}} R_P(h)$ (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 $P(h \neq h') \doteq P_X(h(X) \neq h'(X))$.

DEFINITION 2. Throughout, a dataset $S$ of $(x,y)$ pairs in $\mathcal{X} \times \mathcal{Y}$ will be treated as a multiset, i.e., allowing duplicates. Therefore, size $|S|$, union $S \cup S'$, and summation $\sum_{(x,y) \in S}$ are understood as for multisets (i.e., viewing duplicates as distinct elements).

Given a dataset $S$ of $(x,y)$ pairs, 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 $\hat{P}_S(h \neq h') \doteq \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}\{h(x) \neq h'(x)\}$.

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

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

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 $\mathcal{H}$, i.e., let $\eta_P(x) = \mathbb{P}_P[Y \mid X = x]$, then the margin condition

$$P_X(x : |\eta_P(x)| \leq \tau) \leq C_{\beta} \tau^\kappa, \forall \tau > 0,$$

implies that $P$ satisfies (2) with $\beta = \kappa/(1 + \kappa)$ for some $C_{\beta}$ (Tsybakov, 2004). Definition 3 is a well-known relaxation of Tsybakov’s margin condition, which admits a broader family of distributions, yet recovers essentially the same optimal rates of convergence as the strict margin condition (Massart and Nédélec, 2006; Koltchinskii, 2006; Bartlett, Jordan and McAuliffe, 2006; van Erven et al., 2015). Indeed, note that unlike the Tsybakov noise assumption, Definition 3 does not require $h_P^*$ to be the Bayes classifier. For instance, a simple example illustrating this is the case $\mathcal{X} = [0,1], \mathcal{H} = \{x \mapsto h_t(x) \doteq \text{sign}(x - t) : t \in [0,1]\}$ (threshold classifiers), $P_X$ uniform on $[0,1]$, and $P(Y = 1 \mid X) = \mathbb{1}\{X \in [1/3, 3/4]\}$. Clearly the Bayes classifier is not a threshold (it is a two-sided interval). Moreover, $h_P^*$, the best function in $\mathcal{H}$, is unique, and is $h_{1/3}$: a threshold at $1/3$. Notice that any $h_t$ with low $\mathcal{E}_P(h_t)$ must be close to $h_P^*$ and in particular, one can easily verify that Equation (2) holds with $\beta = 1$ and $C_{\beta} = 4$, i.e., $\mathcal{E}_P(h_t)$ is within a constant factor of the disagreement $P(h_t \neq h_P^*) = |t - 1/3|$.

The importance of considering $\beta$ becomes 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 $\beta$, in that a smaller $\beta$ makes adaptation more challenging.
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 \( \mathcal{E}_P(h) \) w.r.t. any of these distributions. W.l.o.g. we fix attention to a single target distribution \( \mathcal{D} = P_{N+1} \), and therefore reduce the setting to that of multisource\(^1\).

**Definition 4.** Fix sample sizes \( n_t \in \mathbb{N} \). A (multisource) learner is any function \( \hat{h} : \prod_{t \in [N+1]} (\mathcal{X} \times \mathcal{Y})^{n_t} \rightarrow \mathcal{Y}^{\mathcal{X}} \). That is, given a multisample \( Z = \{ Z_t \}_{t \in [N+1]}, |Z_t| = n_t \), \( \hat{h} \) returns a classifier. 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 \( Z \). Much of the rest of the paper will therefore focus on multisource for fixed target \( \mathcal{D} = P_{N+1} \).

3.1. Relating Sources to Target. Clearly, how well one can do in multitask depends on how distributions relate to each other, as relevant to the chosen hypothesis class \( \mathcal{H} \) (a family of predictors). The following parametrizes this relationship, for given \( \mathcal{H} \).

**Definition 5.** Let \( \rho > 0 \) (up to \( \rho = \infty \)), and \( C_\rho \geq 2 \). We say that distribution \( P \) has transfer exponent \( (C_\rho, \rho) \) w.r.t. distribution \( \mathcal{D} \), if

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

This happens to be a simple way to express a desirable behavior in transfer learning where \( \mathcal{E}_P \rightarrow 0 \implies \mathcal{E}_\mathcal{D} \rightarrow 0 \). Notice that the above always holds with at least \( \rho = \infty \).

We have shown in earlier work (Hanneke and Kpotufe, 2019) 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 hypotheses \( h^*_P, h^*_\mathcal{D} \) 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 \( \rho \) and \( \beta \) (a Bernstein class parameter on \( P \) and \( \mathcal{D} \)). In other words, \( \rho \) captures an effective data size \( n_P^{1/\rho} \) contributed by the source to the target: this decreases as \( \rho \rightarrow \infty \), 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 as discussed at the end of Example 2 below).

 Altogether, the transfer exponent \( \rho \) 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 (Hanneke and Kpotufe, 2019). 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 \( d_A, d_Y \) 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 \( h^*_\tau \) thresholded at \( \tau \in [0, 1] \), \( 2P_X(h^*_\tau \neq h^*) = D_X(h^*_\tau \neq h^*) \), where for \( \tau > 1 \), \( P_X(h^*_\tau \neq h^*) = \frac{1}{2}(\tau - 1/2) \geq \frac{1}{2}D_X(h^*_\tau \neq h^*) = \frac{1}{4} \). Thus, the exponent \( \rho = 1 \) with \( C_\rho = 2 \), so we have fast transfer at the same rate \( 1/n_P \) as if sampling from \( \mathcal{D} \).

---

\(^1\)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.
On the other hand, recall that the $d_A$-divergence takes the form:

$$d_A(P, \mathcal{D}) \triangleq \sup_{h \in \mathcal{H}} |P_X(h \neq h^*) - \mathcal{D}_X(h \neq h^*)|,$$

while the $\mathcal{Y}$-discrepancy $d_Y(P, \mathcal{D}) \triangleq \sup_{h \in \mathcal{H}} |\mathcal{E}_P(h) - \mathcal{E}_D(h)|$.

The two coincide when $Y = h^*(X)$.

Now, take $h_{\tau}$ as the threshold at $\tau = 1/2$, and $d_A = d_Y = \frac{1}{4}$ which would wrongly suggest a slow transfer rate $\mathcal{E}_D(\hat{h}) \leq o_P(1) + \frac{1}{4}$; we can in fact make this situation worse, i.e., let $d_A = d_Y \to \frac{1}{2}$ 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 (Hanneke and Kpotufe, 2019).

**Example 2. (Minimum $\rho$, and the asymmetry of transfer).** Consider $\mathcal{H}$, the class of linear classifiers in $\mathbb{R}^d$, and w.l.o.g., let $h^*(x) = \text{sign}(v^T x)$. Let $\mathcal{D}_X$ be absolutely continuous, with compact support containing 0 on its interior (i.e., the decision boundary passes through the interior of the support). Suppose the density of $\mathcal{D}_X$ takes the form $f_\mathcal{D} \propto |v^T x|^{\alpha_D - 1}$, $\alpha_D \geq 1$, while $P_X \ll \mathcal{D}_X$ has density $f_P(x) \propto |v^T x|^{\alpha_P}$, $\alpha_P \geq 0$; that is both $P_X, \mathcal{D}_X$ vanish at the decision boundary $\{v^T x = 0\}$ (for $\alpha_D > 1$), while $P_X$ vanishes faster. It can be shown that $1 + \alpha_P/\alpha_D$ is the smallest possible transfer exponent (see Appendix H). Notice that $\alpha_D = 1$ recovers the common assumption of $(\mathcal{D}_X)$ density bounded away from zero.

As a simple illustration in $\mathbb{R}$, let $\mathcal{H}$ be the class of one-sided thresholds on the line, $h^*$ 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 $h_{\tau}$ at threshold $\tau > 0$, we have $P_X(h_{\tau} \neq h^*) = \int_0^\tau f_P(x) \propto \tau^\rho$, while $\mathcal{D}_X(h_{\tau} \neq h^*) \propto \tau$. Notice that for any fixed $\epsilon > 0$,

$$\lim_{\tau > 0, \tau \to 0} \frac{\mathcal{D}_X(h_{\tau} \neq h^*)^{\rho-\epsilon}}{P_X(h_{\tau} \neq h^*)} = \lim_{\tau > 0, \tau \to 0} \frac{\tau^{\rho-\epsilon}}{\tau^\rho} = \infty.$$

We therefore see that $\rho$ is the smallest possible transfer-exponent. Interestingly, consider transferring instead from $\mathcal{D}$ to $P$: we would have $\rho(\mathcal{D} \to P) = 1/\rho \leq 1 \leq \rho(P \to \mathcal{D})$; 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 to the decision boundary. This is *not captured by symmetric notions of distance*, e.g., metrics or semi-metrics such as $d_A$, $d_Y$, MMD, TV, or Wasserstein distance.

