On Some Problems in Transfer Learning

Nicholas R. Galbraith

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024
Abstract
On Some Problems in Transfer Learning
Nicholas R. Galbraith

This thesis consists of studies of two important problems in transfer learning: binary classification under covariate-shift transfer, and off-policy evaluation in reinforcement learning. First, the problem of binary classification under covariate shift is considered, for which the first efficient procedure for optimal pruning of a dyadic classification tree is presented, where optimality is derived with respect to a notion of average discrepancy between the shifted marginal distributions of source and target. Further, it is demonstrated that the procedure is adaptive to the discrepancy between marginal distributions in a neighbourhood of the decision boundary. It is shown how this notion of average discrepancy can be viewed as a measure of relative dimension between distributions, as it relates to existing notions of information such as the Minkowski and Renyi dimensions. Experiments are carried out on real data to verify the efficacy of the pruning procedure as compared to other baseline methods for pruning under transfer.

The problem of off-policy evaluation for reinforcement learning is then considered, where two minimax lower bounds for the mean-square error of off-policy evaluation under Markov decision processes are derived. The first of these gives a non-asymptotic lower bound for OPE in finite state and action spaces over a model in which the mean reward is perturbed arbitrarily (up to a given magnitude) that depends on an average weighted chi-square divergence between the behaviour and target policies. The second provides an asymptotic lower bound for OPE in continuous state-space when the mean reward and policy ratio functions lie in a certain
smoothness class. Finally, the results of a study that purported to have derived a policy for sepsis treatment in ICUs are replicated and shown to suffer from excessive variance and therefore to be unreliable; our lower bound is computed and used as evidence that reliable off-policy estimation from this data would have required a great deal more samples than were available.
# Table of Contents

Acknowledgments .................................................. vii

Chapter 1: Introduction .............................................. 1

1.1 Covariate Shift .................................................. 2

1.1.1 Binary Classification Under Transfer ....................... 3

1.1.2 Off-Policy Evaluation ........................................ 4

Chapter 2: Classification Tree Pruning Under Covariate Shift ......... 6

2.1 Introduction .................................................... 6

2.1.1 Background and Related Work ................................ 8

2.1.2 Outline ....................................................... 10

2.2 Problem Setup and Notation .................................... 10

2.2.1 Assumptions on the Regression Function ..................... 12

2.2.2 Relating the Source to the Target ............................ 12

2.3 Overview of Results ............................................ 15

2.3.1 Tree-Based Classification ................................... 16

2.3.2 Oracle Rates ................................................ 17

2.3.3 Adaptive Rates .............................................. 18

2.3.4 Computational Considerations ............................... 21
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.5</td>
<td>Main Result: Localizing to Decision Boundary</td>
<td>21</td>
</tr>
<tr>
<td>2.4</td>
<td>Relation to Notions of Dimension</td>
<td>24</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Relation to transfer exponent</td>
<td>26</td>
</tr>
<tr>
<td>2.5</td>
<td>Analysis Details</td>
<td>30</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Main Tools</td>
<td>30</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Overview of Proofs</td>
<td>32</td>
</tr>
<tr>
<td>2.6</td>
<td>Experiments On Real Data</td>
<td>46</td>
</tr>
<tr>
<td>2.6.1</td>
<td>Datasets</td>
<td>46</td>
</tr>
<tr>
<td>2.6.2</td>
<td>Results</td>
<td>47</td>
</tr>
<tr>
<td>2.7</td>
<td>Appendix to Chapter 2</td>
<td>48</td>
</tr>
<tr>
<td>2.7.1</td>
<td>Properties of the Aggregate Transfer Exponent</td>
<td>48</td>
</tr>
<tr>
<td>2.7.2</td>
<td>Details for Figure 2.2</td>
<td>52</td>
</tr>
<tr>
<td>2.7.3</td>
<td>Calculations for Example 2</td>
<td>53</td>
</tr>
<tr>
<td>2.7.4</td>
<td>Proofs of Auxiliary Lemmas</td>
<td>55</td>
</tr>
<tr>
<td>2.7.5</td>
<td>Supplemental Experiments</td>
<td>63</td>
</tr>
<tr>
<td>2.7.6</td>
<td>A Pathological Example</td>
<td>66</td>
</tr>
</tbody>
</table>

Chapter 3: On the Fundamental Limits of Off-Policy Evaluation
Some Results, Open Questions, and Future Work | 70   |

3.1 Introduction | 70   |

3.2 A Real-World Case Study | 71   |
| 3.2.1 Data, Methodology, and Results of Original Study | 72   |
| 3.2.2 Evaluating the AI-Clinician using Alternative Methods | 74   |
| 3.2.3 Estimating a Lower Bound | 77   |
3.3 A Local Minimax Lower Bound. ........................................... 77
  3.3.1 Proof of Proposition 36. ........................................... 81
  3.3.2 Proof of Proposition 37. ........................................... 84
  3.3.3 Auxiliary Lemmas. .................................................. 91
3.4 A Lower Bound over Smoothness Classes. ................................. 92
  3.4.1 Problem Setup. ...................................................... 92
  3.4.2 Result. ............................................................. 93
  3.4.3 Proof. ............................................................. 93
3.5 Open Questions and Further Work. ....................................... 98
  3.5.1 Discussion of Results. ............................................. 98
  3.5.2 Further Work. ..................................................... 100
  3.5.3 Other Open Questions. ........................................... 100
3.6 Appendix to Chapter 3. .................................................. 102
  3.6.1 Kallus and Uehara’s Double Machine Learning Method. .......... 102

Conclusion or Epilogue ...................................................... 103

References ................................................................. 105
List of Figures

2.1 Two source densities $P_X$ are shown, which yield equal transfer exponents w.r.t $Q_X = U([0, 1]^2)$, despite the fact that the right-hand source density vanishes over a larger region of high target mass. This is captured by the aggregate transfer exponent: while both densities have transfer exponent $\rho = 3$ with respect to $Q_X$, the left-hand density has an aggregate exponent of $\gamma = \rho$, while the right-hand density has an aggregate exponent of $\gamma = \rho + 1$. The effects of this distinction on transfer will be felt most strongly when, as is the case here, the density ratios vanish along the decision boundary (shown in dark blue). This will be elucidated in Section 2.3.5 when we introduce a notion of the aggregate exponent localized to the decision boundary. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

2.2 Risk estimates for a regular dyadic tree, with depth selected by 2-fold cross-validation over 100 target samples ($n_Q = 100$), for datasets generated according to the distributions outlined in Appendix 2.7.2; we use 1000 source samples ($n_P = 1000$). Error bars give the standard deviation over 10 runs. The Bayes risk is 0.18. Top: empirical error curves for five distributions with varying aggregate transfer exponents but identical exponents; the construction of the marginal distributions mirrors the construction from Figure 2.1. Bottom: empirical error curves for five distributions with varying transfer exponents, but with identical aggregate exponents. Again the construction from Figure 2.1 is used, except in this case the singularity strengths $\nu$ decrease as they occur over larger regions, so as to keep the aggregate transfer exponents identical across the five distributions. Taken together, these two simulations support our claim that the aggregate exponent more adequately captures the difficulty of transferring between distributions. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
2.3 Target risk estimates for four tree-pruning methods used on dyadic trees applied respectively to the Algerian Forest Fire data \((d = 11)\), the MAGIC data \((d = 10)\), the SUSY data \((d = 18)\), and the Biodegradation data \((d = 41)\). Error bars show the standard errors over 10 iterations (except on the Forest Fire data, where we use 50 iterations). AD (blue) indicates the adaptive method, CV (yellow) is 2-fold cross validation performed by taking the hold-out risk on target samples only, FCV (green) is 2-fold cross validation performed by taking the hold-out risk on the combined sample, and IWCV is the importance weighted CV method, where the hold-out risk is taken over all points, and at source points it is weighted by the density ratio of the features at those points. LOO denotes leave-one-out cross-validation. 

2.4 When the support of the target distribution lies on a lower-dimensional manifold in \([0, 1]^D\), the dyadic partition of the latter can lead to an irregular partition of the former; this problem is mitigated by taking envelopes. Above we see a dyadic grid, along with \(X_Q\) in blue. One cell from the partition is in red, and its envelope in light red. The black boxes centred on \(x\) are \(B(x, r)\) and \(B(x, 2r)\), demonstrate that the envelope \(\tilde{A}(x)\) satisfies the conditions for being a grid element (see 7). 

2.5 Risk estimates for our adaptive method (AD), compared to the baseline cross-validation and the pruning method of [18] (SN, SN-Q), applied to the Algerian Forest Fire data (left) and the MAGIC data (right). Error bars show standard errors over 50 and 20 iterations, respectively. 

2.6 Target risk estimates for four tree-depth selection methods applied to CART trees for the Wine Quality dataset \((d = 11)\); left), and the Steel Plates Faults dataset \((d = 27)\); right). Error bars show the standard errors over 10 iterations. 

2.7 Illustration of the support of the measures \(P, Q\) described above on the unit square \([0, 1]^2\). 

3.1 OPE estimates produced by respectively, DM, DR, MIS, and WMIS for our replication of the ‘AI-clinician’ of [51]. Box-and-whisker plots are shown over the bootstrap distribution; \(n_{boot} = 1000\). 

3.2 An estimate of the lower bound on the mean-square estimation error (see Theorem 35) for the target policies \(\pi_\alpha^e = \alpha \pi^e + (1 - \alpha) \pi^b\) as a function of \(\alpha\). For values of \(\alpha\) close to 1, the estimated lower bound is itself subject to large variance; we have omitted this display.
List of Tables

2.1 Datasets. ................................................. 69
Acknowledgements

I should like to first acknowledge my advisors - Professors Samory Kpotufe and Hongseok Namkoong - without whose guidance, support, advice, patience, wisdom, and insights, I would not have been able to complete this work. More generally, I would like to thank all those who contributed to making my time at Columbia so enriching: all of the staff, students, and faculty of the Statistics department; my thanks is also extended to the visitors to the department who shared in and contributed to the culture - speakers, visiting students and faculty, etc. I recall with particular fondness the countless hours spent engaged in discussion on myriad matters around the table on the 10th floor of SSW - to those who participated in making these hours so edifying, I thank you all.
Chapter 1: Introduction

Recent years have witnessed extraordinary advances, both technological and methodological, that have greatly increased the ubiquity of statistical models in routine everyday use. This proliferation has proceeded with astonishing rapidity, and the harnessing of such developments has brought immeasurable utility to millions; however, deleterious effects have made themselves felt as well. Many diverse systems are now automated to make use of the outputs of such statistical models, often on the basis of user-specific information or input. Such systems are (habitually) incapable of distinguishing whether the subject in question is represented by the population on which the underlying model was built - this has traditionally largely been taken for granted, both in theory and in practice. It has become increasingly clear in recent years that this is no longer an admissible assumption in many cases. Datasets are now frequently so large that even subtle distributional mismatch between training and testing populations can induce serious deterioration in performance [1], while models are deployed so widely that differing testing populations become unavoidable. Examples of what can go wrong when these issues are not properly addressed abound - the infamous case of FB’s early human image recognition software comes to mind [2].

Broadly speaking, the efforts to characterize and detect the presence of factors that might cause a method or model to perform more poorly than one might expect from a superficial assessment, with the goal of addressing such shortcomings and preventing harmful effects down the line (that is, of producing results that are dependable and robust), is a sub-discipline known as reliable learning, which in particular encompasses transfer learning (or, domain adaptation) - the subject of this dissertation. Formally, in transfer learning, one is faced with the following setting: there is a reference, or source, distribution $P$, from which data has been collected, and a $^1$ target distribution

---

$^1$Some works consider extensions in which there may be multiple source or target distributions. We will largely stick to the case in which $P$ and $Q$ are single distributions.
$Q$ on which one wishes to eventually deploy a model, perform inference, etc.; typically, data from $Q$ is either unavailable or quite costly to obtain. Naturally, there are myriad problem formulations that correspond to this general paradigm, variations of which have led to large transfer learning literature addressing diverse questions under various sets of assumptions. In the following section, we give a formal presentation of the problem setup under which we have worked, as well as a brief introduction to the two particular problems (namely: binary classification under transfer, and off-policy evaluation in reinforcement learning) that we have studied. Notation will be introduced as necessary.

1.1 Covariate Shift

Let us suppose that we are in the transfer setting, and that we have a source distribution $P$ over the data $(Y, X)$, where we have the outcome $Y \in \mathcal{Y}$ and features $X \in \mathcal{X}$ for some spaces $\mathcal{Y}, \mathcal{X}$. Obviously, in order to progress, we must place some assumptions on the target distribution $Q$ if anything is to be learned from $P$. A common assumption that is very natural in many transfer learning problems is that of covariate shift [3], which, as the name implies, assumes that only the marginal feature distributions, denoted respectively $P_X$ and $Q_X$, change$^2$. The conditional distributions of the outcomes given the features are unchanged: the distributions $P_{Y|X}$ and $Q_{Y|X}$ are identical. This greatly simplifies the problem, because the statistician can safely rely on the information he has gleaned from source samples about the feature-outcome relationship and be wary only of biases induced by the covariate shift. Common examples in which covariate shift is often taken to hold include medical diagnostics and image classification. In the former, patient characteristics might vary across populations, but the underlying mechanisms influencing the dependent variable are unvarying; in the latter, it might be reasonable to suppose, say, that different ‘objects’ occur with variable frequencies across populations, without their actual nature changing (a rocking chair remains a rocking chair irrespective of whether it is observed on an Alabama porch or in a Manhattan high-rise, although one supposes it to be more characteristic of the former).

$^2$Subscripts will be used to denote derived distributions of $P, Q$, so that, for example, the conditional distribution of the outcomes given the features, under $P$, is denoted $P_{Y|X}$.
Due simultaneously to the fact that it is often reasonable to assume and renders analysis more tractable, much of the theoretical transfer learning literature has focused on the covariate shift setting. Additionally, as we will see, many problems that are not explicitly thought of as being transfer learning problems can nonetheless be comprised within the covariate shift framework, including the problem of off-policy evaluation in reinforcement learning [4]. We will now briefly introduce the two problems that we have studied.

1.1.1 Binary Classification Under Transfer

Binary classification is, of course, a canonical problem in statistics/machine learning. Traditionally, one supposes that one has access to a sample \((X, Y)_{n} = (X_1, Y_1), \ldots, (X_n, Y_n) \sim P\), the goal being to learn a function \(f : \mathcal{X} \to \mathcal{Y}\) that minimizes the risk for a new independent data-point \((X, Y) \sim P\). We consider the scenario in which we have the access to a combined sample \((X, Y)_{nP} \sim P\) and \((X, Y)_{nQ} \sim Q\) (possibly \(n_Q = 0\)), where the goal is now to find a classifier minimizing the risk for a new independent sample \((X, Y) \sim Q\). The literature on this problem is growing rapidly; noteworthy contributions include, for example, [5, 3, 6, 7]. Building on these ideas, in Chapter 2 we will present some results for the problem of binary classification under transfer; this Chapter is joint work with Prof. Samory Kpotufe. We make the following contributions:

i) We introduce a quantity - that we have termed an *aggregate transfer exponent* that serves in some sense to parameterize the ‘amount of shift’ in the marginal feature densities. We situate this quantity in the context of the existing literature, establishing links to quantities that have been introduced in past works on this problem, and also elucidate some of its basic properties. It is shown that the quantity characterizes the minimax rates of transfer in a certain nonparametric setting.

ii) We propose a method for pruning dyadic decision trees that is at once a) adaptive to unknown parameters, achieving optimal convergence rates of the excess risk in the transfer
setting without knowledge that would be a priori unavailable to the statistician, and b) computationally feasible, unlike some other instances of rate-optimal procedures proposed in similar contexts.

1.1.2 Off-Policy Evaluation

Let us briefly recall the sequential decision-making framework as commonly studied in reinforcement learning before introducing the problem of off-policy evaluation. Here and subsequently, we shall endeavour to use the notation of [8]. We posit the existence of a finite-horizon decision process (which is may or may not be assumed to satisfy the Markov property), characterized by a sequence of state and action spaces \( S_t, \mathcal{A}_t \) for \( t = 0, 1, \ldots, T \), an initial state distribution \( s_0 \sim P_0 \), a family of transition kernels \( P_t(s_t | \mathcal{H}_{t-1}) \), and reward distributions \( R_t \sim P_r(r_t | \mathcal{H}_t) \), where \( \mathcal{H}_t = (s_0, a_0, s_1, a_1, \ldots, s_t, a_t) \) is the trajectory history up to action \( a_t \). Along with a policy \( \pi_t(a_t | s_t, \mathcal{H}_{t-1}) \), the decision process induces a distribution over state-action-reward trajectories, which we shall denote \( P_\pi \) (expectations under \( P_\pi \) will simply be written as \( \mathbb{E}_\pi [\cdot] \)). A fundamental quantity in reinforcement learning is the value of the policy \( \pi \), \( \rho^\pi \), defined as the expected sum of rewards under \( P_\pi \):

\[
\rho^\pi := \mathbb{E}_\pi \left[ \sum_{t=0}^T R_t \right].
\]

The problem of off-policy evaluation (OPE) occurs when we have access to a sample consisting of trajectories drawn from \( P_{\pi^b} \) for some policy \( \pi^b \) - the behaviour\(^3\) policy (which may or may not be known) - while the goal is to estimate the value of a target policy, \( \pi^e \); that is, to estimate \( \rho^{\pi^e} \). The practical considerations that have motivated the study of this problem are precisely those that are characteristic of transfer learning problems - namely, the collection of data under the target policy is prohibitively “expensive (e.g. in robotics, educational settings, or healthcare), or dangerous (e.g. in autonomous driving or healthcare)” [4]. Furthermore, since only changes in the policy are

\(^3\)Source policy would be appropriate nomenclature that consistent with the transfer learning literature, but ‘behaviour’ is the term used in RL, so we will use it so as to avoid confusion.
considered (the initial distribution, transition kernels, and reward distributions are unchanged in
passing from the source to the target), this can be thought of as a covariate shift problem. Further
discussion is deferred to Chapter 3, where we present some results on off-policy evaluation; this is
joint work with Prof. Hongseok Namkoong. We make the following contributions:

i) We present a critique of a real-world study that used OPE for evaluating an ‘AI-clinician’
(that is, a learned policy to assign treatments to patients in a clinical setting) in which the
difficulties of performing inference for OPE are well-illustrated.

ii) We give finite-sample minimax lower bounds for the risk of OPE estimation, which reveals
a natural ‘importance weight’ style quantity that serves to characterize the difficulty of the
problem.

iii) We present a fully nonparametric minimax lower bound for OPE estimation in the single
period setting.
Chapter 2: Classification Tree Pruning Under Covariate Shift

This Chapter is taken from [9], with Prof. Samory Kpotufe, to appear in the IEEE Transactions on Information Theory.

2.1 Introduction

Decision trees are one of the most enduring tools for classification, given their ability to adapt to complex patterns in data. Such adaptability requires proper pruning, our main concern in this work, that is, the selection of a most appropriate subtree, or tree level, that best fits unknown patterns. In particular, much of the existing theory, and resulting pruning procedures, concerns the ideal case of i.i.d. data, although in fact, it is now common in practice to combine training data from multiple inhomogeneous sources. For instance, in modern application domains such as medicine, computer speech and vision, ideal i.i.d. target data is hard to come by, so one relies on related but different data sources.

A common assumption in these settings with inhomogeneous data—which we adopt in this work—is that of so-called covariate-shift, whereby only marginal distributions are shifted from source to target data distributions. For intuition, consider for instance a drug study targeting a particular country $Q$ with a given social makeup, while much data from another population $P$ might be available. In other words, most samples $(X, Y)_P$ would be drawn from a distribution $P_{X,Y}$, in addition to some samples $(X, Y)_Q$ from an ideal distribution $Q_{X,Y}$; the two are different but related, in that, $Q_X \neq P_X$, i.e., signifying different social makeup, while $Q_{Y|X} = P_{Y|X}$, i.e., drug responses conditioned on social variables $X$ remain the same. We are interested in reducing error under target $Q$, beyond what is possible using $(X, Y)_Q$ alone, by harnessing information from $(X, Y)_P$; the extent to which this is possible depends not only on good model selection, but also on
how far $P_X$ is from $Q_X$, which we also aim to appropriately capture.

Unfortunately, existing model selection approaches, specifically for tree pruning, do not readily extend to this setting with mixed data. For instance, cross-validation approaches, often penalizing by subtree-size, might aim to use the limited samples from $Q$, since a priori this is most faithfully indicative of future performance under $Q$; however, failing to integrate the potentially much larger $P$ data can induce too large a variance to guarantee an optimal choice. Our main contribution is to derive a pruning approach that integrates both $P$ and $Q$ data to guarantee performance commensurate with the aggregate amount of data and the distance between $P_X$ and $Q_X$, appropriately captured.

As a first study of tree pruning under covariate-shift, we focus on the case of dyadic trees, which are known to be amenable to analysis [10, 11], and derive a practical pruning procedure, not only guaranteed to adapt automatically to the amount of shift from $P_X$ to $Q_X$, but yielding significant practical improvements over baselines of available cross-validation approaches. Our empirical studies not only confirm our theoretical results, but demonstrate wider applicability of our procedure beyond the case of dyadic trees.

Our approach for tree-pruning instantiates a so-called Intersecting Confidence Intervals (ICI) approach, a generic strategy for adaptation to unknown distributional parameters [12, 13]. Here in particular, in the covariate-shift setting, we parametrize the unknown level of shift $P_X \rightarrow Q_X$, via a relaxation of a recently proposed notion of transfer-exponent [3] which captures (a worst-case notion of) the mass $P_X$ assigns to regions of large $Q_X$-mass. Our relaxation, which we term aggregate-transfer-exponent, instead captures an average sense of the mass $P_X$ assigns to regions of large $Q_X$-mass near the decision boundary, and can be more easily interpretable, as it readily ties into traditional notions of information such as the Minkowski or Renyi dimensions.

Our main technical result is a guarantee that ICI, as instantiated for tree-pruning, achieves a rate adaptive to the unknown level of shift, as expressed in terms of aggregate-transfer-exponent $P_X \rightarrow Q_X$ (Theorems 3, 5). As such, transfer rates are more tightly controlled by this milder measure of shift, yielding at times faster rates than expressible via the transfer-exponent of [3].
(see Remark 4 in Section 2.4.1). A main difficulty in establishing this result lies in the fact that, unlike in usual analyses of tree-based classification, our covariate-shift setting precludes the usual simplifying assumption in tree-based prediction that the marginal measure is uniform on its support $X$. Without the uniform-measure assumption, which allows control of the mass of points in regions of space, and hence of prediction variance, we instead have to proceed via careful integration of the risk at appropriate scales over regions of space, resulting in the first proof that (dyadic) tree-based classification attains the minimax lower-bound given by [14] for general distributions; in particular, this reveals a choice of optimal tree level, qualitatively different than prescribed by the usual theory under a uniform-measure assumption (see Remark 1). As such, our results are in fact of independent interest even for the usual i.i.d. setting where $P_X = Q_X$.

2.1.1 Background and Related Work

Model selection is arguably the most important challenge in the context of decision trees, and thus naturally the focus of much of the literature on the subject. In its most general form, model selection encompasses all decisions on partitioning the $X$ space, giving rise to both greedy methods such as CART [15] which iteratively refine a space partitioning guided by labeled data $(X, Y)$, and pruning methods which select a subtree (using $Y$ labels) of an initial hierarchical partition tree built primarily using $X$ data alone. While both approaches work well in practice, the greedy approaches admit fewer theoretical guarantees in the literature. Our focus in this work is on pruning methods as discussed above, although, interestingly, our approach may serve to refine even greedily obtained trees such as CART (see Appendix 2.7.5).

Various pruning approaches have been proposed in the i.i.d. settings $P_X = Q_X$, the simplest ones consisting of selecting a tree level by cross-validation, where a level corresponds to a partition into cells of a given maximum diameter. These methods are efficient in that there are typically $O(\log n)$ levels to explore for a sample of size $n$, as opposed to visiting all possible subtrees, of which there are $O(2^n)$ [16]. Quite importantly, such simple cross-validation admits strong guarantees in the i.i.d. setting, by extending from traditional guarantees on the performance at each
level—typically for dyadic trees [17]. However, in our context where $P_X \neq Q_X$, such guarantees no longer apply, while furthermore, cross-validation is limited by the amount $n_Q \ll n_P$ of $Q$ data, making it unlikely to obtain guarantees that integrate both data sizes as obtained here.

Methods more elaborate than cross-validation have been proposed in the i.i.d. setting, which in particular can return any possible subtree by relying on clever dynamic programs to efficiently explore the very large $O(2^n)$ search space. In particular, these dynamic programs work by splitting the (estimated) classification error over cells of a partition (leaves of a subtree) to minimize this error globally. Sharp guarantees are achievable in the i.i.d. setting (see e.g. [18]) by relating local error estimates in cells to expected target error. Unfortunately, it is unclear how to extend such estimates of target error to the covariate-shift setting where relatively little target data is available locally in each cell. Initial experiments with such approaches indicate their inadequacy in the covariate-shift setting (see Appendix 2.7.5).

Covariate-shift [3, 7, 19, 20, 21], and more general changes in distribution [6, 5, 22, 23, 24, 25], have been receiving much renewed attention as it has become clear that the i.i.d. setting fails to capture many contemporary applications. Many interesting measures of shift have thus been proposed, the closest to the present work being the transfer-exponent of [3] and its recent relaxations in [21], [7] which similarly aim to capture an average sense of local deviations $P_X \rightarrow Q_X$. Our initial notion of aggregate-transfer-exponent appears most appropriate for partition-trees as it is defined over $r$-grids of the $X$ space, and appears more readily interpretable as it relates to existing notions of information. Our main adaptivity guarantee in Theorem 5 further localizes this aggregate measure to the unknown decision boundary so as to more tightly capture the relevant shift from $P_X$ to $Q_X$.

While all the above works propose model-selection approaches under covariate-shift, they are invariably concerned with classifiers other than decision trees; in particular, the works of [3, 7] implement an ICI approach, e.g., for nearest neighbor classifiers, which not only do not extend to our case, but result in computationally impractical methods due to the nature of the classifiers considered (see Section 2.3.4). An important exception is the approach of [19], which implements
importance-weighted risk estimates based on estimates of density ratios $d_{Q_X}/d_{P_X}$. Such density ratio estimates are generally attractive, and may be applied in our context of decision trees, but not only can result in expensive pre-processing steps, but may also fail to yield accurate $Q$-risk estimates when $d_{Q_X}/d_{P_X}$ does not exist (i.e. $Q_X$ is singular w.r.t. $P_X$) which is a practical scenario allowed in this work (and also in [3, 21, 7]).

2.1.2 Outline

This Chapter is organized as follows. In Section 2.2, we formally state the problem, and introduce the assumptions under which our analysis proceeds, including our condition relating the source distribution $P$ to the target $Q$. In Section 2.3 we define the decision tree models that we study and present a sequence of results that build up to our main adaptive result, Theorem 5. Computational considerations are discussed in this section as well. Section 2.4 relates the aggregate transfer exponent to various notions of information, such as the Minkowski and Renyi dimensions; proofs of results from this Section are in Appendix 2.7.1. Our analysis is presented in Section 2.5, with proofs of auxiliary results given in the Appendix 2.7.4. Our results are supported experimentally via implementations on real data in Section 2.6; results of some supplemental experiments given in Appendix 2.7.5.

2.2 Problem Setup and Notation.

We consider the binary classification problem, with data $(X, Y)$ lying in a set $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{0, 1\}$, and we assume that $\mathcal{X} = [0, 1]^D$. Unless otherwise specified, we write $\|\cdot\|, \|\cdot\|_{\infty}$ for the Euclidean and $\ell_{\infty}$ norms respectively, and let $B(x, r)$ denote the open $\ell_{\infty}$ ball of radius $r$ about $x$; that is $B(x, r) = \{y \in [0, 1]^D : \|y - x\|_{\infty} < r\}$.

In transfer learning, we simultaneously consider a source distribution $P$ and a target distribution $Q$, both on $\mathcal{X} \times \mathcal{Y}$. We let the marginal feature distributions be denoted by $P_X$ and $Q_X$ respectively, and suppose that these are supported on $\mathcal{X}_P, \mathcal{X}_Q \subset \mathcal{X}$. We assume that a sample $(X_i, Y_i)_{i=1}^n$ is
observed, where \( n = n_P + n_Q \) and we write

\[
(X, Y)_P = (X_1, Y_1), \ldots, (X_{n_P}, Y_{n_P}) \sim P \quad \text{and} \quad (X, Y)_Q = (X_{n_P+1}, Y_{n_P+1}), \ldots, (X_{n_P+n_Q}, Y_{n_P+n_Q}) \sim Q,
\]

with all observations being independent, and we let \( (X, Y) = (X, Y)_P \cup (X, Y)_Q \) denote the full sample.

The goal is to produce classifiers \( f : X \mapsto Y \) with good performance under the target distribution, as captured by its risk, defined as follows.

