Tailored Overlap for Learning Under Distribution Shift

David Bruns-Smith Alex D'Amour Avi Feller Steve Yadlowsky

Abstract

Distributional overlap is a critical determinant of learnability in domain adaptation. The standard theory quantifies overlap in terms of χ^2 divergence, as this factors directly into variance and generalization bounds agnostic to the functional form of the Y-X relationship. However, in many modern settings, we cannot afford this agnosticism; we often wish to transfer across distributions with disjoint support, where these standard divergence measures are infinite. In this note, we argue that "tailored" divergences that are restricted to measuring overlap in a particular function class are more appropriate. We show how χ^2 (and other) divergences can be generalized to this restricted function class setting via a variational representation, and use this to motivate balancing weight-based methods that have been proposed before, but, we believe, should be more widely used.

1 Introduction

Learning to predict in a target domain using labeled data from a potentially-different source domain is a fundamental task in reliable machine learning. Successful learning requires that the source domain is sufficiently representative of the target. This concept is called *overlap* in the causal inference literature and typically takes the form of conditions on the likelihood ratio or the χ^2 divergence between the source and target. The χ^2 divergence occurs naturally in statistical causal inference results [Tsiatis, 2006], but also in generalization bounds under covariate shift [Cortes et al., 2010].

An important technical challenge is that overlap breaks down in high dimensional settings: the χ^2 divergence between the source and target densities is likely to be unbounded [D'Amour et al., 2021]. And yet, we regularly consider applied problems with disjoint support, and in many cases supervised learning models can transfer without much difficulty. Consider the shift between ImageNet and ImageNet-R [Hendrycks et al., 2021]. Transfer is still possible even though overlap is violated by design, as every image in the target domain is a drawing. We therefore need a different notion of overlap to help explain why the training set is nonetheless informative for the test set.

In this paper, we connect two strands of the literature to make progress on this question. The first strand derives generalization bounds under covariate shift that remain meaningful even when the source and target populations are non-overlapping. In particular, we build on Johansson et al. [2022], who use integral probability metrics (IPMs), which are distances between probability distributions with respect to a specific function class. The second strand focuses on leveraging variational representations for training Generalized Adversarial Networks (GANs). Specifically, Glaser et al. [2021], Birrell et al. [2022a] develop IPM-analogs of the χ^2 divergence by considering a variational representation restricted to range over a given function class. Here we show that the function-class-specific χ^2 divergence is exactly the quantity that controls the generalization error in Johansson et al. [2022]. We then argue that this quantity is the appropriate measure of overlap in high dimensional settings.

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Problem Setup and Notation

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the covariate (or feature) space, and $\mathcal{Y} \subset \mathbb{R}$ denote the outcome (or label) space. Let P and Q be joint distributions on $(\mathcal{X}, \mathcal{Y})$ representing the source and target domains respectively. The likelihood ratio dQ/dP is a key measure of the difference between P and Q.

We assume that both covariates and outcomes are observed in the source sample, but that in the target sample, we only observe the covariates. Our goal is either to estimate the prediction function $\mu(x) := \mathbb{E}_Q[Y|X = x]$; or the mean missing outcome $\mathbb{E}_Q[Y]$. In this work, we will assume that only the covariate distribution shifts (in various literatures called covariate shift, ignorability, exogeneity, conditional mean independence, or no unobserved confounders).

2.1 Overlap and Learning Under Covariate Shift

The representativeness of the source population for learning the target population is called overlap:[1]

Definition 1 (Overlap)**.** *Distributions P and Q satisfy* overlap *if $dQ/dP(x) < \infty$ a.s.*

This is typically called a *continuity* assumption in the domain adaptation literature. We can quantify overlap by computing moments of the likelihood ratio. For example:

Definition 2 (Weak Overlap, χ^2 Divergence)**.** *Distributions P and Q satisfy* weak overlap *if*

$$\chi^2(Q||P) := \mathbb{E}_P[(dQ/dP(X) - 1)^2] < \infty.$$

Such assumptions regularly appear in generalization bounds under covariate shift using importance weighting.