Interestingly, no smallest $\rho$ may exist, i.e., all $\rho > \rho^\beta$, for some $\rho^\beta$ may be admissible as transfer exponents, while $\rho^\beta$ itself is not admissible. Finally, note that all above examples extend naturally to nondeterministic labels by considering the two measures $\int |\mathcal{E}_\mu(Y|x)| \ d\mu(x)$, for $\mu$ taken as $P$ or $\mathcal{D}$, in place of $P_X$ and $\mathcal{D}_X$. This is all discussed in Appendix H.

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

**Definition 6 (Multisource class).** We consider classes

$$\mathcal{M} \triangleq \mathcal{M} \left( \mathcal{C}_\rho, \{\rho_t\}_{t \in [N]} \cup [N+1], C_\beta, \beta \right)$$

of product distributions of the form $\Pi = \prod_{t \in [N]} P_t^{n_t} \times \mathcal{D}^{n_{\mathcal{D}}}$, $n_t \geq 1$, and $n_{\mathcal{D}} \triangleq n_{N+1} \geq 0$, satisfying:
(A1). There exists \( h^* \in \mathcal{H} \), \( \forall t \in [N + 1], h^* \in \text{argmin}_{h \in \mathcal{H}} \mathcal{E}_{P_t}(h) \),

(A2). Sources \( P_t \)'s have transfer exponent \((C_{\rho}, \rho_t)\) w.r.t. the target \( D \),

(A3). All sources \( P_t \) and target \( D \) admit a Bernstein class condition with parameter \((C_{\beta}, \beta)\).

For notational simplicity, we will often let \( P_{N+1} \equiv D \). Also, although not a parameter of the class, we also refer to \( \rho_{N+1} \equiv 1 \), as \( D \) 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 \( h^*_t \in \text{argmin}_{h \in \mathcal{H}} \mathcal{E}_{P_t}(h) \). Furthermore for any \( t \in [N] \) such that \( \rho_t < \infty \), we also have \( \mathcal{E}_{D}(h^*_t) = 0 \), implying \( D(h^* \neq h^*_t) = 0 \) for \( \beta > 0 \).

### 3.3. Additional Notation.

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

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

We will often be interested in order statistics \( \rho(1) \leq \rho(2) \cdots \leq \rho(t) \) 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 \( Z^{(t)} = Z^{(1), \ldots, (t)} \).

- **Average transfer exponent.** Define
  \[
  \bar{\rho}_t = \sum_{s \in [t]} \frac{\rho(s)n(s)}{\sum_{r \in [t]} n(r)}, \quad t \in [N + 1].
  \]

- **Aggregate ERM.** For any \( Z^I, I \subset [N + 1] \), we let \( \hat{h}_{Z^I} = \text{argmin}_{h \in \mathcal{H}} \mathcal{E}_{Z^I}(h) \), 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 write \( a \wedge b = \min \{a, b\} \), \( a \vee b = \max \{a, b\} \).

- **Positive Logarithm.** For any \( x \geq 0 \), define \( \log(x) = \max \{ \ln(x), 1 \} \).

- **1/0 Convention.** We adopt the convention that \( 1/0 = \infty \).

- **Asymptotic Order.** We often write \( a \preceq b \) 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 \( \rho_t \) 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 in 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 learner with knowledge of \( \mathcal{M} \). We have:**

\[
\inf_{\hat{h}} \sup_{\Pi \in \mathcal{M}} \mathbb{E}_\Pi \left[ \mathcal{E}_D(\hat{h}) \right] \asymp \min_{t \in [N+1]} \left( \sum_{s = 1}^{t} n(s) \right)^{-1/(2-\beta)\bar{\rho}_t}.
\]
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 \( \hat{\rho}_t \) with simply \( \rho(t) \) and the result would still be true. In other words, although intuitively \( \hat{\rho}_t \) 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 (see Remark 7 of the supplementary material). The case \( \rho_t \in (0, 1) \) represents a kind of super transfer, where the source samples are actually more informative than target samples.

We will see that, for any cutoff \( t \in [N + 1] \), the rate \( (\sum_{s=1}^{t} n(s))^{-1/(2-\beta)} \rho_t \) is achieved by ERM over the aggregate sample \( Z^{(t)} \) (Theorem 7). Therefore, combined with Theorem 1, it follows that it is sufficient to select one of \( N + 1 \) possible aggregations – defined by the ranking \( \rho(1) \leq \rho(2) \leq \cdots \leq \rho(N+1) \) – to nearly achieve the minimax rate. This is interesting given there are over \( 2^N \) possible ways of aggregating datasets.

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 \( D_X(h \neq h') \). For illustration, considering the case \( \beta = 1 \), for any \( h, h' \) that are sufficiently far under \( D_X \) so that \( \mathcal{E}_D(h'; h) \gtrsim \epsilon \), the lower-bound construction is such that

\[
(3) \quad \text{KL}(\Pi_{h} \| \Pi_{h'}) \lesssim \sum_{t=1}^{N+1} n_t e^{\rho t} \lesssim 1,
\]

ensuring that \( \Pi_{h}, \Pi_{h'} \) are hard to distinguish from finite sample. Thus, the largest \( \epsilon \) satisfying the second inequality above, say \( \epsilon_{N+1} \), 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 \( \beta \). Now, again for illustration, let \( \beta = 1 \), and consider the ERM \( \hat{h} = \hat{h}_Z \) combining all \( N + 1 \) datasets. Let \( \mathcal{E}_\alpha \equiv \sum_{t=1}^{N+1} \alpha_t \mathcal{E}_t(h), \alpha_t = n_t / (\sum s n_s) \), then the concentration arguments described above ensure that \( \mathcal{E}_\alpha(h) \lesssim 1 / (\sum s n_s) \). Now notice that, by definition of \( \rho_t \), \( \mathcal{E}_\alpha(h) \gtrsim \sum t \alpha_t \mathcal{E}_D^{\rho_t}(h) \), in other words, \( \mathcal{E}_D^{\rho_t}(h) \) satisfies the second inequality in (3), and must therefore be at most of order \( \epsilon_{N+1} \). This establishes the tightness of \( \epsilon_{N+1} \) 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 \( \beta \), 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 \( \rho_t \)'s w.r.t. any target \( \mathcal{D} = P_s, s \in [N + 1] \)). 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 D.

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

Let \( h_Z, h_{Z_P}, h_{Z_D} \) denote the ERMs over \( Z, Z_P, Z_D \) respectively. Suppose \( 1 \leq n_D^2 \leq \frac{1}{8} n_P^2 (2^{-3-\beta}) / (2^-\beta) \). Then

\[
\mathbb{E} \left[ \mathcal{E}_D(h_{Z_P}) \right] \wedge \mathbb{E} \left[ \mathcal{E}_D(h_Z) \right] \geq \frac{1}{4}, \text{ while } \mathbb{E} \left[ \mathcal{E}_D(h_{Z_D}) \right] \lesssim n_D^{-1/(2-\beta)};
\]

however, \( \mathbb{E} \left[ \mathcal{E}_P(h_{Z_P}) \right] \vee \mathbb{E} \left[ \mathcal{E}_P(h_Z) \right] \lesssim n_p^{-1/(2-\beta)} \).

**Remark 3 (Suboptimality of pooling).** A common practice is to pool all datasets together and return an ERM as in \( h_Z \). 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., without additional conditions, such as the situations considered in this section. 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 \( \rho_t \)'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 \( 1 - \delta \),

\[
\mathcal{E}_D(h_Z) \leq \min_{t \in [N+1]} C_\rho \left( c C_\beta \frac{d_H \log \left( \frac{1}{a_H} \sum_{s=1}^{N+1} n_s \right) + \log(1/\delta)}{\sum_{s=1}^{N} n(s)} \right)^{1/\rho_t}.
\]

The theorem is proven in Section 7.1. We also state a general bound on \( \mathcal{E}_D(h_Z) \), holding for any \( \beta \), in Corollary 2 of Appendix C; the implied rates are not always optimal, though interestingly they are near-optimal in the case that \( \sum_{t=1}^{N} n(t) \propto \sum_{t=1}^{N} n_t \), where \( n^{*} \) 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 Theorem 3 are larger, as they are 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, \( C_\beta, \beta \), 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 $\varepsilon(m, \delta) = \frac{d\log}{m} \log \left( \frac{m}{d\log} + \frac{1}{\delta} \right)$, for $\delta \in (0, 1)$ and $m \in \mathbb{N}$. Let $n_t = \sum_{s=1}^{t} n(s)$ for each $t \in [N+1]$, and recall that $\hat{h}_{Z(t)}$ denotes the ERM over the aggregate sample $Z(t)$.