**Definition 1.** For a function \( f : X \to Y \), the **risk** of \( f \) under the target \( Q \), denoted \( R(f) \), is defined as

\[
R(f) := \mathbb{E}_Q 1\{f(X) \neq Y\},
\]

which simply gives the probability that \( f \) incorrectly classifies a point drawn from \( Q \). Let \( R^* = \inf_f R(f) \), and define the **excess risk** of \( f \) as

\[
E(f) := R(f) - R^*.
\]

Note that in the case of binary classification (\( Y = \{0, 1\} \)), we have \( R(f^*) = R^* \) for \( f^*(x) = 1 \{\eta_Q(x) \geq 1/2\} \), where \( \eta_Q(x) := Q(Y = 1 \mid X = x) \) is the regression function. We study the particular case of covariate shift, which supposes that the class probabilities given the features are common to the source and target, so that \( P(Y = y \mid X) = Q(Y = y \mid X) \); we may therefore without ambiguity write \( \eta \) for the regression function common to \( P \) and \( Q \).

Let \( \hat{f}_n \) denote a generic classifier learned from the sample \( (X, Y) \). This work is concerned with understanding rates of convergence of \( \mathbb{E}_{(X, Y)}[E(\hat{f}_n)] \to 0 \) in terms of source and target samples in various regimes, and developing efficient algorithms which adaptively (i.e. without recourse to problem-specific knowledge) attain these rates. Here and elsewhere \( \mathbb{E}_{(X, Y)} \) denotes expectation over the sample, while we use \( \mathbb{E} \) for expectation over the sample and an independent target point.
(that is, over $P^n \times Q^n \times Q$). Further notation will be introduced as necessary.

2.2.1 Assumptions on the Regression Function.

We restrict the complexity of the class of regression functions under consideration by making the standard assumption that this regression function belongs to a class of Hölder-continuous functions.

**Definition 2.** For $\alpha \in (0, 1)$, $L > 0$, we say that the regression function $\eta : [0, 1]^D \to \mathbb{R}$ is $(L, \alpha)$-Hölder continuous if

$$|\eta(x) - \eta(x')| \leq L \|x - x'\|_\infty^\alpha \quad \text{for all } x, x' \in X.$$ (2.1)

We adopt the following classical noise condition, which captures the difficulty of classification under $Q$, by parametrizing the likelihood of queries $x$ with low margin between class probabilities (see [14] for a detailed discussion).

**Definition 3.** We say that $Q$ satisfies Tsybakov’s noise condition with parameters $C_\beta, \beta \geq 0$, if

$$\mathbb{Q}_X(|\eta(X) - 1/2| \leq t) \leq C_\beta t^\beta.$$ (2.2)

2.2.2 Relating the Source to the Target.

We require a condition on the support of $Q$; let us quickly recall a standard notion.

**Definition 4.** A set $Z \subset X_Q$ is called an $r$-cover of $X_Q$ if $X_Q \subset \cup_{z \in Z} B(z, r)$. The size of the smallest $r$-cover is called the $r$-covering number, denoted $N(S, r)$.

**Definition 5.** We say that $X_Q$ is of metric dimension (no greater than) $d$ if there are constants $C_d, d > 0$ such that for all $r > 0$ we have

$$N(X_Q, r) \leq C_d r^{-d}.$$ (2.3)
Since $X_Q \subset [0, 1]^D$, the above is always satisfied with $d = D$.

In a previous work [3] on transfer learning under covariate shift, the authors introduce the transfer exponent between the source and target measures, which we now recall:

**Definition 6.** We call $\rho > 0$ a **transfer exponent** from $P_X$ to $Q_X$ if there exists $C_\rho > 0$ such that

$$\forall r \leq \text{diam}(X_Q) \text{ and } \forall x \in X_Q, \quad \frac{Q_X(B(x, r))}{P_X(B(x, r))} \leq C_\rho r^{-\rho}.$$  

We will consider pairs of joint distributions over $X \times Y$ with marginals $(P_X, Q_X)$ satisfying the following condition, which is a simple relaxation of the above in which we control the aggregate scaling of the ball-mass ratios rather than the worst-case scaling over the support of the target $X_Q$, which will allow us to derive tighter rates in a many scenarios; for an example, see Figure 2.1. There are a number of ways to define this relaxation, which we will consider in Section 2.4. For now we will use stick to one particular version, for which we first require a definition.

**Definition 7.** A collection $\Xi$ of subsets of $X$ is an $r$-grid of $X_Q$ if $\{X_Q \cap E : E \in \Xi\}$ is a partition of $X_Q$, and for each $E \in \Xi$ such that $E \cap X_Q \neq \emptyset$ there is an $x_E \in E \cap X_Q$ such that $B(x_E, r) \subset E \subset B(x_E, 2r)$.

We now relate $P$ to $Q$ as follows.

**Definition 8.** We say that $\gamma \in [0, \infty]$ is an **aggregate transfer exponent** from $P_X$ to $Q_X$ with constant $C_\gamma > 0$ if for any $0 < r \leq 1$, we have that for all $r$-grids $\Xi$ of $X_Q$ in $X$,

$$\sum_{E \in \Xi} \frac{Q_X(E)}{P_X(E)} \leq C_\gamma \cdot r^{-\gamma}.$$  

(2.4)

Note that the above always holds for $\gamma = \infty$, so an aggregate transfer exponent always exists. In Section 2.4, we will show that (2.4) is, under regularity conditions, equivalent to an integrated version of the transfer exponent (this quantity was in fact independently introduced recently by [21]). Note that we could have defined (2.4) by taking the sum over elements of the dyadic partition of the ambient space $X = [0, 1]^D$. In Appendix 2.7.6, we show by means of an example that this
Figure 2.1: Two source densities $P_X$ are shown, which yield equal transfer exponents w.r.t $Q_X = U([0, 1]^2)$, despite the fact that the right-hand source density vanishes over a larger region of high target mass. This is captured by the aggregate transfer exponent: while both densities have transfer exponent $\rho = 3$ with respect to $Q_X$, the left-hand density has an aggregate exponent of $\gamma = \rho$, while the right-hand density has an aggregate exponent of $\gamma = \rho + 1$. The effects of this distinction on transfer will be felt most strongly when, as is the case here, the density ratios vanish along the decision boundary (shown in dark blue). This will be elucidated in Section 2.3.5 when we introduce a notion of the aggregate exponent localized to the decision boundary.

would lead to a measure of discrepancy between $P_X$ and $Q_X$ that is not in general equivalent to the aggregate exponent or integrated exponent, and could lead to worse rate unless stronger assumptions are imposed on $P_X$ and $Q_X$.

We will see later how to refine the relation of $P$ to $Q$ in such a way as to account for the conditional distribution of the labels $Y$ given the features, although for simplicity we will first develop results using the aggregate transfer exponent as we have just defined it. These results will allow us to build towards the main result in Section 2.3.5, in which we provide stronger guarantees by localizing the aggregate transfer exponent to the decision boundary.

In Section 2.4, we will consider some properties of the aggregate transfer exponent. In particular, we explicitly relate the aggregate exponent to the transfer exponent of [3] and demonstrate that considering (2.4) refines the rates that can be derived for the convergence of the excess expected risk to zero. In Figure 2.2, we provide a simulation that demonstrates that the aggregate transfer
exponent can more effectively capture the difficulty of transfer between $P$ and $Q$; a simple example of a situation in which this is the case is given in Figure 2.1. We also show that the aggregate transfer exponent can be related to various notions of dimension for measures, for instance, the Renyi dimension; in particular we will argue that the aggregate transfer exponent can be interpreted as a relative dimension of $P_X$ with respect to $Q_X$.

Formally, we consider the following classes of pairs of distributions.

**Definition 9.** Let $T = T(C_\alpha, \alpha, C_\beta, \beta, C_d, d, C_\gamma, \gamma)$ denote the transfer class of distribution pairs $(P, Q)$ such that

i) The conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ are equal.

ii) The regression function $\eta$ is $(C_\alpha, \alpha)$-Hölder continuous.

iii) $Q_X$ satisfies a $(C_\beta, \beta)$-Tsybakov noise condition.

iv) $Q_X$ has metric dimension $d$ with constant $C_d$.

v) $P_X$ has aggregate transfer exponent $(C_\gamma, \gamma)$, w.r.t. $Q_X$.

Note that under weak conditions on $Q_X$ of metric dimension $d$ we have $\gamma \geq d$, as shown in Proposition 10. As we shall see, this implies that in most scenarios of interest it is impossible to improve on the learning rate attainable using target samples if only source samples are available.

We are now ready to state our results.

### 2.3 Overview of Results.

In this section, we present our results. These will build towards our main theorem to be presented in Section 2.3.5: an excess risk bound for a fully adaptive procedure which features a refined notion of the aggregate exponent that is localized to the decision boundary. Our first result towards this end is an oracle rate for a regular dyadic tree with respect to the (non-localized) aggregate
transfer exponent. We then present an efficient scheme for adaptively selecting a tree level, yielding a procedure that attains the oracle rate up to log factors, and then finally we show that this adaptive procedure in fact adapts to the localized aggregate exponent.

We begin by introducing the procedure from which the oracle rate is derived.

2.3.1 Tree-Based Classification

We briefly introduce some further notation in the following definitions:

**Definition 10.** For \( r \in \{2^{-i}\}_{i \in \mathbb{N}} \), let \( \Pi_r \) denote a regular partition of \([0, 1]^D\) into hypercubes of side length \( r\). A **dyadic partition tree** refers to a collection of nested partitions \( \{\Pi_r\}_{r \in \mathbb{R}} \), where we say that \( \Pi_r \) is nested in \( \Pi_{r'} \), \( r < r' \), if every cell (partition element) of \( \Pi_r \) is properly contained in a cell of \( \Pi_{r'} \). We denote cells in any given \( \Pi_r \) by \( A \), and in particular we let \( A_r(x) \) denote the cell of \( \Pi_r \) containing \( x \).

**Definition 11.** For a set \( A \in \Pi_r \), the **r-envelope** of the cell \( A \), denoted \( \tilde{A} \) is defined as

\[
\tilde{A} := \bigcup_{x' \in A} B(x', r),
\]

so that \( \hat{A}_r(x) \) is the \( r \)-envelope of \( A_r(x) \).

Recall that we have observed \((X, Y) = ((X, Y)^{n_P}, (X, Y)^{n_Q})\), where \((X, Y)^{n_P} \sim P^{n_P}\) and \((X, Y)^{n_Q} \sim Q^{n_Q}\). We consider a regular classification tree defined as follows:

**Definition 12.** We consider **tree-based classifiers** \( \hat{f}_r = 1\{\hat{\eta}_r(x) \geq 1/2\} \), for regression estimates \( \hat{\eta}_r \) defined over levels \( r \) of \( T \) as follows: let

\[
|\tilde{A}_r(x) \cap X| := |\tilde{A}_r(x) \cap \{X_1, \ldots, X_{n_P+n_Q}\}|
\]

denote the number of sample points falling in \( \tilde{A}_r(x) \). We set \( \hat{\eta}_r(x) = 0 \) if \( |\tilde{A}_r(x) \cap X| = 0 \), otherwise we set

\[
\hat{\eta}_r(x) = \frac{1}{|\tilde{A}_r(x) \cap X|} \sum_{i=1}^{n} Y_i 1\{X_i \in \tilde{A}_r(x)\}.
\]
2.3.2 Oracle Rates.

Our first result is an oracle bound on the risk of such an estimator.

**Theorem 1 (Oracle Rates).** Let \( \hat{f}_r \) be the plug-in classifier based on the dyadic tree regression (2.5) at level \( r \). For \( r = \min \left\{ n_p^{-1/(2\alpha+\alpha\beta+\gamma)}, n_Q^{-1/(2\alpha+\alpha\beta+d)} \right\} \) there is a constant \( C = C(T) \) such that

\[
\sup_{(P,Q) \in \mathcal{T}} \mathbb{E}[\mathcal{E}(\hat{f}_r)] \leq C \min \left\{ n_p^{-\alpha(\beta+1)/(2\alpha+\alpha\beta+\rho+d)}, n_Q^{-\alpha(\beta+1)/(2\alpha+\alpha\beta+d)} \right\}.
\]

**Remark 1.** To the best of our knowledge, this is the first result that gives an upper bound for the excess risk of dyadic trees, under the smoothness and noise assumptions that we have considered, that does not impose a uniform-measure assumption on the marginal distribution \( Q_X \), and is therefore of interest even in the case \( P_X = Q_X \). Under the standard theory with a uniform-measure assumption (that is, that \( Q_X(B(x,r)) \geq C_0 r^d \) for all \( r > 0, x \in X_Q \)), we would find the optimal level to be \( r^* = n_Q^{-1/(2\alpha+d)} \). We remark that [7] work under an interesting assumption that allows them to consider both uniform-type measures and measures which may not be absolutely continuous with respect to the ambient Lebesgue measure, and the rates that they give are optimal in both settings.

Note the role played by the quantity \( \gamma \) in (2.6). We see that the form of the rate of convergence to zero of the excess expected risk in terms of the number of source samples is the same as the rate in terms of target samples, with the exception of the substitution of the aggregate exponent \( \gamma \) for the metric dimension \( d \) - this supports our interpretation of \( \gamma \) as a notion of relative dimension, an interpretation which is further justified in Section 2.4. Note that previously, [3] obtained analogous results for a nearest-neighbour procedure, with the same bound on the excess expected risk as in (2.6), but with \( \rho+d \) in place of \( \gamma \), that is, they found that \( \mathbb{E}_{(X,Y)}[\mathcal{E}(\hat{f}_r)] \leq C n_p^{-\alpha(\beta+1)/(2\alpha+\alpha\beta+\rho+d)} \) for pairs \( (P,Q) \) with a transfer exponent of \( \rho \). In Section 2.4 we demonstrate that the rate in Theorem 1 is tighter by showing that \( \gamma \leq \rho + d \) under some regularity conditions, and we argue that strict inequality holds here in all but edge cases. Figure 2.2 shows the results of a simple simulation.
which shows that the aggregate exponent can capture the difficulty of transferring from $P$ to $Q$ while the transfer exponent fails to do so. Using the same idea as in Figure 2.1, we construct two sequences of marginal distributions $P_X$: the top plot of Figure 2.2 shows distributions for which the aggregate exponent increases while the transfer exponent stays fixed, while the bottom plot shows the reverse - the aggregate exponent is fixed while the transfer exponent increases. In this instance, the transfer exponent between source and target does not give a proper indication of the capacity for transfer, while we see that the aggregate exponent distinguishes a spectrum of easy-to-difficult transfer. Full details of the construction of the distributions used here are provided in Appendix 2.7.2.

The above upper-bound is indeed tight, as a Corollary to a lower-bound of [3], stated below for completion. This follows from our Proposition 7 below, via which we can show that the class $\mathcal{T}$ contains a class of distribution pairs with bounded transfer exponent, over which we invoke the lower bound of [3]. The result is as follows (the formal argument is in Appendix 2.7.4):

**Proposition 2.** Let $\mathcal{T}$ be fixed such that $d < \gamma$, $\alpha \beta \leq d$, and that $\gamma < \infty$. Then there is an absolute constant $C = C(\mathcal{T})$ such that for any classifier $\hat{h}$ learned on a sample from $P^n_P \times Q^n_Q$, we have

$$
\sup_{(P, Q) \in \mathcal{T}} \mathbb{E}_{(X, Y)} \left[ \mathcal{E}(\hat{h}) \right] \geq C' \min \left\{ n_P^{d/2\alpha + \alpha \beta + \gamma}, n_Q^{d/2\alpha + \alpha \beta + d} \right\}.
$$

We note that, as pointed out by [14], the assumption that $\alpha \beta \leq d$ is not overly restrictive, as the complementary case $\alpha \beta > d$ contains only examples in which the decision boundary does not intersect the support of the target. The minimax rate for this problem in the non-transfer setting is well-known and goes back to [14], and their result is implied by Theorem 2 by setting $n_P = 0$.

2.3.3 Adaptive Rates.

Our next result concerns an adaptive procedure - that is, a procedure that can be implemented without prior knowledge of the structural parameters $(\alpha, \beta, \gamma, d)$ - based on the dyadic tree from which we derived the oracle rate. We demonstrate that this method can attain the optimal rate
Figure 2.2: Risk estimates for a regular dyadic tree, with depth selected by 2-fold cross-validation over 100 target samples \((n_Q = 100)\), for datasets generated according to the distributions outlined in Appendix 2.7.2; we use 1000 source samples \((n_P = 1000)\). Error bars give the standard deviation over 10 runs. The Bayes risk is 0.18. **Top:** empirical error curves for five distributions with varying aggregate transfer exponents but identical exponents; the construction of the marginal distributions mirrors the construction from Figure 2.1. **Bottom:** empirical error curves for five distributions with varying transfer exponents, but with identical aggregate exponents. Again the construction from Figure 2.1 is used, except in this case the singularity strengths \(\nu\) decrease as they occur over larger regions, so as to keep the aggregate transfer exponents identical across the five distributions. Taken together, these two simulations support our claim that the aggregate exponent more adequately captures the difficulty of transferring between distributions.

of Theorem 1 up to a logarithmic factor. This procedure is based on the Intersecting Confidence Intervals method of [12], which features in the work of Lepski (e.g. [13]). The key idea is given a concise exposition in [26].

The method proceeds by locally constructing confidence bounds for \(\eta(x) = P(Y = 1 \mid X = x)\) based on the data-dependent quantity

\[
\hat{\sigma}_r(x) := \frac{C}{\sqrt{|A_r(x) \cap X|}}, \tag{2.7}
\]

which requires no problem-specific information to be computable (the constant \(C\) is an input), and serves as a proxy for the local estimation error \(|\hat{\eta}_r(x) - \tilde{\eta}_r(x)|\) at level \(r\), where \(\tilde{\eta}_r(x) = \mathbb{E}[\hat{\eta}_r(x) \mid X]\). We begin by looking deep in the tree, and move up the branch containing \(x\) until a level is reached that achieves an approximate balance of estimation and approximation error,
thus giving optimal rates of convergence. Details are given in Algorithm 1 below, and the result is
as follows:

**Theorem 3** (Adaptive Rates). Let \( \hat{f} \) denote the classifier determined by Algorithm 1. Then there
is a universal \( C = C(T) \) not depending on \( n_P, n_Q \) such that

\[
\sup_{(P,Q) \in T} \mathbb{E}(X,Y)[\mathcal{E}(\hat{f})] \leq C \min \left\{ \left( \frac{\log n_P}{n_P} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\gamma}}, \left( \frac{\log n_Q}{n_Q} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+d}} \right\}.
\]

As far as we are aware, Theorem 3 is the first adaptivity result for decision tree classifiers in the
covariate-shift setting. Comparable rates have been derived for K-nearest neighbour procedures,
as well as kernel methods, in [3] and [21]. In the one-sample setting, [18] have derived adaptive
rates under conditions related to ours, but that are not directly comparable. Their method, based on
structural risk minimization, does not seem to extend to multiple distribution settings (as we argue
in Appendix 2.7.5).

**Algorithm 1:** Local Depth Selection.

**input:** A point \( x \in \mathcal{X}_Q \), a depth \( r_0 \), a constant \( C > 0 \).

**output:** A classification estimate \( \hat{f}(x) = 1\{\hat{\eta}(x) \geq 1/2}\).

Set \( r = r_0 \). Set \( \hat{\eta}^+ = \hat{\eta}_r(x) + 2\hat{\sigma}_r(x) \) and \( \hat{\eta}^- = \hat{\eta}_r(x) - 2\hat{\sigma}_r(x) \); set \( \hat{\eta} = \hat{\eta}_r(x) \).

while \( \hat{\eta}^- \leq 1/2, \hat{\eta}^+ \geq 1/2, r < 1/2 \) do

\[
\begin{align*}
& r \leftarrow 2r, \\
& \hat{\eta}^- \leftarrow \max\{\hat{\eta}^-, \hat{\eta}_r(x) - 2\hat{\sigma}_r(x)\}, \\
& \hat{\eta}^+ \leftarrow \min\{\hat{\eta}^+, \hat{\eta}_r(x) + 2\hat{\sigma}_r(x)\},
\end{align*}
\]

if \( \hat{\eta}^+ \leq \hat{\eta}^- \), then

stop

\[
\hat{\eta} \leftarrow (\hat{\eta}^+ + \hat{\eta}^-)/2
\]

end if

\( \hat{\eta} \leftarrow \hat{\eta}_r \)

end while

return \( \hat{f}(x) = 1\{\hat{\eta}(x) \geq 1/2}\).
2.3.4 Computational Considerations

Although many minimax-optimal methods for classification and regression using the Intersecting Confidence Interval method [12] have been proposed, we note that the underlying dyadic-tree structure makes Algorithm 1 particularly appealing due to its low computational cost at test-time. Once the tree is initialized, for a given input point \( x \), it takes at most \( -\log_2 r_0 = O(\log(n)) \) steps to compute \( \hat{f}(x) \), where \( n \) is the combined sample size. The initialization of the tree can be achieved in \( O(n \log n) \) time as well (see e.g. [27, 28]), and for a further cost of \( O(n \log n) \) operations, the output \( \hat{f}(x) \) can be computed for all further \( x \) at once, reducing the burden to \( O(1) \) operations at test time. On the other hand, using the ICI method with a nearest-neighbour procedure (as in [3, 7]) implies locally finding \( k(x) \) neighbours for each query point \( x \) at testing time, which requires at least \( O(n \log n) \) operations since \( k(x) \) will be at least a root of \( n \). We note also that once the base tree is initialized, it can be updated in online fashion at cost only \( O(\log n) \) for each new data point, since to update the tree to include a new point \((y, x)\) requires only one pass along the branch that ends in the leaf containing \( x \).

2.3.5 Main Result: Localizing to Decision Boundary

Below we consider how to refine the rates presented in the previous section. Throughout, we suppose that we have measures \((P, Q) \in T\), where \( T = T(\alpha, C_\alpha, \beta, C_\beta, \gamma^*, C_{\gamma^*}, d, C_d) \). Note that in this section we use \( \gamma^* \) to denote the aggregate transfer exponent of \( P_X \) relative to \( Q_X \) in order to distinguish it from local exponents defined below.

To begin, recall that we suppose \( X_P, X_Q \subset X \), and for \( \epsilon > 0 \) let us define the decision boundary margin region \( G^{-\epsilon} := \{ x \in X : |\eta(x) - \frac{1}{2}| \leq \epsilon \} \) and \( G^{+\epsilon} := \{ x \in X : |\eta(x) - \frac{1}{2}| > \epsilon \} \). The first step is to define an aggregate transfer exponent restricted to \( G^{-\epsilon} \), which contains the decision boundary.

**Definition 13.** Let \((P, Q) \in T\). The aggregate transfer exponent **localized** to the region \( G^{-\epsilon} \), denoted \( \gamma(\epsilon) \), is defined to be the least positive number such that for all \( r \)-grids \( \xi(\epsilon) \) of \( X_P \cap G^{-\epsilon} \)
we have
\[ \sum_{E \in \xi(\epsilon)} \frac{Q_X(E)}{P_X(E)} \leq C_{\gamma^*} r^{-\gamma(\epsilon)} \]
for all \( 0 < r < 1 \), where \( C_{\gamma^*} \) is the parameter from \( T \).

Our first result is an excess risk bound for an oracle classifier based on an arbitrary and fixed margin value.

**Theorem 4.** There is a universal constant \( C = C(T) \) such that, for any \( \epsilon > 0 \), there is a classifier \( \hat{f}_\epsilon \) such that
\[
\mathbb{E} \mathcal{E}(\hat{f}_\epsilon) \leq C \left[ \min \left( \frac{\alpha(\beta+1)}{2a_1 + a_2 \beta + \gamma(\epsilon)}, \frac{\alpha(\beta+1)}{2a_1 + a_2 \beta + 1} \right)^{\frac{1}{2}} + \min \left( \frac{1}{n_p}, \frac{1}{n_Q} \right) \right].
\] (2.8)

Our main result demonstrates that, up to log terms, the above oracle bound is attained by the adaptive classifier of Algorithm 1, for an optimal value of the margin parameter \( \epsilon \). Choosing \( \epsilon \) to approximately balance the terms in (2.8) (adjusted to include log-terms) gives the finite-sample rate that is achievable at sample size \( n = (n_p, n_Q) \), determined by the behaviour of the aggregate exponent in the vicinity of the decision boundary.

**Theorem 5 (Adaptive Localized Rate).** Consider the following critical level
\[
\epsilon_n = \inf \left\{ \epsilon > 0 : \frac{1}{n_p} \epsilon^{-(1+\gamma^*/\alpha)} \leq \left( \frac{\log n_p}{n_p} \right)^{-\alpha(\beta+1) + \gamma(\epsilon)} \right\},
\]
and let \( \gamma_n = \gamma(\epsilon_n) \). Let \( \hat{f} \) denote the adaptive classifier of Algorithm 1, and suppose that \( \alpha + d > 1 \). Then there are constants \( C_1, C_2 \), depending only on \( T \), such that the following holds. If \( \max(n_p, n_Q) \geq C_1 \), then for any \( \epsilon > 0 \) we have
\[
\mathbb{E}(X,Y)[\mathcal{E}(\hat{f})] \leq 2C \cdot \min \left\{ \left( \frac{\log n_p}{n_p} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\gamma n}}, \left( \frac{\log n_Q}{n_Q} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+d}} \right\},
\]

with \(C\) as in Theorem 4.

**Remark 2.** Note that the critical margin level \(\epsilon_n\) does not depend on \(n_Q\). This is because the level is chosen to minimize the source-data contribution to the intermediate bound (2.8).

From the above, we see that if the relative dimension of \(P\) to \(Q\) is small on a low-margin region (that is, close to the decision boundary), then for sufficiently large sample size, the expected excess risk will converge to zero at a rate determined by this localized value of the relative dimension.

**Remark 3.** Suppose that for some \(\gamma > 0\) there is a corresponding \(\epsilon_0 > 0\) such that \(\gamma(\epsilon_0) = \gamma\). Then, it is immediate by rearranging that for

\[
\frac{n_p}{\log n_p} \geq \left( \frac{2\alpha+\alpha\beta+\gamma}{\alpha(\alpha+\gamma)} \right)^{(\alpha+\gamma^*)},
\]

we have

\[
\frac{1}{n_p} \epsilon_0^{-(1+\gamma^*/\alpha)} \leq \left( \frac{\log n_p}{n_p} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\gamma(\epsilon_0)}},
\]

which by construction implies that \(\epsilon_n \leq \epsilon_0\) from which we get \(\gamma_n = \gamma(\epsilon_n) \leq \gamma(\epsilon_0) = \gamma\), since \(\gamma(\cdot)\) is non-decreasing. It then follows from Theorem 5 that we have

\[
\mathbb{E}(X,Y)[\mathcal{E}(\hat{f})] \leq C \min \left\{ \left( \frac{\log n_p}{n_p} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\gamma}}, \left( \frac{\log n_Q}{n_Q} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+d}} \right\},
\]

where \(C\) is as in Theorem 5. This illustrates the following behaviour: supposing that for some margin value \(\epsilon_0\) we have a low value \(\gamma(\epsilon_0) = \gamma < \gamma^*\), then when enough source samples have been collected, the expected excess risk will decrease to zero at the faster rate (2.10). Until this
threshold has been achieved however, the convergence will be at the slower rate, determined by $\gamma^*$, and unfortunately, the requisite sample size is exponential in the global aggregate transfer exponent $\gamma^*$. However, we emphasize that with stronger assumptions on the behaviour of $P_X, Q_X$ away from the decision boundary this would no longer be the case - see for example Proposition 3.7 of [14], where it is shown that exponential convergence rates can be derived for error away from the boundary under the strong density condition.

**Example 1.** The following construction in one dimension is an example of a case in which the aggregate transfer exponent becomes smaller when localized to the decision boundary, giving a faster rate of convergence. Suppose that $d = 1, X_Q = X_P = [0, 1]$, and take $Q_X \sim U([0, 1])$. Let $P_X$ have density $p(x) \propto x^{\gamma}$ for some $\gamma > 1$, and let $\eta(x) = x$. One can show that that $(P, Q)$ are in the transfer class with parameters $\alpha = 1, \beta = 1, \gamma^* = \gamma$, and so Theorem 3 tells us that

$$\mathbb{E}[\mathcal{E}(\hat{f})] \leq C \left( \frac{\log n_P}{n_P^2} \right)^{\frac{2}{3+\gamma}}. \tag{2.11}$$

On the other hand, in this example the singularity in the density ratio occurs away from the decision boundary, and one readily finds for any $\epsilon < 1/2$, we have $\gamma(\epsilon) = 1$, and so by the argument given above we find that for $n_P/\log n_P \geq 4^\gamma$ we have

$$\mathbb{E}[\mathcal{E}(\hat{f})] \leq C \left( \frac{\log n_P}{n_P} \right)^{\frac{1}{2}},$$

a faster rate than (2.11).

### 2.4 Relation to Notions of Dimension.

In this section we will consider how the aggregate transfer exponent defined in (2.4) relates to traditional notions of dimension of a measure. In particular, it is immediate (as argued below) that, in the limits, $\gamma_* = \inf \{\gamma\}$ (see Definition 14) can be viewed as a relative version of the Minkowski
dimension. Similarly, we can argue that, under regularity conditions, \( \gamma_\ast \) is equivalent to an *average* version of the *transfer exponent* of [3]. This average version in turn defines a relative version of the Renyi dimension.

Note that the ensuing discussion applies to general measures supported on a common metric space \((X, d)\), with open balls of radius \( r \) denoted \( B(\cdot, r) \), although the rest of our work considers chiefly the case \( X = [0, 1]^D \) with the \( \ell_\infty \) metric. We use \( P_X, Q_X \) throughout this section to denote generic measures supported on \( X_P \subset X \) respectively.

