Proceedings of Machine Learning Research vol 98:1–24, 2019 30th International Conference on Algorithmic Learning Theory

A Generalized Neyman-Pearson Criterion for Optimal Domain Adaptation

Clayton Scott CLAYSCOT@UMICH.EDU
Division of Electrical and Computer Engineering and Department of Statistics
University of Michigan
Ann Arbor, MI 48109 USA

Editors: Aurélien Garivier and Satyen Kale

Abstract

In the problem of domain adaptation for binary classification, the learner is presented with labeled examples from a source domain, and must correctly classify unlabeled examples from a target domain, which may differ from the source. Previous work on this problem has assumed that the performance measure of interest is the expected value of some loss function. We study a Neyman-Pearson-like criterion and argue that, for this optimality criterion, stronger domain adaptation results are possible than what has previously been established. In particular, we study a class of domain adaptation problems that generalizes both the covariate shift assumption and a model for feature-dependent label noise, and establish optimal classification on the target domain despite not having access to labelled data from this domain.

Keywords: Domain Adaptation, Neyman-Pearson Classification, Feature-Dependent Label Noise, Covariate Shift, Immunity

1. Introduction

In the problem of domain adaptation for binary classification, the learner is given labeled examples from a source distribution, and must design a classifier that performs well on a potentially different target distribution. We consider the semi-supervised setting where, in addition to labeled training data from the source distribution, the learner has access to an unlabeled sample from the target distribution. To gain traction on this problem, it is necessary to make some assumptions relating the source and target distributions, and several types of assumptions have been considered previously in the literature, such as covariate shift, target shift, and various forms of label noise.

Previous work on domain adaptation has focused almost exclusively on a particular class of performance measures, namely, those expressible as the expected value of some loss function, with particular attention being paid to the 0-1 loss. We argue that the difficulty of a domain adaptation problem depends on the performance measure being optimized, and the focus on loss-based criteria has limited the contributions of prior work. The present work was motivated by the problem of classification with feature- (or instance-) dependent label noise (FDLN), where previous efforts to minimize the expected 0-1 loss (probability of error) require excessively strong assumptions on the nature of the label noise. Our work also bears on the covariate shift model, where prior work requires source and target distributions to be rather similar in order to make strong performance guarantees.

We examine an optimality criterion for binary classification that we call the controlled discovery rate (CDR), which is a special case of a more general class of generalized Neyman-Pearson criteria. We show that it is possible to optimize CDR over a broad class of domain adaptation problems that we refer to as *covariate shift with posterior drift*. We do this by showing that the CDR criterion is *immune* to this class of domain adaptation problems, meaning one can train a classifier as if the source and target distributions were the same, and still optimize the CDR criterion when they are different. Thus, no particularly novel algorithms are required to achieve optimal domain adaptation. Our results lead to more general statements of optimality for covariate shift and FDLN than have previously been established.

1.1. Notation

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the feature space and $\{0, 1\}$ the label space. Let Q be a probability distribution on $\mathcal{X} \times \{0, 1\}$. If the pair (X, Y) are jointly distributed according to Q, let Q_y, $y \in \{0, 1\}$, denote the conditional distribution of X given $Y = y$. Q_0 and Q_1 are referred to as the "class-conditional distributions." Denote by π_Q the marginal probability that $Y = 1$, and by $\eta_Q(x)$ the conditional probability that $Y = 1$ given $X = x$. In classification, Y may be viewed as an unknown parameter that must be predicted from X, and in this spirit we refer to π_Q and $\eta_Q(x)$ as the "prior" and "posterior" probabilities associated to Q. Finally, let $Q_X := \pi_Q Q_1 + (1 - \pi_Q)Q_0$ be the marginal distribution of X.

Throughout this work we assume that Q_0 and Q_1 have densities q_0 and q_1, defined w.r.t. some dominating measure μ, and related to $\eta_Q(x)$ via Bayes rule:

$$\eta_Q(x) = \frac{\pi_Q q_1(x)}{(1 - \pi_Q)q_0(x) + \pi_Q q_1(x)}. \tag{1}$$

We will often refer to a second distribution P on $\mathcal{X} \times \{0, 1\}$ in addition to Q. The associated quantities $P_0, P_1, \pi_P, \eta_P, P_X, p_0$ and p_1 are defined analogously. In this case the densities p_0, p_1, q_0, q_1 are assumed to have a common dominating measure. The choice $\mu = P_0 + P_1 + Q_0 + Q_1$ is always valid, but typically μ is either the Lebesgue or counting measure.

1.2. Objective

In domain adaptation there are two distributions, P and Q, referred to as the *source* and *target* distributions. We consider the semi-supervised setting where the learner observes $(X_1, Y_1), \ldots, (X_m, Y_m) \sim P$ and $X_{m+1}, \ldots, X_{m+n} \sim Q_X$, and must design a classifier whose performance/optimality is assessed with respect to Q. The focus of this paper is to consider a particular optimality criterion, the CDR criterion, such that optimal classification is possible under a class of domain adaptation problems now described.

1.3. Covariate Shift with Posterior Drift

The class of domain adaptation problems considered is a combination of two fundamental classes that have been separately considered in prior work. The first, *covariate shift*, assumes

(CS) $\eta_P = \eta_Q$.

In particular, under **(CS)**, the source and target posteriors are the same, while P_X and Q_X are allowed to differ. Covariate shift has been studied extensively and related work is discussed in Section 3. It arises, for instance, when there is a *sample selection bias* that causes source and target feature vectors to follow different distributions (Heckman, 1979). For example, in developing a classifier for a certain disease, source subjects may have volunteered for a clinical study, while testing subjects are drawn from the general public. These two populations are different and hence $P_X \neq Q_X$, but presumably $\eta_P = \eta_Q$.

The second type of domain adaptation, which we call *posterior drift*, assumes

(PD) $P_X = Q_X$, and there exists a strictly increasing function ϕ such that for all x, $\eta_P(x) = \phi(\eta_Q(x))$.

Posterior drift is a model for FDLN. In this work, label noise refers to a corruption of the labels of the training data, and is in addition to any uncertainty in the optimal label arising from overlap of Q_0 and Q_1. Posterior drift may be viewed as a model for so-called "annotator" noise, which models the way a human might (noisily) assign labels to unlabeled data (Urner et al., 2012). In particular, let $(X, Y, \tilde{Y})$ be jointly distributed. Let Q be the distribution of (X, Y), where X is the feature vector and Y the true label. Let P be the distribution of $(X, \tilde{Y})$, where $\tilde{Y}$ is a noisy label assigned by the annotator. Clearly $P_X = Q_X$ in this setting. Furthermore, η_Q is the true probabilistic labeller, while η_P is the probabilistic labeller associated to the annotator. **(PD)** asserts that as the probability of the true label being 1 increases, so too does the probability of the annotator's label being 1. See Section 3 for more discussion of FDLN.

Finally, it is natural to combine these two assumptions, leading to the following.

(CSPD) There exists a strictly increasing function ϕ such that for all x, $\eta_P(x) = \phi(\eta_Q(x))$.

In this model, the marginal distribution of X is allowed to shift, as in **(CS)**, while the posterior is simultaneously allowed to drift, as in **(PD)**.

1.4. Contributions

To our knowledge, this work is the first to study the **(CSPD)** class of domain adaptation problems, making it the largest class of domain adaptation problems for which immunity (and hence optimal performance) has been established. Relative to prior work on covariate shift, we are the first to establish optimal domain adaptation without requiring a high degree of similarity between P and Q (see related work below). Relative to prior work on classification with FDLN, our work is the first to establish optimal performance without overly restrictive assumptions on the label noise (again, see related work). We also introduce a new family of optimality criteria that has not previously been considered in machine learning. Finally, we introduce two algorithms for optimizing CDR in the semi-supervised setting, including the first analysis of a level set method based on kernel logistic regression.

1.5. Outline

In the next section we introduce a family of generalized Neyman-Pearson criteria for binary classification. Section 3 discusses related work. In Section 4, consistent estimators for the CDR criterion are established, and in Section 5, we synthesize the results of prior sections to explain how optimal domain adaptation is achieved under covariate shift with posterior drift. The final section concludes, and proofs are found in an appendix.

2. A Generalized Neyman-Pearson Criterion

We introduce a family of constrained criteria for classifier design, indexed by parameters $0 \leq \theta_0 < \theta_1 \leq 1$ and $0 \leq \alpha \leq 1$, and defined with respect to a distribution Q as described in Section 1.1. The Neyman-Pearson (NP) criterion corresponds to the special case $\theta_1 = 1$ and $\theta_0 = 0$. After this section, we will be particularly interested in the case $\theta_1 = 1$ and $\theta_0 = \pi_Q$ in the context of the domain adaptation problems mentioned previously.

A classifier is a function $g : \mathcal{X} \to [0, 1]$. We view classifiers as potentially randomized, where x is classified as 1 with probability $g(x)$, independent of all other random variables. The *power* $B_Q(g)$ of a classifier g is the probability that the predicted label is 1, given that the true label is one. That is,

$$B_Q(g) := \mathbb{E}_{Q_1}[g(X)] = \int g(x)q_1(x)d\mu(x).$$

The power is also referred to as 1 - Type II error, detection rate, true positive rate, sensitivity, or recall. The *size* $A_Q(g)$ of a classifier g is the probability that a predicted label is 1, given that the true label is zero. That is,

$$A_Q(g) := \mathbb{E}_{Q_0}[g(X)] = \int g(x)q_0(x)d\mu(x).$$

Size is also known as the Type I error, false alarm rate, false positive rate, or 1 - specificity.