**Rank-based Procedure $\tilde{h}$:** Let $\delta_t = \delta/(6t^2)$, and let $C_0$ be the numerical constant from Lemma 1. For any $t \in [N+1]$, define:

$$
\mathcal{H}(t) = \left\{ h \in \mathcal{H} : \hat{\mathcal{E}}_{Z(t)}(h; \hat{h}_{Z(t)}) \leq C_0 \sqrt{\mathbb{P}_{Z(t)}(h \neq \hat{h}_{Z(t)}) \varepsilon(n_t, \delta_t) + C_0 \varepsilon(n_t, \delta_t)} \right\}.
$$

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 (Rank-based Procedure).** Let $\delta \in (0, 1)$. Let $\tilde{h}$ 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 $1 - \delta$,

$$
\mathcal{E}_D(\tilde{h}) \leq \min_{t \in [N+1]} C_\rho \left( c C_\beta \frac{d\log}{m} \left( \frac{1}{d\log} \sum_{s=1}^{t} n(s) \right) + \log(1/\delta) \right)^{1/(2-\beta)\rho_t}.
$$

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 $D$. In this case, in contrast to the lower bounds proven by (Bartlett, 1992), 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 $\rho_t$ distributions by the sequence of $\rho_t$ values, we can describe scenarios where the traditional upper bound analysis of drifting distributions from (Bartlett, 1992; Ben-David, Benedek and Mansour, 1989; Barve and Long, 1997) 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 $\rho_t$ values. While we make no claim as to how well these scenarios actually capture practice, they offer some additional flexibility to be wrong in the ranking as is likely in practice (since the ranking cannot be learned, as a consequence of our non-adaptivity results of Theorem 5).

One immediate observation is that, since Definition 5 does not require $\rho_t$ to be minimal, any larger value of $\rho_t$ would also be valid; therefore, for any permutation $\sigma : [N+1] \to [N+1]$, there always exists some valid choices of $\rho_t$ 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 $\rho_t$ values) for the rank-based procedure applied with this $\sigma(t)$ ordering of the tasks. For instance, this means that we can use an ordering $\sigma(t)$ that only swaps the order of some $\rho_t$ values that are within some $\epsilon$ 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. However, in general, as stated below, no algorithm can guarantee a rate better than a dependence on the number of samples \( n_D \) from the target task, even when a rank-based procedure can achieve any desired rate \( \epsilon \).

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

- **Let** \( \hat{h} \) **denote any multisource learner with no knowledge of** \( M. \) **There exists a multisource class** \( M_0 \) **with parameters** \( n_{N+1} = n_D, \ n_t = n, \forall t \in [N], \) **and all other parameters satisfying the above, such that**
  \[
  \sup_{\Pi \in M_0} \mathbb{E}_\Pi \left[ \mathcal{E}_D(\hat{h}) \right] \geq c \cdot \left( 1 \wedge n_D^{-1/(2-\beta)} \right) \text{ for a universal constant } c > 0.
  \]

- **On the other hand, there exists a semi-adaptive classifier** \( \tilde{h}, \) **which, given ranking information** \( \rho(1) \leq \rho(2) \cdots \leq \rho(N+1) \) **on** \( M_0, \) **but no other information, achieves the rate**
  \[
  \sup_{\Pi \in M_0} \mathbb{E}_\Pi \left[ \mathcal{E}_D(\hat{h}) \right] \leq \epsilon.
  \]

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 \( \rho_t \) (see Proposition 1) – 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 \( M \) simply means that a minimax analysis is performed over a larger family containing such \( M \)'s, indexed over choices of ranking each corresponding to a fixed \( M. \)

Finally, we note that the result leaves open the possibility of adaptivity under further distributional restrictions on \( M, \) for instance requiring that the number of samples \( n \) per source task be large w.r.t. other parameters such as \( \beta \) and \( N; \) in particular, our current conditions on \( n \) w.r.t. \( \beta \) might simply be a feature of our construction as explained in Section 8. 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 C that the ERM \( \hat{h}_Z \) which pools all datasets achieves a target excess risk \( \mathcal{E}_D(\hat{h}_Z) \) depending on the (weighted) median of the \( \rho_t \) 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 \( \rho_t, \) 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,

\[
\mathcal{E}_D(\hat{h}) > c_1 \min_{t \in [N+1]} \left( \frac{c_2 d_{\mathcal{H}}}{(\sum_{s=1}^{t} n(s)) \log^2 \left( \sum_{s=1}^{t} n(s) \right)} \right)^{1/(2-\beta)\rho_t}
\]

for numerical constants \( c_1, c_2 > 0. \)
Here we briefly sketch the main ideas of the proof, for the special case \( d_H = 1 \). In this case, we construct two distributions \( \Pi_\sigma = \prod_{t \in [N+1]} P^\sigma_t, \sigma \in \{-1, 1\} \), such that any \( h \) can have \( \mathcal{E}_{D^\sigma}(h) < \epsilon \) on at most one of them (for some \( \epsilon \)). Specifically, for \( x_0, x_1 \) any two points such that \( \exists h_0, h_1 \in \mathcal{H} \) with \( h_0(x_0) = h_1(x_0) \) and \( h_0(x_1) \neq h_1(x_1) \), we can define \( P^\sigma_t \) to have marginal probability of \( x_1 \) as \( e^{\rho_t} \) and conditional probability of label 1 given \( x_1 \) as \( \frac{1}{2} + \frac{1}{2} e^{\rho_t(1-\beta)} \), and marginal probability of \( x_0 \) as \( 1 - e^{\rho_t} \) and conditional probability of label \( h_0(x_0) \) given \( x_0 \) as one. In particular, for any \( \sigma \), any \( h \) with \( \mathcal{E}_{D^\sigma}(h) < \epsilon \) must have \( h(x_1) = \sigma \). Furthermore, we argue that \( \text{KL}(\Pi_{-1} || \Pi_1) \leq \epsilon \sum_{t \in [N+1]} n_t e^{\rho_t(2-\beta)} \) for a constant \( c \). Then, choosing any value \( \epsilon \) for which this last expression is bounded by a certain constant yields a lower bound \( \propto \epsilon \). It is at this point in the proof that we have an interesting leap, as we find that it suffices to plug in a value of \( \epsilon \) of roughly the same form as the upper bound of Theorem 7 (differing only in log factors). The details of the proof, and the extension to general \( d_H \), are included in Appendix A of the supplementary material.

6. Upper Bound Analysis. We will in fact establish the upper bound as a bound holding with high probability \( 1 - \delta \), for any \( \delta \in (0, 1) \). Let

\[
\hat{t}^* \doteq \arg\min_{t \in [N+1]} \left( 2^{10} C_0^4 C_\beta \left( \frac{d_H \log \left( \frac{1}{d_H} \sum_{s=1}^t n_{(s)} \right) + \log(1/\delta)}{\sum_{s=1}^t n_{(s)}} \right) \right)^{1/(2-\beta)\hat{\rho}_t},
\]

for \( C_0 \) a numerical constant from Lemma 1 below. The oracle procedure just returns \( \hat{h}_{Z(\hat{t}^*)} \), the ERM over \( Z(\hat{t}^*) \). We have the following theorem.

**Theorem 7.** For any \( t \in [N+1], \Pi \in \mathcal{M}, \) and \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \),

\[
\mathcal{E}_{D}(\hat{h}_{Z(t)}) \leq C \left( \frac{d_H \log \left( \frac{1}{d_H} \sum_{s=1}^t n_{(s)} \right) + \log(1/\delta)}{\sum_{s=1}^t n_{(s)}} \right)^{1/(2-\beta)\hat{\rho}_t},
\]

for \( C = 2^{10} C_0^4 C_\beta \), where \( C_0 \) is a numerical constant from Lemma 1 below. In particular, the upper bound portion of the claim of Theorem 1 follows by taking \( t = \hat{t}^* \) defined above above.

The proof will rely on the following Lemma 1: 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., Koltchinskii, 2006). For completeness, we include a proof of the extension to non-identically distributed data in Appendix B, based on a generalization of Bousquet’s inequality due to Klein and Rio (2005).

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

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

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

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 $\Pi$ as in Theorem 7, for any $I \subseteq [N + 1]$, letting $\mathbf{n}_I = \sum_{t \in I} n_t$ and $\bar{P}_I = \mathbf{n}_I^{-1} \sum_{t \in I} n_t P_t$, for any $\delta \in (0, 1)$, on the event (of probability at least $1 - \delta$) from Lemma 1 (for $S = Z_t'$ there), $\forall h \in \mathcal{H}$ satisfying

\[ \hat{E}_{Z_t'}(h; \hat{h}_{Z_t'}) \leq C_0 \sqrt{\bar{P}_{Z_t'}(h \neq \hat{h}_{Z_t'}) \varepsilon(n_I, \delta) + C_0 \varepsilon(n_I, \delta)}, \]

it holds that $\hat{E}_{\bar{P}_I}(h) \leq 32C_0^2 (C_\beta \varepsilon(n_I, \delta))^{1/(2-\beta)}$.