**Definition 14.** For measures \( P_X, Q_X \), take \( \Xi_r \) to be the set of all \( r \)-grids of \( X_Q \) in \( X \), and define the *minimal aggregate transfer exponent* \( \gamma_\ast = \gamma_\ast(P_X, Q_X) \) as follows:

\[
\gamma_\ast = \inf_{\gamma \geq 0} \left\{ \exists C_\gamma \text{ s.t. } \sum_{E \in \xi} \frac{Q_X(E)}{P_X(E)} \leq C_\gamma r^{-\gamma} \text{ for all } \xi \in \Xi_r \right\}.
\]

(2.12)

Observe that one can express all of the main results in terms of \( \gamma_\ast(P_X, Q_X) \), so that, for instance, Theorem 1 implies that for any \( s > \gamma_\ast(P_X, Q_X) \), the excess risk as a function of the source samples converges to zero at a rate no worse than \( n_p^{-\alpha(1+\beta)/(2\alpha+\alpha+\beta+s)} \).

As it turns out, the minimal \( \gamma_\ast \) can be viewed as a relative version of the Minkowski dimension of a measure, which is defined as follows (see [29]).

**Definition 15.** Let \( S \subset X_Q \). The *Minkowski dimension* of \( S \), denoted \( \dim_M(S) \), is defined as

\[
\dim_M(S) = \limsup_{r \to 0} \frac{\log N(S, r)}{-\log r},
\]

(2.13)

and the Minkowski dimension of the measure \( Q_X \) is defined via

\[
\dim_M(Q_X) = \inf \{ \dim_M(S) : Q_X(S) = 1 \}.
\]

Note that (see [30], Section 3.1) one obtains an equivalent definition of the Minkowski dimension if the covering number \( N(S, r) \) in (2.13) is replaced by the maximal packing number, which in turn is easily seen to equal the maximal size of an \( r \)-grid. The following proposition immediately
follows from this observation.

**Proposition 6** ($\gamma_*$ is relative Minkowski). We have

$$\gamma_*(Q_X, Q_X) = \dim_M(Q_X).$$

**Proof.** Let $M(X_Q, r)$ denote the $r$-packing number of $X_Q$, and let $\Xi$ denote the set of $r$-grids of $X_Q$. Since an $r$-grid gives an $r$-packing, and an $r$-packing can be transformed into a grid by using a Voronoi tessellation, we have

$$\sup_{\xi \in \Xi} \sum_{E \in \xi} Q_X(E) = \sup_{\xi \in \Xi} |\xi| = M(X_Q, r),$$

and the result follows since (2.13) yields the same value for $\dim_M$ if the packing number $M$ is used in place of the cover number $N$.

2.4.1 Relation to transfer exponent.

Recall the transfer exponent of [3], introduced in Definition 6. We now consider an alternative way to obtain an average version of this notion. As we show thereafter, this alternative is, under some conditions, equivalent to the aggregate transfer exponent introduced in Definition 8.

**Definition 16.** We call $\bar{\rho} > 0$, an integrated transfer exponent from $P_X$ to $Q_X$, if there exists $C_{\bar{\rho}} > 0$ such that

$$\forall r > 0, \int_X \frac{1}{P_X(B(x, r))} Q_X(dx) \leq C_{\bar{\rho}} \cdot r^{-\bar{\rho}}. \tag{2.14}$$

We will be interested in the minimal integrated transfer exponent $\bar{\rho}_*$:

$$\bar{\rho}_* = \inf\{\bar{\rho} > 0 : \exists C_{\bar{\rho}} \text{ such that (2.14) holds.}\}.$$ 

Note that the integrated transfer exponent (2.14) was recently introduced and investigated by [21], who use it to derive excess risk bounds similar to our Theorem 1 in the context of nonparamet-
The following proposition gives a simple relation between $\bar{\rho}_*$ and the minimal transfer exponent $\rho_*$. The implication of this result is discussed at the end of the section.

**Proposition 7.** Let $\rho_*$ be the minimal transfer exponent from $P_X$ to $Q_X$, and let $\bar{\rho}_*$ be the minimal integrated transfer exponent from $P_X$ to $Q_X$. Suppose $Q_X$ is of metric dimension $d$. We then have that $\bar{\rho}_* \leq \rho_* + d$.

Note that Proposition 7 is a straightforward extension of Lemma 1 from [21]. Example 2 below shows that we can have the strict inequality $\bar{\rho}_* < \rho_* + d$. As it turns out, $\bar{\rho}_*$ can be interpreted as a relative form of the Renyi dimension, similar to the interpretation of $\gamma_*$ in terms of Minkowski dimension. This is immediate from the following definition (see [31]):

**Definition 17.** The Renyi dimension (of order zero) of a measure $Q_X$, $\dim_R(Q_X)$, is given by

$$\dim_R(Q_X) = \lim \sup_{r \to 0} \frac{\log \int_{Q_X} (B(x, r))^{-1} Q_X(dx)}{-\log r}.$$

One immediately sees that the minimal integrated transfer exponent $\bar{\rho}_*(Q_X, Q_X)$ corresponds to the Renyi dimension of order zero, $\dim_R(Q_X)$. It is known that, for any measure $Q_X$, $\dim_R(Q_X)$ is equal to the Minkowski dimension $\dim_M(Q_X)$ (see [29]); stated differently, we have $\bar{\rho}_*(Q_X, Q_X) = \gamma_*(Q_X, Q_X)$. In what follows, we will show that in the case of two measures $P_X, Q_X$, under some regularity condition we also have that $\bar{\rho}_*(P_X, Q_X) = \gamma_*(P_X, Q_X)$ (Proposition 9). We start with the following condition ([29], [32]) on the measures which in particular allow us to relate conditions on grids in the definition of $\gamma_*$ to conditions on balls in the definition of $\bar{\rho}_*$.

**Definition 18.** A measure $\mu$ is said to be doubling with constant $C$ if for all $r > 0$ and all $x \in \text{supp}(\mu)$, we have

$$\mu(B(x, r)) \geq C \mu(B(x, 2r)).$$

The simplest examples of doubling measures are measures supported on well-behaved compact subsets $\mathbb{R}^d$ with densities that are bounded above and away from zero with respect to the Lebesgue
measure - a regularity condition that is frequently imposed in the nonparametrics literature, for instance in the seminal works [33] and [14] (in the latter this is the strong-density assumption). The strong-density condition is not necessary however; for example, the measure on \([0, 1]\) with density given by \(p(x) \propto x^\nu \, dx\) for any \(\nu > 0\) is doubling. The doubling condition is also commonly seen in the literature on dimensions of measures, where a measure satisfying (2.15) is sometimes referred to as a Federer measure, or diametrically regular [29]; this work also lists larger classes of doubling measures, and further examples are provided by [34]. The survey of nearest-neighbour methods [35] presents connections between neighbour searching and notions of metric space dimension that includes a discussion of doubling measures, and may therefore be of interest.

Under a doubling assumption, we can establish the following result, complementing Proposition 7:

**Proposition 8.** Suppose that the minimal transfer exponent satisfies \(\rho^*(P_X, Q_X) \geq \rho\). If \(Q_X\) is a doubling measure, then any integrated transfer exponent \(\bar{\rho}\) satisfies \(\bar{\rho} \geq \rho\).

Note that the doubling assumption allows us to relate the measures of sets that are of comparable size; this assumption could be dropped if, say, we were to define the transfer exponent condition as needing to hold for all sets of aspect ratio bounded by some constant \(\phi > 1\) rather than all balls (that is to say, if we required that \(P_X(A) \geq C Q_X(A) r^\rho\) whenever there exists an \(x \in X_Q\) such that \(B(x, r) \subset A \subset B(x, \phi r)\)).

We have the following relation.

**Proposition 9.** Let \(\gamma_*, \bar{\rho}_*\) be the minimal aggregate and integrated transfer exponents from \(P_X\) to \(Q_X\), respectively. If \(P_X, Q_X\) are doubling measures, then \(\gamma_* = \bar{\rho}_*\).

The condition that \(P_X, Q_X\) be doubling measures is sufficient but not necessary; in fact the equivalence can be shown to hold in greater generality, though we do not pursue this here. We note however that a simpler definition of the aggregate transfer exponent, in which the sum in (2.4) is taken over a regular dyadic partition of \([0, 1]^D\) rather than over grids, can give a different value for the relative dimension. In Appendix 2.7.6, we give an example of non-doubling \(P_X, Q_X\) for which
\( \gamma_\ast = \rho_\ast \), and yet these differ from the value one would obtain if a dyadic partition of the ambient space had been used to define the relative dimension.

**Remark 4.** Together, Propositions 7 and 9 show that upper bounds for the excess risk that we derive are at least as sharp as those of [3], and sharper in the case \( \gamma_\ast < \rho_\ast + d \). Whether or not this occurs depends on the geometry of the sub-region of \( X_Q \) where the worst case local scaling of ratios \( Q_X(B(x,r))/P_X(B(x,r)) \) occurs; when this region has the same dimension as \( X_Q \), one obtains \( \gamma_\ast = \rho_\ast + d \). This occurs, for example, in the case in which both \( P_X, Q_X \) are regular of dimension \( d_P, d \) respectively, with \( d_P > d \). In this case we have \( \gamma_\ast(P_X,Q_X) = d_P \) and \( \rho = d_P - d_Q \) (see [3]) so clearly \( \gamma_\ast = \rho + d_Q \). On the other hand, when this worst-case scaling is restricted to a lower-dimensional subset of \( X_Q \), we expect the strict inequality \( \gamma_\ast < \rho_\ast + d \), and in this case the upper bound of [3] is not optimal.

The following example illustrates this.

**Example 2.** Let \( Q_X \) be uniformly distributed on \([0, 1]^d\), and for \( x \in [0, 1]^d \) let \( P_X \) have density \( p(x) \propto \|x\|^\nu \) for some \( \nu > 0 \). In Appendix 2.7.3, we show that there is a constant \( C \) for which \( \gamma = \max(\nu, d) \) is an aggregate transfer exponent of \( P_X \) relative to \( Q_X \). In this example, it is shown in [3] that the transfer exponent is simply \( \rho(P_X, Q_X) = \nu \), and so we have \( \gamma_\ast(P_X, Q_X) = \max(\nu, d) < \rho(P_X, Q_X) + d \) when \( \nu > 0 \). This example can be extended in a simple way so as to yield aggregate exponents between \( \max(\nu, d) \) and \( \nu + d \), and it is precisely this construction that is used in Figure 2.2. Fix an integer \( k \in \{1, \ldots, d - 1\} \), and let

\[
A_k = \{ x \in [0, 1]^d \mid (x_{k+1}, \ldots, x_d) = 0 \};
\]

note that \( A_k \) is a subset of \([0, 1]^d\) of Minkowski dimension \( k \). Letting \( d(x, A) = \inf_{y \in A} \|x - y\| \) denote the standard distance function, we can define distributions \( P_X^k \) on \([0, 1]^d\) via density functions \( p^k(x) \propto d(x, A_k)^\nu \); again, we take \( Q_X \sim U([0, 1]^d) \). With this notation, the example given above corresponds to the choice \( k = 0 \). Now, note that for \( x \in A_k \), we have that \( Q_X(B(x,r))/P_X(B(x,r)) \sim O(r^{-\nu}) \). Since the Minkowski dimension of \( A_k \) is \( k \), there will be \( O(r^{-k}) \) elements of any grid that
intersect $A_k$, yielding a contribution to the aggregate transfer exponent that is $O(r^{-(k+\nu)})$. In the rest of the space, we have $Q_X(B(x,r))/P_X(B(x,r)) \sim O(1)$, giving a contribution to the aggregate exponent of order $O(r^{-d})$. As $r \to 0$ the sum of these contributions grows as $O(r^{-\max(k+\nu,d)})$, giving an aggregate transfer exponent of $\max(k+\nu,d)$; this argument is made precise in the case $k = 0$ in Appendix B; for higher $k$ the calculations are entirely analogous and are therefore omitted. In each of these cases, we have $\rho(P^k_X, Q_X) = \nu$, so we see that the aggregate exponent accounts for both the strength of any marginal singularities and also the ‘size’ of the region over which they occur. An analysis involving only the transfer exponent $\rho$ essentially assumes the worst case: that the singularity strength is of uniform order over the entire support $X_Q$; see the discussion of Figure 2.2.

We close this section with the following Proposition, which demonstrates that, under a regularity condition on the target, the minimal aggregate transfer exponent is bounded below by the Minkowski dimension of the target. This implies that the upper bound on the excess risk from Theorem 1 decreases no faster in $n_P$ than in $n_Q$, and precludes the possibility of super-transfer (that is, transfer with faster rates using source samples alone; see [36]).

**Proposition 10.** Let $P_X, Q_X$ be probability measures on $[0,1]^D$. If there is an $S \subset X_Q$ with $\dim_M(S) = d$ and constants $C, r_0 > 0$ such that for all $x \in S$ and all $0 < r < r_0$ we have $Q_X(B(x,r)) \geq Cr^d$, then $\gamma^*(P_X, Q_X) \geq d$.

**2.5 Analysis Details.**

This section contains an overview of the proofs of our main results; any details not fully provided here are to be found in the Appendices.

2.5.1 Main Tools.

We begin by recalling some well-known results that will give a decomposition of the excess risk of $\hat{f}_r$ under the target distribution $Q$. Let $\hat{f}_r$ denote the plug-in estimator $\hat{f}_r := 1{\hat{\eta}_r \geq 1/2}$
based on an estimate $\hat{\eta}_r$ of $\eta$; let $f^*$ denote the Bayes classifier. Then it is well-known (see [37], Theorem 2.2) that the excess risk $\mathcal{E}_Q(\hat{f}_r)$ can be expressed as

$$
\mathcal{E}_Q(\hat{f}_r) = R_Q(\hat{f}_r) - R^* = 2 \int |\eta(x) - 1/2| \cdot 1\{\hat{f}_r(x) \neq f^*(x)\} Q_X(dx).
$$

(2.16)

It will be convenient to also consider the expectation of $\hat{\eta}$ conditional on the features $X$, which we denote $\tilde{\eta}_r = \mathbb{E}_{Y|X} [\hat{\eta}_r]$. Note that by (2.5) we have $\tilde{\eta}_r(x) = 0$ if $|\tilde{A}_r(x) \cap X| = 0$ and

$$
\tilde{\eta}_r(x) = \frac{1}{|\tilde{A}_r(x) \cap X|} \sum_{i=1}^{n} \eta(X_i) 1\{X_i \in \tilde{A}_r(x)\}
$$

(2.17)

otherwise. Observe now that on $\{\hat{f}_r(x) \neq f^*(x)\}$, we have $|\eta(x) - 1/2| \leq |\hat{\eta}_r(x) - \eta(x)|$. A standard bias-variance decomposition of the error $|\hat{\eta}_r - \eta|$ then yields

$$
|\hat{\eta}_r(x) - \eta(x)| \leq |\hat{\eta}_r(x) - \tilde{\eta}_r(x)| + |\tilde{\eta}_r(x) - \eta(x)|.
$$

The following is a simple consequence of equation 2.16 and the inequality $1\{A \leq B + C\} \leq 1\{A \leq 2B\} + 1\{A \leq 2C\}$:

**Lemma 11.** Let $\hat{f}_r$ be as in equation 2.16, and write $M(x) := |\eta(x) - 1/2|$ for the margin function. Then

$$
\mathcal{E}_Q(\hat{f}_r) \leq 2 \int M(x) \cdot 1\{M(x) \leq 2|\hat{\eta}_r(x) - \tilde{\eta}_r(x)|\} Q_X(dx) = A_1
$$

(2.18)

$$
+ 2 \int M(x) \cdot 1\{M(x) \leq 2|\tilde{\eta}_r(x) - \eta(x)|\} Q_X(dx) = A_2
$$

(2.19)
2.5.2 Overview of Proofs.

Upper Bounds for Dyadic Trees

In this section we outline the proof of our oracle upper bound. The tools that we employ are standard, although our approach makes it possible to derive the minimax optimal rate, unspoiled by a log-term, with minimal assumptions on the distributions $P_X, Q_X$.

As a first step towards bounding the terms in Lemma 11 in expectation over the sample, we derive some simple bounds on the error terms $|\tilde{\eta}_r(x) - \tilde{\eta}_r(x)|$ and $|\tilde{\eta}_r(x) - \eta(x)|$. Let us consider the following event:

**Notation 12.** Let

$$S_r(x) := \{|\tilde{A}_r(x) \cap X| > 0\}$$

**denote the event that the envelope of the cell containing** $x$ **is non-empty, and likewise let** $S_r(X) := \{|\tilde{A}_r(X) \cap X| > 0\}$. The following bounds are immediate:

**Lemma 13 (Bias).** Let $\tilde{\eta}, S_r(x)$ be as above. The bias term can be bounded as follows for all $x \in X_Q$:

$$|\tilde{\eta}_r(x) - \eta(x)| \mathbb{1}_{S_r(x)} \leq C_\alpha r^\alpha.$$  

**Lemma 14 (Variance).** Let $\tilde{\eta}, \tilde{\eta}, S_r(x)$ be as above. For all $x \in X_Q$, we have:

$$\mathbb{E}_{Y|X} \left[ |\tilde{\eta}_r(x) - \tilde{\eta}_r(x)|^2 \mid X \right] \mathbb{1}_{S_r(x)} \leq \frac{1}{4 |\tilde{A}_r(x) \cap X|} := \frac{1}{4} \tilde{V}_r(x).$$  

Using Lemma 13, we obtain a bound on the expectation over $(X, Y)$ of (2.19) when we restrict to $S_r(X)$ (that is, when $X$ is in the region corresponding to non-empty cells in $\Pi_r$):
**Proposition 15** (Bounding $A_2$ in expectation.). Let $(P, Q) \in T$. Let $\tilde{\eta}_r$ be as in (2.17), let $S_r(X)$ be as in Notation 12, and let $M(x) = |\eta(x) - \frac{1}{2}|$. Then for some $C > 0$ free of $n_P, n_Q$ we have

$$
\mathbb{E} \left[ M(X) \cdot 1 \{ M(X) \leq 2|\tilde{\eta}_r(X) - \eta(X)| \} \cdot 1 \{ S_r(X) \} \right] \leq C r^{\alpha(\beta+1)},
$$

where the expectation is over the full sample $(X, Y)$ and the target $X \sim Q_X$.

**Proof.** By Lemma 13, we have

$$
\mathbb{E} \left[ M(X) \cdot 1 \{ M(X) \leq 2|\tilde{\eta}_r(X) - \eta(X)| \} \cdot 1 \{ S_r(X) \} \right] \leq \mathbb{E}_{Q_X} \left[ M(X) \cdot 1 \{ M(X) \leq 2C \alpha r^\alpha \} \right]
$$

$$
\leq 2C \alpha r^\alpha \cdot Q_X(0 < |\eta(X) - \frac{1}{2}| < 2C \alpha r^\alpha)
$$

$$
\leq 2C \alpha r^\alpha \cdot C_\beta (2C \alpha r^\alpha)^\beta
$$

$$
= C r^{\alpha(\beta+1)},
$$

where the final inequality applies the noise condition (2.2). \qed

To obtain a corresponding bound in expectation for the first term of Lemma 11, we consider the event in which the quantity $\hat{V}_r(x)$ in (2.20) is close to its population counterpart $V_r(x)$, defined below.

**Notation 16.** Let $x \in X_Q$, and let

$$
V_r(x) := \frac{1}{n_P P_X(\tilde{A}_r(x)) + n_Q Q_X(\tilde{A}_r(x))}.
$$

Consider the event

$$
\Phi_r(x) := \{ \hat{V}_r(x) < 2V_r(x) \} = \{ |\tilde{A}_r(x) \cap X| > \frac{1}{2} \mathbb{E} |\tilde{A}_r(x) \cap X| \},
$$

where $\mathbb{E}$ denotes the expectation over the full sample $(X, Y)$.
that is, the event that the number of samples in $\tilde{A}_r(x)$ concentrates (multiplicatively) to its expectation. We further set $\Phi_r(X) := \{\tilde{V}_r(X) < 2V_r(X)\}$

Now, since $\Phi_r(x) \subset S_r(x)$, restricting attention to $\Phi_r(x)$ for $x \in \mathcal{X}_Q$ gives (via Lemma 14) the bound

$$\mathbb{E}(X,Y) |\hat{\eta}(x) - \tilde{\eta}(x)|^2 \cdot \mathbbm{1}\{\Phi_r(x)\} \leq V_r(x)/2,$$

and we use this to bound the first term in Lemma 11 in terms of $\mathbb{E}_Q[X]\left[V_r(X)\right]$. We obtain the following:

**Lemma 17** (Bounding $A_1$ in expectation). Let $(P,Q) \in T$, let $\Phi_r(X)$ be as in Notation 16 and let $M(x) = |\eta(x) - \frac{1}{2}|$. Then there is a constant $C > 0$, independent of $n_P, n_Q$, such that for all $0 < t < 1$ we have

$$\mathbb{E} [M(X) \cdot \mathbbm{1}\{M(X) \leq 2|\hat{\eta}_r(X) - \tilde{\eta}_r(X)|\} \cdot \mathbbm{1}\{\Phi_r(X)\}] \leq C \left(t^{\beta+1} + \frac{\mathbb{E}_Q[X]\left[V_r(X)\right]}{t}\right),$$

where the expectation is over the full sample $(X,Y)$ and the target $X \sim Q_X$.

**Proof.** Let $\Phi_r$ be as above, let $\hat{M}_r(x) := |\hat{\eta}_r(x) - \tilde{\eta}_r(x)|$, and let

$$\Gamma(X) := M(X) \cdot \mathbbm{1}\{M(X) \leq 2\hat{M}_r(X)\}.$$  

Fixing a $t \in (0, 1)$, we then have

$$\mathbb{E} \Gamma(X) \cdot \mathbbm{1}\{\Phi_r(X)\} = \mathbb{E} \left[\Gamma(X) \cdot \mathbbm{1}\{\Phi_r(X)\} \cdot \mathbbm{1}\{\hat{M}_r(X) \leq t\}\right] + \mathbb{E} \left[\Gamma(X) \cdot \mathbbm{1}\{\Phi_r(X)\} \cdot \mathbbm{1}\{\hat{M}_r(X) > t\}\right],$$

and we proceed to bound these two terms separately. For the first term, note that we have

$$\mathbb{E} \Gamma(X) \cdot \mathbbm{1}\{\Phi_r(X)\} \cdot \mathbbm{1}\{\hat{M}_r(X) \leq t\} \leq 2t Q_X(|\eta(X) - \frac{1}{2}| \leq t),$$

34
and by the noise condition we have

\[ Q_x(|\eta(X) - \frac{1}{2}| \leq t) \leq C_{\beta t^\beta}. \]

therefore we find

\[ \mathbb{E} \left[ \Gamma(X) \cdot 1(\Phi_r(X)) \cdot 1\{\hat{M}_r(X) \leq t\} \right] \leq 2C_{\beta t^{(\beta+1)}}. \]

For the second term, we have

\[ \mathbb{E} \left[ \Gamma(X) \cdot 1(\Phi_r(X)) \cdot 1\{\hat{M}_r(X) > t\} \right] \leq 2 \mathbb{E} \left[ \hat{M}_r(X) \cdot 1\{\hat{M}_r(X) > t\} \cdot 1(\Phi_r(X)) \right], \]

and we further let

\[ \hat{M}_r(X, t) := \hat{M}_r(X) \cdot 1\{\hat{M}_r(X) > t\} \cdot 1\{\Phi_r(X)\}. \]

Then, writing \( \mathbb{P} \) for the joint distribution over the full sample and the target marginal \( Q_x \), we have

\[ \mathbb{E} \hat{M}_r(X, t) = \int_0^1 \mathbb{P}(\hat{M}_r(X, t) > s) \, ds = \int_0^t \mathbb{P}(\hat{M}_r(X, t) > s) \, ds + \int_t^1 \mathbb{P}(\hat{M}_r(X, t) > s) \, ds. \]

Now, since \( \hat{M}_r(X, t) \) takes values on \( \{0\} \cup (t, 1] \), for \( s \in [0, t] \) we have

\[ \mathbb{P}(\hat{M}_r(X, t) > s) = \mathbb{P}(\hat{M}_r(X, t) > t), \]

whence

\[ \int_0^t \mathbb{P}(\hat{M}_r(X, t) > s) \, ds = t \mathbb{P}(\hat{M}_r(X, t) > t) \leq \frac{1}{t} \mathbb{E} \left[ \hat{M}_r(X)^2 1\{\Phi_r(X)\} \right], \]
by Chebyshev’s inequality. By Lemma 14, we have (note that \( \Phi_r(X) \subseteq S_r(X) \))

\[
\mathbb{E} \left[ \hat{M}_r(X)^2 \mathbb{1}\{\Phi_r(X)\} \right] \leq \mathbb{E} \left[ \hat{V}_r(X) \mathbb{1}\{\Phi_r(X)\} \right] \leq 2\mathbb{E}_{Q_X}[V_r(X)].
\]

Another application of Chebyshev’s inequality gives

\[
\int_t^{1} \mathbb{P}(\hat{M}_r(X, t) > s) \ ds \leq \mathbb{E}[\hat{M}_r(X)^2 \mathbb{1}\{\Phi_r(X)\}] \int_t^{1} s^{-2} \ ds \\
= \mathbb{E} \left[ \hat{M}_r(X)^2 \mathbb{1}\{\Phi_r(X)\} \right] (t^{-1} - 1) \leq 2 \frac{1}{t} \mathbb{E}_{Q_X}[V_r(X)],
\]

where the final inequality again uses Lemma 14 and the definition of \( \Phi_r(X) \); this completes

the proof. \( \square \)

**Remark 5.** The above Lemma constitutes the critical step to proving Theorem 1. Interestingly,
the proof given above essentially follows a strategy outlined by Audibert and Tsybakov in [14]
(see Lemma 5.2 and the ensuing discussion) but rejected by these authors as being inadequate
to establish upper bounds of this type under the noise condition (2.2) with \( \beta > 0 \). The issue is
resolved by splitting the excess risk into terms controlled by the approximation error \( |\tilde{\eta} - \eta| \) and
the estimation error \( |\hat{\eta} - \tilde{\eta}| \), and applying the technique they outline only to the term featuring
\( |\hat{\eta} - \eta| \), as we do in Lemma 17 above. In the absence of this splitting, we would end up with an
upper bound for the excess risk in terms of \( \mathbb{E}(\hat{\eta}(X) - \tilde{\eta}(X))^2 \); this would be insufficient to derive
the bound we seek, since under the noise condition with \( \beta > 0 \), the optimal plug-in model for
classification differs from the plug-in model that uses an estimate of the regression function that
minimizes the \( L^2 \) regression error.

Choosing \( t \) appropriately to balance the terms above immediately gives:

**Corollary 18.** Under the conditions of Lemma 17, choosing \( t = \mathbb{E}_{Q_X}[V_r(X)]^{\frac{1}{\beta + 1}} \), we have

\[
\mathbb{E} \left[ M(X) \cdot \mathbb{1}\{M(X) \leq 2|\hat{\eta}_r(X) - \tilde{\eta}_r(X)|\} \cdot \mathbb{1}\{\Phi_r(X)\} \right] \leq C \cdot \mathbb{E}_{Q_X}[V_r(X)]^{\frac{\beta + 1}{\beta + 2}}.
\]
In order to bound $\mathbb{E}_{Q_X}[V_r(X)]$, we will make use of the aggregate transfer exponent (see Definition 8) to account for the contribution of the source feature distribution $P_X$. This is done via the following technical Lemma, whose proof may be found in Appendix 2.7.4:

**Lemma 19.** Let $(P, Q) \in \mathcal{T}$, and let $\Pi_r(X_Q) := \{A \in \Pi_r : Q_X(A) > 0\}$. There exists $C > 0$ independent of $r$ such that

$$\sum_{A \in \Pi_r(X_Q)} \frac{Q_X(\tilde{A})}{P_X(\tilde{A})} \leq C r^{-\gamma}.$$ 

Note that in order to obtain Lemma 19, it is essential that the sum on the left-hand side should take the mass-ratios of the *envelopes* of the cells $A \in \Pi_r(X_Q)$; indeed, it is for this reason that we define the algorithm in this way. Equipped with this Lemma, we obtain the following bound on the expectation of $V_r$:

**Lemma 20.** Let $(P, Q) \in \mathcal{T}$, and let $V_r$ be as above. Then for some $C > 0$ free of $r$, we have

$$\mathbb{E}_{Q_X}[V_r(X)] \leq C \min \left(\frac{r^{-\gamma}}{n_p}, \frac{r^{-d}}{n_Q}\right).$$

**Proof.** Since $V_r$ is constant over the cells $A \in \Pi_r$, we have

$$\mathbb{E}_{Q_X}[V_r(X)] = \sum_{A \in \Pi_r(X_Q)} \left\{ \frac{1}{n_p P_X(\tilde{A}) + n_Q Q_X(\tilde{A})} \right\} \cdot Q_X(A) \leq \sum_{A \in \Pi_r(X_Q)} \frac{Q_X(\tilde{A})}{n_p P_X(\tilde{A}) + n_Q Q_X(\tilde{A})} \leq \frac{1}{n_Q} |\Pi_r(X_Q)| \leq C_d \frac{r^{-d}}{n_Q},$$

where the final step uses $|\Pi_r(X_Q)| \leq C_d r^{-d}$, which follows from assumption (2.3). Similarly, we have

$$\mathbb{E}_{Q_X}[V_r(X)] \leq \sum_{A \in \Pi_r(X_Q)} \frac{Q_X(\tilde{A})}{n_p P_X(\tilde{A}) + n_Q Q_X(\tilde{A})} \leq \frac{1}{n_p} \sum_{A \in \Pi_r(X_Q)} \frac{Q_X(\tilde{A})}{P_X(\tilde{A})} \leq C \frac{r^{-\gamma}}{n_p},$$

37
where the final inequality is an application of Lemma 19, and the result follows immediately.