For the *generalized Neyman-Pearson* (GNP) criterion with parameters $0 \leq \theta_0 < \theta_1 \leq 1$ and $0 < \alpha < 1$, a classifier g is optimal if it solves the following optimization problem:

$$\begin{aligned} \max_g \ & \theta_1 B_Q(g) + (1 - \theta_1)A_Q(g) \\ \text{s.t. } & \theta_0 B_Q(g) + (1 - \theta_0)A_Q(g) \leq \alpha. \end{aligned} \tag{2}$$

where the max is over all classifiers. Notice that B_Q is an accuracy measure, whereas A_Q is an error quantity. The condition $\theta_0 < \theta_1$ ensures that the relative emphasis on acruracy in the objective, and error in the constraint, lead to a meaningful criterion for classification. Indeed, the optimal classifier is obtained by thresholding $\eta_Q(x)$. Equivalently, the optimal classifier is a likelihood ratio test (LRT), since $\eta_Q(x)$ and $q_1(x)/q_0(x)$ are monotonically related according to (1).

Theorem 1 *Given $0 \leq \theta_0 < \theta_1 \leq 1$, and $0 < \alpha \leq 1$, there exist $t_{Q,\alpha} \in [0, 1]$, $q_{Q,\alpha} \in [0, 1)$, such that a solution to (2) is*

$$g_{Q,\alpha}(x) := \begin{cases} 1, & \eta_Q(x) > t_{Q,\alpha}, \\ q_{Q,\alpha}, & \eta_Q(x) = t_{Q,\alpha}, \\ 0, & \eta_Q(x) < t_{Q,\alpha}. \end{cases}$$

The proof uses an argument of Blanchard et al. (2016) to show that the GNP criterion can be viewed as a conventional NP criterion with respect to two different contaminated versions of Q. Then, the NP lemma is used to show that the optimal classifier is a LRT, and this result is transformed back to the GNP criterion.

In this paper we are primarily concerned with the special case where $\theta_1 = 1$ and $\theta_0 = \pi_Q$. The expression in the constraint becomes $D_Q(g) := Q_X(g(X) = 1)$, which we refer to as the *discovery rate* of g. In this case, we aim to solve

$$\begin{aligned} \max_g \ & B_Q(g) \\ \text{s.t. } & D_Q(g) \leq \alpha, \end{aligned}$$

which yields the most powerful classifier that predicts at most a fraction α of test instances as positive. We refer to this specific criterion as the *controlled discovery rate* (CDR) criterion. The CDR criterion is desirable in applications where positively classified examples from the target domain will be subjected to further scrutiny, and there is a limited budget to conduct follow-up investigations. For example, in information retrieval it is common that only the top $100\alpha\%$ of the test instances will be inspected by a user. In this context, the CDR criterion seeks the classifier with maximum recall that assigns a positive label to $100\alpha\%$ of the test instances. Thus, CDR is similar in spirit to criteria that aim to measure "accuracy at the top" (Boyd et al., 2012). Previous work relating to the CDR criterion is discussed in the next section.

We show in this work that the CDR criterion can be optimally learned under **(CSPD)**. The intuition behind this fact, and the *primary insight* of this paper, is as follows. Consider the infinite sample setting where P and Q_X are known. Since P is known, we know η_P, which is monotonically equivalent to η_Q under **(CSPD)**. By this monotone equivalence, the optimal classifier (for the target domain) has the form $g(x) = \mathbf{1}_{\{\eta_P(x) \geq t\}}$ for some t. This threshold t can be set to ensure that $D_Q(g) = \alpha$ (which must be satisfied by the optimal classifier) because $D_Q(g)$ *depends on* Q *only through* Q_X. In the finite sample case, our algorithms naturally rely on estimates of η_P and D_Q. The details of this argument are worked out in the sequel.

3. Related Work

Target Shift: A kind of dual of covariate shift is *target shift*, where $P_0 = Q_0$ and $P_1 = Q_1$, but $\pi_P \neq \pi_Q$. This form of domain adaptation arises frequently in applications where training and testing data are gathered according to different sampling plans. For example, training data gathered prospectively may have a user-determined π_P, while testing data analyzed retrospectively may have a π_Q that is beyond the user's control.

Target shift is a class of problems that satisfy neither **(CS)** nor **(PD)**, but do satisfy **(CSPD)**. To see this, just note that $P_X = \pi_P P_1 + (1-\pi_P)P_0 \neq \pi_Q P_1 + (1-\pi_Q)P_0 = Q_X$, so **(PD)** is violated, and

$$\eta_P(x) = \frac{1}{1 + \frac{1-\pi_P}{\pi_P}\frac{p_0(x)}{p_1(x)}} \neq \frac{1}{1 + \frac{1-\pi_Q}{\pi_Q}\frac{p_0(x)}{p_1(x)}} = \eta_Q(x),$$

so **(CS)** is violated. Yet clearly η_P and η_Q are monotonically equivalent, so **(CSPD)** holds.

Previous work on target shift has focused on estimating π_Q in the semi-supervised setting (Hall, 1981; Titterington, 1983; Latinne et al., 2001; Du Plessis and Sugiyama, 2012; Sanderson and Scott, 2014). Since target shift is a special case of **(CSPD)**, our methods optimize the CDR criterion for such problems, notably *without* needing to estimate π_Q. In fact, all GNP criteria are immune to target shift.

Immunity: An optimality criterion is immune to a class of domain adaptation problems if the optimal classifier is the same for both the source distribution and the target distribution (see Appendix A for a more formal definition). Practically speaking, immunity implies that the learner can ignore the possibility of domain adaptation (i.e., assume $P = Q$) and still be optimal when $P \neq Q$. As an example, consider the probability of error as a performance measure (i.e., the risk with 0-1 loss). It is well known that the probability of error is immune to symmetric, feature-independent label noise (Angluin and Laird, 1988; Kearns, 1993; Jabbari, 2010). To see this, suppose Q is the "clean" distribution on (X, Y), and P is the contaminated distribution on $(X, \tilde{Y})$, such that a realization of $(X, \tilde{Y})$ is obtained by drawing (X, Y) from Q, and replacing Y with $1 - Y$ with

probability $\nu < \frac{1}{2}$, independent of X. It follows that $\eta_P(x) = (1-\nu)\eta_Q(x) + \nu(1-\eta_Q(x))$. This implies $\eta_P(x) - \frac{1}{2} = (1-2\nu)(\eta_Q(x) - \frac{1}{2})$, and therefore the optimal classifiers for P and Q coincide. Thus, training a classifier to optimize probability of error on noisy training data leads to an optimal classifier with respect to Q.

Immunity has been established for other types of label noise. The probability of error is immune to symmetric, feature-dependent label noise, while the AUC is immune to a type of feature-dependent annotator noise that implies **(PD)** (Menon et al., 2018). The balanced error rate (BER) is immune to asymmetric label-dependent (but feature-independent) label noise (Menon et al., 2015). Menon et al. (2015) also argue that BER is the only performance measure that is immune to label-dependent label noise. The class of performance measures they study does not include the CDR criterion, so there is no contradiction with our results which apply to label-dependent label noise (see below).

Other instances of the GNP family also possess immunity for certain domain adaptation problems. For example, consider the target shift problem described above. Any GNP criterion is trivially immune to target shift (when trained only on labeled training data from the source distribution) because it does not depend on the prior class probability in the first place. The same is obviously true for other criteria that don't involve the class priors, such as the balanced error rate or the min-max criterion. The Neyman-Pearson criterion has further been shown to be immune to classification with one-sided, label-dependent label noise, also known as learning with positive and unlabeled examples (Blanchard et al., 2010). In Appendix A we argue that any GNP criterion with $\theta_0 = 0$ is immune to one-sided, feature-dependent label noise. The immunity of NP for target shift has been described by Xia et al. (2018).

In this work we show that, in the semi-supervised setting, the CDR criterion is immune to **(CSPD)**. To our knowledge, this is the most general class of problems for which immunity has been established for some binary classification optimality criterion. For further discussion of immunity, see Appendix A.

Covariate Shift and General Domain Adaptation: Previous work on covariate shift (Shimodaira, 2000) has focused on performance measures that can be expressed as risks, that is, as the expectation of a loss function with respect to P or Q. Because of this, many papers have focused on the problem of estimating the ratio $q_X(x)/p_X(x)$, where q_X and p_X are the densities of Q_X and P_X, respectively (Zadrozny, 2004; Huang et al., 2007; Cortes et al., 2008; Sugiyama et al., 2008; Bickel et al., 2009; Kanamori et al., 2009). Unfortunately, this introduces an intermediate (and potentially quite challenging) estimation problem into the learning pipeline. In contrast, learning with respect to the CDR criterion avoids estimation of the density ratio.

Several previous works have theoretically studied, under covariate shift as well as more general domain adaptation settings, when a good classifier on the target domain can be learned. For example, several papers have shown that the target risk can be bounded in terms of the source risk and some notion of "discrepancy" between P and Q (and possibly other terms) (Ben-David et al., 2007, 2010; Blitzer et al., 2008; Mansour et al., 2009; Cortes et al., 2015; Germain et al., 2016), which has led to the conclusion that in order "for generalization to be possible . . . Q and P must not be too dissimilar" (Mansour et al., 2009). Ben-David and Urner (2012) argue that covariate shift alone is insufficient to ensure good performance on the target domain. In particular, they argue that under covariate shift, good performance on the target domain cannot be guaranteed even if the supports are equal and densities q_X and p_X are mutually bounded.