PROOF. On the event from Lemma 1 for $S = Z_t'$, $\forall h, h' \in \mathcal{H}$, (since $\hat{E}_{\bar{P}_I}(h; h') = \mathbb{E} [\hat{E}_{Z_t'}(h; h')]$) the following two inequalities hold:

\[ \hat{E}_{\bar{P}_I}(h; h') \leq \hat{E}_{Z_t'}(h; h') + C_0 \sqrt{\min \{ \bar{P}_I(h \neq h'), \bar{P}_{Z_t'}(h \neq h') \} \varepsilon(n_I, \delta) + C_0 \varepsilon(n_I, \delta)} \]

(8) $\hat{E}_{\bar{P}_I}(h; h') \leq \hat{E}_{Z_t'}(h; h') + C_0 \sqrt{\min \{ \bar{P}_I(h \neq h'), \bar{P}_{Z_t'}(h \neq h') \} \varepsilon(n_I, \delta) + C_0 \varepsilon(n_I, \delta)}$

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

\[ \hat{E}_{\bar{P}_I}(h_{Z_t'}) \leq C_0 \sqrt{\bar{P}_I(h_{Z_t'} \neq h^*) \varepsilon(n_I, \delta) + C_0 \varepsilon(n_I, \delta)}. \]

Combining this with (10) (for $h = \hat{h}_{Z_t'}$) implies

\[ \hat{E}_{\bar{P}_I}(h_{Z_t'}) \leq C_0 \sqrt{C_\beta (\hat{E}_{\bar{P}_I}(h_{Z_t'}))^\beta \varepsilon(n_I, \delta) + C_0 \varepsilon(n_I, \delta)}, \] which in turn implies

(11) $\hat{E}_{\bar{P}_I}(h_{Z_t'}) \leq 2C_0 (C_\beta \varepsilon(n_I, \delta))^{1/(2-\beta)}$.

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

\[ \hat{E}_{\bar{P}_I}(h; h_{Z_t'}) \leq 2C_0 \sqrt{\bar{P}_{Z_t'}(h \neq \hat{h}_{Z_t'}) \varepsilon(n_I, \delta) + 2C_0 \varepsilon(n_I, \delta)}, \]

and (9) implies the right hand side is at most (after some rearranging)

(13) $2C_0 \sqrt{2\bar{P}_I(h \neq \hat{h}_{Z_t'}) \varepsilon(n_I, \delta) + 2C_0(1 + \sqrt{C_0}) \varepsilon(n_I, \delta)}$.

By the triangle inequality and (10), together with (11), we have

\[ \bar{P}_I(h \neq \hat{h}_{Z_t'}) \leq \bar{P}_I(h \neq h^*) + \bar{P}_I(\hat{h}_{Z_t'} \neq h^*) \leq C_\beta (\hat{E}_{\bar{P}_I}(h))^\beta + C_\beta (\hat{E}_{\bar{P}_I}(\hat{h}_{Z_t'}))^\beta \leq C_\beta (\hat{E}_{\bar{P}_I}(h))^\beta + \bar{P}_I(\hat{h}_{Z_t'}) \leq C_\beta (\hat{E}_{\bar{P}_I}(h))^\beta + C_\beta (2C_0)^\beta (C_\beta \varepsilon(n_I, \delta))^{\beta/(2-\beta)}. \]

Combining this with (13) and (12) implies (after simplifying the resulting expression)

\[ \hat{E}_{\bar{P}_I}(h; h_{Z_t'}) \leq 2C_0 \sqrt{2C_\beta (\hat{E}_{\bar{P}_I}(h))^\beta \varepsilon(n_I, \delta) + 8C_0^2 (C_\beta \varepsilon(n_I, \delta))^{1/(2-\beta)}}. \]

Since $\hat{E}_{\bar{P}_I}(h) = \hat{E}_{\bar{P}_I}(h; h_{Z_t'}) + \hat{E}_{\bar{P}_I}(h_{Z_t'})$, together with (11) this implies

\[ \hat{E}_{\bar{P}_I}(h) \leq 2C_0 \sqrt{2C_\beta (\hat{E}_{\bar{P}_I}(h))^\beta \varepsilon(n_I, \delta) + 10C_0^2 (C_\beta \varepsilon(n_I, \delta))^{1/(2-\beta)}}. \]

This inequality immediately implies the claimed inequality in the lemma. \qed
PROOF OF THEOREM 7. Fix any \( t \in [N + 1] \), and define \( n_t = \sum_{s=1}^{t} n_{(s)} \). For brevity, let \( \tilde{\rho} = \hat{\rho}_t \) and \( \hat{h} = \hat{h}_{Z^{(t)}} \). Let \( \tilde{P} = n_t^{-1} \sum_{s=1}^{t} n_{(s)} P_{(s)} \). Note that
\[
\mathcal{E}_{\tilde{P}}(\hat{h}) \geq \frac{1}{n_t} \sum_{s=1}^{t} n_{(s)} \left(C_{\rho}^{-1} \mathcal{E}_D(\hat{h})\right)_{\rho(s)} \geq \left(C_{\rho}^{-1} \mathcal{E}_D(\hat{h})\right)^{\tilde{\rho}},
\]
where the final inequality is due to Jensen’s inequality. In particular, this implies \( \mathcal{E}_D(\hat{h}) \leq C_{\rho} \mathcal{E}_{\tilde{P}}^{1/\tilde{\rho}}(\hat{h}) \), so that it suffices to upper bound \( \mathcal{E}_{\tilde{P}}(\hat{h}) \). Toward this end, note that \( \hat{h} \) trivially satisfies (7) for \( I = \{(1), \ldots, (t)\} \). Therefore, Lemma 2 implies that with probability at least \( 1 - \delta \), it holds that \( \mathcal{E}_{\tilde{P}}(\hat{h}) \leq 32 \bar{C}_0^2 (C_\beta \varepsilon(n_t, \delta))^{1/(2-\beta)} \), and the theorem follows. \( \square \)

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.

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 \( D \), whenever \( \beta = 1 \).

PROOF OF THEOREM 3. For any \( t \in [N + 1] \), define
\[ n_t = \sum_{s=1}^{t} n_{(s)} \] and let \( \tilde{P}_t = (n_t)^{-1} \sum_{s=1}^{t} n_{(s)} P_{(s)} \).

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, \( \mathcal{E}_{P_{N+1}}(\hat{h}_{Z}) \leq 32 \bar{C}_0^2 C_\beta \cdot \varepsilon(n_{N+1}, \delta) \).

Combining this with the definition of \( \rho_t \), we have
\[
\mathcal{E}_{\tilde{P}_t}(\hat{h}_{Z}) \leq 32 \bar{C}_0^2 C_\beta \cdot \varepsilon(n_{N+1}, \delta).
\]

Since the left hand side is monotonic in \( \mathcal{E}_D(\hat{h}_{Z}) \), we have that \( \mathcal{E}_D(\hat{h}_{Z}) \) is upper-bounded by any value of \( \varepsilon \) such that
\[
\sum_{t \in [N+1]} n_t \left(C_{\rho}^{-1} \mathcal{E}_D(\hat{h}_{Z})\right)^{\rho_t} \geq n_{N+1} \cdot 32 \bar{C}_0^2 C_\beta \cdot \varepsilon(n_{N+1}, \delta).
\]

In particular, let us take \( \varepsilon = \min_{t \in [N+1]} C_{\rho} \left(n_{N+1} \cdot 32 \bar{C}_0^2 C_\beta \varepsilon(n_{N+1}, \delta)\right)^{1/\rho_t} \) and let \( t^* \) denote the value of \( t \in [N + 1] \) achieving the minimum. Then, verify that, by Jensen’s,
\[
\sum_{t \in [N+1]} n_t \left(C_{\rho}^{-1} \mathcal{E}_D(\hat{h}_{Z})\right)^{\rho_t} \geq n_{t^*} \sum_{t=1}^{t^*} \frac{n_{(t)}}{n_{t^*}} \left(C_{\rho}^{-1} \mathcal{E}_D(\hat{h}_{Z})\right)^{\rho_{(t)}} \geq n_{t^*} \left(C_{\rho}^{-1} \mathcal{E}_D(\hat{h}_{Z})\right)^{\rho_{(t^*)}} = 32 \bar{C}_0^2 C_\beta n_{N+1} \varepsilon(n_{N+1}, \delta). \]