Combining Lemma 14, Corollary 18, and Lemmas 19 and 20 gives:

**Proposition 21.** Let \((P, Q) \in \mathcal{T}\), let \(\Phi_r(X)\) be as in Notation 16. Then there is a constant \(C > 0\), independent of \(n_P, n_Q\), such that

\[
\mathbb{E} [M(X) \cdot 1\{M(X) \leq 2|\hat{\eta}_r(X) - \tilde{\eta}_r(X)|\} \cdot 1\{\Phi_r(X)\}] \leq C \left( \min \left( \frac{r^{-\gamma} r^{-d}}{n_P}, \frac{r^{-d}}{n_Q} \right) \right)^{\frac{\theta+1}{\theta+2}},
\]

where the expectation is over the full sample \((X, Y)\) and the target \(X \sim Q_X\).

Up until now, we have established bounds on the excess expected risk of \(\hat{f}_r\) conditioned on the events \(S_r(X)\) and \(\Phi_r(X)\), and it therefore remains to show that these events occur with high enough probability. Since clearly \(\Phi_r(X) \subset S_r(X)\), we must show that \(\Phi_r(X)^c\) is sufficiently rare. In order to see this, we employ a multiplicative Chernoff bound (Lemma 34 in Appendix 2.7.4), followed by an application of Lemma 19. This gives:

**Lemma 22.** Let \(\mathbb{P}\) denote the joint probability over the sample and the target, and let \(\Phi_r(X)\) be as in Notation 16. There is a constant \(C > 0\) free of \(r, n_P, n_Q\) such that we have

\[
\mathbb{P}(\Phi_r(X)^c) \leq C \min \left( \frac{r^{-\gamma} r^{-d}}{n_P}, \frac{r^{-d}}{n_Q} \right)
\]

Proof. Recall that \(\Phi_r(X) = \{\hat{V}_r(X) < 2V_r(X)\}\), and note that \(\hat{V}_r^{-1}(X) = [\hat{A}_r(x) \cap X]\) is a sum of \(n_P + n_Q\) independent Bernoulli variables, and has mean \(V_r(X)^{-1}\). Fixing \(x \in X_Q\) and applying the multiplicative Chernoff bound (Lemma 34) gives

\[
\mathbb{P}(\Phi_r(x)^c) = \mathbb{P}(\hat{V}_r^{-1}(x) \leq \frac{1}{2} V_r^{-1}(x)) \leq \exp\{-\frac{1}{8} V_r^{-1}(x)\}.
\]

It follows then that
\[
\mathbb{P}(\Phi_r(X)^C) = \sum_{A \in \Pi_r} \mathbb{P}(\Phi_r(X)^C \cap \{X \in A\}) = \sum_{A \in \Pi_r} \mathbb{P}(\Phi_r(X)^C \mid X \in A)Q_X(A)
\]

\[
\leq \sum_{A \in \Pi_r} \exp \left\{-\frac{1}{8}V_r(A)^{-1}\right\} \cdot Q_X(A) = \sum_{A \in \Pi_r} V_r(A)^{-1} \exp \left\{-\frac{1}{8}V_r(A)^{-1}\right\} V_r(A) \cdot Q_X(A)
\]

\[
\leq \frac{8}{e} \sum_{A \in \Pi_r} V_r(A)Q_X(A) \leq 3\mathbb{E}_{Q_X}[V_r(X)],
\]

where the second inequality uses \(xe^{-ax} \leq 1/ae\), valid for all real \(x\). The proof is completed by applying Lemma 19.

Having now fully established a bound on the expected excess risk of \(\hat{f}_r\), the proof of Theorem 1 is completed upon collecting and balancing terms.

**Proof. (Theorem 1)** By Lemma 11, Propositions 15, 21 and Lemma 22, we can choose \(C > 0\) such that

\[
\mathbb{E}[\mathcal{E}(\hat{f}_r)] \leq C \left(r^{a(\beta+1)} + \min \left(\frac{r^{-\gamma}r^{-d}}{n_p}, \frac{r^{-d}}{n_Q}\right)\right) + C \min \left(\frac{r^{-\gamma}r^{-d}}{n_p}, \frac{r^{-d}}{n_Q}\right).
\]

Some algebra shows that choosing

\[
r_n^* := \min \left(n_p^{-\frac{1}{2a+2a'}}, n_Q^{-\frac{1}{2a+2a'}}\right)
\]

balances the first two terms, and the third term is clearly strictly smaller than the second. This gives a final bound of

\[
\mathbb{E}[\mathcal{E}(\hat{f}_n^*)] \leq C \min \left(n_p^{-\frac{a(\beta+1)}{2a+2a'}}, n_Q^{-\frac{a(\beta+1)}{2a+2a'+d}}\right),
\]

completing the proof.
Adaptive Rates.

In this section we outline the proof of Theorem 3, establishing an upper bound on the excess risk of the classifier output by Algorithm 1 that locally determines a depth at which to estimate $\eta$ at a point $x$. We demonstrate that it achieves the minimax rate plus a log-term. Note that ‘log-spoiled’ adaptive rates are usually unavoidable (see the discussion in [13] for more detail on this point) without any knowledge of the unknown parameters. The method proceeds by making an initial choice of tree-depth, and then for an input point $x$, proceeds up the branch containing $x$ until a level is found at which the local error due to bias and variance are approximately balanced (see Algorithm 1). The level of the tree which achieves this will depend on the problem parameters and is therefore unknown, and the crux of the argument lies in demonstrating that this can nonetheless be achieved with only an estimate of the local variance.

We introduce the following notation: fix $\epsilon = (n_P + n_Q)^{-1/2}$, and let the set of admissible levels be given by

$$\hat{R}_n := \{2^{-i} \lceil \log_2(1/\epsilon) \rceil \}_{i=0}.$$ (2.21)

The key to this approach is to establish control of the error $|\hat{\eta}_r(x) - \tilde{\eta}_r(x)|$ uniformly for $x \in X_Q$, and at all levels $r \in \hat{R}_n$, simultaneously. Using McDiarmid’s bounded differences inequality [38], we can derive the following bound in high probability. Recall that $A_r(X)$ denotes the cell containing $x$ at level $r$, $\tilde{A}_r$ is its $r$-envelope, $|\tilde{A}_r(X) \cap X|$ denotes the number of sample points falling in $\tilde{A}_r(X)$, and we let $\hat{V}_r(X) = |\tilde{A}_r(X) \cap X|^{-1}$.

**Lemma 23.** Let $P, Q \in T$. Fix $\delta > 0$, and let $\hat{\eta}_r, \tilde{\eta}_r$ be as in (2.5) and (2.17), respectively. Then with probability at least $1 - \delta$ over the joint sample $(X, Y)$, we have, for all $x \in X_Q$ and each $r \in \hat{R}_n$, that

$$|\hat{\eta}_r(x) - \tilde{\eta}_r(x)| \leq \frac{1}{2} \sqrt{\hat{V}_r(x)} \left( 1 + 2 \sqrt{\log(n_P + n_Q)/\delta} \right).$$ (2.22)
Using this, the triangle inequality, and the fact that the regression function is $\alpha$-Hölder continuous immediately gives

**Lemma 24.** Let $P, Q \in T$. Fix $\delta > 0$, and let $\hat{\eta}_r, \eta_r$ be as in (2.5) and (2.17), respectively. Then with probability at least $1 - \delta$ over the joint sample $(X, Y)$, we have, for all $x \in X_Q$ and each $r \in \hat{R}_n$, that

$$|\hat{\eta}_r(x) - \eta(x)| \leq \frac{1}{2} \sqrt{\hat{V}_r(x)} \left(1 + 2 \sqrt{\log(n_P + n_Q)/\delta}\right) + C_\alpha r^n.$$

Note that above we need not worry about controlling for empty cells in applying the Hölder condition above, as the bound is trivial in this case (as $|\tilde{A}_r(x) \cap X| = 0$). Now, let $C_{n,\delta} = \left(1 + 2 \sqrt{\log(n_P + n_Q)/\delta}\right)$ and set $\hat{\sigma}_r(x) = C_{n,\delta} \hat{V}_r(x)^{1/2}$, and set $r_0 = 2^{-[\log(n_P + n_Q)/2]}$ (the smallest value in $\hat{R}_n$), and choose $\hat{r}(x)$ by implementing Algorithm 1 with these specifications. To show that this procedure yields the correct rates we appeal to the following Lemma, after which the proof follows the same argument as for Theorem 1.

Let

$$r_n := \min \left\{ \left(\frac{\log n_P}{n_P}\right)^{\frac{1}{2a+2\beta+\gamma}}, \left(\frac{\log n_Q}{n_Q}\right)^{\frac{1}{2a+2\beta+d}} \right\};$$

the optimal depth choice implied by Theorem 1 with an adjustment due to the logarithmic dependence of $C_{n,\delta}$ on $n = n_P + n_Q$. The critical result is as follows:

**Lemma 25.** There is a universal constant $C$ such that the following holds. Let $r_n$ be as in (2.23). For a fixed $x$, let $\hat{f}(x)$ denote the output of Algorithm 1 when run from level $r_0 = (n_P + n_Q)^{-1/2}$, with $\hat{\sigma}_r(x)$ as above and $C_{n,\delta} = \frac{1}{2} \left(1 + 2 \sqrt{\log(n_P + n_Q)/\delta}\right)$. Then with probability at least $1 - \delta$ over the full sample, simultaneously at all $x \in X_Q$, we have:

$$\mathbb{I}\{f^*(x) \neq \hat{f}(x)\} \leq \mathbb{I}\{||\eta(x) - \frac{1}{2}|| \leq C \left(C_{n,\delta} \hat{V}_{r_n}(x)^{1/2} + C_\alpha r_n^\alpha \right) \}.$$  

(2.24)
This allows us to write the following bound, in analogy with Lemma 11 from the previous subsection:

**Lemma 26.** Let $C_{n,\delta}$ and $\hat{f}$ be as above, and take $r_n$ as in (2.23); set $M(x) = |\eta(x) - \frac{1}{2}|$. With probability at least $1 - \delta$ over the full sample $(X, Y)$, we have

$$E_Q(\hat{f}) \leq 2 \int M(x) \cdot \mathbb{1}\{M(x) \leq 2C_{n,\delta} \hat{V}_{r_n}(x)^{1/2}\} Q_X(dx) + 2 \int M(x) \cdot \mathbb{1}\{M(x) \leq 2C_n r_n^\alpha\} Q_X(dx).$$

The rest of the proof of Theorem 3 can be completed in the exact same manner as Theorem 1, except that Lemmas 13, 14 are superseded by Lemma 24. Note that the presence of the additional term involving $C_{n,\delta} = \sqrt{\log(nP + n_Q)}$ is accounted for in the final balancing step by choosing $r_n$ as in (2.23). The full argument is given in Appendix 2.7.4.

**Localization to the Decision Boundary**

In order to prove Theorem 4, we consider an oracle classification model, based on the dyadic-tree classifier introduced above that, for fixed $\epsilon$, estimates the regression function at different levels of the tree based on whether the input is in $G_\epsilon^-$ or $G_\epsilon^+$.

**Definition 19.** Let $\hat{f}_r$ denote the dyadic-tree classifier of Definition 12, built from the tree from Definition 10 at level $r$. For fixed $\epsilon$, let $r_\epsilon = r_\epsilon(\epsilon)$, $r_+ = r_+(\epsilon)$ be functions of $\epsilon$ (to be specified later), and let $\hat{f}_\epsilon$ be the hybrid dyadic-tree classifier that estimates at level $r_\epsilon$ when $x \in G_\epsilon^-$ and at level $r_+$ when $x \in G_\epsilon^+$, that is,

$$\hat{f}(x) = \hat{f}_{r_\epsilon}(x) \mathbb{1}\{x \in G_\epsilon^-\} + \hat{f}_{r_+}(x) \mathbb{1}\{x \in G_\epsilon^+\}.$$ 

Using the techniques used to prove Theorem 1, we can bound the contribution to the excess risk of $\hat{f}$ of the region $G_\epsilon^-$. 

42
Notation 27. For a set \( A \subset X \) we let

\[
\mathcal{E}(\hat{f}; A) := 2 \int_A |\eta(x) - \frac{1}{2} \mathbb{1}\{\hat{f}(x) \neq f^*(x)\}| Q_X(dx),
\]

(2.25)

so that \( \mathcal{E}(\hat{f}; \mathcal{G}_e^-) \) denotes the excess risk of \( \hat{f} \) from the region \( \mathcal{G}_e^- \).

We have the following bound.

Lemma 28 (Risk contribution from \( \mathcal{G}_e^- \)). There is a universal constant \( C = C(T) \) such that the following holds. Fix \( \epsilon > 0 \), and let \( \hat{f}_\epsilon \) and \( \mathcal{G}_e^- \) be as above, and set \( r_- = \min\left(n_p^{-1/(2\alpha+\alpha\beta+\gamma(\epsilon))}, n_Q^{-1/(2\alpha+\alpha\beta+d)}\right) \). Then we have

\[
\mathbb{E} \mathcal{E}(\hat{f}_\epsilon; \mathcal{G}_e^-) \leq C \min\left\{n_p^{-\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\gamma(\epsilon)}}, n_Q^{-\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+d}}\right\}.
\]

Proof. This follows immediately from Theorem 1, since by (2.25) we need only integrate over \( \mathcal{G}_e^- \), and in this region we have the local aggregate exponent \( \gamma(\epsilon) \). \hfill \Box

Lemma 29 (Risk contribution from \( \mathcal{G}_e^+ \)). There is a universal constant \( C = C(T) \) such that the following holds. Fix \( \epsilon > 0 \), and let \( \hat{f}_\epsilon \) and \( \mathcal{G}_e^+ \) be as above, and suppose that \( r_+ < (\epsilon/2C_\alpha)^{1/\alpha} \). Then we have

\[
\mathbb{E} \mathcal{E}(\hat{f}_\epsilon; \mathcal{G}_e^+) \leq C \frac{1}{\epsilon} \min\left(\frac{1}{n_p r_+^{-\gamma^*}}, \frac{1}{n_Q r_+^{-d}}\right).
\]

Notice that the above expression is minimized when \( r_+ \) is taken as large as possible. This immediately implies:

Corollary 30. Setting \( r_+ := (\epsilon/4C_\alpha)^{1/\alpha} \), we have

\[
\mathbb{E} \mathcal{E}(\hat{f}; \mathcal{G}_e^+) \leq C \min\left(\frac{1}{n_p \epsilon^{-(1+\gamma^*/\alpha)}}, \frac{1}{n_Q \epsilon^{-(1+d/\alpha)}}\right).
\]

Let \( \hat{f}_\epsilon \) be as above, where

\[
r_- = \min\left(n_p^{-1/(2\alpha+\alpha\beta+\gamma(\epsilon))}, n_Q^{-1/(2\alpha+\alpha\beta+d)}\right)
\]

43
and \( r_+ = (\epsilon/4C_\alpha)^{1/\alpha} \). Combining the above results gives

**Corollary 31.** There is a universal constant \( C = C(\mathcal{T}) \) such that, for any \( \epsilon > 0 \), if \( \hat{f}_\epsilon \) is set as above then we have

\[
\mathbb{E} \mathcal{E}(\hat{f}_\epsilon) \leq C \left[ \min \left( \frac{1}{n_P} \epsilon^{-(1+\gamma^*/\alpha)} , \frac{1}{n_Q} \epsilon^{-(1+d/\alpha)} \right) \right].
\]

**Theorem 32.** Let \( \hat{f} \) denote the adaptive classifier of Algorithm 1, and suppose that \( \alpha + d > 1 \). Then there are constants \( C_1, C_2 \), depending only on \( \mathcal{T} \), such that the following holds. If \( \max(n_P, n_Q) \geq C_1 \), then for any \( \epsilon > 0 \) we have

\[
\mathbb{E} [\mathcal{E}(\hat{f})] \leq C_2 \left[ \min \left( \frac{\log n_P}{n_P} \epsilon^{\alpha/(2\alpha+\alpha\beta+\gamma \epsilon)} , \frac{\log n_Q}{n_Q} \epsilon^{\alpha/(2\alpha+\alpha\beta+\gamma n)} \right) + \min \left( \frac{1}{n_P} \epsilon^{-(1+\gamma^*/\alpha)} , \frac{1}{n_Q} \epsilon^{-(1+d/\alpha)} \right) \right].
\]  

(2.26)

**Remark 6.** Note that in the proof of Theorem 32, we in fact consider only \( \epsilon \) bounded away from zero: indeed, we assume that \( \epsilon \geq 4C_\alpha (n_P + n_Q)^\alpha \). We show however that this introduces no loss of generality, since taking \( \epsilon \) below this threshold gives a vacuous excess risk bound.

The proof of Theorem 32 can now be completed using techniques used to prove Theorem 3; details are in Appendix B. Theorem 5 now follows immediately.

**Proof.** [Theorem 5] It is easy to see that the function \( \gamma(\epsilon) \) is right-continuous, and therefore by continuity we have

\[
\frac{1}{n_P} (\epsilon n_p)^{-(1+\gamma^*/\alpha)} \leq \left( \frac{\log n_P}{n_P} \right)^{\alpha/(2\alpha+\alpha\beta+\gamma n)}.
\]

Letting \( C \) be the constant appearing in Theorem 32 and bounding the minimums in (2.26) by their first arguments yields

\[
\mathbb{E}(X,Y) [\mathcal{E}(\hat{f})] \leq 2C \left( \frac{\log n_P}{n_P} \right)^{\alpha/(2\alpha+\alpha\beta+\gamma n)}.
\]
Figure 2.3: Target risk estimates for four tree-pruning methods used on dyadic trees applied respectively to the Algerian Forest Fire data \((d = 11)\), the MAGIC data \((d = 10)\), the SUSY data \((d = 18)\), and the Biodegradation data \((d = 41)\). Error bars show the standard errors over 10 iterations (except on the Forest Fire data, where we use 50 iterations). AD (blue) indicates the adaptive method, CV (yellow) is 2-fold cross validation performed by taking the hold-out risk on target samples only, FCV (green) is 2-fold cross validation performed by taking the hold-out risk on the combined sample, and IWCV is the importance weighted CV method, where the hold-out risk is taken over all points, and at source points it is weighted by the density ratio of the features at those points. LOO denotes leave-one-out cross-validation.

Taking \(\epsilon = 1\) in Theorem 32 and bounding the minimums by their second arguments gives

\[
\mathbb{E}_{(X,Y)}[\mathbb{E}(\hat{f})] \leq 2C \left( \frac{\log n_Q}{n_Q} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+d}},
\]

and combining these bounds gives the result.
2.6 Experiments On Real Data

In this section we carry out a brief sequence of experiments using real-world datasets to illustrate the effectiveness of our proposed adaptive method for tree-pruning as opposed to alternative methods under covariate-shift based on empirical risk minimization. In particular, in each experiment, we grow a slight variant of the dyadic tree (12) in which nodes are split by randomly choosing a feature along which the node in question is longest, so that the ratio of the longest to shortest edge never exceeds 2). We contrast the effectiveness of locally selecting the tree depth adaptively via Algorithm 1 (AD) with three alternative strategies in which the depths are chosen by a) cross validation (CV) in which the hold-out sets are made by splitting the target sample points, over which we compute the empirical risk (in one experiment we also implement the leave-one-out cross-validation (LOO), where we fit $n_{\Omega}$ times in the model-selection stage), b) cross validation performed as usual over the full (source and target) sample (FCV), and c) importance-weighted cross validation (IWCV), in which the hold-out sets use the whole sample, but the empirical risk is calculated by assigning a weight to the risk contribution of each source point according to the estimated density ratio of the source/target distributions at that point; see [19] for details and discussion. To estimate the density ratios required by the IWCV method, we use the RuLSIF method (see [39]) as implemented in the associated densratio package. Note that density ratios for the IWCV method are pre-computed using both full samples (that is, using all the data at our disposal, even when we report error rates for the methods learned from smaller subsamples) to get the best possible estimates. Note that all the datasets are preprocessed by scaling the features so that they lie in $[0, 1]^d$.

2.6.1 Datasets.

Each of the datasets that we use can be found on the UCI machine learning repository [46]. See Table 2.1 for details and references.
2.6.2 Results.

Figure 2.3 display the empirical target risk for the aforementioned methods at various source/target sample size configurations for the datasets in Table 2.1. At each of the chosen sample size settings, we use the three cross-validation based methods to perform model selection, and then grow a single tree from which the predictions for all four models are computed for the target test set. This is averaged over 10 iterations for each setting (except for the Algerian Forest Fire data, where we use 50 iterations); for each iteration, a source/target sample of the required size is taken uniformly at random from all samples available.

Immediately we observe that in all configurations, on each dataset, the adaptive method performs at least as well as the empirical-risk based methods, and in certain settings considerably better. On the Algerian Forest Fire data (Figure 2.3, top left), the MAGIC data (Figure 2.3, bottom left) and the SUSY data (Figure 2.3, bottom right) the gulf in performance is considerable, while for the Biodegradation data (Figure 2.3, top right), the adaptive method is roughly matched in performance by the cross-validation approaches that use the full sample to compute the empirical risk (FCV and IWCV), although we note that even in this setting where the adaptive method does not seem to offer much improvement, it nevertheless performs no worse than the baselines. It should of course be remembered that the AD method is computationally much cheaper to implement, since it does not require multiple trees to be grown for model selection purposes, nor does it require the calculation of any empirical risks.

All the experiments above use a value of $C = 1/4$ for the AD method, which was found to perform reasonably well across a wide variety of settings. At the price of increasing computation time by a constant factor, this quantity could be straightforwardly selected in a data-dependent way by simply doing cross-validation over hold-out sets of the target sample, although we do not pursue this here ([7] demonstrate that, for a similar nearest-neighbours based procedure, this strategy does not affect the attained rates). Finally, we note as well that the error bars tend to be considerably smaller for the AD method than for the cross-validation based methods, especially when the ratio of source/target samples is large.
Finally, we remark that while the CV method generally exhibits worse performance than the IWCV method (as expected), there is almost no difference between the models selected with IWCV and with the full pooled sample cross-validation (FCV). We suspect that this is an artefact of the datasets we have chosen; [19] demonstrate examples where FCV performs considerably more poorly than IWCV. Our results nevertheless indicate that using ICI for tree pruning may be fruitful even in the single population case $P_X = Q_X$.

In Appendix 2.7.5, we provide further empirical results indicating that even when the trees are grown using a greedy procedure, the adaptive depth selection (AD) method can still yield a uniform improvement over the cross-validation based model selection procedures, albeit by a seemingly smaller margin.

### 2.7 Appendix to Chapter 2.

#### 2.7.1 Properties of the Aggregate Transfer Exponent.

In this section we prove some of the claims from Section 2.4, and collect some further simple results concerning the relative dimension and its relation to previously studied notions of dimension for measures. We begin with Proposition 9.

**Proof.** (Proposition 7.) Using the definition of transfer exponent (6), we see that

$$\int \frac{1}{P(B(x, r))} dQ(x) = \int \left( \frac{Q(B(x, r))}{P(B(x, r))} \right) \frac{1}{Q(B(x, r))} dQ(x) \leq C \rho r^{-\rho} \int \frac{1}{Q(B(x, r))} dQ(x).$$

Let $Z$ be a minimal $r/2$ cover of $X_Q$, so that by assumption (2.3) we have $|Z| \leq C_d r^{-d}$. Now, if $x \in B(z, r/2)$, then clearly $B(z, r/2) \subset B(x, r)$, so we have

$$\int \frac{1}{Q(B(x, r))} dQ(x) \leq \sum_{z \in Z} \int_{B(z, r/2)} \frac{1}{Q(B(x, r))} dQ(x) \leq \sum_{z \in Z} \int_{B(z, r/2)} \frac{1}{Q(B(z, r/2))} dQ(x) = |Z| \leq C_d r^{-d}.$$
so $\rho + d$ is an integrated transfer exponent from $P$ to $Q$, and the result follows. \qed

We now provide a simple Lemma giving an equivalent definition of the relative Minkowski dimension, which will lead us to a proof of Proposition 9.

**Lemma 33.** Let $\gamma_*(P, Q)$ be as in (2.12). Let $\Xi(r)$ denote the set of $r$-grids for $X_Q$ (see Definition 7). For an $r$-grid $\xi \in \Xi(r)$, let

$$\Lambda(\xi, r) = \sum_{E \in \xi} \frac{Q(E)}{P(E)},$$

and let $\Lambda(r) = \sup_{\xi \in \Xi(r)} \Lambda(\xi, r)$. Then

$$\gamma_*(P, Q) = \limsup_{r \to 0} \frac{\log \Lambda(r)}{-\log r}.$$

**Proof.** If $\gamma \in \{ \gamma \geq 0 : \exists C \gamma \text{ s.t. } (2.4) \text{ holds.}\}$, then $\Lambda(r) \leq C_r r^{-\gamma}$ and clearly then

$$\limsup_{r \to 0} \frac{\log \Lambda(r)}{-\log r} \leq \gamma,$$

whence $\limsup_{r \to 0} -\log \Lambda(r)/\log(r) \leq \gamma_*$. On the other hand, if

$$\limsup_{r \to 0} \frac{\log \Lambda(r)}{-\log r} = s,$$

then if $\epsilon > 0$ we have for small enough $r$ that $-\log \Lambda(r)/\log r \leq \gamma + \epsilon$, so there is some $C$ such that $\Lambda(r) \leq Cr^{-(s+\epsilon)}$, so $\gamma_* \leq s + \epsilon$, and $\epsilon$ is arbitrary so

$$\gamma_*(P, Q) \leq \limsup_{r \to 0} \frac{\log \Lambda(r)}{-\log r},$$

and this completes the proof. \qed

**Proof.** (Proposition 8) Fix an arbitrary $\epsilon > 0$, so that $\rho - \epsilon$ is not a transfer exponent from $P$ to $Q$. 49
In particular, for each \( C > 0 \), there is an \( x \in \mathcal{X}_Q \) and an \( r \in (0, 1) \) such that

\[
\frac{Q(B(x, r))}{P(B(x, r))} \geq Cr^{-(\rho-\epsilon)}.
\]

Set \( C = 1 \), and pick a sequence \((x_n, r_n)\) \( \in \mathcal{X}_Q \times (0, 1) \) with \( r_n \to 0 \) such that

\[
\frac{Q(B(x_n, r_n))}{P(B(x_n, r_n))} \geq r_n^{-(\rho-\epsilon)}.
\]

Note that we can certainly ensure that we can pick \( r_n \to 0 \), since if that were not the case it would imply that there was an \( r_0 > 0 \) such that for all \( x \in \mathcal{X}_Q \) and \( r \in (0, r_0) \) we had \( P(B(x, r)) \geq r^{(\rho-\epsilon)}Q(B(x, r)) \). We could then let

\[
C := \sup_{r \in [r_0, 1), x \in \mathcal{X}_Q} r^{(\rho-\epsilon)} \frac{Q(B(x, r))}{P(B(x, r))} \geq 1
\]

we would have \( P(B(x, r)) \geq C^{-1}r^{-(\rho-\epsilon)}Q(B(x, r)) \) for all \( x \in \mathcal{X}_Q, r \in (0, 1) \) (of course, we have \( C < \infty \)), and this would contradict the fact that \( \rho - \epsilon \) was not a transfer exponent from \( P \) to \( Q \).

Now, let \( C_{2Q} \) be the doubling constant of \( Q \), so that \( Q(B(x, r)) \geq C_{2Q}Q(B(x, 2r)) \) for all \( x \in \mathcal{X}_Q \) and \( r > 0 \). For any \( n \geq 1 \), we have

\[
\varphi(r_n/2) = \int \frac{1}{P(B(x, r_n/2))} Q(dx) \geq \int_{B_{r_n/2}(x_0)} \frac{1}{P(B(x, r_n/2))} Q(dx)
\]

\[
\geq \int_{B_{r_n/2}(x_0)} \frac{1}{P(B(x_n, r_n))} Q(dx) = \frac{Q(B(x_n, r_n/2))}{P(B(x_0, r_n))} = \frac{Q(B(x_n, r_n))}{P(B(x_n, r_n))} \geq C_{2Q} r_n^{-(\rho-\epsilon)},
\]

and it follows that \( \bar{\rho}^* \geq \rho - \epsilon \). Since \( \epsilon > 0 \) was arbitrary, the claim follows. \( \square \)
Proof. (Proposition 9.) Pick an arbitrary $r$-grid $\xi$ of $X_Q$, and let

$$\varphi(r) = \int P(B(x, r))^{-1} dQ(x).$$

Since for $E \in \xi$ there is an $x_E \in X$ such that $E \subset B(x_E, 2r)$, clearly $X \subset \bigcup_{E \in \xi} B(x_E, 2r)$ and so

$$\varphi(4r) = \int \frac{1}{P(B(x, 4r))} dQ(x) \leq \sum_{E \in \xi} \int_{B(x_E, 2r)} \frac{1}{P(B(x, 4r))} dQ(x),$$

and for $x \in B(x_E, 2r)$ we have $B(x_E, 2r) \subset B(x, 4r)$, and therefore

$$\varphi(4r) \leq \sum_{E \in \xi} \int_{B(x_E, 2r)} \frac{1}{P(B(x_E, 2r))} dQ(x) = \sum_{E \in \xi} \frac{Q(B(x_E, 2r))}{P(B(x_E, 2r))}.$$ 

Now, since $Q$ is a doubling measure there is a constant $C_{2Q}$ such that $Q(B(x, r)) \geq C_{2Q} Q(B(x, 2r))$ for all $x \in X_Q$, so

$$\frac{Q(B(x_E, 2r))}{P(B(x_E, 2r))} \leq \frac{C_{2Q}^{-1} Q(B(x_E, r))}{P(E)} \leq \frac{C_{2Q}^{-1} Q(E)}{P(E)},$$

and so we have $C_{2Q} \varphi(4r) \leq \Lambda(\xi, r) \leq \Lambda(r)$, so $\bar{\varphi}_*(P, Q) \leq \gamma_*(P, Q)$ follows after taking logarithms, dividing by $\log(1/r)$ and letting $r \to 0$ by Lemma 33.