In the present work, we show that optimal domain adaptation is possible assuming that **(CSPD)** holds, that the support of P_X contains the support of Q_X, and two relatively benign nonparametric conditions. In particular, optimal domain adaptation is possible even though P_X and Q_X (and hence P and Q) might be vastly different. Our results are not incompatible with previous results because the settings are somewhat different. First, as mentioned previously, we consider a different optimality criterion. Second, our focus is statistical consistency, whereas previous work often considers a fixed hypothesis space. Third, our analysis concerns the error of a classifier *relative* to the best possible classifier, whereas some previous work has addressed making the risk small in an *absolute* sense.

Classification with Feature-Dependent Label Noise: Classification with label noise is a form of domain adaptation, although it has not always been described as such. In this setting, $(X, Y, \tilde{Y})$ are jointly distributed. Q is the distribution of (X, Y), where Y is the true label of X, and P is the distribution of $(X, \tilde{Y})$, where $\tilde{Y}$ is a corrupted version of Y. We reiterate that in this discussion, label noise is in addition to any uncertainty in the optimal label arising from overlap of the supports of Q_0 and Q_1.

In the case of label-dependent label noise (LDLN), the probability that a training label is flipped depends only on the true label. The label-dependent case is fairly well understood (Blanchard et al., 2016; Natarajan et al., 2018; van Rooyen and Williamson, 2018) in the two-class setting. In essence, the difference between the source and target domains can be reduced to two parameters, $\rho_i := \Pr(\tilde{Y} \neq i \mid Y = i)$, $i \in \{0, 1\}$, the label flip probabilities for each class. Given knowledge of these proportions (which can be estimated), it is not difficult to modify a learning algorithm to successfully adapt to the target domain. We also note that LDLN is a special case of **(PD)** provided $\rho_0 + \rho_1 < 1$, see Appendix A.

A more challenging setting is feature-dependent label noise (FDLN), where the distribution of the noisy label can *also* depend on the feature vector. In this case, the label noise is characterized by functions $\rho_i(x) = \Pr(\tilde{Y} \neq i \,|\, Y = i, X = x)$, $i \in \{0, 1\}$, which give the probability that a training label is flipped, depending on the true class label and the feature vector x. These two functions are potentially quite complex, and prior work has made strong assumptions on these functions or the target distribution Q. Thus, Bootkrajang (2016) employs a parametric model for $\rho_0(x)$ and $\rho_1(x)$, while Ghosh et al. (2015) provide a class of nonconvex losses that are robust to FDLN when the Bayes Risk for Q is *zero*.

Menon et al. (2018) established immunity for the probability of error criterion under the condition of symmetric FDLN, that is, $\rho_0(x) = \rho_1(x)$ for all x, which is a strong assumption in practice. Cannings et al. (2018) extend this result by establishing immunity when $\rho_0(x)$ and $\rho_1(x)$ are approximately symmetric in a certain sense, approaching perfect symmetry near the decision boundary.

Menon et al. (2018) make two other contributions to the study of FDLN problems. They introduce a type of annotator noise called boundary-consistent noise (BCN) wherein $\rho_0(x)$ and $\rho_1(x)$ obey certain monotonicity properties, and show that this noise model implies **(PD)**. Under BCN, they show that the area under the ROC curve (AUC) is immune to FDLN. It should be noted, however, that AUC is a criterion for ranking and not for binary classification. They also study a type of generalized linear model under BCN and show that the Isotron algorithm is consistent in this setting.

Cheng et al. (2017) assume that $\rho_0(x)$ and $\rho_1(x)$ are bounded by a number < 0.5. This seems an unlikely model for annotator noise, since one would expect $\rho_1(x) \to 1$ as $\eta_Q(x) \to 0$, and

$\rho_0(x) \to 1$ as $\eta_Q(x) \to 1$. Leveraging ideas from Northcutt et al. (2017), they describe a procedure to find a subset of examples where the label is known to be correct. Knowledge of the bounds on $\rho_0(x)$ and $\rho_1(x)$ are required as input to their algorithm. Their theory analyzes a method that requires knowledge of $\eta_Q(x)$, and a more practical algorithm requires access to, or an estimate of, the same density ratio that arises in covariate shift.

Our contribution to the study of FDLN is as follows. We are the first to establish both *consistency* and *immunity* of a learning algorithm, with respect to some optimality criterion, under a realistic nonparametric model of annotator noise (namely, **(PD)**) and under general nonparametric assumptions on the data distribution. Furthermore, our approach avoids the need to estimate $\rho_0(x)$ or $\rho_1(x)$, or the density ratio mentioned previously.

Other Classes of Domain Adaptation: We mention two other types of domain adaptation. Zhang et al. (2013) study an assumption that is dual to **(CSPD)** in a sense. Whereas **(CSPD)** allows the marginal of X to shift arbitrarily, and the conditional of $Y|X$ to shift in a monotone fashion, they allow the marginal of Y to shift arbitrarily, and the conditional of $X|Y$ to undergo a location-scale shift. Tasche (2017) introduces problems with an "invariant density ratio," where the likelihood ratios of P and Q are equal. This problem is a special case of **(CSPD)** and a generalization of target shift.

Optimality Criteria for Binary Classification: There has been interest in recent years in cataloging different performance measures and optimality criteria for binary classification (Koyejo et al., 2014; Narasimhan et al., 2014; Kotlowski and Dembczyski, 2016; Dembczyński et al., 2017), and establishing consistent learning rules for them. The GNP criteria are evidently a new family of optimality criteria, thus expanding this literature. Tasche (2018) studies a different family of constrained optimization problems that also includes the CDR criterion as a special case, providing an alternate proof of Theorem 1 in the case of CDR. The fact that the CDR criterion is optimized by thresholding η_Q was noted by Clémencon and Vayatis (2007), see Remark 2.

NP Classification: We anticipate that several existing algorithms for Neyman-Pearson classification (Scott and Nowak, 2005; Tong et al., 2016) and similar constrained criteria extend naturally to CDR. To illustrate this point, later we present an adaptation of an algorithm of Lei (2014). In the reverse direction, our algorithm and analysis based on kernel logistic regression should naturally yield algorithms and analysis for Neyman-Pearson classification as well as other classification and level-set criteria.

4. Estimators for the CDR Criterion

In this section we address consistent estimators for the optimal CDR classifier. Our goal is to estimate the set

$$G_{Q,\alpha} := \{x : \eta_Q(x) \geq t_{Q,\alpha}\} \tag{3}$$

where $t_{Q,\alpha}$ is the threshold associated to the CDR criterion at level α. In other words, $Q_X(G_{Q,\alpha}) = \alpha$. Note that this assumes the optimal classifier $g_{Q,\alpha}$ is deterministic, which is formalized in our distributional assumptions below. Also, we view deterministic classifiers and subsets of $\mathcal{X}$ interchangeably by viewing the classifier as an indicator on the subset.

For greater generality that will be needed in the context of domain adaptation, we actually consider the problem of estimating

$$G_{P,Q,\alpha} := \{x : \eta_P(x) \geq t_{P,Q,\alpha}\} \tag{4}$$

where $t_{P,Q,\alpha}$ is such that $Q_X(G_{P,Q,\alpha}) = \alpha$. Note that taking $P = Q$ reduces to (4) to (3).

To preview Section 5, in the context of domain adaptation, $G_{P,Q,\alpha}$ can be estimated since we have data drawn from P and Q_X. Furthermore, under **(CSPD)**, it is not hard to see that $G_{P,Q,\alpha} = G_{Q,\alpha}$, meaning it is possible to consistently estimate the optimal CDR classifier on the target domain.

After formalizing our distributional assumptions and the estimation problem, we present two estimators with associated convergence results. The first assumes access to a sup-norm consistent estimator of η_P, while the second uses kernel logistic regression to estimate η_P. Throughout this section we assume $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$.

4.1. Distributional Assumptions

In addition to **(CSPD)**, our analysis makes the following nonparametric assumptions on P and Q. These assumptions allow P and Q to be quite different from one another according to essentially any commonly used notion of distance or divergence between two distributions.

Define $F_{P,Q}(t) := Q_X(\{x : \eta_P(x) \leq t\})$, the cumulative distribution function of the random variable $\eta_P(X)$ when $X \sim Q_X$. We adopt the following two assumptions:

(A) There exists $t_{P,Q,\alpha} \in (0, 1]$ such that

$$Q_X(\{x : \eta_P(x) \geq t_{P,Q,\alpha}\}) = \alpha.$$

(B) There exist positive constants δ_0, b_1, b_2 and κ such that for all $\delta \in [-\delta_0, \delta_0]$,

$$b_1|\delta|^\kappa \leq |F_{P,Q}(t_{P,Q,\alpha} + \delta) - F_{P,Q}(t_{P,Q,\alpha})| \leq b_2|\delta|^\kappa.$$

(A) ensures that randomized classifiers are not needed. **(B)** states that $F_{P,Q}$ has local growth (in a neighborhood of $t_{P,Q,\alpha}$) characterized by the exponent κ, which characterizes the difficulty of the estimation problem. The lower bound in **(B)** implies that $t_{P,Q,\alpha}$ is unique, while the upper bound implies that $F_{P,Q}$ is continuous at $t_{P,Q,\alpha}$. Under **(A)** and **(B)**, $G_{P,Q,\alpha}$ is well-defined, i.e., the threshold $t_{P,Q,\alpha}$, which must satisfy $Q_X(G_{P,Q,\alpha}) = \alpha$, exists and is unique.