If Q and P have disjoint support, then overlap is violated and the χ^2 divergence is infinite. We introduce versions of the χ^2 divergence and the likelihood ratio that are tailored to a particular function class and remain meaningful without shared support. In a similar vein, some recent theoretical work [Kpotufe and Martinet, 2018, Pathak et al., 2022] considers likelihood ratios over local neighborhoods of the underlying metric space. However, for high dimensional tasks like image classification, the source and target populations are likely to be disjoint even with respect to ℓ_2 or ℓ_∞ balls around the data points — and yet transfer is still possible. Our approach is to choose a notion of locality tailored to the specific task at hand.

3 Tailored Overlap in High Dimensions

We want to define some notion of overlap that: (1) captures the underlying learnability of the target using data from the source; and (2) remains useful in high dimensional settings with disjoint support. Our starting point is the observation that transfer will still be possible under overlap violations *if the outcome function is insensitive to these violations*. This might mean that the relationship between covariates and outcomes only depends on a particular lower-dimensional representation (e.g. semantic features rather than pixels in an image classification task), or has some functional form restriction (e.g. linear or belonging to some RKHS). Integral Probability Metrics (IPMs) formalize the concept of a difference in distributions with respect to a particular outcome function class:

Definition 3 (Integral Probability Metric)**.** *The IPM between P and Q with respect to a function class $\mathcal{F}$ is:*

$$IPM_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \Big\{ |\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \Big\}.$$

Johansson et al. [2022] gives generalization bounds under covariate shift in terms of an IPM but between a *reweighted* source distribution and the target distribution:

Theorem (Johansson et al. 2022, Informal)**.** *Let $\hat{h}$ be any predictor and assume that the loss of $\hat{h}$ belongs to some function class $\mathcal{L}$. Let the conditional variance of Y be uniformly bounded by σ^2. Let P and Q be the empirical distribution of the source and target samples. For any weights w, let*

[1] The terminology of “weak overlap" comes from its use in the causal inference literature [Tsiatis, 2006].

$Err(w)$ be the difference between the true risk of $\hat{h}$ on the target population and the empirical risk of $\hat{h}$ on the source population reweighted by w. Then,

$$Err(w) \leq c_1 \underbrace{IPM_{\mathcal{L}}(wP, Q)}_{bias} + c_2 \underbrace{\sigma^2 Var_P[w]}_{variance} + \textit{additional terms}, \tag{1}$$

where the additional terms and the scalars c_1 and c_2 are unrelated to the covariate shift.

We will show that the sum of the bias and variance terms are exactly equivalent to a variational representation of the χ^2 divergence that is tailored to a function class $\mathcal{F}$.

3.1 Defining the Likelihood Ratio and χ^2 Divergence with Respect to a Function Class

The χ^2 divergence can be written in a variational representation as an optimization problem over continuous functions [Nguyen et al., 2010, Agrawal and Horel, 2021, Birrell et al., 2022b]. This representation also implicitly defines the likelihood ratio dQ/dP, which is the unique function that achieves the optimum. We can define analogous versions of these measures that are tailored to a particular function class $\mathcal{F}$ by restricting the optimization problem as in Birrell et al. [2022a].

Definition 4 ($\mathcal{F}$-χ^2 Divergence and $\mathcal{F}$-Likelihood Ratio with Parameter α)**.** *The $\mathcal{F}$-χ^2 Divergence with parameter α is:*

$$\chi^2_{\mathcal{F},\alpha}(Q||P) := \min_{f \in \mathcal{F}} \Big\{ \frac{1}{4\alpha} Var_P[f(X)] - (\mathbb{E}_Q[f(X)] - \mathbb{E}_P[f(X)]) \Big\}. \tag{2}$$

Denoting the minimizer f^*, we define the $\mathcal{F}$-Likelihood Ratio with Parameter α as:

$$(dQ/dP)_{\mathcal{F},\alpha} := \frac{1}{2\alpha}(f^* - \mathbb{E}_P[f^*]) + 1. \tag{3}$$