Let now $\xi$ be an $r$-grid of $X_Q$, and let $x_E$ be chosen so that $B(x_E, r) \subset E \subset B(x_E, 2r)$. Then using that $P, Q$ are doubling and that $x \in B(x_E, r)$ implies $B(x, r) \subset B(x_E, 2r)$ gives

$$\varphi(r) = \int \frac{1}{P(B(x, r))} dQ(x) \geq \sum_{E \in \xi} \int_{B(x_E, r)} \frac{1}{P(B(x, r))} dQ(x)$$

$$\geq \sum_{E \in \xi} \int_{B(x_E, r)} \frac{1}{P(B(x_E, 2r))} dQ(x) = \sum_{x_E \in \xi} \frac{Q(B(x_E, r))}{P(B(x_E, 2r))}$$

$$\geq \sum_{E \in \xi} \frac{C_{2Q} Q(E)}{C_{2p} P(E)} = \frac{C_{2Q}}{C_{2p}} \Lambda(\xi, r),$$

and since $\xi$ is arbitrary we have $\varphi(r) \geq C \Lambda(r)$, which gives $\gamma_*(P, Q) \leq \bar{\varphi}_*(P, Q)$. \qed
Proof. (Proposition 10.) For small $r$, choose an $r$ packing of $S$ of size at least $r^{-d}$, which can be done by (2.3); denote the centres of the chosen balls by $Z$, and let $ξ$ be an $r$ grid that extends the packing, in the sense that for each $z ∈ Z$ there is a $E_z ∈ ξ$ such that $B(x, r) ⊂ E_z$. By solving a constrained optimization using Lagrange multipliers [47], we see that

$$\Lambda(ξ, r) = \sum_{E ∈ ξ} \frac{Q(E)}{P(E)} ≥ \left( \sum_{E ∈ ξ} Q(E)^{1/2} \right)^2,$$

and using the assumption on $S$ yields

$$\left( \sum_{E ∈ ξ} Q(E)^{1/2} \right)^2 ≥ \left( \sum_{z ∈ Z} Q(B(z, r)^{1/2}) \right)^2 ≥ |Z|(Cr^d)^{1/2} ≥ r^{-2d} Cr^d = Cr^{-d},$$

which implies $\Lambda(r) ≥ Cr^{-d}$, so taking logarithms and a limit as $r → 0$ gives

$$\gamma_*(P, Q) ≥ d.$$  

2.7.2 Details for Figure 2.2

Here we provide details of the constructions of the distributions used for Figure 2.2. Recall the setup of Example 2 from Section 2.4 below, with $d = 5$. We choose five marginal distributions $P_1^k$, $k = 0, 1, 2, 3, 4$, having density $p^k(x) ∝ d(x, A_k)^ν$ for some $ν ≥ 0$, where $d(x, A) = \inf_{y ∈ A} \|x - y\|$ is the standard distance function and

$$A_k = \{x ∈ [0, 1]^5 : (x_{k+1}, \ldots, x_5) = 0\}$$

are subsets of $[0, 1]^5$ of dimension $k$. We set $Q_X$ to be uniform on $[0, 1]^5$, and it follows from the discussion in Section 2.4 that we have $ρ(P_1^k, Q_X) = ν$ (so $ρ + d = ν + d$), while $γ(P_1^k, Q_X) = ν + k$.

The top plot in Figure 2.2 shows the results of applying a regular dyadic tree procedure, with depth chosen by 2-fold cross-validation, for transferring from $P_i$ to $Q$, where $P_i$ and $Q$ are determined from $P_1^k$ and $Q_X$ and the regression function $η(x) = \frac{1}{2}(1 + \sin(\pi\|x\|_1))$. As can be seen, transfer performance from $P$ to $Q$ deteriorates as $γ(P_X, Q_X)$ increases, as our results predict. Since
each of the source measures have the same transfer exponent $\nu$, this naturally illustrates the drawback of using that quantity as an indicator of potential for transfer. The bottom plot of Figure 2.2 shows the results of the same procedure applied to the distributions $P_X^k$ for $k = 0, 1, 2, 3, 4$, where the density is now given by $p^k(x) \propto d(x, A_k)^{\nu-k}$, so that now $\rho(P_X^k, Q_X) = \nu - k$, while $\gamma(P_X^k, Q_X) = \max(\nu, d)$ is constant for each distribution, and again $\nu$ is set to 5; the regression function is unchanged. As we can clearly see in the figure, in this case there is little to tell between the distributions in terms of how effectively we can transfer, which is in line with what we expect, since each distribution has the same aggregate transfer exponent with respect to the target measure. On the other hand, the transfer exponents do change, but in this instance the higher transfer exponents do not translate to inferior transfer.

2.7.3 Calculations for Example 2.

Recall that we take $Q_X$ to be uniformly distributed on $[0, 1]^d$, and for $x \in [0, 1]^d$ we let $P_X$ have density $p(x) \propto \|x\|^\nu$ for some $\nu > 0$. Note that in this case, both $P_X$ and $Q_X$ are doubling measures, so we may use the integral form (2.14) to calculate $\gamma$. For $x \in B(0, 2r)$, we have (up to a constant) $P_X(B(x, r)) \geq P_X(B(0, r)) = Cr^d + \nu$ for some constant $C$. For $x \notin B(0, 2r)$ and $y \in B(x, r)$, we have $\|y\| \geq (1 - \frac{r}{\|x\|}) \|x\| \geq \frac{1}{2} \|x\|$, and therefore $P_X(B(x, r)) \geq 2^{-\nu}\|x\|^{\nu}V_d(r)$, where $V_d(r) = C(d)r^d$ is the volume of an $r$-ball in dimension $d$. This implies that we have

$$\int_{[0,1]^d \cap B(0,2r)} P_X(B(x,r))^{-1} Q_X(dx) \leq \int_{B(0,2r)} C_1 r^{-(\nu+d)} dx = C_1 C(d) 2^dr^{-\nu},$$  \hspace{1cm} (2.27)

while

$$\int_{[0,1]^d \setminus B(0,2r)} P_X(B(x,r))^{-1} Q_X(dx) \leq 2^\nu C(d)^{-1} r^{-d} \int_{[0,1]^d \setminus B(0,2r)} \|x\|^{-\nu} dx.$$

(2.28)
Now, we have

\[
\int_{[0,1]^d \setminus B(0,2r)} \|x\|^{-\nu} \, dx \leq 2^{-d} \int_B \|x\|^{-\nu} \, dx,
\]  

(2.29)

and switching to spherical coordinates and integrating over the angular variables gives

\[
\int_{B(0,\sqrt{d}) \setminus B(0,2r)} \|x\|^{-\nu} \, dx = \int_{[0,\pi/2]^{d-1}} \int_{2r}^{\sqrt{d}} s^{-\nu} C(\varphi) s^{d-1} \, d\varphi \, ds
\]

\[
= C \int_{2r}^{\sqrt{d}} s^{-(\nu-d-1)} \, ds.
\]  

(2.30)

(2.31)

We distinguish three cases: firstly, if \( \nu < d \), then \(- (\nu - d - 1) > -1\) and we have

\[
\int_{2r}^{\sqrt{d}} s^{-(\nu-d-1)} \, ds = O(1)
\]  

(2.32)

as \( r \to 0 \), and therefore in this case by (2.28)-(2.32) we have

\[
\int_{[0,1]^d \setminus B(0,2r)} P_X(B(x,r))^{-1} Q_X(dx) \leq \int_{[0,1]^d \setminus B(0,2r)} P_X(B(x,r))^{-1} Q_X(dx) \leq C r^{-d},
\]

and so \( d = \max(\nu, d) \) is an aggregate exponent in this case. If \( \nu > d \), then we have

\[
\int_{2r}^{\sqrt{d}} s^{-(\nu-d-1)} \, ds = O(r^{-(\nu-d)})
\]  

(2.33)

as \( r \to 0 \), and so by (2.27), (2.28), (2.29) and (2.33) we have

\[
\int_{[0,1]^d} P_X(B(x,r))^{-1} Q_X(dx) \leq C r^{-\nu},
\]
and so \( \nu = \max(\nu, d) \) is an aggregate exponent. If \( \nu = d \), then we find that

\[
\int_{2r}^{\sqrt{d}} s^{-(\nu-d-1)} ds = O(\log(1/r)) \tag{2.34}
\]

as \( r \to 0 \); in this case for any \( \epsilon > 0 \) we have that \( d+\epsilon \) is an aggregate exponent, while \( \gamma_*(P_X, Q_X) = d = \max(\nu, d) \).

2.7.4 Proofs of Auxiliary Lemmas.

Here we fill in the proofs of the Lemmas that were used to prove the main results. For completeness, we also state the version of the multiplicative version of the Chernoff bound that we employ to prove Lemma 22.

**Multiplicative Chernoff Bound.**

Here we recall the Angluin-Valiant form of the relative Chernoff bound, which can be found, for example, in [48].

**Lemma 34** (Angluin-Valiant). Let \( X_1, \ldots, X_n \) be independent Bernoulli random variables with mean \( EX_i = p_i \). Let \( X = \sum X_i \), and \( \mu = EX \). Then for \( 0 < \beta < 1 \) we have

\[
\mathbb{P}(X \leq (1 - \beta)\mu) \leq \exp\left\{-\frac{1}{2} \beta^2 \mu\right\}.
\]

**Upper Bounds.**

**Proof.** (Lemma 13.) Recall that \( S_r(x) = \{|\tilde{A}_r(x) \cap X| > 0\} \), so we may write

\[
|\tilde{\eta}_r(x) - \eta(x)| \mathbb{1}\{S_r(x)\} = \frac{1}{|\tilde{A}(x) \cap X|} \sum_{i:X_i \not\in \tilde{A}(x)} (\eta(X_i) - \eta(x)) \mathbb{1}\{S_r(x)\}\]

\[
\leq \frac{1}{|\tilde{A}(x) \cap X|} \sum_{i:X_i \not\in \tilde{A}(x)} |\eta(X_i) - \eta(x)| \mathbb{1}\{S_r(x)\} \leq C_\alpha r^\alpha,
\]

where the final inequality is an application of the H"older continuity condition (2.1) on \( \eta \). \( \square \)
Proof. (Lemma 14.) Using the definitions of $\hat{\eta}$ (2.5) and $\tilde{\eta}$ (2.17), we see that we have

$$
\mathbb{E}_{Y|X}[|\hat{\eta}_r(x) - \tilde{\eta}_r(x)|^2 | X] \leq \Phi_r(X) \Rightarrow \frac{1}{|\tilde{A}_r(x) \cap X|^2} \left( \sum_{i : X_i \in \tilde{A}_r(x)} (Y_i - \eta(X_i))^2 \right)^2 | X] \leq \Phi_r(X)
$$

$$
= \frac{1}{|\tilde{A}_r(x) \cap X|^2} \sum_{i : X_i \in \tilde{A}_r(x)} \eta(X_i)(1 - \eta(X_i)) \leq \frac{1}{4|\tilde{A}_r(x) \cap X|} = \frac{1}{4} \tilde{V}_r(x),
$$

where the second equality is the Pythagorean theorem, and the inequality is simply $\eta(x)(1 - \eta(x)) \leq 1/4$, since $\eta(x) \in [0, 1]$. Since $x \in X$ was arbitrary, this completes the proof.

Proof. (Lemma 19.) Recall that the cells $A \in \Pi_r$ are hypercubes of side length $r$. Observe that by construction each of the $r$-envelopes of sets $A$ with $Q_X(A) > 0$ (that is, the cells $A \in \Pi_r(X_Q)$) has a point $x \in A$ with $B(x, r) \cap X_Q \subset \tilde{A}_r$, and of course $\tilde{A}_r \subset B(x, 2r)$; see Figure 2.4. We aim to show that there is a constant $K = K(D)$ such that one can find $K$ grids $\xi_1, \ldots, \xi_K$ of $X_Q$ so that for each $A \in \Pi_r(X_Q)$, the envelope $\tilde{A}_r$ belongs to one of the grids. In order for it to be possible for the envelopes of two such cells $A, A'$ to belong to the same grid, one must be able to find $x, x' \in X_Q$ such that $B(x, r) \cap B(x', r) = \emptyset$; a sufficient condition ensuring that this can be done is that $d(x, x') \geq 2r$ for all $x \in A, x' \in A'$, that is, the dyadic cells $A, A'$ are ‘separated’ by at least two other cells. For a given cell $A$, this fails to hold for only the $5^D$ cells that constitute the $2r$ envelope of $A$. Now, consider a graph with a vertex corresponding to each cell $A \in \Pi_r(X_Q)$, with an edge between the vertices for $A, A'$ if $A' \in \tilde{A}_2 r$. Observe that a colouring of this graph - that is, an assignment of colours to each vertex such that no two vertices of the same colour share an edge - corresponds to a group of cells that can be assigned to the same grid. A simple greedy argument shows that one can always colour a graph $G$ using $\max_{v \in G} \deg(v) + 1$ colours, and the above reasoning implies that the maximal degree of our graph is $5^D$, whence we can choose.
Figure 2.4: When the support of the target distribution lies on a lower-dimensional manifold in 
$[0, 1]^D$, the dyadic partition of the latter can lead to an irregular partition of the former; this prob-
lem is mitigated by taking envelopes. Above we see a dyadic grid, along with $X_Q$ in blue. One cell 
from the partition is in red, and its envelope in light red. The black boxes centred on $x$ are $B(x, r)$ 
and $B(x, 2r)$, demonstrate that the envelope $\hat{A}(x)$ satisfies the conditions for being a grid element 
(see 7).

$K = 5^D + 1$. The rest of the argument is immediate:

$$\sum_{A \in \Pi_r(x_Q)} \frac{Q_X(\hat{A}_r)}{P_X(\hat{A}_r)} \leq \sum_{i=1}^K \sum_{\hat{A}_r \in \xi_i} \frac{Q_X(\hat{A}_r)}{P_X(\hat{A}_r)} \leq K C \gamma^{-\gamma},$$

where the final step uses (2.4).

**Proof. Proposition 2.** Consider the transfer class $T(\alpha, \beta, \gamma, d)$ for some $\alpha, \beta, d \geq 0$ and $\gamma \geq d$.

Then, by Propositions 7 and 9, we see that $T$ contains the restricted class $\tilde{T}$, where $(P, Q) \in \tilde{T}$ 
if assumptions i)-iv) of Definition 9 hold, and $P_X$ has transfer exponent $\gamma - d$ with respect to $Q_X$, 
and $P_X, Q_X$ are doubling measures. It is easy to check that all measures used in the construction 
of [3] are doubling, and so it follows by their Theorem 1 that we have

$$\inf_{\hat{f}} \sup_{(P,Q) \in \tilde{T}} \mathbb{E} \mathcal{E}(\hat{f}) \geq C' \min \left\{ \frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\gamma}, \frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+d} \right\}.$$  

Since $\tilde{T} \subset T$, the result follows. 

57
Adaptive Rates.

Proof. (Lemma 23.) For fixed \( x, r \), let \( Y_{x,r} \) denote the \( Y \)-values corresponding to those \( X \)'s that fall in \( \tilde{A}_r(x) \). Supposing that the \( X \)'s are held fixed, let \( \phi(Y_{x,r}) = |\tilde{\eta}_r(x) - \tilde{\eta}(x)| \), and note that in the proof of Lemma 14 it is shown that \( \mathbb{E}[\phi(Y_{x,r})] \leq \frac{1}{2\sqrt{|\tilde{A}_r(x) \cap X|}} = \frac{1}{2} \tilde{V}_r(x)^{1/2} \), where the expectation is over \( Y^n \mid X^n \). Since changing one \( Y \) value can change \( \phi(Y_{x,r}) \) by at most \( \frac{1}{2\sqrt{|\tilde{A}_r(x) \cap X|}} \), for a fixed \( t_{x,r} > 0 \) an application of McDiarmid’s bounded differences inequality gives

\[
P(\phi(Y_{x,r}) > \mathbb{E}[\phi(Y_{x,r})] + t_{x,r}) \leq \exp\left(-2t_{x,r}^2|\tilde{A}_r(x) \cap X|\right),
\]

for any \( x \) such that \( |\tilde{A}_r(x) \cap X| > 0 \), where the probability on the left hand side is again over \( Y^n \) with the \( X^n \) held fixed. Note that the above holds trivially on empty cells, since \( \phi(Y_{x,r}) \equiv 0 \) there. For a fixed level \( r \), there are at most \( n_P + n_Q \) empty cells, and by construction (see (2.21)) there are no more than \( \log(n_P + n_Q) \) levels, and so setting \( t_{x,r} \) appropriately, (2.35) reads that with probability at least \( 1 - \delta/(n_P + n_Q)^2 \), we have

\[
\phi(Y_{x,r}) \leq \frac{1}{2\sqrt{|\tilde{A}_r(x) \cap X|}} + \sqrt{\log((n_P + n_Q)/\delta)} = \frac{1}{2} \tilde{V}_r(x)^{1/2} \left(1 + 2 \sqrt{\log(n_P + n_Q)/\delta}\right).
\]

Taking a union bound over each non-empty cell at each depth yields (2.22) with probability at least \( 1 - \delta \) over the conditional distribution \( Y^n \mid X^n \), and we conclude by taking an expectation over \( X^n \). \( \square \)

Proof. (Lemma 25.) Fix \( x \in X_Q \). Recall that we let \( \hat{\sigma}(x) = C_{n,\delta} \hat{V}_r(x)^{1/2} \), where \( \hat{V}_r(x) = |\tilde{A}_r(x) \cap X|^{-1} \); define

\[
r^*(x) := \max \left\{ r \in \hat{R} : \hat{\sigma}(x) \geq C_n r^\alpha \right\}.
\]

Clearly there is an \( N \) for which this set is non-empty provided \( n_P + n_Q > N \), and so \( r^* \neq -\infty \). Set \( \hat{r}(x) \) to be the depth at which Algorithm 1 stops, that is

\[
\hat{r}(x) := \min \left\{ r \in \hat{R} : \hat{\eta}_r^+ \leq \hat{\eta}_r^- \text{, or } \hat{\eta}_r^+ \leq 1/2 \text{, or } \hat{\eta}_r^- \geq 1/2 \text{, or } r = 1 \right\},
\]

58
where $\hat{\eta}_r^+$ and $\hat{\eta}_r^-$ denote the values of $\hat{\eta}^+$, $\hat{\eta}^-$ at the iteration at which the depth is set to $r$. Though $r^*$ is of course unknown, we will argue that $\hat{r}$ cannot behave too differently. Again, by Lemma 24 we have

$$|\hat{\eta}_r(x) - \eta(x)| \leq \frac{1}{2\sqrt{|A_r(x) \cap X|}} \left( 1 + 2\sqrt{\log(n_P + n_Q)/\delta} \right) + C_\alpha r^\alpha = \hat{\sigma}_r(x) + C_\alpha r^\alpha$$ (2.36)

for each $r \in \hat{R}$. We proceed by arguing by cases. Suppose first that $\hat{r} \leq r^*$. In this case, since $\hat{\sigma}_r(x)$ is non-increasing in $r$, while $C_\alpha r^\alpha$ is increasing, it must be that $\hat{\sigma}_\hat{r} \geq C_\alpha \hat{r}^\alpha$, and it follows from (2.36) that $\eta(x) \in [\hat{\eta}_r(x) - 2\hat{\sigma}_r(x), \hat{\eta}_r(x) + 2\hat{\sigma}_r(x)]$; in fact this holds for any $r < r^*$, and so we must have $\eta(x) \in [\hat{\eta}_r^-, \hat{\eta}_r^+]$, because by Algorithm 1 we have

$$[\hat{\eta}_r^-, \hat{\eta}_r^+] = \bigcap_{r \in \hat{R}, \hat{r} \leq r} [\hat{\eta}_r(x) - 2\hat{\sigma}_r(x), \hat{\eta}_r(x) + 2\hat{\sigma}_r(x)].$$

It must be that $\hat{\eta}_r^- \leq \hat{\eta}_r^+$, and so the algorithm must have stopped due to either $\hat{\eta}_r^+ \leq 1/2$ or $\hat{\eta}_r^- \geq 1/2$. In either case $\hat{\eta}_r(x)$ is on the same side of $1/2$ as $\eta(x)$, so $\hat{f}(x) = f^*(x)$ and (2.24) holds since the left hand side is equal to zero.

Suppose on the other hand that $\hat{r} > r^*$. Let

$$\hat{\sigma} := \min_{r \in \hat{R}} \{ \max(\hat{\sigma}_r(x), C_\alpha r^\alpha) \}. \quad (2.37)$$

Again, since $r \mapsto \hat{\sigma}_r$ is non-increasing while $r \mapsto C_\alpha r^\alpha$ is increasing, we see that the minimum in the above expression must be attained at either $r^*$ or $2r^*$. If the minimum occurs at $r^*$, we have $\hat{\sigma} = \hat{\sigma}_{r^*} \geq \hat{\sigma}_{2r^*}$, and in the latter case we have $\hat{\sigma}_{2r^*} \leq C_\alpha (2r^*)^\alpha = \hat{\sigma}$; it follows that we have $[\hat{\eta}_{2r^*}^-(x), \hat{\eta}_{2r^*}^+(x)] \subseteq [\hat{\eta}_{2r^*}^-(x) - 2\hat{\sigma}, \hat{\eta}_{2r^*}^+(x) + 2\hat{\sigma}]$. Now, $\hat{r} > r^*$ means $\hat{r} \geq 2r^*$, and so by construction we have

$$[\hat{\eta}_\hat{r}^-(x), \hat{\eta}_\hat{r}^+(x)] \subseteq [\hat{\eta}_{2r^*}^-(x), \hat{\eta}_{2r^*}^+(x)],$$
and and $\eta(x) \in [\hat{\eta}_{2r} - 2\sigma, \hat{\eta}_{2r} + 2\sigma]$ holds by (2.36), and so it follows that

$$|\hat{\eta}_r(x) - \eta(x)| \leq 2\sigma.$$  

(2.38)

Now, since obviously $\epsilon < r_n$, there exists an $r \in \hat{R}$ such that $r \leq r_n < 2r$, and so (2.37) and (2.38) imply

$$|\hat{\eta}_r(x) - \eta(x)| \leq 4 \left(\hat{\sigma}_{r_n}(x) + C_\alpha r_n^\alpha\right),$$

and the final claim follows immediately because

$$\{\hat{f}(X) \neq f^*(X)\} \subset \{|\eta(X) - \frac{1}{2}| \leq |\hat{\eta}_r(X) - \eta(X)|\},$$

and $x \in X_Q$ was arbitrary.  

$\Box$

**Proof. (Theorem 3.)** We proceed in the same way as for Theorem 1. Let $\hat{f}$ denote the classifier given by Algorithm 1. Recall that we have set $\hat{V}_r(x) := |\hat{A}_r(x) \cap X|^{-1}$, and $V_r(x) = \left(n_P P_X(\tilde{A}_r(x)) + n_Q Q_X(\tilde{A}_r(x))\right)^{-1}$, and we let $\Phi := \Phi_{r_n}(X) = \{\hat{V}_r(X) < 2V_{r_n}(X)\}$. Further, let

$$\Omega_\lambda := \{V_{r_n}(X) \leq \lambda\}.$$

Applying Lemma 26 with expectations taken on the event $\Phi$, we see that (recalling that we write $M(x) = |\eta(x) - \frac{1}{2}|$)

$$\mathbb{E}[\mathcal{E}(\hat{f})] \leq 2\mathbb{E} \left[ M(X) \cdot \mathbb{1} \left\{ M(X) \leq 2 \frac{C_\eta}{\sqrt{|\hat{A}_{r_n}(X) \cap X|}} \right\} \cdot \mathbb{1}\{\Phi\} \right] + 2\mathbb{E} \left[ M(X) \cdot \mathbb{1} \left\{ M(X) \leq 2 C_\alpha r_n^\alpha \right\} \cdot \mathbb{1}\{\Phi\} \right] + \mathbb{E} \left[ \mathbb{1}\{\Phi^c\} \mathbb{1}\{\Omega_\lambda\} + \mathbb{1}\{\Omega_\lambda^c\} \right].$$

By Lemmas 15, 17 and 22, we can bound the four terms in the above display, yielding
\[ \mathbb{E}[\bar{E}(\hat{f})] \leq C \left( r_n^{\alpha(\beta+1)} + C_{\delta,\eta}^2 \min \left( \frac{r_n^{-\gamma}}{n_p}, \frac{r_n^{-d}}{n_Q} \right)^{\frac{\beta+1}{\beta+2}} + \exp\{-\lambda^{-1}/8\} + \frac{1}{\lambda} \min \left( \frac{r_n^{-\gamma}}{n_p}, \frac{r_n^{-d}}{n_Q} \right) + \delta. \]

Setting \( \delta := \frac{1}{n_p+n_Q} \) gives \( C_{\delta,\eta} \leq \sqrt{d \log(n_p+n_Q)} \), and now again we can set \( \lambda = r_n^{\alpha} \) and ignore the third term, which gives a final bound of

\[ \mathbb{E}[\bar{E}(\hat{f})] \leq C \min \left\{ \left( \frac{\log n_p}{n_p} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\delta}}, \left( \frac{\log n_Q}{n_Q} \right)^{\frac{\alpha(\beta+1)}{2\alpha+\alpha\beta+\delta}} \right\}. \]

for some \( C > 0 \) free of \( n_p, n_Q \), and this completes the proof.