The following assumption is widely adopted in the study of covariate shift.

(C) The support of Q_X is contained in the support of P_X.

A strengthened form of this assumption is employed in the analysis of our second algorithm (Yu and Szepesvari, 2012).

(C') There exists $c_0 > 0$ such that $Q_X \leq c_0 P_X$. Equivalently, Q_X is absolutely continuous with respect to P_X, and $\partial Q_X / \partial P_X$ is essentially bounded by c_0.

4.2. The Estimation Problem

We focus on estimating $G_{P,Q,\alpha}$ given the following data:

$$(X_1, Y_1) \ldots, (X_m, Y_m) \overset{iid}{\sim} P$$
$$X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} Q_X.$$

The two samples are assumed to be independent of each other. Let $\widehat{G}_{P,Q,\alpha}$ be an estimate of $G_{P,Q,\alpha}$. We further focus on the performance measure

$$Q_X(\widehat{G}_{P,Q,\alpha} \Delta G_{P,Q,\alpha}),$$

where $G \Delta G' := (G \backslash G') \cup (G' \backslash G)$ is the *symmetric difference* of G and G'.

According to the following result, convergence with respect to the above measure implies convergence of the objective and constraint functions for GNP criteria.

Proposition 2 *Let g and g' be two deterministic classifiers, and let $G = \{x : g(x) = 1\}$ and $G' = \{x : g'(x) = 1\}$ be the associated sets. For any $\epsilon \in [0,1]$ and any Q,*

$$\left|\epsilon B_Q(g) + (1-\epsilon) A_Q(g) - [\epsilon B_Q(g') + (1-\epsilon) A_Q(g')]\right| \leq \left(\frac{\epsilon}{\pi_Q} + \frac{1-\epsilon}{1-\pi_Q}\right) Q_X(G \Delta G').$$

In what follows, let P^m denote the product measure governing $(X_1, Y_1) \ldots, (X_m, Y_m)$, and Q_X^n denote the product measure governing $X_{m+1}, \ldots, X_{m+n}$. We use $\Pr$ to denote the product measure $P^m \times Q_X^n$ on $(\mathcal{X} \times \{0,1\})^m \times \mathcal{X}^n$, which governs the combined draw of the two samples. The goal is to show $\Pr(Q_X(\widehat{G}_{P,Q,\alpha} \Delta G_{P,Q,\alpha})) \to 0$ in probability as $m, n \to \infty$.

4.3. A result based on sup-norm consistent estimation of the posterior

The CDR criterion is sufficiently similar to NP classification and related problems that we can easily modify existing algorithms and theory to our setting. To illustrate this, we begin by establishing a consistent CDR estimator based on a sup-norm consistent estimate of η_P. The results in this subsection translate ideas from Lei (2014), where a different generalization of the Neyman-Pearson criterion was considered. Let $\widehat{\eta}_P$ denote an estimate, based on $(X_1, Y_1), \ldots, (X_m, Y_m)$, of the posterior η_P associated to the joint distribution P. Let δ_m, τ_m be two sequences of positive reals numbers tending to 0.

Definition 3 *An estimator $\widehat{\eta}_P$ is (δ_m, τ_m)-accurate if $P^m(\|\widehat{\eta}_P - \eta_P\|_\infty \geq \delta_m) \leq \tau_m$ as $m \to \infty$.*

Specific examples of (δ_m, τ_m)-accurate estimators are provided by Lei (2014), with explicit rates (tending to 0) for δ_m and τ_m. In particular, Audibert and Tsybakov (2007) and van de Geer (2008) give explicit rates for local polynomial regression and ℓ_1-penalized logistic regression, respectively. These estimators in turn yield explicit rates of convergence in our setting. We refer the reader to Lei (2014) for details.

Remark: (δ_m, τ_m)-accurate estimators of η_P may require additional distributional assumptions on P beyond what we have assumed so far. This is the case for the two examples mentioned above. This does not change our conclusion that P and Q can still be substantially different. Also, our goal in this subsection is to demonstrate an estimator with a rate of convergence, but other consistent estimators that do not require additional assumptions could also be adapted to CDR estimation.

Define $\widehat{G}_{P,Q,\alpha} = \{x : \widehat{\eta}_P(x) \geq \widehat{t}_{P,Q,\alpha}\}$, where $\widehat{t}_{P,Q,\alpha}$ is the $\lfloor n(1-\alpha) \rfloor$th smallest value among $\{\widehat{\eta}_P(X_{m+1}), \ldots, \widehat{\eta}_P(X_{m+n})\}$.

Theorem 4 *Let P and Q be joint distributions, and let $(X_1, Y_1) \ldots, (X_m, Y_m) \overset{iid}{\sim} P$ and $X_{m+1}, \ldots, X_{m+n} \overset{iid}{\sim} Q_X$. Assume **(A)**, **(B)**, and **(C)** hold, and that $\widehat{\eta}_P$ is a (δ_m, τ_m)-accurate*

estimator of η_P. For each $r > 0$, there exists a positive constant c such that for m and n large enough, with probability at least $1 - \tau_m - n^{-r}$ with respect to the draw of the training data,

$$Q_X(\widehat{G}_{P,Q,\alpha} \Delta G_{P,Q,\alpha^*}) \leq c \left\{ \delta_m^\kappa + \left(\frac{\log n}{n} \right)^{1/2} \right\}.$$

When this result is instantiated with the (δ_m, τ_m)-accurate estimator of Audibert and Tsybakov (2007), and $\kappa = 1$, the rate above matches or is similar to known rates for related set estimation and classification problems. See Lei (2014) for additional discussion.

4.4. A result for kernel logistic regression

In this section, we examine an estimator based on kernel logistic regression (KLR), which is perhaps a more practical estimator for η_P than the methods mentioned in the previous subsection. Although KLR is not known to be sup-norm consistent, we are able to establish an asymptotic convergence result for our estimator based on theory developed by Steinwart (2003). We believe this is the first such result for a set estimator based on KLR.

Let $\widehat{\eta}_P$ be the estimate of η_P resulting from KLR with symmetric, positive definite kernel k and regularization parameter λ, based on $(X_i, Y_i), i = 1, \ldots, m$. That is,

$$\widehat{\eta}_P(x) = \frac{1}{1 + \exp(-\widehat{f}_P(x))}$$

where $\widehat{f}_P$ solves

$$\min_{f \in \mathcal{H}} \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 + \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-(2Y_i - 1) f(X_i))).$$

Here $\mathcal{H}$ is a reproducing kernel Hilbert space of functions over $\mathbb{R}^d$ associated to kernel k. Later, we will assume that k is a *universal kernel*, which means that $\mathcal{H}$ has nice approximation properties (Steinwart and Christmann, 2008).

For a set G define $\widehat{Q}_X(G) = \frac{1}{n} \sum_{i=m+1}^{m+n} \mathbf{1}_{\{X_i \in G\}}$, the empirical measure with respect to the second training sample. Let α be the user-specified constant defining the CDR criterion. Now define the empirical estimate of $t_{P,Q,\alpha}$, with tuning parameters β and γ, as

$$\widehat{t}_{P,Q,\alpha} = \inf\{t \,|\, \widehat{Q}_X(\{x : \widehat{\eta}_P(x) \geq t + \beta\}) \leq \alpha + \gamma + \epsilon_n\} \tag{5}$$

where $\epsilon_n = 4(\log(n+1)/n)^{1/2}$, and define the estimator of $G_{P,Q,\alpha}$ to be

$$\widehat{G}_{P,Q,\alpha} = \{x : \widehat{\eta}_P(x) \geq \widehat{t}_{P,Q,\alpha}\}. \tag{6}$$

Theorem 5 *Assume **(A)**, **(B)**, and **(C')** hold. Let k be a universal kernel and let $\lambda = \lambda_m$ such that $\lambda \to 0$ and $m\lambda^2 \to \infty$. For all $\epsilon > 0$, there exist β and γ such that*

$$Q_X(\widehat{G}_{P,Q,\alpha} \Delta G_{P,Q,\alpha}) \leq \epsilon$$

in probability as $m, n \to \infty$.

The proof hinges on a result of Steinwart (2003), who effectively shows that $\widehat{\eta}_P$ is uniformly close to η_P, to arbitrary accuracy, on an event with probability tending to 1 as $m \to \infty$. We then use **(B)** to translate accuracy of $\widehat{\eta}_P$ to accuracy of the associated set estimate. The proof gives constructive choices for β and γ depending on ϵ and the constants appearing in **(B)**. Concrete rates of convergence are not available because the same is true of the result of Steinwart (2003) that we leverage.

This result does not show consistency of a specific algorithm, since β and γ depend on ϵ. Nonetheless it demonstrates the theoretical capacity of a KLR-based estimator to deliver arbitrarily accurate estimates of $G_{P,Q,\alpha}$. In practice, of course, the threshold on $\widehat{\eta}_P$ would be determined in a data-driven fashion (Tong et al., 2018).