In the special case where Q and P have common support and $\mathcal{F}$ is unrestricted, we recover the χ^2 divergence: for any α, $(dQ/dP)_{\mathcal{F},\alpha} = dQ/dP$, so the second term in (2) is zero, and $\chi^2_{\mathcal{F},\alpha}(Q||P) = \alpha\chi^2(Q||P)$. However, when the infimum is restricted to $\mathcal{F}$, $\chi^2_{\mathcal{F},\alpha}(Q||P)$ will not generally be linear in α and the second term will not generally be zero. Instead, for each α, the infimum will be achieved by different witness functions that are not equal up to constants. In other words, the constrained optimization problem now must trade off the first and second terms in the objective with the strength of the trade-off governed by α. Notice that, even if Q and P have disjoint support, the minimum in (2) will still be finite when $\mathcal{F}$ is sufficiently constrained.

3.2 The Tailored χ^2 Divergence Controls Generalization Under Covariate Shift

The sum of the bias and variance terms in (1) is exactly equal to the tailored χ^2 divergence defined in (2). In fact, with proper tuning of α, $w = (dQ/dP)_{\mathcal{L},\alpha}$ gives the tightest possible bound. To see this, consider the following result, proved in Bruns-Smith and Feller [2022] and Birrell et al. [2022a]:

Proposition 1 (Bruns-Smith and Feller, 2022)**.**

$$\chi^2_{\mathcal{F},\alpha}(Q||P) = \inf_{w:\mathbb{E}_P[w]=1} \Big\{ IPM_{\mathcal{F}}(wP, Q) + \alpha Var_P[w] \Big\}. \tag{4}$$

Furthermore, $(dQ/dP)_{\mathcal{F},\alpha}$ is the unique w which achieves the infimum.

Then we have the following connection to Johansson et al. [2022]:

Proposition 2. *Consider the setting in (1). Let $c_1, c_2 = 1$, otherwise the same result holds for a different scaling factor. Then:*

$$Err(w) \leq \chi^2_{\mathcal{L},\sigma^2}(Q||P) + \textit{ additional terms},$$

where the additional terms are unrelated to covariate shift. Furthermore, $\chi^2_{\mathcal{L},\sigma^2}(Q||P)$ is the tightest possible bound over all choices of w in (1) given these assumptions.

Proof. The inequality follows from (1) with $w = (dQ/dP)_{\mathcal{L},\sigma^2}$. Then from Proposition 1, these weights give the smallest bound over all w. □

And so, with proper tuning of α, the tailored χ^2 divergence in Definition 4 precisely controls the difficulty of learning under covariate shift. Indeed, Proposition 1 shows that the tailored χ^2 divergence is equivalent to the *balancing weights* optimization problem used as an estimator in Johansson et al. [2022]. The resulting optimization problem is strictly convex for $\alpha > 0$, and so can be solved efficiently. We consider implications of this in Section 4.

3.3 Example: A Toy Problem

Here we introduce a simple toy problem that shows how restrictions on $\mathcal{F}$ can flexibly incorporate dimensionality reduction and functional form restrictions. Let $\mathcal{X} = \{1, 2, 3\}$. The source and target populations are shown in lefthand panel of Figure 1.

$$P(x) = \begin{cases} 2/5 & \text{if } x = 1 \\ 0 & \text{if } x = 2 \\ 3/5 & \text{if } x = 3 \end{cases}$$

$$Q(x) = \begin{cases} 0 & \text{if } x = 1 \\ 4/5 & \text{if } x = 2 \\ 1/5 & \text{if } x = 3 \end{cases}$$

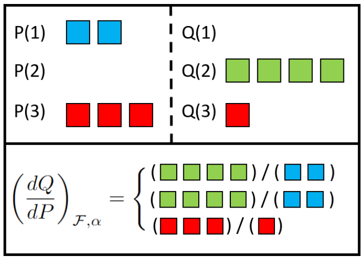

Figure 1: Source and target distributions and tailored likelihood ratio for the toy example.