□

**Localizing to the Decision Boundary**

**Lemma 29.** First, note that we have (recall that \( \{\hat{f} \neq f^*\} \subset \{|\eta(x) - \frac{1}{2}| \leq |\hat{\eta}_r(x) - \eta(x)|\} \))

\[ \mathbb{E}[\bar{E}(\hat{f}; \mathcal{G}^+_{\epsilon})] = 2 \mathbb{E}\left[ M(X) \mathbb{1}\{\hat{f} \neq f^*\} \mathbb{1}\{X \in \mathcal{G}^+_{\epsilon}\}\right] \leq 2 \mathbb{E}\left[ M(X) \cdot \mathbb{1}\{M(X) \leq |\hat{\eta}_r(X) - \eta(X)|\} \cdot \mathbb{1}\{X \in \mathcal{G}^+_{\epsilon}\}\right] \]

\[ \leq 2 \mathbb{E}\left[ M(X) \cdot \mathbb{1}\{M(X) \leq 2|\hat{\eta}_r(X) - \eta_r(X)|\} \cdot \mathbb{1}\{X \in \mathcal{G}^+_{\epsilon}\}\right] + 2 \mathbb{E}\left[ M(X) \cdot \mathbb{1}\{M(X) \leq 2|\hat{\eta}_r(X) - \eta(X)|\} \cdot \mathbb{1}\{X \in \mathcal{G}^+_{\epsilon}\}\right]. \]

Now, let \( \Phi_r(X) \) be the event introduced in Notation 16. On \( \Phi_r(X) \), we have (see Lemma 13) \( |\hat{\eta}_r(X) - \eta(X)| \leq C_r r_n^\alpha \), and so \( r_+ < (\epsilon/2C_r)^{1/\alpha} \) implies that

\[ \mathbb{1}\{M(X) \leq 2|\hat{\eta}_r(X) - \eta(X)|\} \mathbb{1}\{X \in \mathcal{G}^+_{\epsilon}\} \cdot \mathbb{1}\{\Phi_r(X)\} = 0. \]
We then have

$$\mathbb{E}(\hat{f}_\epsilon; G_\epsilon^+) \leq 2 \mathbb{E} \left[ M(X) \cdot 1 \{ M(X) \leq 2|\hat{\eta}_r(X) - \bar{\eta}_r(X)| \} 1\{X \in G_\epsilon^+\} \right] + \mathbb{P}(\Phi_r(X)^c).$$

By Lemma 22, we have

$$\mathbb{P}(\Phi_r(X)^c) \leq C \min \left( \frac{1}{n_P} r_{+}^{- \gamma}, \frac{1}{n_Q} r_{+}^{- d} \right) \leq \left( \frac{1}{\epsilon} \right) C \min \left( \frac{1}{n_P} r_{+}^{- \gamma}, \frac{1}{n_Q} r_{+}^{- d} \right)$$

since $\epsilon < 1$. The remaining term is bounded using the same steps as in the proof of Lemma 17 with $t = \epsilon$, where the first term vanishes because we integrate over $G_\epsilon^+$. □

**Proof.** [Theorem 32] Fix $\epsilon > 0$. By Theorem 3, there is a universal $C = C(T)$ such that

$$\mathbb{E}(\hat{f}; G_\epsilon^-) \leq C \min \left( \frac{\log(n_P)}{n_P} \frac{\alpha(\beta + 1)}{2\alpha + \alpha \beta + \gamma(\epsilon)}, \left( \frac{\log(n_Q)}{n_Q} \right)^{\frac{\alpha(\beta + 1)}{2\alpha + \alpha \beta + d}} \right).$$

Now, set $r_n^+ = (\epsilon/4C_a)^{1/\alpha}$. Assume for now that $\epsilon \geq 4C_a (n_P + n_Q)^{-\alpha}$. Then we have $\frac{1}{n_P + n_Q} \leq r_n^+$, and so an application of Lemma 25 at level $r_n^+$ (with $\delta = (n_P + n_Q)^{-1}$ and $C_n = C_{n, \delta}$) yields that, for $x \in G_\epsilon^+$, we have with probability at least $1 - (n_P + n_Q)^{-1}$ that

$$1\{\hat{f}(x) \neq f^*(x)\} \leq 1 \left\{ M(x) \leq 2 \left( C_n |\tilde{A}_{r_n^+}(x) \cap X|^{-1/2} + C_a (r_n^+)^\alpha \right) \right\},$$

and then splitting the indicator on the right-hand side gives

$$1\{\hat{f}(x) \neq f^*(x)\} \leq 1 \left\{ |\eta(x) - \frac{1}{2}| \leq 4C_n |\tilde{A}_{r_n^+}(x) \cap X|^{-1/2} \right\} + 1 \left\{ |\eta(x) - \frac{1}{2}| \leq 4C_a (r_n^+)^\alpha \right\}.$$
Now, by construction we have

\[ 4C_\alpha (r_n^*)^\alpha = \epsilon, \]

and therefore

\[ 1 \{ x \in G^+ \} 1 \{ |\eta(x) - \frac{1}{2}| \leq 4C_\alpha (r_n^*)^\alpha \} = 0, \]

and so by Proposition 2.16 it remains only to bound

\[ \mathbb{E} \left[ M(X) \cdot 1 \{ M(X) \leq 4C_\alpha |\tilde{A}_{r_n^*}(X) \cap X|^{-1/2} \} \right], \]

which we can achieve by following the exact steps used to prove Theorem 3. It remains only to consider the case \( \epsilon < 4C_\alpha (n_P + n_Q)^{-\alpha} \). Assume \( n_P \geq n_Q \). In this case, we have

\[
\min \left( \frac{1}{n_P} e^{-(1+\gamma^*/\alpha)}, \frac{1}{n_Q} e^{-(1+d/\alpha)} \right) \geq \frac{1}{n_P} e^{-(1+d/\alpha)} \geq (4C_\alpha)^{-(1+d/\alpha)} n_P^{-1+\alpha+d},
\]

where we use \( \gamma^* \geq d \). If \( \alpha + d > 1 \), this bound exceeds 1 provided that \( n_P \geq (4C_\alpha)^{a(a+d-1)} \). If \( n_Q \geq n_P \), then using the same argument we get a bound that exceeds 1 when \( n_Q \geq (4C_\alpha)^{a(a+d-1)} \), and so the result holds provided

\[
\max(n_P, n_Q) \geq (4C_\alpha)^{a(a+d-1)}.
\]

### 2.7.5 Supplemental Experiments

In this section we present the results of some supplemental experiments. First we carry out a comparison of our method against the dyadic-tree pruning algorithm introduced in [18]. We then investigate the performance of our method against the baselines considered in Section 2.6, except we use Algorithm 1 and the baselines to prune trees grown using the CART algorithm [15], which are non-dyadic. Our theoretical results do not apply to this setting, but the results nonetheless
demonstrate that Algorithm 1 can improve over these baselines in the context of non-dyadic trees.

Another Baseline Comparison

Decision trees have long been studied in the context of classification, and there exist numerous methods for pruning which employ techniques from structural risk minimization. A noteworthy contribution in this direction is that of [18], who devised an ingenious penalty term for selecting dyadic trees that provably attain minimax optimal learning rates for the expected excess risk under conditions similar to ours in the one-sample case.

Their method consists of penalizing a tree $T$ with leaves $\pi(T)$ according to an empirical analog of

$$
\Phi_n(T) = \sum_{A \in \pi(T)} \sqrt{2p_A \left[\log 2 + \log(2/\delta)\right] \log 2 + \log(2/\delta) \frac{n}{n}}
$$

(2.39)

where $p_A$ is the mass of the leaf under the sampling distribution, $\delta \in (0, 1)$ is the usual confidence parameter, and $\left[\cdot\right]$ is chosen such that $\sum_{A \in \pi(T)} 2^{-\left[\left[\cdot\right]\right]} \leq 1$, which they show can be done via the Kraft inequality for prefix codes [49]. The penalty (2.39) induces unbalanced trees, which can produce better approximations of the decision boundary than balanced trees [18]. The local depth selection of Algorithm 1 is similar in that it mimics unbalanced trees by choosing different levels of the tree for different regions of feature space when making predictions, so we provide a comparison against this baseline in order to verify that the advantage of our method is not simply that it yields unbalanced trees, but rather that it does so in a way that correctly aggregates the information added by the source data under covariate shift.

Note that the main result of [18] relies on an application of a relative Chernoff bound to derive (c.f. [18], Theorem 2) that with probability at least $1 - 2\delta$, one has

$$
|R(T) - \hat{R}_n(T)| \leq \Phi_n(T)
$$

(2.40)

for all $T$ in a suitable class of dyadic decision trees, where $R, \hat{R}_n$ are respectively the risk and
empirical risk of the tree $T$. Unfortunately, in the transfer learning setting in which we have access to samples from a source $P$ and a target $Q$, extending the penalty (2.39) in the obvious way does not lead to a bound of the form (2.40) with the target risk $R_Q(T)$, but rather would feature the risk under a convex combination of source and target $R_{\alpha P + (1-\alpha)Q}(T)$, where $\alpha = n_P/(n_P + n_Q)$. It may be possible to derive optimal rates for transfer using a penalty akin to (2.39) by employing the risk minimization strategy outlined in [36], although the latter studies a different notion of discrepancy between $P$ and $Q$, and so resulting rates are not directly comparable. We do not pursue this further here.

Figure 2.5 shows the result of using the SN penalty to prune a dyadic tree, against the uniform-depth CV baseline, and against Algorithm 1. We try two strategies for the SN method: the first (SN) simply pools the data and uses the penalty of [18] on the combined sample, the second (SN-Q) uses only the available target data to prune the tree, but uses the full data to calculate the estimate at each leaf. As we see in Figure 2.5, the pruning method does not achieve the performance of choosing the depth via 2-fold cross-validation over the target sample (note that, as pointed out in [18], this method is quite sensitive to the choice of dampening constant for the penalty; the results that we display are for a pre-tuned value). Of course, as pointed out above, in general neither SN nor SN-Q achieves the rate of Algorithm 1 in the covariate shift setting.

**Beyond Dyadic Trees**

Here we repeat the experiments shown in Section 2.6, but we implement the various depth-selection methods on trees grown using the vanilla CART algorithm of [15] (that is, splits are chosen to minimize the so-called Gini impurity at each node). For the sake of variety, we consider two further datasets: Wine Quality, and Steel Plates Faults, described in Table 2.1. Note that the Wine Quality dataset has 10 classes indicating various quality levels. We label wines with quality at least 6 as 1 and the rest as 0 to get a binary classification problem.

Although we can no longer claim the rates of Theorem 1 when Algorithm 1 is applied to these trees, as we see in Figure 2.6, the adaptive (AD) performs at least as well and sometimes signif-
Figure 2.5: Risk estimates for our adaptive method (AD), compared to the baseline cross-validation and the pruning method of [18] (SN, SN-Q), applied to the Algerian Forest Fire data (left) and the MAGIC data (right). Error bars show standard errors over 50 and 20 iterations, respectively.

Anticantly better than the cross-validation based depth-selection methods, suggesting that it could be useful as a pruning procedure even when trees are grown using more elaborate splitting criteria. We leave further exploration of this possibility for future work.

2.7.6 A Pathological Example.

In this section we provide an example of non-doubling measures $P, Q$ for which the aggregate and integrated transfer exponents are equal - $\gamma^* = \rho^*$ - but that these do not agree with the value that we would obtain if we defined the exponent by taking the sum of mass-ratios over the dyadic partition of the ambient space. Suppose that we have measures $P, Q$ supported on $[0, 1]^d$, and define

$$\Lambda(P, Q, n) = \sum_{C \in D^n} \frac{Q(C)}{P(C)},$$

where $D^n$ is the regular dyadic partition of $[0, 1]^d$ of order $n$. In analogy with how we have defined the aggregate exponent between $P$ and $Q$, let $\gamma_D(P, Q) = \gamma_D > 0$ be any constant for which there
exists a $C > 0$ such that for all $n = 0, 1, 2, \ldots$ we have

$$\Lambda(P, Q, n) \leq C 2^{n\gamma_D},$$

and let $\gamma_D^*$ be the minimal $\gamma_D$ such that the above holds. Let $d = 2$, and consider the following construction: consider two countable sequence of boxes \{B_i; \ i = 1, 2, \ldots\} and \{B'_i; \ i = 1, 2, \ldots\}, where $B_i$ is a square with side length $2^{-i}$ with top-right corner at $x_i = (\sum_{k=0}^{i-1} 2^{-i-k}, 2^{-i})$, and $B'_i$ is a square of side-length $2^{-2i}$ with bottom-left corner at $x_i$; see Figure 2.7 below. We take $Q = \sum_{i=1}^{\infty} Q_i$ and $P = P_0 + \sum_{i=1}^{\infty} P_i$, where $Q_i, P_i$ are supported on $B_i \cup B'_i$ for each $i \geq 1$. Let $Q_i$ uniformly assign mass $3 \cdot (1/4)^i$ to $B_i \cup B'_i$, yielding density $q_i$. Let $P_i$ be equal to $Q_i$ on $B_i$. On $B'_i$, let $P_i$ have density $q_i \|x - x_i\|^{\nu}$ for some $\nu > 0$. Finally, let $P$ have an atom at $x_0 = (0, 1)$ with all remaining mass. Note that, by calculations identical to those from Example 2, we have $\gamma^*(P, Q) = \max(2, \nu)$. We proceed to estimate $\Lambda(P, Q, n)$ for $n \geq 1$. Observe that by construction, there is a cube $C \in \mathcal{D}^n$ such that $Q(C) = Q(B'_n)$, and $P(C) = P(B'_n)$. Further, there is a constant $c_0$ such that $\cup_{i=1}^{n} B_i$ is exactly covered by $c_0 2^{2n}$ cubes in $\mathcal{D}^n$, and by construction for each of these cubes we have
\[
Q(C)/P(C) = 1. \text{ Therefore we have }
\]
\[
\Lambda(P, Q, n) = \sum_{C \in \mathcal{D}^n} \frac{Q(C)}{P(C)} \geq c_o 2^{2n} + \frac{Q(B'_n)}{P(B'_n)} = c_0 2^{-2n} + \frac{q_n 2^{4n}}{q_n c(\nu) 2^{2n(2+\nu)}} = c_0 2^{2n} + c(\nu)^{-1} 2^n 2^\nu,
\]
where \(c(\nu) > 0\) is some constant depending only on \(\nu\). It follows immediately that \(\gamma_D^*(P, Q) \geq \max(2, 2\nu)\), and so for \(\nu > 1\) we have \(\gamma_D^*(P, Q) > \gamma^*(P, Q)\). In particular, this shows that if we have used \(\gamma_D\) to derive our results, the induced rates would not be as sharp as those using the aggregate exponent \(\gamma\). Observe also that since \(P\) is not a doubling measure, this example also serves to demonstrate that doubling is not necessary for the equivalence in Proposition 9 to hold, since one can straightforwardly verify that \(\gamma^* = \rho^*\) in this case.

![Illustration of the support of the measures](image)

Figure 2.7: Illustration of the support of the measures \(P, Q\) described above on the unit square \([0, 1]^2\).
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source/Target Split</th>
<th>Test Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algerian Forest Fires ($d = 11$) See [40].</td>
<td>Observations from Sidi bel-Abbas constitute the Source sample, while the Target data are observations from the Bejaia region.</td>
<td>100</td>
</tr>
<tr>
<td>MAGIC Gamma ($d = 10$)  See [41].</td>
<td>Three variable indices (2, 7, and 9) were chosen at random. Observations with normalized values of these variables each above 0.3 go in the Target set with probability $p = 0.95$; the rest go into the Source set with probability $p$.</td>
<td>2500</td>
</tr>
<tr>
<td>SUSY ($d = 18$)  See [42].</td>
<td>Three variable indices (5, 10, and 15) were chosen at random. Observations with normalized values of these variables each above 0.1 go in the Target set with probability $p = 0.95$; the rest go into the Source set with probability $p$.</td>
<td>5000</td>
</tr>
<tr>
<td>Wine Quality ($d = 11$)  See [43].</td>
<td>White wine data is used as Source, red wine as Target.</td>
<td>1000</td>
</tr>
<tr>
<td>Biodegradation ($d = 41$)  See [44].</td>
<td>Observations with more than 75% Carbon atoms go in the Target set with probability $p = 0.95$; those with less go in the Source set with probability $p$.</td>
<td>200</td>
</tr>
<tr>
<td>Steel Plates Faults ($d = 27$)  See [45].</td>
<td>Two variable indices (2 and 13) were chosen at random. Observations with normalized values of these variables each above 0.25 go in the target set with probability $p = 0.95$; the rest go into the source set with probability $p$.</td>
<td>300</td>
</tr>
</tbody>
</table>
Chapter 3: On the Fundamental Limits of Off-Policy Evaluation

Some Results, Open Questions, and Future Work

This Chapter consists of unpublished work with Prof. Hongseok Namkoong.

3.1 Introduction

In this Chapter, we consider the problem of off-policy evaluation in reinforcement learning, with the goal of furthering our understanding of its fundamental hardness, which we take here to mean the degree of accuracy with which the target functional (described below) can be estimated. We were interested in this for two reasons: firstly, as will be seen below through an analysis of a published study, the problem can be deceptively difficult in some situations of applied interest - increasing awareness of fundamental limits and what characterizes them could help minimize the risk for such errors to be repeated. Secondly, as briefly mentioned in Chapter 1, off-policy evaluation can be thought of as a covariate-shift problem, and we had hoped that it might be possible to capture this difference using techniques developed in that area. Before we proceed, let us describe the problem. We suppose that we are in the following sequential decision-making framework: we suppose that we have a finite-horizon Markov Decision Process (see e.g. [50]), characterized by a sequence of state and action spaces $S_t, A_t$ for $t = 0, 1, \ldots, T$, an initial state distribution $S_0 \sim P_0$, a family of transition kernels $S_t \sim P_t(\cdot \mid A_t = a)$, and reward distributions $R_t \sim P_r(\cdot \mid S_t = s, A_t = a)$. Along with a policy - a sequence of mechanisms for randomly selecting actions - $A_t \sim \pi_t(\cdot \mid S_t = s)$, the decision process induces a distribution over state-action-reward trajectories, which we shall denote $\mathbb{P}_\pi$ (expectations under $\mathbb{P}_\pi$ will simply be written as $\mathbb{E}_\pi[\cdot]$). A fundamental quantity in reinforcement learning is the value of the policy $\pi$, which we
denote $\theta(\pi)$, defined as the expected sum of rewards under $P_\pi$:

$$\theta(\pi) := \mathbb{E}_\pi \left[ \sum_{t=0}^{T} R_t \right].$$

The problem of off-policy evaluation (OPE) occurs when we have access to a sample consisting of trajectories drawn from $P_{\pi^b}$ for some policy $\pi^b$ - the *behaviour* policy (which may or may not be known) - while the goal is to estimate the value of a *target* policy, $\pi^e$; that is, to estimate $\theta(\pi^e)$. At a high-level, our guiding question was the following: can we characterize in certain situations, in a minimax sense, the minimum mean-square error that is achievable by an OPE estimator? Can we capture this error in terms of quantities that measure the discrepancy between behaviour and target that are familiar, from the transfer learning literature or elsewhere?

In what follows, we present some work that, while incomplete, provides a point of departure for further work on these questions. The structure of the chapter is as follows: in Section 3.2, we analyze a study (see [51]) published in the journal *Nature Medicine*, which serves as wake-up call illustrating the difficulty of OPE. The following two Sections are devoted to the results we have derived: Section 3.3 gives a finite sample lower bound for the minimax OPE that brings out a term related to the chi-square divergence between the behaviour and target policies, while Section 3.4 gives an asymptotic lower bound over models for $P$ for which the mean reward function and the ratio of the policy density functions belong to a nonparametric smoothness class. Finally, Section 3.5 discusses where we have left the problem - questions that we were unable to answer, and directions for further research.

### 3.2 A Real-World Case Study.

In this section, we will introduce the experimental setting of [51], who developed a policy for dynamic treatment of sepsis patients via RL methods, and then evaluated this policy using OPE. As we shall see via a replication of their results - using state-of-the-art methods for OPE in MDPs, the results reported in the paper are grievously flawed in the sense of suffering from such high variance
as to make them meaningless. We argue that a lack of awareness regarding the theoretical limits of off-policy evaluation led to overconfidence about what such methods could achieve in this case, and, anticipating the lower bound of Section 3.3, we provide evidence that many more samples would be required for this problem in order to be able to estimate the value of the proposed policy with an acceptable mean-square error.

3.2.1 Data, Methodology, and Results of Original Study.

The methodology used in the study [51] was validated on an a database of ICU visits called the Medical Information Mart for Intensive Care, version III (MIMIC-III) [52]; this contains, in particular, information on patients deemed to fulfill the international sepsis-3 criteria, who are the population relevant to the work. Variables associated with the clinical treatment (fluid and vasopressor dosages received, denoted $a$) and outcome (survival or death, denoted $y$) are recorded at 4h time intervals for 72h from the first onset of sepsis, along with 48 other contextual features (demographics, vital signs, etc., denoted $x$). This yields a trajectory for each patient, consisting of the measurements at each time interval, along with an indicator at the last step indicating survival (given as +100 in the original) or death ($-100$) within 90 days of observation. The data is modelled as being a (time-homogeneous) Markov Decision Process over the state-action pairs $(x, a)$. For simplicity, a clustering of the original vector of features is chosen, so that states $x \in \mathbb{R}^{48}$ are mapped to cluster indices $s \in [750]$; death and discharge are added as absorbing states, for a total of 752. The administered fluid and vasopressor doses are discretized into 5 bins each for a total of 25 combinations, yielding state-action pairs $(s, a) \in [752] \times [25]$. The empirical estimate of the MDP transition matrix was used for computing the ‘optimal policy’ via policy iteration; the behaviour policy (‘clinician policy’ in the paper) is again estimated using the simple empirical analog. This procedure is repeated by the authors 500 times over random initializations of the clustering algorithm, and at each iteration value of the optimal policy is computed using WIS over a hold-out test set, which is in turn bootstrapped 100 times; of the 500 policies so-chosen, the authors select the one with the largest 5th percentile of bootstrapped WIS and dub it the ‘AI-Clinician’
(AIC). This paper (published in the journal *Nature Medicine*) received substantial attention - 816 citations and counting, as of Oct. 2023. But how good is the AI-Clinician?

The critical step - the off-policy evaluation of the AIC - is carried out using weighted importance sampling (see e.g. [53, 54], which is cited by the authors as an example of a High-Confidence Off-Policy Estimation method (as per [55]). The high variance and consequently low reliability of off-policy estimators generally is understood and acknowledged (see e.g. [56, 55, 54, 8]), but the relevant literature generally consists of papers that proposed improved methods for estimation, along with experimental results that demonstrate their increased efficacy on simple problems. It is our belief that a sense of the fundamental difficulty of the OPE problem is lacking in the wider community of applied practitioners; as a case-in-point, in justifying their use of WIS for OPE, the authors of [51] write “…we implemented a type of HCOPE method, WIS, and used bootstrapping to estimate the true distribution of the policy value in the test sets. WIS may be a biased although consistent policy estimator, so the bootstrap confidence interval may also be biased, even though the literature suggests that consistency is a more desirable property than unbiasedness. It is accepted that bootstrapping produces accurate confidence intervals with less data than exact HCOPE methods and is safe enough in high-risk applications, such as healthcare.” The final results of the off-policy estimation of the expected cumulative discounted reward, with a discounting factor of $\gamma = 0.99$, is given by the authors as “median of 56.9, interquartile range (IQR) (54.7, 58.8) for the clinician’s policy, and median of 84.5, IQR (84.3, 87.7)” for the AI-Clinician policy (statistics computed over 2000 bootstrap samples). On a separate cohort (an independent dataset), they report a median of “85.1, interquartile range (85.1, 86.0).” This latter result should immediately give one pause, for it seems curious indeed that the 25th percentile and the median values of the bootstrap sampled WIS estimates should be the same value. Scrutinizing the actual reported values is more disconcerting still. Recall that all observed rewards are either 0 (patient monitoring still ongoing), 100 (discharge from hospital) or −100 (death), and that the authors compute cumulative discounted rewards with discount factor $\gamma = 0.99$, and we remark that the patient trajectories have an average length of roughly 15 and a maximum length of $T = 19$. As it turns out, of the four numbers reported
as IQR limits, each of them is exactly (up to two decimal places of precision) equal to $100 \cdot \gamma^t$ for some $t$: $(84.3, 87.7) \approx (100 \cdot \gamma^{17}, 100 \cdot \gamma^{13})$, while $(85.1, 86.0) \approx (100 \cdot \gamma^{16}, 100 \cdot \gamma^{15})$. It is relatively straightforward to understand what has happened here: the importance weights for some trajectories are so large that the estimators are simply dominated by the reward values of those trajectories. This pathology is masked by a) the normalization of the IS weights, which forces the estimates to be within the feasible range $(-100, 100)$, and b) the discounting, which due to the trajectory length lowers the values from $\approx 100$ to $\approx 85$. One might be forgiven for speculating further about the credulity induced by the excitement surrounding applications of AI in healthcare (a score of 85 would correspond to a reduction in mortality from roughly 22% to roughly 7.5%). We remark that this phenomenon is confirmed when the procedure is repeated without discounting - in this case, the bootstrap samples (of WIS values for the optimal policy for the 500 models derived from the random clusterings) frequently have a 5th percentile OPE value in excess of 99.9.

In the following subsection, we present the results for a suite of alternative OPE estimators, including modern, including a doubly-robust, state-of-the-art method catered to MDPs in particular, due to Kallus and Uehara [8].

3.2.2 Evaluating the AI-Clinician using Alternative Methods.

As we have seen, the OPE estimates provided in [51] are not trustworthy. In this section, we evaluate the AI-Clinician using four alternative OPE methods. Below we briefly summarize these methods, following the exposition given in [8]. All action-value functions $q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ are assumed taken with respect to the evaluation policy, $\pi^e$. Let the OPE functional of interest be denoted $\rho$.

**Direct Method:** Given an estimate of the action-value functions at time $t = 0$, $\hat{q}_0$, the direct method simply takes

$$\hat{\rho}_{DM} = \mathbb{E}_{\pi} \left[ \mathbb{E}_{\pi^e} [\hat{q}_0(s_0, a_0) \mid s_0] \right],$$

where the inner expectation is over the (known) evaluation policy, while the outer is taken over the $n$ sample trajectories.
**Double Reinforcement Learning:** Inspired by the Double ML paradigm of [57] and by earlier works proposing doubly-robust RL estimators (see for example [54]), Kallus and Uehara propose a sample splitting, doubly-robust, efficient estimator for OPE in MDPs. We reproduce their description of the method, for completeness, in Section 3.6.1.

**Marginal Importance Sampling:** In this case, we assume access to estimates \( \hat{\mu}_t \) of the marginal density ratios

\[
\mu_t(s_t, a_t) = \frac{\pi^e_t(a_t \mid s_t) P^e_t(s_t)}{\pi^b_t(a_t \mid s_t) P^b_t(s_t)},
\]

where \( P^e_t \) (resp. \( P^b_t \)) is the marginal distribution of state \( t \), having followed the policy \( \pi^e \) (resp. \( \pi^b \)). The MIS estimate is

\[
\hat{\rho}_{MIS} = \mathbb{E}_n \left[ \sum_{t=0}^T \hat{\mu}_t r_t \right].
\]

**Weighted Marginal Importance Sampling:** IS-type estimators are known to suffer from high variance; to mitigate this, the WMIS estimator self-normalizes by the sum of the weights:

\[
\hat{\rho}_{MIS} = \frac{\mathbb{E}_n \left[ \sum_{t=0}^T \hat{\mu}_t r_t \right]}{\mathbb{E}_n \left[ \sum_{t=0}^T \hat{\mu}_t \right]}.
\]

In practice, we recursively estimate the marginal density ratios \( \hat{\mu}_t \) and the action-value functions \( \hat{q}_t \) using XGBoost models, according to the backward induction schemes outlined in [8].

As we can see in Figure 3.1, none of the methods outlined above are sufficient in this case to provide satisfactory estimates. The DR and MIS methods have extreme variance, induced by the large discrepancy in this case between behaviour and evaluation policies; their estimates are useless in this case. The DM and WMIS estimators, on the other hand, seem to have much lower variance, but suffer in this case from poor action-value estimates and extreme importance weights, respectively. The DM estimate looks plausible until one recalls that it yields a lower estimate than the estimated value of the clinician policy, which the evaluation policy is designed to optimize.
Figure 3.1: OPE estimates produced by respectively, DM, DR, MIS, and WMIS for our replication of the ‘AI-clinician’ of [51]. Box-and-whisker plots are shown over the bootstrap distribution; $n_{boot} = 1000$. 
In the following section, we consider a measurement of the fundamental hardness of this problem.

3.2.3 Estimating a Lower Bound.

As we shall see below in Theorem 35, there is a fundamental limit to the mean-square error of OPE\(^1\) given by

\[
\frac{1}{n} \sum_{t=1}^{T} \mathbb{E} \left[ \left( \frac{P_t^e(X_t)}{P_t^*(X_t)} \right)^2 \sum_{a_t \in \mathcal{A}_t} \frac{\pi^e(a_t | X_t)^2}{\pi^b(a_t | X_t)} \sigma_t^2(X_t, a_t) \right].
\]

This quantity is therefore implicitly of great interest to the practitioner, as their sanguinity at the prospect of producing reliable OPE estimates should be in inverse proportion to its magnitude. Consider Figure 3.2 below, which displays an estimate of (3.1) for the sepsis data (where we have scaled the rewards to be ±1 rather than ±100 for ease of interpretation). As we can clearly see, even for moderate values of \(\alpha\), the discrepancy between behaviour and target policies induces a severe deterioration in the quality of estimates (at least, in a minimax sense); in the study [51], the authors perform OPE on \(\pi_{0.99}^e\). Of course, the marginal density ratios \(\pi^e_t / \pi^b_t(\cdot)\) and the behaviour policy itself are unknown, so the expectation in (3.1) cannot be computed exactly; the plot above is produced using Monte Carlo estimates over the trajectory for the outside expectation, with the marginal density ratios being first estimated according to scheme detailed in [8]. Indeed, even were we able to compute the expectation exactly, the qualitative interpretation of Figure 3.2 is unlikely to change: the minimal possible MSE is large enough that any estimates produced are essentially meaningless.

3.3 A Local Minimax Lower Bound.

In this section we derive a finite-sample local minimax lower bound for OPE, which will help give a sense of the inherent difficulty of the off-policy estimation problem. For now, we assume that

\(^1\)When the truth is known to lie within a particular model class. As will be seen, the class in question takes everything to be fixed save the mean reward function, and so roughly corresponds to the setup that the authors of the study assume; namely, that their estimates of the MDP parameters (behaviour policy and state transition dynamics, initial distribution) correspond to the truth.
we are in the Markov setting, so that the data consist of trajectories sampled from the Markov decision process (MDP) specified by the sequence of state and action spaces $X_t, A_t$ for $t = 1, \ldots, T$, an initial state distribution $s_1 \sim P_1$, a family of transition kernels $P_t(s_t \mid a_t)$, and reward distributions $R_t \sim P_{r_t}(r_t \mid s_t, a_t)$. We assume that the data have been collected by following the policy $\pi^b = \{\pi^b_t; t = 1, \ldots, T\}$, and that the true reward distribution satisfies

$$\mathbb{E}[R_t \mid X_t, A_t] = \mu^*_{t}(X_t, A_t),$$

and we let

$$\sigma^2_t(x, a) = \mathbb{E} \left[ (R_t - \mu^*_{t}(X_t, A_t))^2 \mid X_t = x, A_t = a \right].$$

It may therefore be supposed that $R_t = R^0_t + \mu^*_t$, where $\{R^0_t\}$ is a collection of mean-zero variables with variance $\sigma^2_t$, mutually independent after conditioning on the trajectory history. Under
these conditions, the underlying MDP is indexed by \( (\{P_t\}_t, \pi^b, \mu^*, R^0) \). Unless otherwise indicated, all expectations are taken with respect to the sampling distribution. We are interested in the average cumulative reward under \( \pi^e = \{\pi^e_t\} \). In this section, we shall restrict our attention to discrete state/action spaces. Our main result is a local minimax lower bound over a neighbourhood in which the mean rewards functions can vary up to a certain amount: for a set of functions \( \delta_t : \mathcal{X}_t \times \mathcal{A}_t \to \mathbb{R} \), consider the neighbourhoods (where \( \mu^* = \{\mu^*_1, \ldots, \mu^*_T\} \))

\[
N_\delta(\mu^*) := \left\{ \{\mu_t\} \mid |\mu_t(x, a) - \mu^*_t(x, a)| \leq \delta_t(x, a), \forall x \in \mathcal{X}_t, a \in \mathcal{A}_t, t = 1, \ldots, T \right\}.
\]

We shall bound the minimax risk over the class of MDPs indexed by \( M = (\{P_t\}, \pi^b, \mu, R^0) \), where \( \{P_t\}, \pi^b, R^0 \) are fixed, while \( \mu \in N_\delta(\mu^*) \); denote this family of MDPs \( M_{\delta, \sigma} \). We use \( P_t(x), P^e_t(x) \) to denote the marginal state distributions at time \( t \) for trajectories generated by following \( \pi^b \) and \( \pi^e \), respectively.