5. Domain Adaptation for the CDR Criterion

Recall that the goal of domain adaptation with the CDR criterion is to recover

$$G_{Q,\alpha} = \{x : \eta_Q(x) \geq t_{Q,\alpha}\}$$

given realizations of P and of Q_X. In the previous section, we saw that it is possible to consistently estimate

$$G_{P,Q,\alpha} = \{x : \eta_P(x) \geq t_{P,Q,\alpha}\}$$

under assumptions **(A)**, **(B)**, and **(C)** or **(C')**.

The key insight of this paper is that under **(CSPD)**, $G_{Q,\alpha} = G_{P,Q,\alpha}$, and therefore $G_{Q,\alpha}$ can be consistently estimated. To see that $G_{Q,\alpha} = G_{P,Q,\alpha}$ under **(CSPD)**, simply recall the definition of **(CSPD)** which assumes the existence of a strictly increasing function $\phi : [0,1] \to [0,1]$ such that for all x, $\eta_P(x) = \phi(\eta_Q(x))$. Now, $G_{Q,\alpha} = G_{P,Q,\alpha}$ follows by taking $t_{Q,\alpha} = \phi^{-1}(t_{P,Q,\alpha})$. Under **(A)** and **(B)**, $t_{P,Q,\alpha}$ exists and is unique, and therefore the same is true of $t_{Q,\alpha}$.

In light of the above, we have the following:

Corollary 6 *Assume **(CSPD)**, **(A)**, **(B)**, and **(C)** (respectively, **(C')**) hold. Then the estimator of Section 4.3 (resp., Section 4.4) satisfies the conclusion of Theorem 4 (resp., Theorem 5), where now the set being estimated is $G_{Q,\alpha}$.*

6. Conclusions

We have introduced a family of generalized Neyman-Pearson optimality criteria, and shown that a member of this family, the controlled discovery rate criterion, is immune to domain adaptation under the model of covariate shift with posterior drift. Compared with prior work on domain adaptation, we do not require that the source and target distributions be close in some sense in order to obtain optimal performance on the source domain. With respect to prior work on covariate shift, our approach does not require estimating a density ratio, and in fact allows the density ratio to be unbounded under condition **(C)**. Comparing to the literature on feature-dependent label noise, ours is the first work to establish consistency/immunity under a general and flexible model for annotator noise, without requiring knowledge of the specific annotator noise model. These results are enabled by consideration of an optimality criterion different from the usual ones based on expected loss.

Acknowledgments

The author was supported by NSF Grants No. 1422157 and 1838179.

References

D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35:608–633, 2007.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 139–153, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. 2007.

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

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, 10:2137–2155, 2009.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10:2780–2824, 2016.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 129–136. 2008.

Jakramate Bootkrajang. A generalised label noise model for classification in the presence of annotation errors. *Neurocomputing*, 192:61–71, 2016.

Stephen Boyd, Corinna Cortes, Mehryar Mohri, and Ana Radovanovic. Accuracy at the top. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 953–961. 2012.

Timothy I. Cannings, Yingying Fan, and Richard J. Samworth. Classification with imperfect training labels. Technical Report arXiv:1805.11505, 2018.

Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance- and label-dependent label noise. Technical Report arxiv:1709.03768v1, 2017.

S. Clémencon and N. Vayatis. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Algorithmic Learning Theory*, pages 38–53, 2008.

Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 169–178, 2015.

Krzysztof Dembczyński, Wojciech Kotłowski, Oluwasanmi Koyejo, and Nagarajan Natarajan. Consistency analysis for binary classification revisited. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 961–969, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

M. C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In J. Langford and J. Pineau, editors, *Proc. 29th Int. Conf. on Machine Learning*, pages 823–830, 2012.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 859–868, 2016.

Aritra Ghosh, Naresh Manwani, and P.S. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93 – 107, 2015.

P. Hall. On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 43(2):147–156, 1981.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Scholkopf. Correcting sample selection bias by unlabeled data. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 601–608, 2007.

S. Jabbari. PAC-learning with label noise. Master's thesis, University of Alberta, December 2010.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.*, 10:1391–1445, 2009.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, pages 392–401, 1993.

Wojciech Kotlowski and Krzysztof Dembczyski. Surrogate regret bounds for generalized classification performance metrics. In Geoffrey Holmes and Tie-Yan Liu, editors, *Asian Conference on Machine Learning*, volume 45 of *Proceedings of Machine Learning Research*, pages 301–316, 2016.

O. Koyejo, N. Natarajan, P. Ravikumar, and I. Dhillon. Consistent binary classification with generalized performance metrics. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2744–2752, 2014.

P. Latinne, M. Saerens, and C. Decaestecker. Adjusting the outputs of a classier to new a priori probabilities may signicantly improve classication accuracy: Evidence from a multi-class problem in remote sensing. In C. Sammut and A. H. Hoffmann, editors, *Proc. 18th Int. Conf. on Machine Learning*, pages 298–305, 2001.

Jing Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18:1269–83, 1990.

A. Menon, B. Van Rooyen, C. S. Ong, and R. Williamson. Learning from corrupted binary labels via class-probability estimation. In F. Bach and D. Blei, editors, *Proc. 32th Int. Conf. Machine Learning (ICML)*, Lille, France, 2015.

Aditya Krishna Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107:1561–1595, 2018.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1493–1501. 2014.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018. URL http://jmlr.org/papers/v18/15-226.html.

Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

T. Sanderson and C. Scott. Class proportion estimation with application to multiclass anomaly rejection. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Trans. Info. Theory*, 51(11):3806–3819, 2005.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244, 2000.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

Ingo Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4: 1071–1105, 2003.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bŭnau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Dirk Tasche. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18:1–32, 2017.

Dirk Tasche. A plug-in approach to maximizing precision at the top and recall at the top. Technical Report arxiv:1804.03077v1, 2018.

D. M. Titterington. Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 45(1):37–46, 1983.

Xin Tong, Yang Feng, and Anqi Zhao. A survey on Neyman-Pearson classification and suggestions for future research. *WIREs Comput. Stat.*, 8(2), 2016.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman-Pearson classification algorithms and NP receiver operating characteristics. *Science Advances*, 4(2), 2018.

Ruth Urner, Shai Ben David, and Ohad Shamir. Learning from weak teachers. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1252–1260, 2012.

Sara van de Geer. High-dimensional generalized linear models and the LASSO. *The Annals of Statistics*, 36:614–645, 2008.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

L. Xia, R. Zhao, Y. Wu, and X. Tong. Intentional control of Type I error over unconscious data distortion: a Neyman-Pearson approach to text classification. Technical Report arXiv:1802.02558, 2018.

Yao-Liang Yu and Csaba Szepesvari. Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th International Conference on Machine Learning*, pages 607–614, 2012.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

K Zhang, B Scholkopf, Krikamol Muandet, and Z Wang. Domain adaptation under target and conditional shift. In *30th International Conference on Machine Learning, ICML 2013*, pages 1856–1864, 01 2013.

Appendix A. Immunity

This appendix provides supplemental details and observations pertaining to immunity.

The immunity of an optimality criterion with respect to a class of domain adaptation problems is formally defined as follows. We distinguish between the *inductive* setting, where the learner has access only to labeled data from P, and the *semi-supervised* setting, where the learner has an additional unlabeled sample from Q_X. Let $\mathcal{P}$ be some class of distributions of interest, e.g., all distributions on $\mathcal{X}$. A class of domain adaptation problems is a subset $\mathcal{D} \subseteq \mathcal{P} \times \mathcal{P}$ where, for $(P, Q) \in \mathcal{D}$, P is the source domain and Q the target. At times we express a distribution Q as the pair (η_Q, Q_X). The classifier (or set of classifiers) optimizing an optimality criterion for distribution Q is denoted $\mathrm{OPT}(Q)$ in the inductive case, and $\mathrm{OPT}(\eta_Q, Q_X)$ in the semi-supervised case. We say that an optimality criterion is *immune* to $\mathcal{D}$ if, for all $(P, Q) \in \mathcal{D}$, $\mathrm{OPT}(Q) = \mathrm{OPT}(P)$ in the inductive setting, or $\mathrm{OPT}(\eta_Q, Q_X) = \mathrm{OPT}(\eta_P, Q_X)$ in the semi-supervised setting. Except for our result on the CDR criterion, all of the immunity results mentioned in Section 3 are for the inductive setting.

To see that LDLN is a special case of **(PD)** provided $\rho_0 + \rho_1 < 1$, observe

$$\begin{aligned}
\eta_P(x) &= \Pr(\tilde{Y} = 1 \mid X = x) \\
&= \Pr(\tilde{Y} = 1 \mid Y = 1, X = x) \Pr(Y = 1 \mid X = x) \\
&\quad + \Pr(\tilde{Y} = 1 \mid Y = 0, X = x) \Pr(Y = 0 \mid X = x) \\
&= \Pr(\tilde{Y} = 1 \mid Y = 1) \Pr(Y = 1 \mid X = x) + \Pr(\tilde{Y} = 1 \mid Y = 0) \Pr(Y = 0 \mid X = x) \\
&= (1 - \rho_1)\eta_Q(x) + \rho_0(1 - \eta_Q(x)) \\
&= (1 - \rho_0 - \rho_1)\eta_Q(x) + \rho_0.
\end{aligned}$$

We also note that **(CSPD)** is preserved by composition of domain adaptations, because the composition of strictly increasing functions is strictly increasing. For example, consider distributions P, Q, and R. Let $\mathcal{D}$ be the set of (P, R) such that there exists Q for which Q is related to R by target shift, and P is generated from Q by LDLN (with $\rho_0 + \rho_1 < 1$). Then there exist strictly increasing ϕ_1 and ϕ_2 such that, for all x, $\eta_P(x) = \phi_1(\eta_Q(x))$ and $\eta_Q(x) = \phi_2(\eta_R(x))$. Thus $\eta_P(x) = \phi(\eta_R(x))$ where $\phi = \phi_1 \circ \phi_2$, which is strictly increasing, and therefore $\mathcal{D}$ satisfies **(CSPD)**.