Here $dQ/dP(1) = \infty$ and the χ^2 divergence is unbounded. But suppose we were interested in the function class $\mathcal{F} = \{f : f(1) = f(2)\}$; for example, as a model for $\mathbb{E}[Y|X]$ as in Bruns-Smith and Feller [2022] or for the loss of a predictor as in Johansson et al. [2022]. Then for all α, the tailored density ratio is the same, and is illustrated in the righthand panel of Figure 1; the tailored χ^2 divergence is the variance of this ratio. Unlike the unconstrained divergence, the constrained divergence is finite due to the lower-dimensional representation defined by the choice of $\mathcal{F}$.

Even so, this function class does not admit a bias-variance tradeoff because the functions in $\mathcal{F}$ can be arbitrarily large. If we knew instead that $\mathcal{F} = \{f : f(1) = f(2), \|f\|_\infty \leq B\}$, we would then obtain a continuum of tailored ratios for $\alpha \in (0, \infty)$, each defined by the convex optimization problem (2). Thus, adding the norm restriction introduces a bias-variance tradeoff: choosing $\alpha = \sigma^2$ gives the best generalization bound from Johansson et al. [2022]. Other values of α give a different tradeoff.

4 Discussion

While this brief note focuses on a technical connection, we argue this has broader implications for practical learning under distribution shift when source and target have disjoint support.

First, we argue that the finite-sample χ^2 divergence tailored to a function class of interest is a natural diagnostic tool that remains meaningful even in high dimensions. Likewise, the tailored likelihood ratio can be inspected for overlap violations, as in Lei et al. [2022], or used to extract the most anomalous examples with respect to $\mathcal{F}$, as in Rabanser et al. [2019].

Second, the optimization problem expressed in Equation 4 is known as *balancing weights* [Ben-Michael et al., 2021, Johansson et al., 2022]. Above we give an additional interpretation of that objective as the tailored χ^2 divergence for a given $\mathcal{F}$. Thus, balancing weights are an appropriate high-dimensional generalization of directly estimating the likelihood ratio that controls the generalization error under covariate shift even with disjoint support.

Finally, the push to specify an outcome function $\mathcal{F}$ when considering overlap is a departure from current practice. We argue, however, that this is inevitable: while "untailored" χ^2 divergences allow for arbitrary outcome functions, the corresponding bounds are entirely uninformative in many modern ML tasks. Thus, in some sense, we always need to take a stand on $\mathcal{F}$ to make progress. This raises several important points: (1) how do we choose $\mathcal{F}$? (2) what choices of $\mathcal{F}$ are practically useful in various domains? and (3) what if our choice of $\mathcal{F}$ is wrong? These are critical questions for future study. In general, the choice of $\mathcal{F}$ induces a bias-variance trade off that will be different in different settings. We can also conduct formal sensitivity analyses to the choice of $\mathcal{F}$.

References

R. Agrawal and T. Horel. Optimal bounds between f-divergences and integral probability metrics. *The Journal of Machine Learning Research*, 22(1):5662–5720, 2021.

E. Ben-Michael, A. Feller, D. A. Hirshberg, and J. R. Zubizarreta. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021.

J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet. (f, gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022a.

J. Birrell, M. A. Katsoulakis, and Y. Pantazis. Optimizing variational representations of divergences and accelerating their statistical estimation. *IEEE Transactions on Information Theory*, 2022b.

D. A. Bruns-Smith and A. Feller. Outcome assumptions and duality theory for balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 11037–11055. PMLR, 2022.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

P. Glaser, M. Arbel, and A. Gretton. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, 34:8018–8031, 2021.

D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23:1–50, 2022.

S. Kpotufe and G. Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pages 1882–1886. PMLR, 2018.

L. Lei, A. D'Amour, P. Ding, A. Feller, and J. Sekhon. Distribution-free assessment of population overlap in observational studies. 2022.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

R. Pathak, C. Ma, and M. Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR, 2022.

S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

A. A. Tsiatis. Semiparametric theory and missing data. 2006.