Before we state our bound, we require some assumptions, for which we will introduce some further notation. For a random variable \( Y \), define the \((2,4)\)-moment-ratio \( \|Y\|_{2 \to 4} := \sqrt{\mathbb{E}[Y^4]} / \mathbb{E}[Y^2] \).

Further, let

\[
O_{\sigma} = \sqrt{\sum_{t=1}^T \mathbb{E} \left[ \left( \frac{P^e_t(X_t)}{P^*_t(X_t)} \right)^2 \sum_{a_t \in \mathcal{A}_t} \frac{\pi^e(a_t \mid X_t)^2}{\pi^b(a_t \mid X_t)} \cdot \sigma^2_t(X_t, a_t) \right]},
\]

and define \( O_{\delta} \) analogously. Also set

\[
O_{\delta, t} = \sqrt{\mathbb{E} \left[ \left( \frac{P^e_t(X_t)}{P^*_t(X_t)} \right)^2 \sum_{a_t \in \mathcal{A}_t} \frac{\pi^e(a_t \mid X_t)^2}{\pi^b(a_t \mid X_t)} \cdot \delta^2_t(X_t, a_t) \right]},
\]

so that \( O_{\delta} = \sum_t O_{\delta, t} \). We further assume the following:

**A1** For \( t = 1, \ldots, T \), let

\[
Z_t(X_t, A_t) := \delta_t(X_t, A_t) \frac{\pi^e_t(A_t \mid X_t) P^e_t(X_t)}{\pi^b(A_t \mid X_t) P_t(X_t)}.
\]
There exists a constant $M_{2\rightarrow 4} > 0$ such that $\|Z_t\|_{2\rightarrow 4} \leq M_{2\rightarrow 4}$ for all $t = 1, \ldots, T$.

**A2)** For each $t = 1, \ldots, T$, we have the following bound on the neighbourhood size:

$$\sqrt{n} \delta_t(x, a) \geq \frac{\sigma_t^2(x, a) \pi_t^e(a \mid x) P_t^e(x)}{O_\sigma \pi_t^b(a \mid x) P_t(x)} \quad \forall x \in X_t, a \in A_t.$$

**A3)** There exists an $\epsilon > 0$ such that for each $t = 1, \ldots, T$, we have $P_t(x) \leq \epsilon$ for all $x \in X_t$.

We shall remark on each of these assumptions in due course. Our main result is the following, where $C$ denotes a universal constant, $\theta$ denotes the functional of interest, and $\hat{\theta}_n$ denotes any estimator thereof learned from a sample of $n$ trajectories:

**Theorem 35.** Let $\mathcal{M}_\delta$ be the class of MDPs outlined above over discrete state and action spaces $X, A$. Suppose that assumptions A1-A3 hold. We have

$$\inf_{\hat{\theta}_n} \sup_{M \in \mathcal{M}_\delta} \mathbb{E} \left[ \left\| \hat{\theta}_n - \theta(M) \right\| \right] \geq C \frac{1}{\sqrt{n}} (O_\delta + O_\sigma).$$

The salient feature of this bound, similar in flavour to is that it explicitly captures the importance of overlap between the evaluation and behaviour policies. If we assume that the outcome variance and neighbourhood size respectively satisfy $\sigma_t^2(x_t, a_t) \equiv \sigma_t^2(x_t)$ and $\delta_t(a_t, x_t) = \delta_t(x_t)$, then the quantity

$$\sum_{a_t \in A_t} \frac{\pi_t^e(a_t \mid X_t)^2}{\pi_t(a_t \mid X_t)} = 1 + D_{\chi^2} \left( \pi^e(\cdot \mid X_t) \| \pi^b(\cdot \mid X_t) \right),$$

where $D_{\chi^2}(P \| Q)$ denotes the Chi-square divergence between distributions $P$ and $Q$, locally captures the added difficulty of the off-policy estimation problem; averaging a version of this Chi-square term weighted by the marginal state likelihood ratio $\Lambda_t(X_t)$ and the parameters $\delta, \sigma$ gives our lower bound. It is worth noting that while the bound reflects the difficulty imposed by the outcome variances via the $\sigma_t^2(X_t, a_t)$ factors that appear in the $\chi^2$ term, [58] have shown that lower bounds involving the latter are unavoidable even if the outcome was observed noiselessly.
The aim of the next sections will be to prove Theorem 35, which we will achieve by proving the following two intermediate results:

**Proposition 36.** Let \( \mathcal{M}_\delta \) be the class of MDPs outlined above over discrete state and action spaces \( X, \mathcal{A} \). Suppose that \( A2 \) holds. We have

\[
\inf_{\hat{\theta}_n} \sup_{M \in \mathcal{M}_\delta} \mathbb{E} \left[ \left| \hat{\theta}_n - \theta(M) \right| \right] \geq C \frac{1}{\sqrt{n}} O(\sigma).
\]

**Proposition 37.** Let \( \mathcal{M}_\delta \) be the class of MDPs outlined above over discrete state and action spaces \( X, \mathcal{A} \). Suppose that assumptions \( A1 \) and \( A3 \) hold. We have

\[
\inf_{\hat{\theta}_n} \sup_{M \in \mathcal{M}_\delta} \mathbb{E} \left[ \left| \hat{\theta}_n - \theta(M) \right| \right] \geq C \frac{1}{\sqrt{n}} O(\delta).
\]

Proposition 36 is proved in the next subsection, using Le Cam’s two-point method. Following that, we prove Proposition 37 using a technique developed in [58], based on Le Cam’s mixture-vs-mixture method, where it was used to prove a version of the result in the single-step case. Technical or well-known lemmas are collected at the end.

3.3.1 Proof of Proposition 36.

Let us begin by introducing our construction. Letting \( H_t = (A_t, X_t, \ldots, A_1, S_1) \) denote the history, consider a distribution \( M_0 \in \mathcal{M}_n \) for which \( X_1 \sim P_1, A_1 \mid X_1 \sim \pi^b_1(A_1 \mid X_1), R_1 \mid H_1 \sim N(\delta_1(X_1, A_1), \sigma^2(X_1, A_1)) \), for \( A_1 \in \mathcal{A}_1 \), independently, and \( X_2 \mid H_1 \sim P_2(\cdot \mid X_1, a_1), A_2 \mid H_2 \sim \pi^b_2(A_2 \mid X_2), R_2 \mid H_2 \sim N(\delta_2(X_2, A_2), \sigma^2(X_2, A_2)) \), for \( A_2 \in \mathcal{A}_2 \), independently, \( \ldots, X_T(a_{T-1}) \mid H_{T-1} \sim P_T(a_{T-1} \mid X_{T-1}), A_T \mid H_T \sim \pi_T(A_T \mid X_T), \) and \( R_T \mid H_T \sim N(\delta_T(X_T, A_T), \sigma^2(X_T, A_T)) \), independently. Let \( M_1 \) be identical, save that the reward distributions are mean zero Gaussians.

**Lemma 38.** Let \( \Lambda_t(X_t) = P^+_t(X_t) / P_t(X_t) \) denote the likelihood ratio of being in state \( X_t \) at time \( t \) having followed the evaluation policy relative to that had we followed the behaviour policy. We
have

\[ |\theta(M_1) - \theta(M_0)| = \sum_{t=1}^{T} \mathbb{E} \left[ \Lambda_t(X_t) \sum_{a_t \in \mathcal{A}_t} \delta_t(X_t, a_t) \pi_t^b(a_t \mid X_t) \right]. \]

**Proof.** Let \( \bar{R}_t \) denote the reward at \( t \), where we have followed the evaluation policy throughout. We have

\[
\mathbb{E}_1 \bar{R}_t = \mathbb{E} \left[ \mathbb{E} \left[ \bar{R}_t \mid \tilde{X}_t \right] \right]
= \mathbb{E} \left[ \sum_{a_t} \pi_t^e(a_t \mid \tilde{X}_t) \mathbb{E}[\bar{R}_t \mid \tilde{X}_t, \tilde{A}_t = a_t] \right]
= \mathbb{E} \left[ \sum_{a_t} \pi_t^e(a_t \mid \tilde{X}_t) \mathbb{E}[R_t(a_t) \mid \tilde{X}_t] \right]
= \mathbb{E} \left[ \Lambda_t(X_t) \sum_{a_t} \pi_t^e(a_t \mid X_t) \mathbb{E}[R_t(a_t) \mid X_t] \right]
= \mathbb{E} \left[ \Lambda_t(X_t) \sum_{a_t} \pi_t^e(a_t \mid X_t) \delta_t(X_t, a_t) \right].
\]

By the same logic, we find that \( \mathbb{E}_0 \bar{R}_t = 0 \), and the result follows by summing over the time steps. \( \square \)

Now, the result will follow from Le Cam’s Lemma and Pinsker’s Inequality. To apply these, we must calculate the KL divergence between \( P_0 \) and \( P_1 \).

**Lemma 39.** We have

\[
D_{\text{kl}}(M_1 \parallel M_0) = \sum_{t=1}^{T} \mathbb{E} \left[ \sum_{a_t \in \mathcal{A}_t} \frac{\delta_t^2(X_t, a_t)}{2\sigma_t^2(X_t, a_t)} \pi_t(a_t \mid X_t) \right]
\]

**Proof.** Taking the likelihood ratios of a trajectory (where \( h_t = (s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t) \)), we find that

\[
\frac{p_1(h_T, a_T, r_T)}{p_0(h_T, a_T, r_T)} = \prod_{t=1}^{T} \exp \left\{ \frac{r_t^2 - (r_t - \delta_t(x_t, a_t))^2}{2\sigma_t^2(x_t, a_t)} \right\};
\]

82
and taking logs, we proceed to calculate the expectation term-by-term; we have

\[ D_{\text{kl}}(M_1 \| M_0) = \sum_{t=1}^{T} \mathbb{E}_1 \left[ \frac{R_t^2 - (R_t - \delta_t(X_t, A_t))^2}{2\sigma_t^2(X_t, A_t)} \right] \]

\[ = \sum_{t=1}^{T} \mathbb{E}_1 \left[ \frac{\delta_t^2(X_t, A_t)}{2\sigma_t^2(X_t, A_t)} \right] \]

\[ = \sum_{t=1}^{T} \mathbb{E}_1 \left[ \sum_{a_t \in \mathcal{A}_t} \frac{\delta_t^2(X_t, a_t)}{2\sigma_t^2(X_t, a_t)} \pi_t(a_t \mid X_t) \right], \]

where the second inequality follows from the tower law, taking the outside expectation over the distribution of \( R_t \); the calculation is a standard one for computing the KL-divergence between Gaussian distributions.

**Proof.** [Proposition 36.] By Le Cam’s Lemma and Pinsker’s inequality, we have

\[ \inf_{\hat{\theta}} \sup_{M \in \mathcal{M}_n} \mathbb{E} \left[ |\hat{\theta} - \theta(M)| \right] \geq \frac{1}{2} |\theta(M_1) - \theta(M_0)| \times \left( 1 - \sqrt{n D_{\text{kl}}(M_0 \| M_1)} \right). \]

Using Lemmas 38 and 39, we find that

\[ \inf_{\hat{\theta}} \sup_{M \in \mathcal{M}_n} \mathbb{E} \left[ |\hat{\theta} - \theta(M)| \right] \geq \frac{1}{2} \sum_{t=1}^{T} \mathbb{E} \left[ \Lambda_t(X_t) \sum_{a_t} \delta_t(X_t, a_t) \pi_t^e(a_t \mid X_t) \right] \]

\[ \times \left( 1 - \sqrt{\frac{n}{2} \sum_{t=1}^{T} \mathbb{E} \left[ \sum_{a_t} \frac{\delta_t^2(X_t, a_t)}{2\sigma_t^2(X_t, a_t)} \pi_t^b(a_t \mid X_t) \right]} \right). \]

For \( t = 1, \ldots, T \), let

\[ \delta_t(X_t, a_t) = C \cdot \sigma_t^2(X_t, a_t) \Lambda_t(X_t) \frac{\pi_t^e(a_t \mid X_t)}{\pi_t^b(a_t \mid X_t)}, \]
yielding

\[
\inf \sup \mathbb{E} \left[ \left| \hat{\theta} - \theta(M) \right| \right] \geq \frac{1}{2} \sum_{t=1}^{T} C \cdot \mathbb{E} \left[ \Lambda_t(X_t)^2 \sum_{a_t} \sigma_t^2(X_t, a_t) \frac{\pi_t^e(a_t | X_t)^2}{\pi_t(a_t | X_t)} \right] \\
\times \left( 1 - C \cdot \sqrt{\frac{n}{2} \sum_{t=1}^{T} \frac{1}{2} \mathbb{E} \left[ \Lambda_t(X_t)^2 \sum_{a_t} \sigma_t^2(X_t, a_t) \frac{\pi_t^e(a_t | X_t)^2}{\pi_t(a_t | X_t)} \right]} \right).
\]

Setting

\[
1/C = \sqrt{n \sum_{t=1}^{T} \mathbb{E} \left[ \Lambda_t(X_t)^2 \sum_{a_t} \sigma_t^2(X_t, a_t) \frac{\pi_t^e(a_t | X_t)^2}{\pi_t(a_t | X_t)} \right]}
\]

and the result follows. \(\square\)

3.3.2 Proof of Proposition 37.

This part of the proof is based on Le Cam’s mixture-vs-mixture method, an exposition of which can be found in Chapter 15 of [59], and uses techniques from [58]. The construction is based on careful adjustments to the outcome mean model introduced by [58]. For a function \(\zeta_t : X_t \times \mathcal{A}_t \rightarrow \{-1, 1\}\), let

\[
\mu_t^\zeta(x, a) = \mu^*_t(x, a) + \zeta_t(x, a) \cdot \delta_t(x, a)
\]

denote a perturbed outcome function, so that by construction \(\mu^\zeta \in \mathcal{N}_\delta(\mu^*)\). Further, let

\[
\rho_t(x, a) = \begin{cases} 
\frac{\delta_t(x, a) \pi_t^e(x,a) P_t^e(x)}{\pi_t^e(x,a) P_t^e(x)} & \text{if } \delta_t(x, a) \cdot \frac{\pi_t^e(x,a) P_t^e(x)}{\pi_t^e(x,a) P_t^e(x)} \leq C \cdot O_{\delta,t} \\
1 & \text{else,}
\end{cases}
\]

where \(C = 2M_{2 \rightarrow 4}\). Fix an \(s \in (0, 1/C]\) and a sign \(z \in \{-1, 1\}\), and define a distribution over the outcome mean functions:

\[
Q^\zeta = \mathcal{L}(\mu^\zeta), \quad \text{where} \quad \zeta_t \sim \prod_{x \in X_t, a \in \mathcal{A}_t} \text{Ber} \left( \frac{1 + zs \rho_t(x, a)}{2} \right),
\]
independently for each $t = 1, \ldots, t$ (where we take the Bernoulli variables to be supported on $\{-1, 1\}$). The result will be shown by analyzing the perturbation distributions $Q_{-1}^s, Q_1^s$ for a suitably chosen $s$. Two intermediate results are required. Below, we let the parameters of the MDP not including the mean reward functions be given by $\xi^*$, and let the probability distribution of trajectories under the model induced by $\mu, \xi^*$ be written $P_{\mu, \xi^*}$.

**Lemma 40.** The total variation distance between the mixture-of-product distributions is upper bounded by

$$d_{TV} \left( \int P_{\mu, \xi^*} \otimes^n dQ_{-1}^s(\mu), \int P_{\mu, \xi^*} \otimes^n dQ_1^s(\mu) \right) \leq 2s \sqrt{n} + 4 e^{-n/4}.$$

**Lemma 41.** Let $\tau(\mu, \xi)$ denote the functional of interest. For $s > 4 \sqrt{n} \epsilon$, we have

$$P_{\mu \sim Q_1^s} \left( \tau(\xi^*, \mu) \geq \tau(\xi^*, \mu^*) + \frac{s}{4} O_\delta \right) \geq 1 - 2 \cdot e^{-4}, \quad \text{and}$$

$$P_{\mu \sim Q_{-1}^s} \left( \tau(\xi^*, \mu) \leq \tau(\xi^*, \mu^*) - \frac{s}{4} O_\delta \right) \geq 1 - 2 \cdot e^{-4}.$$

**Lemma 41.** Let us consider first the case $z = 1$. Then by first, for each $t$, conditioning on the state, action pair $(x_t, a_t)$ and integrating over the reward, we see that we have (where $\nu_t$ is a dominating
measure over the state-action pair at time $t$)

$$
\mathbb{E}_{\mu \sim \mathcal{Q}^t_i} (\tau(\xi^*, \mu)) = \sum_{t=1}^{T} \int \mu_t^c(x_t, a_t) \pi_t^c(a_t | x_t) P_t^c(x_t) \, dv_t
$$

$$
= \sum_{t=1}^{T} \int (\mu_t^c(x_t, a_t) + \zeta_t(x_t, a_t) \cdot \delta_t(x_t, a_t)) \pi_t^c(a_t | x_t) P_t^c(x_t) \, dv_t
$$

$$
= \tau(\xi^*, \mu^*) + \sum_{t=1}^{T} \int \zeta_t(x_t, a_t) \cdot \delta_t(x_t, a_t) \pi_t^c(a_t | x_t) P_t^c(x_t) \, dv_t
$$

$$(*) \quad \geq \tau(\xi^*, \mu^*) + s \sum_{t=1}^{T} \int \rho_t(x_t, a_t) \cdot \delta_t(x_t, a_t) \pi_t^c(a_t | x_t) P_t^c(x_t) \, dv_t
$$

$$
(\ast\ast) \quad \geq \tau(\xi^*, \mu^*) + \sum_{t=1}^{T} \frac{s}{C O_{\delta, t}} \int \delta_t^2(x_t, a_t) \frac{\{\pi_t^c(a_t | x_t) P_t^c(x_t)\}^2}{\pi_t^b(a_t | x_t) P_t(x_t)} \, dv_t
$$

$$
= \tau(\xi^*, \mu^*) + \sum_{t=1}^{T} \frac{s}{C O_{\delta, t}} \mathbb{E} \left[ \delta_t^2(X_t, A_t) \left( \frac{\pi_t^c(A_t | X_t) P_t^c(X_t)}{\pi_t^b(A_t | X_t) P_t(X_t)} \right)^2 \mathbb{1}\{\Gamma_t\} \right],
$$

where the final expectation is over $(X_t, A_t)$ drawn from a trajectory following the behaviour policy, 

$$(*) \quad \text{follows by again conditioning on state and action and integrating over the variables } \zeta_t, \text{ and } (\ast\ast) \quad \text{follows by (3.4), since we take}
$$

$$
\Gamma' := \left\{ (x, a) \in X_t \times A_t : \delta_t(x, a) \cdot \frac{\pi_t^c(x, a) P_t^c(x)}{\pi_t^b(x, a) P_t(x)} \leq C \cdot O_{\delta, t} \right\}.
$$

An application now of Lemma 45 to the variable $Z_t$ (as in (3.3)) gives

$$
\mathbb{E}_{\mu \sim \mathcal{Q}^t_i} (\tau(\xi^*, \mu)) \geq \tau(\xi^*, \mu^*) + \frac{s}{2C} \sum_{t=1}^{T} O_{\delta, t}
$$

$$
\geq \tau(\xi^*, \mu^*) + \frac{s}{2C} O_\delta,
$$

since $O_\delta^2 = \sum_{t} O_{\delta, t}^2$. Now, the difference between $\tau(\xi^*, \mu)$ and $\mathbb{E}_{\mu \sim \mathcal{Q}^t_i} (\tau(\xi^*, \mu))$ is due to the randomness of the Bernoulli $\zeta$ variables $\zeta_t(x_t, a_t)$; the difference can be expressed as the sum of these variables multiplied by the factors $\delta_t \pi_t^c P_t^c$. Conditioned on all state-action variables, the summands are independent, so by Hoeffding’s inequality then, we have with probability $1 - 2e^{-2r}$,
that

$$
\left| \tau(\xi^*, \mu) - \mathbb{E}_{\mu \sim Q^s} (\tau(\xi^*, \mu)) \right| \leq \sqrt{r \sum_{t=1}^{T} \sum_{x_t \in X_t, a_t \in A_t} \delta_t(x_t, a_t)^2 \pi_t^c(a_t \mid x_t)^2 P_t^c(x_t)^2}
$$

$$
\leq \sqrt{r \sum_{x_t \in X_t, a_t \in A_t} \delta_t(x_t, a_t)^2 \pi_t^c(a_t \mid x_t)^2 P_t^c(x_t)^2} = \sqrt{r \epsilon O_{\delta}}.
$$

Setting \( r = 2 \) and taking \( s > 4C^2 \sqrt{2\epsilon} \) then implies, together with the bound above, that

$$
\mathbb{P}_{\mu \sim Q^s} \left\{ \tau(\xi^*, \mu) \geq \tau(\xi^*, \mu^*) + \frac{s}{4} O_{\delta} \right\} \geq 1 - 2 \cdot e^{-4},
$$

which completes the proof, since the second inequality in the Lemma statement is an entirely symmetric case.

With these two results given, the theorem is easily proved by exactly following the steps on p.37 of [58], which we show here for completion.

**Proof.** [**Theorem 35.**] Let \( X = X_1 \times X_2 \times \cdots \times X_T \), and likewise define \( A \). Consider the following two sets:

$$
\mathcal{E}_1 := \left\{ \mu^\zeta \mid \zeta \in \{-1, 1\}^{X \times A} : \tau(\xi^*, \mu^\zeta) \geq \tau(\xi^*, \mu^*) + \frac{s}{4} O_{\delta} \right\},
$$

$$
\mathcal{E}_{-1} := \left\{ \mu^\zeta \mid \zeta \in \{-1, 1\}^{X \times A} : \tau(\xi^*, \mu^\zeta) \leq \tau(\xi^*, \mu^*) - \frac{s}{4} O_{\delta} \right\}.
$$

Set \( s = 1/16\sqrt{n} \), so that Lemma 40 implies that \( Q^s_z \geq 1 - e^{-4} \) for \( z \in \{-1, 1\} \); we then set \( Q^*_z := Q^s_z \mid \mathcal{E}_z \) for \( z \in \{-1, 1\} \). Let \( S = \text{ supp}(Q^*_1) \) and \( S' = \text{ supp}(Q^*_{-1}) \); clearly then the distributions \( Q^*_1, Q^*_{-1} \) satisfy

$$
\inf_{\mu \in S, \mu' \in S'} |\tau(\xi^*, \mu) - \tau(\xi^*, \mu')| \geq \frac{s}{2} O_{\delta} = \frac{1}{32\sqrt{2}} O_{\delta}.
$$

Now, using Lemma 44 to control the total variation between these restricted mixtures and the
original distributions, which is then bounded using Lemma 40, we have

\[
\begin{align*}
    d_{TV} \left( \int P_{\mu, \xi}^{\otimes n} dQ_{-1}^{s}(\mu), \int P_{\mu, \xi'}^{\otimes n} dQ_{1}^{s}(\mu) \right) \\
    \leq \frac{1}{1 - 2e^{-4}} d_{TV} \left( \int P_{\mu, \xi}^{\otimes n} dQ_{-1}^{s}(\mu), \int P_{\mu, \xi'}^{\otimes n} dQ_{1}^{s}(\mu) \right) + 4e^{-4} \\
    \leq \frac{1}{8} + 4e^{-4} + 4e^{-4} \leq \frac{1}{4}.
\end{align*}
\]

It follows then by Le Cam’s mixture-vs-mixture lemma (see Lemma X in Appendix Y) that we have

\[
\inf_{\hat{\theta}} \sup_{P \in M_n} \mathbb{E} \left[ |\hat{\theta} - \theta(P)| \right] \\
\geq \frac{1}{4} \left\{ 1 - d_{TV} \left( \int P_{\mu, \xi}^{\otimes n} dQ_{-1}^{s}(\mu), \int P_{\mu, \xi'}^{\otimes n} dQ_{1}^{s}(\mu) \right) \right\} \cdot \inf_{\mu \in S, \mu' \in S'} [\tau(\xi^s, \mu) - \tau(\xi^s, \mu')] \\
\geq \frac{c}{\sqrt{n}} O_\delta
\]

for a universal constant \( c > 0 \), and this completes the proof. \( \square \)

It remains to prove Lemma 40.

**Lemma 40.** Again we exploit a device used by [58]. We allow the number of observations to be random, which simplifies the calculations. Let

\[
Q_{\xi}^{s, \otimes n} = \int P_{\mu, \xi}^{\otimes n} dQ_{\xi}^{s}(\mu).
\]

Consider the pair of distributions \((\bar{Q}_{-1}^{s}, \bar{Q}_{1}^{s})\), where we draw \( \nu \sim \text{Poisson}(2n) \) independently of everything else and let

\[
\bar{Q}_{\xi}^{s} = \sum_{k=0}^{\infty} \bar{Q}_{\xi}^{s} \cdot P(\nu = k).
\]
Now, for $E_n = \{ \nu \geq n \}$, it is well-known that $P(E_n) \geq 1 - e^{-n/4}$. By Lemma 44, we have

$$d_{TV} \left( Q_{s,\oplus n}^s, Q_{1}^{s, \oplus n} \right) \leq d_{TV} \left( Q_{-1}^s | E_n, Q_1^s | E_n \right) + 4P(E_n^c)$$

$$\leq \frac{1}{P(E_n)} d_{TV} \left( Q_{-1}^s, Q_1^s \right) + 6P(E_n^c)$$

$$\leq 2d_{TV} \left( Q_{-1}^s, Q_1^s \right) + 6 e^{-n/4},$$

which holds for any $n \geq 4$. It remains to analyze the total variation distance between the ‘Poissonized’ mixtures. First, let the empirical count function be given by

$$M(x, a) = \sum_{i=1}^n \mathbb{1} \{ X_i^j = x_t, A_i^j = a_t ; t = 1, \ldots, T \}.$$

Note that conditional on the value of $\nu$, the vector $\{ M(x, a) : (x, a) \in (X^T, A^T) \}$ is multinomially distributed, and since $\nu \sim \text{Poisson}(2n)$, it follows by the thinning theorem that $M(x, a) \sim \text{Poisson}(2n \cdot \lambda(x, a))$, where

$$\lambda(x, a) = \xi^s(x_1) \pi^b(x_1, a_1) \cdots \cdot P_{T}(x_T, a_{T-1}, x_{T-1}) \pi^b(x_T, a_T).$$

We proceed by considering alternative distributions $Q_{\xi}(x, a)$ for $x \in X_1 \times \cdots \times X_T$, $a \in A_1 \times \cdots \times A_T$, and define $Q_{\xi}(x, a)$ for each $(x, a) \in (X^T, A^T)$ as follows:

a) Sample $M(x, a) \sim \text{Poisson}(2n \cdot \lambda(x, a))$.

b) Sample $\xi(x, a) \sim \prod_{t=1}^T \text{Ber} \left( \frac{1 + s_x \rho_t(x_t, a_t)}{2} \right)$.

c) Generate the $M(x, a)$ outcomes $\tilde{Y}$ independently from the conditional distribution $Y_t | X_t = x_t, A_t = a_t$.

One readily sees that $Q_{\xi}^s = \prod_{(x, a) \in (X^T \times A^T)} Q_{\xi}(x, a)$, and so by Pinkser’s inequality and the inde-
pendence of \( \overline{Q}_s(x, a), \overline{Q}_s'(x', a') \) for \((x, a) \neq (x', a')\) gives

\[
d_{TV}(\overline{Q}_s^{-1}, \overline{Q}_s^1) \leq \sqrt{\frac{1}{2} d_{KL}(\overline{Q}_s^{-1}, \overline{Q}_s^1)} \leq \sqrt{\frac{1}{2} \sum_{(x,a) \in X^T \times A^T} d_{KL}(\overline{Q}_s^{-1}(x, a), \overline{Q}_s^1(x, a))},
\]

Now, \( \overline{Q}_s^{-1}(x, a), \overline{Q}_s^1(x, a) \) only differ in the parameters \( \rho_t \) of the \( T \) independent Bernoulli variables \( \zeta_t \), and these are only observed if \( M(x, a) > 0 \). Using the convexity of KL divergence and the independence of the Bernoulli variables, we have

\[
d_{KL}(\overline{Q}_s^{-1}(x, a), \overline{Q}_s^1(x, a)) \leq \mathbb{P}(M(x, a) > 0) \cdot \sum_{t=1}^{T} d_{KL}(\zeta_{t}^1, \zeta_{t}^{-1}),
\]

where \( \zeta_{t}^1 \sim \text{Ber}\left(\frac{1+\rho_t(x_t, a_t)}{2}\right) \) and \( \zeta_{t}^{-1} \sim \text{Ber}\left(\frac{1-\rho_t(x_t, a_t)}{2}\right) \). It follows that

\[
d_{KL}(\overline{Q}_s^{-1}(x, a), \overline{Q}_s^1(x_t, a_t)) \leq (1 - e^{-2nM(x, a)}) \cdot \sum_{t=1}^{T} 4s^2 \rho_t(x_t, a_t)^2.
\]

Now, since \( \lambda(x, a) = \mathbb{P}(X = x, A = a) \) for \((x, a) \in X^T \times A^T \), we have

\[
\sum_{(x,a)} \lambda(x, a) \rho_t^2(x_t, a_t) = \mathbb{E} \left[ \rho_t^2(X_t, A_t) \right] 
\leq \frac{1}{O_\delta^2} \mathbb{E} \left[ \left( \frac{\xi_t^e(X_t)}{\xi_t^e(X_t)} \right)^2 \sum_{a_t \in A_t} \frac{\pi^e(X_t, a_t)^2}{\pi^b(X_t, a_t)} \cdot \delta_t^2(X_t, a_t) \right].
\]

Summing over the time steps and putting together the pieces gives

\[
d_{TV}(\overline{Q}_s^{-1}, \overline{Q}_s^1) \leq 2s \sqrt{n},
\]

completing the proof. \( \square \)
3.3.3 Auxiliary Lemmas.