The focus of the paper has been immunity of the CDR criterion to **(CSPD)** in the semi-supervised setting. We note that the CDR criterion is also immune to **(PD)** in the *inductive* setting. Since $P_X = Q_X$ under **(PD)**, Q_X is already estimable through the data drawn from P, and an unlabeled sample from Q_X is not needed. Indeed, all of the results for **(CSPD)** in the semi-supervised setting could also be stated for **(PD)** in the inductive setting.

Finally, we remark that a subset of GNP criteria (namely, when $\theta_0 = 0$) are immune to a subclass of **(PD)** corresponding to *one-sided* feature-dependent label noise. In particular, define the domain adaptation class

(PD') $P_X = Q_X$, $\rho_1 \equiv 0$ and there exists a strictly increasing function ψ such that $\rho_0(x) = \psi(\eta_Q(x))$ for all x.

Under **(PD')**, true labels of 1 are never corrupted to become 0. Furthermore we have the following.

Lemma 1 ***(PD')** implies **(PD)***

Proof We need to show that $\eta_P(x)$ is a strictly increasing function of $\eta_Q(x)$. For a posterior $\eta(x)$, define $\bar{\eta}(x) = 1 - \eta(x)$. Arguing as we did previously, under **(PD')**,

$$\begin{aligned}
\bar{\eta}_P(x) &= \Pr(\tilde{Y} = 0 \mid X = x) \\
&= \Pr(\tilde{Y} = 0 \mid Y = 1, X = x) \Pr(Y = 1 \mid X = x) \\
&\quad + \Pr(\tilde{Y} = 0 \mid Y = 0, X = x) \Pr(Y = 0 \mid X = x) \\
&= \rho_1(x)(1 - \bar{\eta}_Q(x)) + (1 - \rho_0(x))\bar{\eta}_Q(x) \\
&= (1 - \rho_0(x) - \rho_1(x))\bar{\eta}_Q(x) + \rho_1(x) \\
&= (1 - \rho_0(x))\bar{\eta}_Q(x) + \rho_1(x) \\
&= (1 - \rho_0(x))\bar{\eta}_Q(x).
\end{aligned}$$

This implies that

$$\eta_P(x) = 1 - (1 - \rho_0(x))(1 - \eta_Q(x)).$$

The result now follows. ■

Then *all* GNP criteria with $\theta_0 = 0$ are immune to **(PD')** in the inductive setting. This follows by similar reasoning as for CDR. First, with $\theta_0 = 0$, the constraint in the GNP criterion depends only on Q_0, and $Q_0 = P_0$ because $\rho_1(x) \equiv 0$. Second, η_P and η_Q are monotonically equivalent. Therefore, the level set of η_P with P_0-measure α is also the level set of η_Q with Q_0-measure α.

Appendix B. Proofs

This appendix contains the proofs.

B.1. Proof of Theorem 1

Denote

$$\tilde{q}_1(x) := \theta_1 q_1(x) + (1 - \theta_1) q_0(x), \tag{7}$$

$$\tilde{q}_0(x) := \theta_0 q_1(x) + (1 - \theta_0) q_0(x). \tag{8}$$

Note that $\tilde{q}_1(x)$ and $\tilde{q}_0(x)$ are densities for the distributions $\tilde{Q}_1 := \theta_1 Q_1 + (1 - \theta_1) Q_0$ and $\tilde{Q}_0 := \theta_0 Q_1 + (1 - \theta_0) Q_0$, respectively. Viewing these as the alternative and null distributions in a hypothesis testing problem, the power and size of a classifier g are

$$\begin{aligned}
\tilde{B}_Q(g) &:= \theta_1 B_Q(g) + (1 - \theta_1) A_Q(g) \\
\tilde{A}_Q(g) &:= \theta_0 B_Q(g) + (1 - \theta_0) A_Q(g).
\end{aligned}$$

Thus, the optimization problem in (2) is equivalent to maximizing the power $\tilde{B}_Q(g)$, subject to the constraint that the size $\tilde{A}_Q(g) \leq \alpha$. By the Neyman-Pearson lemma, the optimal classifier has the form

$$g_\alpha(x) = \begin{cases} 1, & \tilde{\Lambda}(x) > \lambda_\alpha, \\ q_\alpha, & \tilde{\Lambda}(x) = \lambda_\alpha, \\ 0, & \tilde{\Lambda}(x) < \lambda_\alpha. \end{cases}$$

where $\tilde{\Lambda}(x) = \tilde{q}_1(x)/\tilde{q}_0(x)$, and $\lambda_\alpha > 0$ and $q_\alpha \in [0, 1)$ are uniquely determined by

$$\tilde{Q}_0(\tilde{\Lambda}(X) < \lambda_\alpha) + q_\alpha \tilde{Q}_0(\tilde{\Lambda}(X) = \lambda_\alpha) = \alpha.$$

Next, we apply Proposition 1 of Blanchard et al. (2016) which we restate in our notation for convenience. (In their notation, $\pi_0 = 1 - \theta_1$, $\pi_1 = \theta_0$.)

Lemma 2 *Let q_0 and q_1 be probability density functions, let $0 \le \theta_0 < \theta_1 \le 1$, and let $\tilde{q}_1$ and $\tilde{q}_0$ be as in (7)-(8). For all $\gamma \ge 0$ and all x such that $q_0(x) > 0$,*

$$\frac{q_1(x)}{q_0(x)} > \gamma \iff \frac{\tilde{q}_1(x)}{\tilde{q}_0(x)} > \lambda,$$

where

$$\lambda = \frac{1 - \theta_1 + \gamma\theta_1}{1 - \theta_0 + \gamma\theta_0}. \tag{9}$$

The result states that the "pure" and "contaminated" likelihood ratios are monotonically equivalent.

Before applying this result, we make the following observations. First, by inspecting (9), as γ varies from 0 to ∞, λ varies between its extremes,

$$\frac{1 - \theta_1}{1 - \theta_0} \le \lambda \le \frac{\theta_1}{\theta_0}.$$

Second, these extremes also bound the range of the contaminated likelihood ratio, which is evident from the expression

$$\tilde{\Lambda}(x) = \frac{\theta_1 q_1(x) + (1 - \theta_1) q_0(x)}{\theta_0 q_1(x) + (1 - \theta_0) q_0(x)} = \frac{\theta_1 \frac{q_1(x)}{q_0(x)} + 1 - \theta_1}{\theta_0 \frac{q_1(x)}{q_0(x)} + 1 - \theta_0}.$$

Third, given λ in this range, one can solve for γ,

$$\gamma = \frac{\lambda(1 - \theta_0) - (1 - \theta_1)}{\theta_1 - \lambda\theta_0} \in [0, \infty].$$

Putting these observations together, λ_α must satisfy $\frac{1-\theta_1}{1-\theta_0} \le \lambda_\alpha \le \frac{\theta_1}{\theta_0}$, and therefore

$$g_\alpha(x) = \begin{cases} 1, & \Lambda(x) > \gamma_\alpha, \\ q_\alpha, & \Lambda(x) = \gamma_\alpha, \\ 0, & \Lambda(x) < \gamma_\alpha, \end{cases}$$

where $\Lambda(x) = q_1(x)/q_0(x)$ and

$$\gamma_\alpha = \frac{\lambda_\alpha(1 - \theta_0) - (1 - \theta_1)}{\theta_1 - \lambda_\alpha\theta_0} \in [0, \infty].$$

Finally, by

$$\eta_Q(x) = \frac{\pi_Q q_1(x)}{\pi_Q q_1(x) + (1 - \pi_Q) q_0(x)} = \frac{\pi_Q \Lambda(x)}{\pi_Q \Lambda(x) + 1 - \pi_Q},$$

we know that $\eta_Q(x)$ is monotonically equivalent to $\Lambda(x)$. This completes the proof.

B.2. Proof of Proposition 2

By the triangle inequality,

$$\begin{aligned}&\left|\epsilon B_Q(g) + (1-\epsilon)A_Q(g) - [\epsilon B_Q(g') + (1-\epsilon)A_Q(g')]\right| \\ &\quad \le \epsilon \left|B_Q(g) - B_Q(g')\right| + (1-\epsilon)\left|A_Q(g) - A_Q(g')\right|.\end{aligned}$$

We claim that $|B_Q(g) - B_Q(g')| \le Q_1(G \Delta G')$. To see this, observe

$$\begin{aligned}B_Q(g) - B_Q(g') &= Q_1(G) - Q_1(G') \\ &= Q_1(G \backslash G') - Q_1(G' \backslash G) \\ &\le Q_1(G \backslash G') + Q_1(G' \backslash G) \\ &= Q_1(G \Delta G').\end{aligned}$$

A similar argument shows that $B_Q(g) - B_Q(g') \ge -Q_1(G \Delta G')$ which establishes the claim.

Similarly, it can be shown that $|A_Q(g) - A_Q(g')| \le Q_0(G \Delta G')$.