7.2. Available Ranking Information. We present here the proof of Theorem 4. It follows by arguing (based on Lemma 2) that any \( h \) in each \( \mathcal{H}_{(t)} \) has \( \mathcal{E}_D(\hat{h}) \) of order \( \left(\sum_{s \in \mathcal{I}} n_{(s)}\right)^{-1/2-\beta} \), so that the result would hold as long as \( \bigcap_{s=1}^{N+1} \mathcal{H}_{(s)} \) is nonempty; based on our Lemma 1 we find that \( h^* \) is likely contained in this set.
Proof of Theorem 4. For any \( t \in [N+1] \) let \( \tilde{P}_t = \mathbf{n}_t^{-1} \sum_{s=1}^{t} n_{(s)} P_{(s)} \). Let \( t^* \) be as in Theorem 7, and by the same argument from the proof of Theorem 7, \( \mathcal{E}_D(\hat{h}) \leq C_0 P_0^{1/\beta}(\tilde{h}) \) holds. Thus, the theorem will be proven if we can show \( \mathcal{E}_{\tilde{P}_{t^*}}(\hat{h}) \leq (c C_0 \beta^{-1} \varepsilon(n_{(s)}, \delta_{(s)}))^{1/(2 - \beta)} \) 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 \( S = \mathbb{Z}^{(t)} \) (and with \( \delta_t \) in place of \( \delta \) there). In particular, this implies that for every \( t \in [N+1] \),

\[
\hat{\mathcal{E}}_{Z(t)}(h^*; \hat{h}_{Z(t)}) \leq C_0 \sqrt{P_{Z(t)}(h^* \neq \hat{h}_{Z(t)})} \varepsilon(n_t, \delta_t) + C_0 \varepsilon(n_t, \delta_t),
\]

so that there 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 \( \mathcal{E}_{\tilde{P}_{t^*}}(\hat{h}) \geq 32 C_0^2 (C_0 \beta^{-1} \varepsilon(n_{(s)}, \delta_{(s)}))^{1/(2 - \beta)} \), which completes the proof (for instance, taking \( c = 2^{10} C_0^4 \)). \( \Box \)

8. Impossibility of Adaptivity over Multisource Classes \( \mathcal{M} \). Theorem 5 follows from Theorem 8, the main result of this section. In particular, the second part of Theorem 5 follows from the condition that \( \mathcal{M} \) contains enough tasks with \( \rho_t = 1 \), and calling on Theorem 4.

Theorem 8 (Impossibility of Adaptivity). Pick \( C_\rho = 3 \), and any \( 0 \leq \beta < 1 \), \( C_\beta \geq 2 \), a number of samples per source task \( 1 \leq n < 2/\beta - 1 \), and a number of target samples \( n_D \geq 0 \). 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 \( N_Q \geq C_0 \), and suppose \( N_P \) is sufficiently large so that \( N_P \geq 3 N_Q \), and furthermore \( N_P^{2-(n+1)/(2-\beta)} \geq C_1 \cdot N_Q^2 \times 2^{15n} \).

Let \( \mathcal{M} \) denote the family of multisource classes \( \mathcal{M} \) satisfying the above conditions, with parameters \( n_s = n, \forall t \in [N], n_{N+1} = n_D \), 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:

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

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

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

The result builds on the following construction. Omitted proofs are in Appendix E.

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 \( n \geq 1, n_D \geq 0, N_P, N_Q \geq 1, 0 \leq \beta < 1 \), and define \( \varepsilon = (n \cdot N_P)^{-1/(2 - \beta)} \) and \( \varepsilon_0 = 1 \wedge n_D^{-1/(2 - \beta)} \). Let \( \sigma \in \{\pm 1\} \) which we often abbreviate as \( \pm \). In what follows, let \( \eta_{\rho(X)} \) denote the function \( \mathbb{P}_{\rho} [Y = 1 \mid X] \) under distribution \( \mu \).

- **Target** \( \mathcal{D}_\sigma = \mathcal{D}_X \times \mathcal{D}_{Y \mid X}^\sigma \): Let \( \mathcal{D}_X(x_1) = \frac{1}{2} \cdot \varepsilon_0, \mathcal{D}_X(x_0) = 1 - \frac{1}{2} \cdot \varepsilon_0^\beta \); finally \( \mathcal{D}_{Y \mid X}^\sigma \) is determined by \( \eta_{\mathcal{D}_\sigma(X)}(x_1) = 1/2 + \sigma \cdot \varepsilon_0 \cdot 1^{-\beta}, \) and \( \eta_{\mathcal{D}_\sigma(X)}(x_0) = 1, \) for some \( \varepsilon_0 \) to be specified.
• **Noisy** $P_\sigma = P_X \times P_{Y|X}^\sigma$: Let $P_X(x_1) = c_1 e^\beta$, $P_X(x_0) = 1 - c_1 e^\beta$; finally $P_{Y|X}^\sigma$ is determined by $\eta_{P,\sigma}(x_1) = 1/2 + \sigma \cdot e^{-\beta}$, and $\eta_{P,\sigma}(x_0) = 1$, for an appropriate constant $c_1$.

• **Benign** $Q_\sigma = Q_X \times Q_{Y|X}^\sigma$: Let $Q_X(x_1) = 1$; finally, for $Q_{Y|X}^\sigma$, let $\eta_{Q,\sigma}(x_1) = 1/2 + \sigma/2$.

The following proposition follows from definitions, and is readily verified.

**Proposition 1 (Exponents).** $P_\sigma$ and $Q_\sigma$ have transfer-exponents $(3, \rho_P)$ and $(3, \rho_Q)$, respectively with respect to $D_\sigma$, with $\rho_P \geq \frac{\log(c_2(2^{(2-\beta)}))}{\log(c_3(1+\rho_D))}$ and $\rho_Q \leq 1$. Also, $P_\sigma, Q_\sigma, D_\sigma$ satisfy the Bernstein class condition with parameters $(C_\beta, \beta), C_\beta = \max\{(1/2)c_0^\beta, 2\}$.

**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 $Z_i$ in $Z$ with covariates at $x_1$, that is: $\hat{N}_\sigma(Z) = \sum_{t \in [N]} \mathbb{1}\{\forall i \in [n], X_{t,i} = x_1 \land Y_{t,i} = \sigma\}$. Next, let $\hat{n}_+$ and $\hat{n}_-$ denote the total number of $\pm 1$ labels in Z, over occurrences of $x_1$. That is: $\hat{n}_\sigma(Z) = \sum_{t \in [N], i \in [n]} \mathbb{1}\{X_{t,i} = x_1 \land Y_{t,i} = \sigma\}$. We then have (under the randomness of $Z \sim \Gamma_-$, and letting $D_\sigma(Z_{N+1}) = 1$ when $n_D = 0) \mathbb{P}\left(\frac{\Gamma_+(Z)}{\Gamma_-(Z)} > 1\right)$ is at least:

$$\mathbb{P}\left(\hat{N}_+(Z) > \hat{N}_-(Z) \land \hat{n}_+(Z) \geq \hat{n}_-(Z)\right) \cdot \mathbb{P}\left(\frac{D_+(Z_{N+1})}{D_-(Z_{N+1})} \geq 1\right).$$

**Remark 5 (Likelihood 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 hence, 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 $\Gamma_\sigma$ be defined in terms of a mixture – rather than as product of $P_\sigma$ and $Q_\sigma$ 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 \( Q_\sigma \) 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 \( P_\sigma \) and \( Q_\sigma \) vectors in \( Z \), and can therefore simply focus on \( Q_\sigma \) vectors to easily discover \( \sigma \).

We now further reduce the r.h.s. of (14) 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. Bernoulli’s we can exploit. Towards this end, we first consider the events of the following proposition.

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

Define the events (on \( Z \)):

\[
E_P = \{ \mathbb{E} \left[ \hat{N}_P \right] /2 \leq \hat{N}_P \leq 2\mathbb{E} \left[ \hat{N}_P \right] \} \quad \text{and} \quad E_Q = \{ \hat{N}_Q \leq 2\mathbb{E} \left[ \hat{N}_Q \right] \} .
\]

We have \( \mathbb{P}(E_P^c) \leq 2\exp \left( -\mathbb{E} \left[ \hat{N}_P \right] /8 \right) \) and \( \mathbb{P}(E_Q^c) \leq \exp \left( -\mathbb{E} \left[ \hat{N}_Q \right] /3 \right) \).