The above proofs make use of two versions of Le Cam’s Lemma, as well as Pinsker’s inequality, which we include here for completeness.

**Lemma 42.** [Le Cam] For any $2\delta$-separated (w.r.t. a pseudo-metric $\rho$) classes of distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ contained within $\mathcal{P}$, any estimator $\hat{\theta}$ has worst-case risk at least

$$\sup_{M \in M_n} \mathbb{E} \left[ \rho(\hat{\theta}, \theta(M)) \right] \geq \frac{\delta}{2} \sup_{M_0 \in \mathcal{P}_0, M_1 \in \mathcal{P}_1} \{1 - \|M_0 - M_1\|_{TV}\}.$$

We will use this result combined with Pinsker’s inequality to derive our first lower bound.

**Lemma 43** (Pinsker’s Inequality). For probability measures $\mathbb{P}_0, \mathbb{P}_1$, we have

$$\|\mathbb{P}_0 - \mathbb{P}_1\|_{TV} \leq \sqrt{\frac{1}{2} D_{kl}(\mathbb{P}_0 \| \mathbb{P}_1)}.$$

Two elementary auxiliary technical lemmas are invoked to prove Proposition 37, which we state here. Proofs of both can be found in [58].

**Lemma 44.** Let $(\mu, \nu)$ be a pair of probability distributions over the same Polish space $S$, and consider a subset $\mathcal{E} \subset S$ such that $\min \{\mu(\mathcal{E}), \nu(\mathcal{E})\} \geq 1 - \epsilon$ for some $\epsilon \in [0, 1/4]$. Then the conditional distributions $\mu \mid \mathcal{E}, \nu \mid \mathcal{E}$ satisfy the bound

$$d_{TV}(\mu, \nu) - 4\epsilon \leq d_{TV}((\mu \mid \mathcal{E}), (\nu \mid \mathcal{E})) \leq \frac{1}{1 - \epsilon} d_{TV}(\mu, \nu) + 2\epsilon.$$

**Lemma 45.** Let $X$ be a real-valued random variable with finite fourth moment, and define the $(2 - 4)$-moment constant $M_{2 \rightarrow 4} = \sqrt{\mathbb{E}[X^4] / \mathbb{E}[X^2]}$. Then we have the lower bound

$$\mathbb{E} \left[ X^2 1|X| \leq 2M_{2 \rightarrow 4}\sqrt{\mathbb{E}[X^2]} \right] \geq \frac{1}{2} \mathbb{E}[X^2].$$
3.4 A Lower Bound over Smoothness Classes.

The lower bound given in above implicitly addresses the case when the behaviour policy \( \pi_b \) and the marginal state distributions (both on and off-policy) are known - which, naturally, cannot often be taken for granted in practice. In the event that the statistician has minimal knowledge about the problem, there will be a price to be paid for estimating the nuisance components required to make a suitable estimate of the average cumulative reward. In this Section, we model the data-generating mechanism as being comprised of functions belonging to Hölder-smoothness classes, and provide a minimax lower bound for the OPE problem in this model.

3.4.1 Problem Setup.

Consider the problem of off-policy estimation in the continuous setting. In particular, suppose that we observe trajectories \( Z = (S_0, A_0, R_0, S_1, \ldots, A_T, R_T) \), such that \( S_t \in S = [0, 1]^d_S \) with \( S_0 \) having density \( f \), actions \( A_t \in A = [0, 1]^d_A \) are selected by a behaviour policy \( \pi_b \), so that the conditional density of \( A_t \mid S_t \) is given by \( \pi_b^e (\cdot \mid S_t) \), and finally the reward \( R_t \in R \) has conditional mean given by \( R_t \mid S_t, A_t \sim \mu_t (S_t, A_t) \); for simplicity, assume that \( R_t \in \{0, 1\} \). The joint observational distribution may therefore be indexed by the triplet \( P \sim (\mu, \pi^b, f) \), while the problem consists of estimating the functional

\[
\theta(P) = \int \mu_0 \pi_0^e f \, d\nu + \sum_{t=1}^T \int \mu_t \pi_t^e P_t \, d\nu, \tag{3.5}
\]

where \( P_t \) is the induced density of \( S_t \) and \( \nu \) is a dominating measure over \( S \times A \times R \) and \( \pi_t^e \) are fixed policy functions \( \pi_t^e : [0, 1]^d \rightarrow \mathbb{R} \) (where we set \( d = d_S + d_A \)). For a sample \( Z^n = \{(S_0^i, A_0^i, R_0^i, S_1^i, \ldots); i = 1, \ldots, n\} \), we let \( \hat{\theta}_n : Z^n \rightarrow \mathbb{R} \) denote a generic estimator of \( \theta \). Let \( \lambda_t(s,a) = (\pi_t^e / \pi_b^e)(s,a) \) denote the policy importance ratio function, and for \( \beta > 0 \), denote by \( \mathcal{H}(\beta) \) the class of \( \beta \)-smooth Hölder functions.
3.4.2 Result.

The result is as follows:

**Theorem 46.** Let $\mathcal{P}$ denote the model over $\mathbb{Z}$ such that

i) For each $t = 0, 1, \ldots, T$, we have that $\mu_t \in \mathcal{H}(\alpha)$ for some $\alpha > 0$.

ii) For each $t = 0, 1, \ldots, T$, $\lambda_t \in \mathcal{H}(\gamma)$ for some $\gamma > 0$ with $\gamma \geq \alpha$.

Suppose further that the (fixed) evaluation policy satisfies $0 < \underline{\pi} \leq \pi_t^e \leq \bar{\pi}$ for all $t$, for some positive constants $\pi, \bar{\pi}$, and that $\alpha + \gamma < d/2$. For $n$ larger than a constant depending only on $(\alpha, \gamma, d)$, we have

$$
\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ (\hat{\theta}(P) - \theta(P))^2 \right] \geq n^{-(\alpha+\gamma)/(2\alpha+2\gamma+d)}.
$$

3.4.3 Proof.

We begin by stating the supporting lemmas that we’ll need. The first lemma below is found in [60]; we state the version as it appears in [61]. Note that in what follows, we consider the case $T = 1$, which is sufficient to establish the result.

**Lemma 47** (Tsybakov 2009). Let $P_{\lambda}$ and $Q_{\lambda}$ denote distributions in $\mathcal{P}$ indexed by a vector $\lambda = (\lambda_1, \ldots, \lambda_k)$, with $n$-fold products denoted by $P^n_{\lambda}$ and $Q^n_{\lambda}$, respectively. Let $\omega$ denote a prior distribution over $\lambda$. If

$$
H^2 \left( \int P^n_{\lambda} d\omega(\lambda), \int Q^n_{\lambda} d\omega(\lambda) \right) \leq \alpha < 2
$$

and

$$
|\psi(P_{\lambda}) - \psi(Q_{\lambda})| \geq s \geq 0
$$

for a functional $\psi : \mathcal{P} \rightarrow \mathbb{R}$ and for all $\lambda$, then

$$
\inf_{\hat{\psi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left\{ \ell \left( |\hat{\psi} - \psi(P)| \right) \right\} \geq \ell(s/2) \left( 1 - \frac{\sqrt{\alpha(1-\alpha/4)}}{2} \right)
$$

for any monotonic non-negative loss function $\ell$. 

93
Lemma 48 (Robins et al. 2009). Let $P_\lambda$ and $Q_\lambda$ denote distributions indexed by a vector $\lambda = (\lambda_1, \ldots, \lambda_k)$, and let $Z = \bigcup_{j=1}^k Z_j$ denote a partition of the sample space. Assume:

1. $P_\lambda(Z_j) = Q_\lambda(Z_j) = p_j$ for all $\lambda$, and

2. the conditional distributions $1\{Z_j\}dP_\lambda/p_j$ and $1\{Z_j\}dQ_\lambda/p_j$ do not depend on $\lambda_\ell$ for $\ell \neq j$.

For a prior distribution $\omega$ over $\lambda$, let $p = \int p_\lambda d\omega(\lambda)$ and $q = \int q_\lambda d\omega(\lambda)$, and define

$$
\delta_1 = \max_j \sup_\lambda \int_{Z_j} \frac{(p_\lambda - p)^2}{p_\lambda p_j} d\nu,
$$

$$
\delta_2 = \max_j \sup_\lambda \int_{Z_j} \frac{(p_\lambda - q_\lambda)^2}{p_\lambda p_j} d\nu,
$$

$$
\delta_3 = \max_j \sup_\lambda \int_{Z_j} \frac{(p - q)^2}{p_\lambda p_j} d\nu,
$$

for a dominating measure $\nu$. If $p/p_\lambda \leq b < \infty$ and $\max_j p_j \leq b/n$, then

$$
H^2\left(\int P^n_\lambda d\omega(\lambda), \int Q^n_\lambda d\omega(\lambda)\right) \leq Cn^2(\max_j p_j)(\delta_1 \delta_2 + \delta_2^2) + Cn\delta_3
$$

for a constant $C$ depending only on $b$.

Our construction of the mixture distributions $P_\lambda, Q_\lambda$ is inspired by that of [62], with some modifications that are crucial for accounting for the off-policy functional.

Definition 20. Let $P' \subset P$ denote the submodel for which $\mathcal{Y} = \{0, 1\}$. Let $H : \mathbb{R}^d \to \mathbb{R}$ be a $C^\infty$ function supported on $[0, 1]^d = \mathcal{A} \times \mathcal{S}$ such that $\int_{\mathcal{A}} H d\nu(a) = 0$ and $\int H^2 d\nu = 1$. Choose a positive integer $k$, and let $\lambda = (\lambda_1, \ldots, \lambda_k) \in \Lambda = \{-1, 1\}^k$. Set $\omega$ to be uniform on $\Lambda$, so that the $\lambda_\ell$s are i.i.d. Rademacher variables. Let $X_1, \ldots, X_k$ be disjoint translates of $k^{-1/d}[0, 1]^d$ with $X_j \subset [0, 1]^d$ for all $j = 1, \ldots, k$; let $x_1, \ldots, x_k$ denote the bottom left corners of these cubes. For $\beta > 0$, define

$$
\Delta^\beta_{\lambda}(x) = (1/k)^{\beta/d} \sum_{i=1}^k \lambda_i H\left((x - x_i)k^{1/d}\right), \quad x \in [0, 1]^d.
$$
It can be shown (see e.g. [60]) that $\Delta^\beta_\lambda \in \mathcal{H}(\beta)$ with norm uniformly bounded in $k$. We index distributions $P_\lambda, Q_\lambda \subset \mathcal{P}'$ using the triplet $(\mu_\lambda, \pi^b_\lambda, f_\lambda)$ via

$$P_\lambda \sim (1/2 + c_1 \Delta^\alpha_\lambda + c_2 \Delta^\alpha_\lambda \Delta^\gamma_\lambda / \pi^e, \pi^e, 1)$$

$$Q_\lambda \sim (1/2 + c_1 \Delta^\alpha_\lambda, \pi^e + c_3 \Delta^\gamma_\lambda, 1) ,$$

for constants $c_1, c_2, c_3$ to be specified later.

The following lemma is straightforward to establish.

**Lemma 49.** Set $c_1 = \frac{1}{8\Delta_1}, c_2 = \frac{\pi}{8\Delta_1}, c_3 = \frac{\pi}{\Delta_1}$. Then we have $P_\lambda, Q_\lambda \subset \mathcal{P}'$.

**Proof.** Just as in [60], we may assume that there exists a $\Delta_1$ such that $|\Delta^\alpha_1| \leq \Delta_1$. Since $\gamma \geq \alpha$, we have $|\Delta^\gamma_1| \leq \Delta_1$ as well. With $c_1 = 1/8\Delta_1$, we clearly have that $1/2 - 1/8 \leq \mu^Q_\lambda \leq 1/2 + 1/8$. Similarly, $c_3 = \pi/\Delta_1$ implies that

$$\mu^P_\lambda = 1/2 + c_1 \Delta^\alpha_\lambda + c_2 \Delta^\alpha_\lambda \Delta^\gamma_\lambda / \pi^e$$

$$\leq 1/2 + (\pi/8\Delta_1^2) \Delta^\alpha_\lambda \Delta^\gamma_\lambda (\pi)^{-1}$$

$$\leq 1/2 + 1/8 + 1/8 = 3/4,$$

and $\mu^P_\lambda \geq 1/4$ in the same way. That both $\mu^P_\lambda, \mu^Q_\lambda$ are $\alpha$-smooth follows since $\Delta^\alpha, \Delta^\gamma \subset \mathcal{H}(\alpha)$ (where $\Delta^\gamma \subset \mathcal{H}(\alpha)$ because $\gamma \geq \alpha$), and we conclude that they are valid outcome mean functions. The behaviour policy under $Q, \pi^b_\lambda = \pi^e + c_3 \Delta^\gamma_\lambda$, is in $\mathcal{H}(\gamma)$ since both $\Delta^\gamma_\lambda$ and $\pi^e$ are. Using $|\Delta^\gamma_\lambda| \leq \Delta_1$ and the value of $c_3$, we see that $\pi^b_\lambda \geq 0$; that these define valid conditional distributions follows from $\int_H Hd\nu(a) = 0$. The conclusion follows. \[\Box\]

The essence of the proof consists of applying Lemma 48 to get an upper bound on the Hellinger distance between the mixture distributions.
Proposition 50. Let \( P_\lambda, Q_\lambda \) be as in Definition 20, and take \( \delta_1, \delta_2, \delta_3 \) as in Lemma 48. Then we have
\[
\delta_1 \leq 4c_3^2 \pi^{-1} (1/k)^{2\alpha/d}, \quad \delta_2 \leq c_2^2 \pi^{-2} (1/k)^{2\gamma/d}, \quad \delta_3 = 0,
\]
and it follows that
\[
H^2 \left( \int P^n_\lambda d\omega(\lambda), \int Q^n_\lambda d\omega(\lambda) \right) \leq C c_P^{-2} \pi^4 \cdot n^2 \frac{1}{k} \left( (1/k)^{2(\alpha+\gamma)/d} + (1/k)^{4\gamma/d} \right).
\]

Proof. Note that we may assume that there is a constant \( c_P > 0 \) such that \( p_\lambda(z) \geq c_P \). Now, partitioning the sample space as \( \mathcal{Z}_1, \ldots, \mathcal{Z}_k = \{0,1\} \times \mathcal{X}_1, \ldots, \{0,1\} \times \mathcal{X}_k \), it is clear that
\[
\bar{\pi}^{-1} \leq Q_\lambda(\mathcal{Z}_j) = P_\lambda(\mathcal{Z}_j) = p_j \leq \bar{\pi}^{-1}
\]
It follows then that we have
\[
\bar{p}(z) := \int p_\lambda(z) d\omega(\lambda) = \int (1/2 + (2y - 1)(c_1 \Delta_\lambda^\alpha(z) + c_2 \Delta_\lambda^\beta(z)/\pi^e(z))\pi^e(z)) d\omega(\lambda)
\]
\[
= \frac{1}{2} \pi^e(z) + (2y - 1)\Delta_\lambda^{2\alpha+\gamma}(z),
\]
since clearly \( \int \Delta_\lambda^\alpha(z) d\omega(\lambda) = 0 \). Similar reasoning implies that
\[
\bar{q}_\lambda(z) = \frac{1}{2} \pi^e(z) + (2y - 1)\Delta_\lambda^{2\alpha+\gamma}(z),
\]
which immediately gives $\delta_3 = 0$. For $\delta_1$, we have $p_\lambda - \bar{p} = c_3 \Delta^a \pi^e$, and so we obtain

$$
\delta_1 = \max_j \sup_\lambda \int_{Z_j} \frac{(p_\lambda - \bar{p})^2}{p_\lambda p_j} \, dv(z) \\
\leq \frac{1}{\pi} \bar{p} \sup_\lambda \int_{Z_j} \frac{(c_3 \Delta^a \pi^e)^2}{p_\lambda} \, dv(z) \\
\leq c_3^2 \frac{1}{\pi} \bar{p} \sup_\lambda \int_{Z_j} \frac{(\Delta^a \pi^e)^2}{\mu^\pi(z) p_\lambda} \, dv(z) \\
\leq 4c_3^2 \frac{1}{\pi} \bar{p} (1/k)^{2\alpha/d} \left\{ k \max_j \int_{Z_j} \sum_{i=1}^k H^2 \left( (z - z_i) k^{1/d} \right) \, dv(z) \right\} \\
= 4c_3^2 \frac{1}{\pi} \bar{p} (1/k)^{2\alpha/d},
$$

which follows because $\int H^2 \, dv = 1$. For $\delta_2$, we have $p_\lambda - q_\lambda = \frac{1}{2} c_2 \Delta^\gamma$, so

$$
\delta_2 = \max_j \sup_\lambda \int_{Z_j} \frac{(p_\lambda - q_\lambda)^2}{p_\lambda p_j} \, dv(z) \\
\leq \frac{1}{\pi} \bar{p} \sup_\lambda \int_{Z_j} \frac{\left( \frac{1}{2} c_2 \Delta^\gamma \right)^2}{p_\lambda} \, dv(z) \\
\leq c_2^2 \frac{1}{\pi} \bar{p} \sup_\lambda \int_{Z_j} \frac{(\Delta^\gamma)^2}{\mu^\pi(z) p_\lambda} \, dv(z) \\
= c_2^2 \frac{1}{\pi} \bar{p} (1/k)^{2\gamma/d},
$$

where the last equality follows as before by construction of $H$. So, provided that $k$ is chosen such that $\bar{p}/p_\lambda \leq b$ and $\max_j p_j = 1/k \leq b/n$, by Lemma 48 and the above bounds on $\delta_1, \delta_2$, we have

$$
H^2 \left( \int P_\lambda^n \, d\omega(\lambda), \int Q_\lambda^n \, d\omega(\lambda) \right) \leq C c_p^{-2} \pi^4 \cdot n^2 \bar{p} \left( (1/k)^{2(\alpha+\gamma)/d} + (1/k)^{4\gamma/d} \right).
$$

The proof of the theorem now simply consists in putting the pieces together:

*Of Theorem.* Let $P_\lambda, Q_\lambda$ be as above, and choose $k \sim n^{2d/(2(\alpha+\gamma)+d}$ up to a constant, such that by
Proposition 50 we have

\[ H^2 \left( \int P_\lambda^n d\omega(\lambda), \int Q_\lambda^n d\omega(\lambda) \right) \leq n^2 k \left( (1/k)^{2(\alpha+\gamma)/d} + (1/k)^{4\gamma/d} \right) \]

\[ \leq n^2 (1/k)^{(\alpha+\gamma)/d-1} \]

\[ \leq 1. \]

By Lemma 47 we then have

\[ \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \hat{\theta}(P) - \theta(P) \right]^2 \geq n^{-2(\alpha+\gamma)/(2\alpha+2\gamma)+d}. \]

Since \( \mathcal{P}' \subset \mathcal{P} \), the lower bound holds over \( \mathcal{P} \) as well, and the proof is complete. \( \square \)

3.5 Open Questions and Further Work.

The results in this chapter constitute a paltry step towards a more comprehensive understanding of the difficulty of off-policy evaluation. To conclude, this section will recap our results and indicate what is needed to prepare this work for publication. In doing so, we give an overview of the some questions that remain unanswered, and also point to avenues that might prove fruitful for further research. We shall consider how our work on lower bounds might be expanded, both to provide richer or more structured characterisations of what might be achievable and to establish optimality by deriving matching upper bounds. Of particular interest are questions that further elucidate connections between the transfer learning and reinforcement learning literature. We also briefly discuss other open questions that were raised during the course of our research.

3.5.1 Discussion of Results.

We have provided two minimax lower bounds for the mean-square error of off-policy evaluation. The first of these, Theorem 35, gives a bound over a class of MDPs in which the mean reward function at each state-action pair lies within a certain distance from a fixed reference. The bound
is seen to be a function of an expectation of a conditional chi-square divergence term between the behaviour and target policies, weighted by the outcome variance and the magnitude of the reward perturbation, providing a notion of overlap between policies; we also see the Markov structure reflected in the presence of the ratio of marginal state distributions in the bound. This result captures the difficulty of estimating the reward function from i.i.d. trajectories from the MDP, but ignores the difficulty introduced by the estimation of nuisance functions. In the simplest setting, in which we suppose that we know the MDP parameters and the behaviour policy but not the reward, this bound can easily be shown to be tight by taking the MSE of the marginal IS estimator. Our second result, Theorem 46, provides a partial answer to the question of the added difficulty of estimating the reward mean function and behaviour policies when these are known to lie in a certain smoothness class. We obtain a lower bound that depends only on the smoothness of the ratio of the target to the behaviour policy, when this is larger than the smoothness of the reward function. The strength of this result is, however, limited in that it does now address the additional difficulty of estimating the state marginal density ratios in the dynamic setting.

These results point to some salient connections to the transfer learning literature. There are parallels between the chi-square term in Theorem 35 and various quantities that have been studied in transfer learning; for example, [63] derive learning bounds for transfer in terms of the family of Renyi divergences (see [64] for details and properties of such divergences), the \( \chi^2 \)-divergence being directly related to the Renyi divergence of order 2. Improving our understanding of how best to quantify the discrepancy between distributions to cater to particular problems remains an important question in both lines of research. As we have seen in Chapter 2, nonparametric minimax rates for classification under covariate shift can be characterized by a quantity intimately related to the Renyi dimension of a measure, and we believe that it may be possible to improve our result in continuous state-spaces by deriving rates that are characterized by the aggregate transfer exponent from the marginal state distribution of the states under the behaviour policy to that of the state under the target (or between the policies themselves, conditional on the states), though at present such a result eludes us.
3.5.2 Further Work.

In order to complement the results of this Chapter, we have attempted to derive the following results, the resolution of which we feel would move this work to the stage of being publishable. The first step is to leverage the framework of structure-agnostic learning, introduced by [65], in order to glean insights that may be readily extended to many structured settings in an ad-hoc manner. These authors derive minimax lower bounds for the estimation of various functionals that are stated in terms of the estimation rates that are achievable for the nuisance components. We conjecture that it should be possible, using their techniques (which are very closely related to the technique we have used to prove Theorem 46; both are ultimately inspired by the seminal work by Robins et. al., [62]), to derive a minimax lower bound that would immediately elucidate how the mean-square error of OPE is influenced by the error of estimating, respectively, the mean reward functions, the policy ratio functions, and the marginal density ratios. Equipped with such a bound in the setting of OPE in MDPs, the problem of extending the lower bound in Theorem 46 would reduce to the twin problems of individually estimating a smooth function, and estimating density ratios, for which there is a large literature. From here, a natural question would be whether we can be explicit about these rates in terms of the aggregate transfer exponents between the behaviour and target policies, conditional on the states. This may also shed light on procedures for OPE in continuous spaces for which density ratios between policies do not exist, which may occur if, for example, the target policy is deterministic in a continuous action space (we have seen in Chapter 2 that the aggregate transfer exponents are well-defined even when the density ratio may not be).

3.5.3 Other Open Questions.

There are, naturally, many secondary questions besides those mentioned above that have suggested themselves to us that remain unanswered; here we give account of some of them. Firstly, it should be possible to extend Theorem 35 to include a term that captures the semiparametric efficient variance, as [58] have done in the non-dynamic setting. It would then be of interest to derive finite sample upper bounds for the doubly-robust marginal importance sampling estimator that has
been shown in [8] to be semiparametrically efficient in the MDP setting - this would be a result of great interest, as it would help render it possible to make judgements about the optimality of this estimator in a finite-sample sense in addition to asymptotically, and may also facilitate the derivation of improvements to this estimator that might match lower bound rates in certain structural settings. Again, in the non-dynamic setting $T = 1$ this has been done in [58]; it is not clear however whether this problem remains tractable given the added complexity induced by the temporal structure, and we have not attempted this. It would also be of interest to relax the requirement of having discrete state spaces, although in this case the proof technique for Proposition 37 would no longer be valid.

The question of whether the lower bound of Theorem 46 can be matched by an upper bound remains open, although this would require conditions on the density ratios. In the single-period causal inference setting of estimating the average treatment effect, [66] derives a procedure that matches the lower bound for the same problem from [62]. Likewise, [61] provides matching bounds for the estimation of heterogeneous treatment effects. Both of these results, however, require conditions on the density functions that are as far as we can see not readily extendable to the OPE problem for MDPs.
3.6 Appendix to Chapter 3.

3.6.1 Kallus and Uehara’s Double Machine Learning Method.

For completeness, we state here the full description of the DRL method as given in [8]. Given estimates of the $q$-functions $\hat{q}_t$ and marginal density ratio functions $\hat{m}_t$, the DRL (with $K$-fold sample splitting) procedure is as follows:

i) Randomly permute the data indices, and let $D_j = \{\lceil (j-1)n/K \rceil + 1, \ldots, \lceil jn/K \rceil \}$ for $j = 1, \ldots, K$. Let $j_i$ be the fold containing observation $i$, so that $i \in D_{j_i}$ (namely, $j_i = 1 + \lfloor (i-1)K/n \rfloor$).

ii) For $j = 1, \ldots, K$, construct estimators $\hat{\mu}_t^{(j)}$ and $\hat{q}_t^{(j)}$ based on the training data given by all the trajectories excluding those in $D_j$.

iii) Let

$$\hat{\rho}_{DRL} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \hat{\mu}_t^{(j_i)}(s_t^{(i)}, a_t^{(i)}) \left( r_t^{(i)} - \hat{q}_t^{(j_i)}(s_t^{(i)}, a_t^{(i)}) \right)$$

$$+ \hat{\mu}_{t-1}^{(j_i)}(s_{t-1}^{(i)}, a_{t-1}^{(i)}) \int_{a_t^{(i)}} q_t^{(j_i)}(s_t^{(i)}, a_t^{(i)}) \, d\pi_t(a_t' | s_t^{(i)}),$$

where $r_t$ is the reward observed at time $t$. 

102
Conclusion

This thesis has addressed two fundamental problems in transfer learning: binary classification across differing populations, and the off-policy estimation problem in reinforcement learning.

In the former case, we introduced a quantity - the aggregate transfer exponent - that we show to characterize the achievable excess risk rates for classification in the covariate shift setting under standard nonparametric assumptions. We have seen how the aggregate transfer exponent may be interpreted as a notion of relative dimension between the marginal feature distributions, and relate it the literature on dimensions for single measures. Further, we show how one can obtain asymptotic rates of convergence determined only by the aggregate transfer exponent in arbitrarily small regions containing the decision boundary. Finally, we derived an algorithm based on Lepski’s procedure which is shown to adaptively achieve rates of convergence determined by such a localized exponent. This algorithm operates on a fully grown tree via a judicious local choice of tree-depth that is also highly efficient, running with logarithmic complexity\(^2\) at testing time, for a modest initialization cost; it is shown to perform competitively with other model selection approaches for decision trees in the transfer setting.

In the latter case, we considered the question of the fundamental difficulty of off-policy evaluation, and presented some incremental results that may serve as the starting point for a more complete characterization of hardness of OPE in terms of mean-square error. Extending some results on functional estimation for causal inference, we provided two minimax lower bounds for

\(^2\)That is, in \(O(\log n)\) for a tree grown from \(n\) samples.
mean-square error of OPE for Markov Decision Processes, both of which immediately generalize to NMDPs. First, we consider the case of MDPs with discrete state spaces, and prove a finite-sample lower bound that is seen to be a weighted version of the chi-square divergence between the behaviour and target policies, thus reaffirming the connections between this problem and transfer learning more generally. Secondly, for continuous state spaces, we provide a minimax lower bound over Hölder smoothness classes of mean reward and policy functions, thus partially addressing the difficulty of OPE when the behaviour policy is unknown and therefore must be estimated as well. Finally, to underline the importance of increasing practitioner’s understanding about the fundamental difficulty of OPE, we analyze a study that carried out misguided inference, and give an indication of the problem’s difficulty by estimating our finite state space lower bound, suggesting that the task undertaken was too ambitious. Many questions remain open. How to correctly account for the Markov dynamics in the bounds is of particular interest; work on deriving matching upper bounds should help shed light on this problem.
References