Since $Q_X = \pi_Q Q_1 + (1-\pi_Q)Q_0$, we know $Q_X \ge \pi_Q Q_1$ and $Q_X \ge (1-\pi_Q)Q_0$ and therefore $Q_1 \le \frac{1}{\pi_Q} Q_X$ and $Q_0 \le \frac{1}{1-\pi_Q} Q_X$. Combining the above facts establishes the result.

B.3. Proof of Theorem 4

Since the support of Q is contained in the support of P, $\widehat{\eta}_P$ is (δ_m, τ_m)-accurate on the support of Q.

Let $\widehat{F}_{P,Q}(t)$ be the empirical CDF of the random variable $\eta_P(X)$, $X \sim Q_X$, based on $X_{m+1}, \ldots, X_{m+n}$. For $r > 0$, introduce the event

$$E_r = \left\{ \|\widehat{\eta}_P - \eta_P\|_\infty \le \delta_m, \sup_t |F_{P,Q}(t) - \widehat{F}_{P,Q}(t)| \le c_r (\log n / n)^{1/2} \right\}.$$

By the DKW inequality (Massart, 1990), there exists c_r such that E_r occurs with probability at least $1 - \tau_m - n^{-r}$.

Remark: The advantage of having the theorem hold for arbitrary $r > 0$ is that for some estimators, e.g., the ℓ_1-penalized logistic regression estimator studied by van de Geer (2008), r needs to be sufficiently large for the estimator to be (δ_m, τ_m)-accurate with specific rates for δ_m and τ_m.

The proof hinges on the following lemma.

Lemma 3 *There exists $c_{r,\kappa} > 0$ such that for m and n large enough, on E_r,*

$$|\widehat{t}_{P,Q,\alpha} - t_{P,Q,\alpha}| \le \delta_m + c_{r,\kappa} \left(\frac{\log n}{n} \right)^{1/2\kappa}.$$

Proof Introduce the sets $L_P(t) = \{x : \eta_P(x) \le t\}$ and $\widehat{L}_P(t) = \{x : \widehat{\eta}_P(x) \le t\}$. Observe that for any $t \in [0,1]$,

$$\widehat{Q}_X(\widehat{L}_P(t)) \le \widehat{Q}_X(L_P(t+\delta_m)) = \widehat{F}_{P,Q}(t+\delta_m) \le F_{P,Q}(t+\delta_m) + c_r \left(\frac{\log n}{n} \right)^{1/2}.$$

Now let $t'_{P,Q,\alpha} := t_{P,Q,\alpha} - \delta_m - \{2c_r b_1 (\log n/n)^{1/2}\}^{1/\kappa}$, where b_1 is from **(B)**. For m and n large enough, we have $\delta_m + \{2c_r b_1 (\log n/n)^{1/2}\}^{1/\kappa} \leq t_{P,Q,\alpha}$ (so that $t'_{P,Q,\alpha} \in [0,1]$), $1/n < c_r(\log n/n)^{1/2}$, and $\{2c_r b_1 (\log n/n)^{1/2}\}^{1/\kappa} \leq \delta_0$ where δ_0 is from **(B)**. It follows that

$$\begin{aligned}
\widehat{Q}_X(\widehat{L}_P(t'_{P,Q,\alpha})) &\leq F_{P,Q}(t_{P,Q,\alpha} - \{2c_r b_1 (\log n/n)^{1/2}\}^{1/\kappa}) + c_r \left(\frac{\log n}{n}\right)^{1/2} \\
&\leq F_{P,Q}(t_{P,Q,\alpha}) - c_r \left(\frac{\log n}{n}\right)^{1/2} \\
&= 1 - \alpha - c_r \left(\frac{\log n}{n}\right)^{1/2} \\
&< 1 - \alpha - n^{-1} \\
&\leq \lfloor n(1-\alpha) \rfloor / n \\
&\leq \widehat{Q}_X(\widehat{L}_P(\widehat{t}_{P,Q,\alpha})),
\end{aligned}$$

where the second inequality follows from **(B)**. It follows that $\widehat{t}_{P,Q,\alpha} \geq t'_{P,Q,\alpha} = t_{P,Q,\alpha} - \delta_m - c^1_{r,\kappa}(\log n/n)^{1/2\kappa}$ where $c^1_{r,\kappa} = \{2c_r b_1\}^{1/\kappa}$.

The reverse inequality is similar with one slight change, in that we redefine $L_P(t) = \{x : \eta_P(x) < t\}$ and $\widehat{L}_P(t) = \{x : \widehat{\eta}_P(x) < t\}$. Similar to before, for any $t \in [0,1]$,

$$\widehat{Q}_X(\widehat{L}_P(t)) \geq \widehat{Q}_P(L(t - \delta_m)) = \widehat{F}_{P,Q}(t - \delta_m) \geq F_{P,Q}(t - \delta_m) - c_r \left(\frac{\log n}{n}\right)^{1/2}.$$

Now let $t'_{P,Q,\alpha} := t_{P,Q\alpha} + \delta_m + \{2c_r b_2 (\log n/n)^{1/2}\}^{1/\kappa}$, where b_2 is from **(B)**. For m and n large enough, we have $\delta_m + \{2c_r b_2 (\log n/n)^{1/2}\}^{1/\kappa} \leq 1 - t_{P,Q,\alpha}$ (so that $t'_{P,Q,\alpha} \in [0,1]$), $1/n < c_r(\log n/n)^{1/2}$, and $\{2c_r b_2 (\log n/n)^{1/2}\}^{1/\kappa} \leq \delta_0$ where δ_0 is from **(B)**. It follows that

$$\begin{aligned}
\widehat{Q}_X(\widehat{L}_P(t'_{P,Q,\alpha})) &\geq F_{P,Q}(t_{P,Q,\alpha} + \{2c_r b_2 (\log n/n)^{1/2}\}^{1/\kappa}) - c_r \left(\frac{\log n}{n}\right)^{1/2} \\
&\geq F_{P,Q}(t_{P,Q,\alpha}) + c_r \left(\frac{\log n}{n}\right)^{1/2} \\
&= 1 - \alpha + c_r \left(\frac{\log n}{n}\right)^{1/2} \\
&> 1 - \alpha + n^{-1} \\
&\geq \lfloor n(1-\alpha) \rfloor / n \\
&\geq \widehat{Q}_X(\widehat{L}_P(\widehat{t}_{P,Q,\alpha}))
\end{aligned}$$

where the second inequality follows from **(B)**. The modified definitions of L_P and $\widehat{L}_P$ are needed in the final step. It follows that $\widehat{t}_{P,Q,\alpha} \leq t'_{P,Q,\alpha} = t_{P,Q,\alpha} + \delta_m + c^2_{r,\kappa}(\log n/n)^{1/2\kappa}$ where $c^2_{r,\kappa} = \{2c_r b_2\}^{1/\kappa}$.

The result now follows by combining the above inequalities and taking $c_{r,\kappa} = \max\{c^1_{r,\kappa}, c^2_{r,\kappa}\}$. ■

To prove the theorem, observe that on E_r,

$$\begin{aligned} Q_X(\widehat{G}_{P,Q,\alpha} \backslash G_{P,Q,\alpha}) &= Q_X(\widehat{\eta}_P(X) \geq \widehat{t}_{P,Q,\alpha}, \eta_P(X) < t_{P,Q,\alpha}) \\ &\leq Q_X\left\{ t_{P,Q,\alpha} - \delta_m - c_{r,\kappa}\left(\frac{\log n}{n}\right)^{1/2\kappa} < \eta_P(X) < t_{P,Q,\alpha} \right\} \\ &= F_{P,Q}(t_{P,Q,\alpha}) - F_{P,Q}\left\{ t_{P,Q,\alpha} - \delta_m - c_{r,\kappa}\left(\frac{\log n}{n}\right)^{1/2\kappa} \right\} \\ &\leq b_2\left\{ \delta_m + c_{r,\kappa}\left(\frac{\log n}{n}\right)^{1/2\kappa} \right\}^{\kappa} \\ &\leq 2^{\kappa} b_2\left\{ \delta_m^{\kappa} + c_{r,\kappa}^{\kappa}\left(\frac{\log n}{n}\right)^{1/2} \right\}, \end{aligned}$$

where the next-to-last inequality follows from **(B)** and holds when m and n are large enough that $2\delta_m + c_{r,\kappa}(\log n/n)^{1/2\kappa} \leq \delta_0$. The other term is handled similarly.

B.4. Proof of Theorem 5

The following result follows from a result of Steinwart (2003).

Lemma 4 *Let k be a universal kernel and let $\lambda = \lambda_m$ such that $\lambda \to 0$ and $m\lambda^2 \to \infty$. For all $\beta, \gamma, \nu \in (0,1)$, for m sufficiently large,*

$$P_X(\{x : |\eta_P(x) - \widehat{\eta}_P(x)| \geq \beta\}) \leq \gamma$$

with probability at least $1 - \nu$ with respect to the draw of $(X_i, Y_i), i = 1, \ldots, m$.

In words, the P_X-measure of the set where $\widehat{\eta}_P$ deviates from η_P by more than β can be made arbitrarily small, with arbitrarily high probability, by taking m large enough.