**Proof.** The proposition follows from multiplicative Chernoff bounds.

The expectations in the above events, for \( Z \sim \Gamma_\sigma \), are given by

\[
\mathbb{E} \left[ \hat{N}_P(Z) \right] = N_P P^0_{\pm} (x_1) (\eta_{P,\pm} (x_1) + \eta_{P,\pm} (x_1)) \quad \text{and} \quad \mathbb{E} \left[ \hat{N}_Q(Z) \right] = N_Q.
\]

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

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

(i) \( \mathbb{P}(\hat{N}_+ > \hat{N} \mid \hat{N}_P, \hat{N}_Q) \mathbb{1}\{E_P \cap E_Q\} \geq \delta_1 \cdot \mathbb{1}\{E_P \cap E_Q\} \).

(ii) \( \mathbb{P}(\hat{n}_+ > \hat{n}_- \mid \hat{n}_P, \hat{n}_Q) \geq \delta_2 \), further assuming that \( n > 1 \) so that \( \hat{n}_\pm \) are not both 0.

We then have that

\[
\mathbb{P}(\hat{N}_+ \geq \hat{N} \land \hat{n}_+ \geq \hat{n}_-) \geq \delta_1 \cdot (\delta_2 - \mathbb{P}(E_P^c) - \mathbb{P}(E_Q^c)) .
\]

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

\[
\mathbb{P}(\tilde{A} \cap B) = \mathbb{P}(\tilde{A} \mid \hat{N}_P, \hat{N}_Q, B) \mathbb{P}(B \mid \hat{N}_P, \hat{N}_Q) \geq \delta_1 \mathbb{E} \left[ \mathbb{P}(\tilde{A} \mid \hat{N}_P, \hat{N}_Q) \mathbb{1}\{E_P \cap E_Q\} \right] 
\]

\[
\geq \delta_1 \cdot \mathbb{E} \left[ \mathbb{P}(\tilde{A} \mid \hat{N}_P, \hat{N}_Q) - \mathbb{1}\{E_P^c \cup E_Q^c\} \right] \geq \delta_1 \cdot (\delta_2 - \mathbb{P}(E_P^c) - \mathbb{P}(E_Q^c)) .
\]

In the first equality, we used that since all vectors in \( Z \) are independent, \( \tilde{A} \) is independent of \( B \) given any value of \( \{ \hat{N}_P, \hat{N}_Q \} \); in the first inequality on the second line, we used that \( \forall p \in [0,1], p \cdot \mathbb{1}\{E\} = p - p \cdot \mathbb{1}\{E^c\} \geq p - \mathbb{1}\{E^c\} \). 

\( \square \)
The following is a well-known counterpart of Chernoff bounds, following from Slud’s inequalities (Slud, 1977) for Binomials. We include a brief proof of this form of the bound in Appendix E of the supplementary material.

**Lemma 3 (Anticoncentration (Slud’s Inequality)).** Let \( \{X_i\}_{i \in [m]} \) denote i.i.d. Bernoulli’s with parameter \( 0 < p \leq 1/2 \). For any \( 0 \leq m_0 \leq m(1 - 2p) \):

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

We make use of the above anti-concentration in much of what follows.

**Proposition 5 (\( \delta_1 \) from Proposition 4).** Pick any \( 0 \leq \beta < 1 \), \( 1 \leq n < 2/\beta - 1 \), and \( N_Q \geq 1 \). Let \( 0 < c_1 \leq 1/64 \) (construction of \( \eta_{P,a} \), \( \sigma = \pm \)). Let \( N_P \) 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 the event \( E = E_P \cap E_Q \) as defined in Proposition 3 for \( Z \sim \Gamma_- \). Using the notation of Proposition 4, we have \( \Pr \left( \hat{N}_+ > \hat{N}_- \mid \hat{N}_P, \hat{N}_Q \right) \geq \frac{c}{12} \cdot \mathbb{1}\{E\} \).

**Proof.** Consider \( N_H = \{ \hat{N}_P, \hat{N}_Q \} \) such that \( E = E_P \cap E_Q \) holds.

Let \( \tilde{B} \doteq \{ \hat{N}_+ > \frac{1}{2} N_P + \mathbb{E} \left[ \hat{N}_Q \right] \} \), and notice that \( \tilde{B} \subset \{ \hat{N}_+ > \hat{N}_- \} \) under \( E_Q \doteq \{ \mathbb{E} \left[ \hat{N}_Q \right] \geq N_Q/2 \} \). Therefore we only need to bound \( \Pr \left( \tilde{B} \mid N_H \right) = \Pr \left( \tilde{B} \mid \hat{N}_P \right) \).

Now, for \( \sigma = \pm \), let \( Z_{P,\sigma} \) denote the set of homogeneous vectors in \( \{ Z_t : \forall t \in [n], X_{ti}x_1 \land Y_{ti} = \sigma \} \) generated by \( P_- \), and note that, conditioned on \( \hat{N}_P \), we have \( \hat{N}_+ \sim \text{Binomial}(\hat{N}_P, p) \), where

\[
p = \Pr \left( Z_t \in Z_{P,+} \mid Z_t \in Z_{P,+} \cup Z_{P,-}, \hat{N}_P \right) = \frac{\eta_{P,-}(x_1)}{\eta_{P,+}(x_1) + \eta_{P,-}(x_1)}.
\]

Therefore, applying Lemma 3, \( \Pr \left( \hat{N}_+ > \hat{N}_- \cdot p + \sqrt{\hat{N}_- \cdot p(1-p)} \mid \hat{N}_P \right) \geq \frac{c}{12} \). We now just need to show that the event under the probability implies \( \tilde{B} \), in other words, that

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

(17)

Next we upper bound the r.h.s of (17) and lower-bound its l.h.s. Under \( E_P \) and using (15):

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

(18)

where we used the fact that, for \( a \geq b > 0 \), we have \( a^n - b^n = (a - b) \sum_{k=0}^{n-1} a^{n-1} \cdot (\frac{b}{a})^k \).

Now, by conditions on \( N_P, \eta_{P,-}(x_1) \geq 1/4 \), and we have that \( \hat{N}_P \cdot p(1-p) \) is at least

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

Combining (18) and (19), we see that \( \tilde{B} \) is implied by

\[
2^{-(2n)} \sqrt{N_P P_X^n(x_1)} \geq N_P P_X^n(x_1) \cdot n \cdot (\eta_{P,+}(x_1) - \eta_{P,-}(x_1)) + N_Q.
\]
which is in turn implied by
\[
\begin{aligned}
n^{-1}2^{-(4n)\epsilon_1^n} (n \cdot N_P)^{\frac{2-(n+1)\sigma}{2-\sigma}} &\geq 2c_1^{2n} (n \cdot N_P)^{\frac{2-2\sigma n}{2-\sigma}} + 2N_Q^2.
\end{aligned}
\]
The conditions of the proposition ensure that the above inequality holds. \(\square\)

We obtain a bound on \(\mathbb{P}(N_+ > \hat{N}_-\) as a corollary.

**Corollary 1.** *Under the conditions of Proposition 5, we have:
\[
\mathbb{P}(\hat{N}_+ > \hat{N}_-) \geq \mathbb{E}[\mathbb{P}(\hat{N}_+ > \hat{N}_- | N_P, \hat{N}_Q) \mathbb{I}(E)] \geq \frac{1}{12}\mathbb{P}(E).
\]

We now turn to bounding \(\delta_2\) of Proposition 4.

**Proposition 6 (\(\delta_2\) 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 \(\hat{N}_P\) is sufficiently large so that
\[
(i) \quad N_P \geq 3N_Q \quad \text{and} \quad (ii) \quad (n \cdot N_P)^{\frac{2-2\sigma}{2-\sigma}} \geq 4096 \cdot n^2 \cdot c_1^{-1}.
\]

*Let \(Z \sim \Gamma_-, \) and \(\hat{n}_\sigma = \hat{n}_\sigma(Z), \sigma = \pm\) as in Proposition 4. We have \(\mathbb{P}(\hat{n}_+ > \hat{n}_-) \geq \frac{1}{18}\).