Proof Denote

$$E_m(\beta) = \{x : |\eta_P(x) - \widehat{\eta}_P(x)| \geq \beta\}.$$

Define $f_P(x) = \log(1 - \eta_P(x))/\eta_P(x)$ and observe that $\eta_P(x) = \tau(f_P(x))$ and $\widehat{\eta}_P(x) = \tau(\widehat{f}_P(x))$, where $\tau(f) = (1 + \exp(-f))^{-1}$. Also define

$$F_m(\beta) = \{x : |f_P(x) - \widehat{f}_P(x)| \geq \beta\}.$$

Notice that $E_m(\beta) \subseteq F_m(\beta)$ because τ is 1-Lipschitz. The result now follows from Theorem 35 of Steinwart (2003) (see also Theorem 22 and Remark 24). ■

For any $\beta, \gamma \in (0,1)$, let $\Theta_m(\beta, \gamma)$ be the event on which $P_X(\{x : |\widehat{\eta}_P(x) - \eta_P(x)| \geq \beta\}) \leq \gamma$. By Lemma 4, $\Pr(\Theta_m(\beta, \gamma))$ can be made arbitrarily close to 1 by taking m sufficiently large.

Now consider the family of sets $\mathcal{C} = \{C_t \,|\, t \geq 0\}$ where $C_t = \{x : \widehat{\eta}(x) \geq t\}$. This family has a shatter coefficient $S(\mathcal{C}, n) = n + 1$. By the VC inequality (Devroye et al., 1996),

$$|Q_X(C_t) - \widehat{Q}_X(C_t)| \leq \sqrt{\frac{8(\log S(\mathcal{C}, n) + \log n)}{n}} \leq 4\sqrt{\frac{\log(n+1)}{n}} = \epsilon_n \tag{10}$$

with probability at least $1 - 1/n$. This follows by applying the VC inequality to the conditional distribution of $X_{m+1}, \ldots, X_{m+n}$ given $(X_1, Y_1), \ldots, (X_m, Y_m)$, and then marginalizing over $(X_1, Y_1), \ldots, (X_m, Y_m)$.

Let Ω_n denote the event on which the bound of (10) holds. Thus, $\Pr(\Omega_n) \geq 1 - 1/n$.

Lemma 5 *Fix $\beta, \gamma > 0$, and assume **(A)**, **(B)**, and **(C)** hold. On the event $\Theta_m(\beta, \gamma/c_0) \cap \Omega_n$*

$$\widehat{t}_{P,Q,\alpha} \leq t_{P,Q,\alpha}.$$

Furthermore, if γ satisfies $(3\gamma/b_1)^{1/\kappa} < \delta_0$, then for n sufficiently large, on the event $\Theta_m(\beta, \gamma/c_0) \cap \Omega_n$

$$t_{P,Q,\alpha} - \widehat{t}_{P,Q,\alpha} \leq 2\beta + (3(\gamma + \epsilon_n)/b_1)^{1/\kappa}.$$

Proof Assume $\Theta_m(\beta, \gamma/c_0) \cap \Omega_n$ occurs. Recall

$$\widehat{t}_{P,Q,\alpha} := \inf\{t \,|\, \widehat{Q}_X(\{x : \widehat{\eta}_P(x) \geq t + \beta\}) \leq \alpha + \gamma + \epsilon_n\},$$

and

$$t_{P,Q,\alpha} := \inf\{t \,|\, Q_X(\{x : \eta_P(x) \geq t\}) \leq \alpha\}.$$

To see that $\widehat{t}_{P,Q,\alpha} \leq t_{P,Q,\alpha}$ on $\Theta_m(\beta, \gamma/c_0) \cap \Omega_n$, from the definition of $t_{P,Q,\alpha}$ we have $Q_X(\{x : \eta_P(x) \geq t_{P,Q,\alpha}\}) \leq \alpha$. By $\Theta_m(\beta, \gamma/c_0)$ and **(C')**, it follows that $Q_X(\{x : \widehat{\eta}_P(x) \geq t_{P,Q,\alpha} + \beta\}) \leq \alpha + \gamma$, and by Ω_n, we have that $\widehat{Q}_X(\{x : \widehat{\eta}_P(x) \geq t_{P,Q,\alpha} + \beta\}) \leq \alpha + \gamma + \epsilon_n$. The result follows by definition of $\widehat{t}_{P,Q,\alpha}$.

For the reverse direction, let γ be small enough such that $(3\gamma/b_1)^{1/\kappa} < \delta_0$. Assume n is large enough that $(3(\gamma + \epsilon_n)/b_1)^{1/\kappa} \leq \delta_0$.

Let $\tilde{t}_{P,Q,\alpha} := t_{P,Q,\alpha} - q$, where $q = 2\beta + (3(\gamma + \epsilon_n)/b_1)^{1/\kappa}$. On $\Theta_m(\beta, \gamma) \cap \Omega_n$, we have that

$$\begin{aligned}
\widehat{Q}_X(\{x : \widehat{\eta}_P(x) \geq \tilde{t}_{P,Q,\alpha} + \beta\}) &= \widehat{Q}_X(\{x : \widehat{\eta}_P(x) \geq t_{P,Q,\alpha} - q + \beta\}) \\
&\geq Q_X(\{x : \widehat{\eta}_P(x) \geq t_{P,Q,\alpha} - q + \beta\}) - \epsilon_n \\
&\geq Q_X(\{x : \eta_P(x) \geq t_{P,Q,\alpha} - q + 2\beta\}) - \gamma - \epsilon_n \\
&= 1 - F_{P,Q}(t_{P,Q,\alpha} - q + 2\beta) - \gamma - \epsilon_n \\
&\geq \alpha + 2\gamma + 2\epsilon_n,
\end{aligned}$$

where the last step follows from **(B)** and $F_{P,Q}(t_{P,Q,\alpha}) = \alpha$. We conclude that $\widehat{t}_{P,Q,\alpha} \geq \tilde{t}_{P,Q,\alpha}$, and therefore $t_{P,Q,\alpha} - \widehat{t}_{P,Q,\alpha} \leq q = 2\beta + (3(\gamma + \epsilon_n)/b_1)^{1/\kappa}$.

■

To prove the theorem, let $\epsilon > 0$ and $\xi > 0$. We will show that for β, γ sufficiently small, $\Pr(Q_X(\widehat{G}_{P,Q,\alpha} \Delta G_{P,Q,\alpha}) \leq \epsilon) \geq 1 - \xi$ for m and n sufficiently large. Thus, select β and γ such that (i) $3\beta + (3\gamma/b_1)^{1/\kappa} < \delta_0$, and (ii) $b_2\beta^\kappa + b_2(3\beta + (3\gamma/b_1)^{1/\kappa})^\kappa + 2\gamma < \epsilon$.

Having fixed β and γ, let n be sufficiently large such that (i) the conclusion of Lemma 5 holds, (ii) $1/n < \xi/2$, and (iii) $b_2\beta^\kappa + b_2(3\beta + (3(\gamma + \epsilon_n)/b_1)^{1/\kappa})^\kappa + 2\gamma < \epsilon$. Also let m be sufficiently large that $\Pr(\Theta_m(\beta, \gamma/c_0)) \geq 1 - \xi/2$. Thus, $\Theta_m(\beta, \gamma/c_0) \cap \Omega_n$ occurs with probability at least $1 - \xi$.

Observe

$$\begin{aligned} Q_X(G_{P,Q,\alpha}\Delta\widehat{G}_{P,Q,\alpha}) = {} & Q_X(\eta_P(X) \geq t_{P,Q,\alpha}, \widehat{\eta}_P(X) < \widehat{t}_{P,Q,\alpha}) \\ & + Q_X(\eta_P(X) < t_{P,Q,\alpha}, \widehat{\eta}_P(X) \geq \widehat{t}_{P,Q,\alpha}). \end{aligned}$$

The first term may be bounded on $\Theta_m(\beta,\gamma) \cap \Omega_n$ as

$$\begin{aligned} Q_X(\eta_P(X) \geq t_{P,Q,\alpha}, \widehat{\eta}_P(X) < \widehat{t}_{P,Q,\alpha}) &\leq Q_X(\eta_P(X) \geq t_{P,Q,\alpha}, \widehat{\eta}_P(X) < t_{P,Q,\alpha}) \\ &\leq Q_X(t_{P,Q,\alpha} \leq \eta_P(X) \leq t_{P,Q,\alpha} + \beta) + \gamma \\ &= F_{P,Q}(t_{P,Q,\alpha} + \beta) - F_{P,Q}(t_{P,Q,\alpha}) + \gamma \\ &\leq b_2\beta^\kappa + \gamma, \end{aligned}$$

where the first step follows from Lemma 5, the second from $\Theta_m(\beta, \gamma/c_0)$, and the last from **(B)**.

As for the second term, let $q = 2\beta + (3(\gamma + \epsilon_n)/b_1)^{1/\kappa}$. Then

$$\begin{aligned} Q_X(\widehat{\eta}_P(X) \geq \widehat{t}_{P,Q,\alpha}, \eta_P(X) < t_{P,Q,\alpha}) &\leq Q_X(\widehat{\eta}_P(X) \geq t_{P,Q,\alpha} - q, \tilde{\eta}_P(X) < t_{P,Q,\alpha}) \\ &\leq Q_X(t_{P,Q,\alpha} - q - \beta \leq \eta_P(X) \leq t_{P,Q,\alpha}) + \gamma \\ &= F_{P,Q}(t_{P,Q,\alpha} - q - \beta) - F_{P,Q}(t_{P,Q,\alpha}) + \gamma \\ &\leq b_2(3\beta + (3(\gamma + \epsilon_n)/b_1)^{1/\kappa})^\kappa + \gamma. \end{aligned}$$

with similar reasoning as the first case. The result now follows from the selected properties of β, γ, m and n.