**Proof.** Under the notation of Proposition 4, let \(N_{P,-} \equiv \hat{N}_P - \hat{N}_+, \) and for homogeneity of notation herein, let \(N_{P,+} = \hat{N}_+.\) Fix \(\delta > 0\) to be defined, and notice that if \(\delta > n \cdot (N_{P,+} - N_{P,-})\), then the event
\[
\hat{A}_\delta \equiv \{\hat{n}_+ + n \cdot N_{P,+} \geq \hat{n}_- + n \cdot N_{P,-} + \delta\} \text{ implies } \{\hat{n}_+ > \hat{n}_-\}.
\]
As a first step, we want to upper-bound \((N_{P,+} - N_{P,-})\). Let \(V_\delta\) 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 \(1 - e^{-t^2/4}\),
\[
(N_{P,+} - N_{P,-}) \leq 4n \cdot \sqrt{V_\delta}, \quad \text{whereby, for } \sqrt{V_\delta} \geq 4, \text{ the event of (20) (with } t = 4) \text{ happens with probability at least } 1 - 1/48.
\]

Hence, we set \(V_\delta = 16 \vee \mathbb{E}[N_{P,+} - N_{P,-}] \geq 16 \vee \text{Var}(N_{P,+} - N_{P,-}), \) where \(\mathbb{E}[N_{P,+} + N_{P,-}] = \mathbb{E}[\hat{N}_P]\) is given in equation (15). We now proceed with a lower-bound on \(\mathbb{P}(\hat{A}_\delta)\).

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, \(n_+ \equiv \hat{n}_+ + n \cdot N_{P,+}\) 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 \(P_-\). From Lemma 3, and integrating over \(\hat{N}_Q\):
\[
\mathbb{P}(\hat{n}_+ > \hat{n}_{x_1}, p + \sqrt{\hat{n}_{x_1}} \cdot p(1-p)) \geq \frac{1}{12}.
\]

Now notice that \(\hat{A}_\delta\) holds whenever \(\hat{n}_+ \geq \frac{1}{2} (\hat{n}_{x_1} + \delta)\), since \(\hat{n}_{x_1} = (\hat{n}_+ + n \cdot N_{P,+}) + (\hat{n}_- + n \cdot N_{P,-})\). Under the event of (21), we have \(\hat{n}_+ > \frac{1}{2} (\hat{n}_{x_1} + \delta)\), whenever it holds that
\[
\hat{n}_{x_1} \cdot p(1-p) \geq \frac{1}{2} (\hat{n}_{x_1} + \delta) - \hat{n}_{x_1} \cdot p = \hat{n}_{x_1} \left(\frac{1}{2} - p\right) + \frac{1}{2} \delta.
\]
Next we bound \(\hat{n}_{x_1}\) with high probability. Consider any value of \(\hat{N}_Q\) such that \(E_Q\) (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
\[
\mathbb{E}[\hat{n}_{x_1} | \hat{N}_Q] = n(N - \hat{N}_Q) \cdot P_X(x_1) \geq \frac{1}{2} n \cdot N_P \cdot P_X(x_1) = \frac{1}{2} c_1 (n \cdot N_P)^{\frac{2-2\sigma}{2-\sigma}}.
\]
where for the first inequality we used the fact that \( N_P \geq 3N_Q \). Hence, by a multiplicative Chernoff bound,
\[
P \left( \frac{1}{2}E [\hat{N}_x | \hat{N}_Q] \right. \leq \hat{n}_x, \leq 2E [\hat{N}_x | \hat{N}_Q] \left. \right) \geq \left( 1 - \frac{1}{48} \right) 1_{\{E_Q\}},
\]
whenever \( c_1 (n \cdot N_P)^{\frac{2-\beta}{2-\sigma}} \geq 40 \). Now, by Proposition 3, \( E_Q \) holds with probability at least \( 1 - 1/48 \) whenever \( E [\hat{N}_Q] = N_Q > 16 \). Thus, integrating over \( \hat{N}_Q \), we get that, with probability at least \( 1 - 1/24 \),
\[
(23) \quad \frac{1}{4} n \cdot N_P \cdot P_X(x_1) \leq \hat{n}_x, \leq 2n \cdot N \cdot P_X(x_1) \leq \frac{8}{3} n \cdot N_P \cdot P_X(x_1).
\]

Thus, bounding both sides of (22), \( \hat{A}_\delta \) holds whenever a) the events of (21) and (23) hold, and b) the following inequality is satisfied:
\[
\sqrt{\frac{1}{4} n \cdot N_P \cdot P_X(x_1) \cdot p(1-p)} \geq \frac{8}{3} n \cdot N_P \cdot P_X(x_1) \left( \frac{1}{2} - p \right) + 2n \cdot \sqrt{V_\delta}
\]
which holds whenever
\[
(24) \quad \frac{1}{4} \sqrt{c_1} \frac{1}{2} (n \cdot N_P)^{\frac{1-\beta}{2-\sigma}} \geq \frac{8}{3} c_1 (n \cdot N_P)^{\frac{1-\beta}{2-\sigma}} + 2n \cdot \left( 4 \vee n^{-1/2} c_1^{n/2} \right) \cdot \frac{n \cdot N_P^{1-(n+1)\beta/2}}{\sqrt{2-\sigma}}.
\]

Finally we bound the \( D_{\pm} \) term in the likelihood equation (14).

**Proposition 7 (\( D_+ / D_- \)).** Let \( n_D > 0 \). Again let \( Z \sim \Gamma_\sigma \), and let \( 0 < c_0 \leq 1/4 \) in the construction of \( \eta_{D,\sigma}, \sigma = \pm \).
\[
P \left( \left| \frac{D_+(Z_{N+1})}{D_-(Z_{N+1})} \right| \geq 1 \right) \geq \frac{1}{84}.
\]

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

We can now combine all the above analysis into the following 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 \( \eta_{P,\sigma}, \eta_{D,\sigma}, \sigma = \pm \). Suppose \( N_P \) is sufficiently large so that \( N_P \geq 3N_Q \), and also

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

(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 \( Z \sim \Gamma_\sigma, \sigma = \pm \):
\[
\inf_{\hat{h}} \sup_{\sigma \in \{\pm\}} P_{T,\sigma} (\mathcal{E}_{D}(\hat{h}) \geq c_0 \cdot \left( 1 \wedge n_D^{-1/(2-\beta)} \right)) \geq \frac{1}{12} \cdot \frac{1}{96} \cdot \frac{1}{84}.
\]
PROOF. Following the above propositions, again assume w.l.o.g. that $Z \sim \Gamma_\ast$. Let $E_P, E_Q$ as defined in Proposition 3 over $Z \sim \Gamma_\ast$, 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 $n > 1$, we have that $\mathbb{P}_{\Gamma_{-}}(\Gamma_{+}(Z) > \Gamma_{-}(Z))$ is at least $\delta_1 (\delta_2 - 1/96) \frac{1}{12}$ by Propositions 2, 4, and 7. 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 $\frac{1}{12} \mathbb{P}(E_P \cap E_Q) \cdot \frac{1}{84}$, 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 $\mathcal{D}_{\mathcal{H}}(\hat{h}) \geq \mathcal{D}_X(x_1) \cdot (\eta_{D, +} - \eta_{D, -}) = c_0 \epsilon_0$. By Remark 5, for any $\hat{h}$, the probability that $\hat{h}$ picks $\sigma = +$ is bounded below by $\mathbb{P}_{\Gamma_{-}}(\Gamma_{+}(Z) > \Gamma_{-}(Z))$. \hfill \Box

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 $E_{Q, \hat{d}}$ denote the event that $\hat{N}_Q \geq \mathbb{E}[\hat{N}_Q]/2 = N_Q/2$. Let $E_c = \left\{ \mathcal{D}_{\mathcal{H}}(\hat{h}) \geq c_0 \cdot \epsilon_0 \right\}$. By Proposition 8, for some $\sigma \in \{\pm\}$, we have that $\mathbb{P}(E_c)$ is bounded below.

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

$$E[\mathbb{P}(E_c | \zeta) | E_{Q, \hat{d}}] \geq E[\mathbb{P}(E_c | \zeta) \cdot 1\{E_{Q, \hat{d}}\}]$$

$$\geq E[\mathbb{P}(E_c | \zeta)] - \mathbb{P}(E_{Q, \hat{d}}) = \mathbb{P}(E_c) - \mathbb{P}(E_{Q, \hat{d}}) \geq c,$$

where we can bound $\mathbb{P}(E_{Q, \hat{d}})$ however small for $N_Q$ sufficiently large (by a multiplicative Chernoff). Now conclude by noting that the above conditional expectation is a projection of the measure $\alpha_N$ onto $\mathcal{M}$ (via the injection $\zeta \mapsto \Pi \in \mathcal{M}$) and is bounded below, implying $\sup_{\zeta \in \mathcal{M}} \mathbb{P}(E_c | \zeta)$ must be bounded below. The second part of the theorem follows from results of Section 7. \hfill \Box

9. Final Remarks and Open Questions. Our main focus in this work was to show that minimax adaptive rates – aggregating dataset sizes – are not possible in multitask, even in apparently easy situations where a non-adaptive learner, i.e., one with partial information on distributions, achieves fast rates in terms of aggregate sample sizes.

However, our result on impossibility of adaptivity relied on a construction of distribution space satisfying certain critical conditions, e.g., that the number of samples per task $n < \frac{\beta}{2} - 1$, which is restrictive outside of the commonly studied case of $\beta = 0$ (no noise condition). While it is clear that some such condition relating $n$ to $\beta$ is required for non-adaptivity – since near-minimax adaptivity is always possible for $\beta = 1$ as per Theorem 3 – it is left open whether these conditions can be further relaxed. In particular, it might be the case that adaptivity is in fact possible for $n$ sufficiently large – i.e., with $n \gg \frac{\beta}{2} - 1$ – or at the very least, that the gap between minimax and adaptive rates tightens with larger $n$.

Furthermore, as stated above, $\beta = 1$ is a regime where near-minimax adaptivity is always possible; this leaves open whether a more refined parametrization of the problem in terms of different noise conditions $\beta_t$, for every task $t \in [N+1]$, might reveal a richer picture on adaptivity. Unfortunately this appears as a hard question: the obvious extensions of our proof techniques to this problem do not yield a complete picture, as for instance, they result in gaps between upper and lower-bounds on the rates as functions of $\beta_t$’s (see Appendix G).
REFERENCES

ACHILLE, A., PAOLINI, G., MBENG, G. and SOATTO, S. (2019). The information complexity of learning tasks, their structure and their distance. arXiv:1904.03292.

ANDO, R. K. and ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6 1817–1853.

ANTHONY, M. and BARTLETT, P. (1999). Learning in Neural Networks: Theoretical Foundations.

ARJOVSKY, M., BOTTOU, L., GULRAJANI, I. and LOPEZ-PAZ, D. (2019). Invariant risk minimization. arXiv:1907.02893.

ARORA, S., KHANDEPARKAR, H., KHODAK, M., PLEVRAKIS, O. and SAUNSHI, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. arXiv:1902.09229.

BARTLETT, P. L. (1992). Learning with a slowly changing distribution. In Proceedings of the 5th Annual Workshop on Computational Learning Theory.

BARTLETT, P., JORDAN, M. I. and McCULIFFE, J. (2006). Convexity, Classification, and Risk Bounds. Journal of the American Statistical Association 101 138–156.

BARVE, R. D. and LONG, P. M. (1997). On the complexity of learning from drifting distributions. Information and Computation 138 170–193.

BAXTER, J. (1997). A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling. Machine Learning 28 7–39.

BEN-David, S., BENEDEK, G. M. and MANSOUR, Y. (1989). A parametrization scheme for classifying models of learnability. In Proceedings of the 2nd Annual Workshop on Computational Learning Theory.

BEN-David, S. and BORBELY, R. S. (2008). A notion of task relatedness yielding provable multiple-task learning guarantees. Machine Learning 73 273–287.

BEN-David, S., BLITZER, J., CRAMMER, K. and PEREIRA, F. (2007). Analysis of representations for domain adaptation. In Advances in Neural Information Processing Systems.

BEN-David, S., BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F. and VAUGHAN, J. W. (2010a). A theory of learning from different domains. Machine Learning 79 151–175.

BEN-David, S., LU, T., LIU, T. and PAL, D. (2010b). Impossibility theorems for domain adaptation. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics.

BLUM, A., HAGHTALAB, N., PROCACCIA, A. D. and QIAO, M. (2017). Collaborative PAC learning. In Advances in Neural Information Processing Systems.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford university press.

CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. The Annals of Statistics 49 100–128.

CARUANA, R. (1997). Multitask Learning. Machine Learning 28 41–75.

CORTES, C., MOHRI, M., RILEY, M. and ROSTAMIZADEH, A. (2008). Sample selection bias correction theory. In International Conference on Algorithmic Learning Theory.

CRAMMER, K., KERNS, M. and WORTMAN, J. (2008). Learning from multiple sources. Journal of Machine Learning Research 9 1757–1774.

DU, S. S., HU, W., KAKADE, S. M., LEE, J. D. and LEI, Q. (2020). Few-shot learning via learning the representation, provably. arXiv:2002.09434.

GRETTON, A., SMOLA, A., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. and SCHÖLkopf, B. (2009). Covariate shift by kernel mean matching. In Dataset Shift in Machine Learning 131–160.

HANNEKE, S. and KPUTFE, S. (2019). On the Value of Target Data in Transfer Learning. In Advances in Neural Information Processing Systems.

HANNEKE, S. and YANG, L. (2019). Statistical Learning under Nonstationary Mixing Processes. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics.

JALALI, A., SANGHAVI, S., RUAN, C. and RAVIKUMAR, P. (2010). A dirty model for multi-task learning. In Advances in Neural Information Processing Systems.

KLEIN, T. and RIO, E. (2005). Concentration around the mean for maxima of empirical processes. The Annals of Probability 33 1060–1077.

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

KOLTCHINSKII, V. (2011). Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d’Été de Probabilités de Saint-Flour XXXVIII-2008 2033.

KONSTANTINOV, N., FRANTAR, E., ALISTAIR, D. and LAMBERT, C. H. (2020). On the Sample Complexity of Adversarial Multi-Source PAC Learning. arXiv:2002.10384.

KPUTFE, S. and MARTINET, G. (2018). Marginal Singularity, and the Benefits of Labels in Covariate-Shift. arXiv:1805.01833.
LOUNICI, K., PONTIL, M., VAN DE GEER, S., TSYBAKOV, A. B. et al. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics* 39 2164–2204.

MAHLOUFI-FAR, S., MAHMOODY, M. and MOHAMMED, A. (2019). Universal multi-party poisoning attacks. In *Proceedings of the 36th International Conference on Machine Learning*.

MANSOUR, Y., MOHRI, M. and ROSTAMI-ZADEH, A. (2009). Multiple source adaptation and the Rényi divergence. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*.

MÁSSART, P. and NÉDÉLEC, E. (2006). Risk bounds for statistical learning. *The Annals of Statistics* 34 2326–2366.

MAURER, A., PONTIL, M. and ROMERA-PAREDES, B. (2013). Sparse coding for multitask and transfer learning. In *International Conference on Machine Learning*.

MAURER, A., PONTIL, M. and ROMERA-PAREDES, B. (2016). The benefit of multitask representation learning. *The Journal of Machine Learning Research* 17 2853–2884.

MCNAMARA, D. and BALCAN, M.-F. (2017). Risk bounds for transferring representations with and without fine-tuning. In *International Conference on Machine Learning*.

MOHRI, M. and MEDINA, A. M. (2012). New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*.

MOUSAVI, N. (2010). How tight is Chernoff bound? [https://ece.uwaterloo.ca/~nmousavi/Papers/Chernoff-Tightness.pdf](https://ece.uwaterloo.ca/~nmousavi/Papers/Chernoff-Tightness.pdf).

MUANDET, K., BALDUZZI, D. and SCHÖLKOPF, B. (2013). Domain generalization via invariant feature representation. In *International Conference on Machine Learning*.

NEGABAN, S. N. and WAINWRIGHT, M. J. (2011). Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block $\ell_q/\ell_\infty$-Regularization. *IEEE Transactions on Information Theory* 57 3841-3863.

PENTINA, A. and LAMPERT, C. (2014). A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*.

QIAO, M. (2018). Do Outliers Ruin Collaboration? *arXiv:1805.04720*.

REDKO, I., HABRARD, A. and SEBBAN, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

REEVE, H. W. J., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Adaptive transfer learning. *The Annals of Statistics To Appear*.

SAUER, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory, Series A* 13 145–147.

SCOTT, C. and ZHANG, J. (2019). Learning from Multiple Corrupted Sources, with Application to Learning from Label Proportions. *arXiv:1910.04665*.

SHEN, J., QU, Y., ZHANG, W. and YU, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *32nd AAAI Conference on Artificial Intelligence*.

SLUD, E. V. (1977). Distribution inequalities for the binomial law. *The Annals of Probability* 404–412.

TATE, R. F. (1953). On a double inequality of the normal distribution. *The Annals of Mathematical Statistics* 24 132–134.

TRIPURANENI, N., JORDAN, M. I. and JIN, C. (2020). On the Theory of Transfer Learning: The Importance of Task Diversity. *arXiv:2006.11650*.

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* 32 135–166.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

VAN ERVEN, T., GRÜNWALD, P., MEHTA, N., REID, M. and WILLIAMSON, R. (2015). Fast rates in statistical and online learning. *Journal of Machine Learning Research* 16 1793–1861.

VAPNIK, V. and CHERVONENKIS, A. (1971). On the uniform convergence of relative frequencies of events to their expectation. *Theory of Probability and its Applications* 16 264-280.

YANG, L., HANNEKE, S. and CARBONELL, J. (2013). A theory of transfer learning with applications to active learning. *Machine learning* 90 161–189.